International Journal of Intelligent Data and Machine Learning (IJIDML)
eISSN: 3087-4262
Volume03 Issue09 | Page No: - 18-24

Received: 22 Aug. 2026, Revised: 29 Aug. 2026, Accepted: 08 September 2026, Published: 17 September 2026

DOI: https://doi.org/10.55640/ijidml-v03i09-03

Intelligent Healthcare Systems Powered by Large Language Models: Applications, Limitations, and Emerging Research Perspectives

Chinedu Emmanuel Okafor [1]

[1] Department of Computer Science, West African Institute of Technology, Lagos, Nigeria

ABSTRACT: Large language models (LLMs) are increasingly being incorporated into healthcare systems to support clinical communication, medical reasoning, information synthesis, patient interaction, documentation, and emerging forms of predictive analytics. Their adoption introduces a significant shift from conventional task-specific artificial intelligence toward general-purpose language-based systems capable of processing heterogeneous clinical information. However, the usefulness of these systems depends not only on linguistic performance but also on factual reliability, clinical fidelity, transparency, confidentiality, evaluation methodology, and regulatory suitability. This paper presents a structured review and analytical framework for understanding intelligent healthcare systems powered by LLMs, based exclusively on the provided literature. The analysis examines major application domains, including clinical decision support, patient-facing communication, clinical summarization, diagnostic reasoning, health trajectory forecasting, and digital-twin-oriented systems. Particular attention is given to the discrepancy between apparent model capability and clinically meaningful performance. The literature indicates that LLMs can provide substantial value in information-intensive healthcare workflows, but hallucinations, inconsistent reasoning, evaluation limitations, privacy concerns, and regulatory uncertainty constrain their safe deployment. The paper proposes a lifecycle-oriented perspective in which LLM-based healthcare systems should be evaluated through task-specific clinical benchmarks, human oversight, safety monitoring, confidentiality controls, and regulatory alignment. The findings position LLMs as augmentation technologies rather than autonomous replacements for healthcare professionals and identify research priorities involving robust evaluation, clinically grounded reasoning, patient safety, and trustworthy system architectures.

KEYWORDS: Large Language Models, Healthcare Artificial Intelligence, Clinical Decision Support, Medical Reasoning, Patient Interaction, Clinical NLP, Digital Health, AI Safety, Healthcare Systems.

1. Introduction

Healthcare is increasingly characterized by large volumes of heterogeneous information generated through electronic clinical documentation, medical literature, diagnostic reports, patient communications, and longitudinal health records. Conventional artificial intelligence systems have generally been designed around narrowly defined prediction or classification tasks. Large language models (LLMs), in contrast, provide a general-purpose computational architecture capable of generating, transforming, summarizing, and interpreting natural-language information across multiple healthcare activities. This capability has stimulated substantial interest in developing intelligent healthcare systems that integrate language-based reasoning with clinical workflows.

The emergence of LLMs in medicine represents both a technological opportunity and an evaluation challenge. The potential value of these systems arises from their ability to transform complex information into accessible language, support clinical documentation, assist patient communication, and facilitate reasoning-oriented tasks. The broader medical landscape is consequently moving toward systems in which LLMs may function as interfaces between clinicians, patients, and large bodies of health information (Clusmann et al., 2023). At the same time, the effectiveness of an LLM cannot be inferred solely from its ability to produce fluent or plausible text. Recent research has demonstrated that evaluation practices themselves can create an overly optimistic impression of model capability, emphasizing the importance of clinically meaningful assessment (Agrawal et al., 2025).

Patient-facing applications illustrate this tension particularly well. LLMs can respond to medical questions

in accessible language and potentially improve access to health information. However, performance on real-world user-generated queries requires evaluation under conditions that differ substantially from curated academic benchmarks. Research assessing GPT-4 using real-world medical queries highlights the importance of examining how models perform in realistic advice-delivery settings rather than relying exclusively on controlled test conditions (Jo et al., 2024). This issue is central to safe deployment because a response that is linguistically convincing but medically inappropriate can create risks for patients.

The clinical reasoning dimension introduces another layer of complexity. LLMs may generate apparently coherent diagnostic explanations, yet coherence does not necessarily imply faithful or clinically valid reasoning. Evidence concerning the influence of LLMs on diagnostic reasoning and the fidelity of medical reasoning suggests that system outputs require careful interpretation and independent validation (Goh et al., 2024; Bedi et al., 2025). Similarly, clinical summarization systems can achieve strong performance and, in some circumstances, outperform medical experts on specific textual summarization tasks, demonstrating meaningful potential for workflow augmentation (Van Veen et al., 2024).

Against this background, this paper investigates how LLM-powered intelligent healthcare systems can be conceptualized, evaluated, and responsibly integrated into healthcare environments. Its objectives are to synthesize the provided literature on applications and limitations, identify the principal technical and clinical dimensions of trustworthy deployment, and establish emerging research directions for future LLM-based healthcare architectures. The scope encompasses clinical reasoning, patient interaction, summarization, predictive health applications, evaluation, confidentiality, hallucination management, and regulatory considerations.

2. Literature Review

The provided literature demonstrates that LLM research in healthcare has progressed from broad conceptual exploration toward increasingly task-specific clinical evaluation. Clusmann et al. (2023) characterize the future landscape of LLMs in medicine, establishing a broad foundation for considering their applications across medical contexts. Maity et al. (2025) similarly synthesize opportunities, challenges, and future directions, emphasizing that the significance of LLMs extends beyond isolated language-generation tasks toward broader healthcare and medical applications.

A major theme emerging from the literature is the diversity of potential applications. Clinical text summarization represents one comparatively well-defined use case. Van Veen et al. (2024) show that adapted LLMs can outperform medical experts in clinical text summarization, indicating that LLMs may offer practical benefits in reducing information-processing burdens. Such applications are particularly attractive because summarization can be positioned as an assistive task in which clinicians retain responsibility for interpreting and validating the generated information.

Patient interaction presents a different risk profile. Johri et al. (2025) propose an evaluation framework for the clinical use of LLMs in patient interaction tasks, emphasizing the need for structured assessment before deployment. Jo et al. (2024) provide complementary evidence by evaluating GPT-4 using real-world user-generated medical queries. Their work is important because patient questions are often ambiguous, context-dependent, and heterogeneous. Consequently, performance on real-world queries offers a more clinically relevant perspective than standardized language benchmarks alone.

The literature also reveals concerns regarding hallucination and misinformation. Hatem et al. (2023) identify AI hallucinations as a significant issue for healthcare professionals and discuss the importance of mitigation strategies. The problem is particularly serious in medicine because fabricated clinical information can appear plausible while influencing decisions or patient behavior. The systematic review by Truhn et al. (2025) further situates patient-care applications within a broader taxonomy of design and output-related risks.

Medical reasoning constitutes another critical research domain. Goh et al. (2024) investigate the influence of LLMs on diagnostic reasoning through a randomized clinical trial, while Bedi et al. (2025) focus on the fidelity of medical reasoning in LLMs. Together, these studies demonstrate that evaluating whether a model reaches a correct answer is insufficient. Researchers must also investigate the reliability, consistency, and clinical validity of the reasoning process underlying outputs.

Agrawal et al. (2025) reinforce this concern through the concept of an evaluation illusion in medical LLM research. Their work indicates that evaluation design can substantially influence perceived model effectiveness. This observation has important theoretical implications: healthcare AI evaluation must measure clinically meaningful outcomes rather than merely linguistic quality or benchmark accuracy.

LLMs are also being explored in longitudinal and predictive applications. Makarov et al. (2025) examine how LLMs can forecast patient health trajectories and support digital twins, suggesting a transition from retrospective information processing toward prospective health modeling. Such applications have potentially

substantial value but also introduce greater requirements for data quality, longitudinal validation, uncertainty estimation, and clinical oversight.

Finally, responsible deployment requires consideration of confidentiality and regulation. Perlis and Fihn (2023) highlight evaluation standards and confidentiality issues associated with using LLMs with clinical data. The U.S. Food and Drug Administration materials concerning artificial intelligence and machine learning software as a medical device, AI-enabled medical devices, and AI-supported regulatory decision-making further demonstrate the growing importance of regulatory alignment. Therefore, the literature collectively suggests that the future of LLM-powered healthcare depends on balancing computational capability with evaluation rigor, clinical safety, privacy protection, and governance.

3. Methodology

3.1 Research Design

This study adopts a structured analytical review methodology based exclusively on the references supplied for the research topic. Rather than treating the literature as a collection of independent studies, the methodology organizes the evidence around a healthcare LLM lifecycle consisting of application, clinical interaction, evaluation, risk management, governance, and future integration.

The analytical approach is designed to identify recurring technical and clinical themes across the literature. Each reference is conceptually positioned according to its primary contribution: application development, clinical evaluation, patient interaction, reasoning assessment, risk identification, predictive modeling, confidentiality, or regulation. This enables a cross-study synthesis rather than a simple chronological literature summary.

3.2 Conceptual Architecture of an Intelligent Healthcare LLM System

An intelligent healthcare system powered by an LLM can be represented as a layered architecture comprising five functional levels.

The data and context layer contains patient records, clinical notes, medical queries, reports, and other authorized healthcare information. The primary requirement at this level is controlled access to relevant information. Clinical confidentiality must be preserved throughout data acquisition and processing, making privacy and governance fundamental architectural properties rather than optional additions (Perlis & Fihn, 2023).

The language-processing layer contains the LLM responsible for understanding and generating natural-language responses. Depending on the application, the model may summarize clinical documentation, answer patient questions, structure information, or support reasoning-oriented tasks. Clinical summarization provides an example in which model adaptation can produce highly useful outputs (Van Veen et al., 2024).

The clinical intelligence layer translates generated language into application-specific functionality. For example, the system may identify relevant clinical information, assist differential reasoning, organize longitudinal patient information, or forecast health trajectories. However, the system should not equate language generation with clinical authority. Research on reasoning fidelity demonstrates that apparent reasoning quality requires independent evaluation (Bedi et al., 2025).

The human oversight layer introduces clinicians or qualified healthcare professionals into decision-critical workflows. Human review is particularly important for patient-facing medical advice, diagnostic reasoning, and other high-consequence applications. Real-world evaluation of GPT-4 medical advice illustrates why outputs should be assessed under realistic conditions before being relied upon clinically (Jo et al., 2024).

The final layer is governance and monitoring. This encompasses safety evaluation, confidentiality, model monitoring, documentation, risk management, and regulatory compliance. The FDA materials supplied in the literature demonstrate that AI-enabled medical technologies increasingly require consideration within formal regulatory frameworks.

3.3 Application Domains

LLM-powered healthcare systems can support several distinct functional domains.

In clinical summarization, the objective is to reduce information overload by transforming lengthy clinical material into concise representations. The advantage is not merely text compression; a useful system must preserve clinically significant information while minimizing omissions and distortions. Van Veen et al. (2024) provide evidence supporting the feasibility of high-performing adapted LLMs for this purpose.

In patient interaction, LLMs can act as conversational interfaces for health information and communication. The theoretical benefit is improved accessibility because patients can interact through natural language rather than specialized interfaces. Nevertheless, patient-facing systems have a higher safety burden because users may interpret generated content as personalized medical advice. Consequently, evaluation must incorporate real-world queries and clinically relevant safety criteria (Jo et

al., 2024; Johri et al., 2025).

In diagnostic reasoning, an LLM can potentially organize symptoms, generate candidate explanations, or assist clinicians in evaluating diagnostic possibilities. However, diagnostic systems must distinguish between correct conclusions and faithful reasoning. Randomized evaluation of LLM influence on diagnostic reasoning and research on medical reasoning fidelity indicate that this distinction is essential (Goh et al., 2024; Bedi et al., 2025).

In longitudinal health modeling, LLMs may contribute to forecasting patient trajectories and supporting digital-twin-oriented applications. Makarov et al. (2025) demonstrate the potential of this direction. Such systems could theoretically integrate longitudinal information and generate prospective representations of patient health, although the consequences of prediction errors require careful investigation.

3.4 Evaluation Framework

A central methodological principle emerging from the literature is that LLM evaluation should be multidimensional. A proposed evaluation framework should therefore measure:

Task performance: whether the system performs the intended clinical or administrative task accurately.

Clinical validity: whether generated information is medically appropriate and useful.

Reasoning fidelity: whether the reasoning represented by the system is reliable rather than merely persuasive.

Safety: whether the system produces hallucinated, misleading, or potentially harmful information.

Human factors: whether clinicians and patients can appropriately interpret and use system outputs.

Confidentiality: whether sensitive healthcare information is adequately protected.

Generalizability: whether performance remains reliable across different populations, clinical settings, and real-world inputs.

The evaluation illusion described by Agrawal et al. (2025) demonstrates why a single benchmark score cannot adequately establish clinical readiness. Similarly, the patient-interaction framework proposed by Johri et al. (2025) supports the need for application-specific evaluation rather than generic model comparison.

3.5 Hallucination and Risk-Control Mechanisms

Hallucination is particularly problematic because LLM-generated errors may be linguistically fluent and therefore difficult for non-experts to recognize. Hatem et al. (2023) emphasize the need for healthcare professionals to understand and mitigate these risks. A practical system should therefore incorporate source grounding where possible, confidence or uncertainty indicators, structured validation, escalation mechanisms, and human review for high-risk outputs.

The risk-control architecture should be proportional to the consequences of errors. A system summarizing administrative information may require less intensive oversight than a system generating diagnostic recommendations. This risk-sensitive approach is consistent with the broader literature on clinical LLM applications and patient-care risks (Truhn et al., 2025).

4. Results / Findings

The synthesis of the provided literature identifies several consistent findings concerning intelligent healthcare systems powered by LLMs. First, the evidence supports a broadening of healthcare LLM applications from general language generation toward specialized clinical workflows. Clinical summarization, patient interaction, diagnostic reasoning, and longitudinal health forecasting represent distinct application classes with different performance requirements. This indicates that healthcare LLMs should not be evaluated as a single technological category; rather, effectiveness must be determined according to the clinical function performed.

Second, the strongest near-term opportunity appears to be augmentation of information-intensive workflows. Clinical summarization demonstrates that LLMs can provide measurable value when the task is clearly defined and human professionals remain responsible for interpretation (Van Veen et al., 2024). This suggests that systems designed to reduce documentation and information-processing burdens may provide a more manageable deployment pathway than autonomous decision-making systems.

Third, patient-facing applications demonstrate a substantial gap between general conversational capability and clinically trustworthy interaction. Real-world medical-query evaluation indicates that models must be tested against the complexity and variability of actual user-generated questions rather than relying exclusively on curated benchmarks (Jo et al., 2024). The implication is that clinical evaluation should incorporate ambiguity, incomplete information, inappropriate assumptions, and safety-sensitive scenarios.

Fourth, the literature indicates that correctness alone is insufficient for assessing medical reasoning. Research examining diagnostic reasoning and reasoning fidelity shows that an LLM may produce a plausible answer without necessarily providing dependable clinical

reasoning (Goh et al., 2024; Bedi et al., 2025). Therefore, evaluation frameworks should distinguish outcome accuracy from reasoning reliability.

Fifth, hallucination represents a cross-cutting risk affecting multiple application domains. Because generated errors can appear authoritative, conventional accuracy measures may underestimate practical risk. Hallucination mitigation must consequently become an architectural component of healthcare LLM systems rather than an after-the-fact correction strategy (Hatem et al., 2023).

Sixth, the literature reveals an important evaluation problem. Agrawal et al. (2025) demonstrate that evaluation methodology itself can produce misleading impressions of model performance. This finding implies that healthcare LLM research requires standardized, clinically meaningful, and application-specific evaluation protocols.

Seventh, LLM applications are expanding toward predictive and longitudinal systems. The use of LLMs for health trajectory forecasting and digital twins indicates an emerging transition from language-based assistance to dynamic patient modeling (Makarov et al., 2025). However, this domain involves greater uncertainty and potentially greater consequences from erroneous predictions.

Finally, confidentiality and regulatory considerations emerge as fundamental constraints. Clinical data cannot be treated as ordinary language-model input, and deployment requires explicit consideration of privacy, evaluation standards, medical-device governance, and regulatory decision-making. Overall, the findings indicate that the greatest value of healthcare LLMs is likely to emerge from clinically supervised augmentation, supported by rigorous evaluation and risk-aware system design.

5. Discussion

The findings indicate that the development of intelligent healthcare systems should move beyond the question of whether LLMs are capable of performing medical tasks and instead address the more consequential question of under what conditions their capabilities can be trusted. This distinction has both theoretical and practical significance. Theoretically, healthcare LLMs challenge conventional assumptions that model performance can be represented through a single accuracy measure. The evaluation illusion identified by Agrawal et al. (2025) demonstrates that benchmark selection and evaluation design can materially influence perceived capability. Thus, clinical AI research requires a theory of evaluation that accounts for context, risk, uncertainty, and downstream consequences.

From a practical perspective, the literature supports a human-AI collaboration model rather than unrestricted automation. Clinical summarization is a strong example because the LLM can perform a labor-intensive information-transformation task while the clinician remains responsible for interpretation (Van Veen et al., 2024). In contrast, autonomous diagnostic reasoning introduces significantly greater risk because incorrect conclusions may directly influence clinical decisions. Evidence concerning diagnostic reasoning and reasoning fidelity therefore supports retaining human oversight for high-consequence decisions (Goh et al., 2024; Bedi et al., 2025).

Patient interaction presents an especially important trade-off. LLMs can improve accessibility and produce understandable responses, but accessibility can simultaneously increase risk if users overestimate model authority. Real-world medical-query assessment demonstrates the importance of evaluating systems under authentic interaction conditions (Jo et al., 2024). A clinically responsible architecture should therefore distinguish informational assistance from personalized medical decision-making and establish explicit escalation mechanisms.

Hallucination represents a related contradiction. The linguistic fluency that makes LLMs attractive for healthcare communication can also make erroneous information more persuasive. Consequently, conventional user satisfaction cannot serve as an adequate safety indicator. Hallucination mitigation, source verification, uncertainty communication, and human review should be integrated into the system architecture (Hatem et al., 2023).

The emergence of health trajectory forecasting and digital twins expands the research frontier but also increases the importance of longitudinal validation. Predictive systems must account for changing patient states, incomplete observations, and uncertainty rather than presenting forecasts as deterministic outcomes (Makarov et al., 2025). This suggests that future healthcare LLM research should integrate language modeling with robust temporal validation and clinically interpretable uncertainty assessment.

Confidentiality and regulation further demonstrate that technical performance cannot be separated from deployment context. Clinical data protection must be incorporated throughout the system lifecycle, while regulatory considerations should be addressed during system design rather than after development (Perlis & Fihn, 2023). FDA materials concerning AI/ML medical devices and AI-supported regulatory decision-making reinforce the need for governance-oriented development.

Overall, the principal implication is that trustworthy healthcare LLMs require a socio-technical architecture.

Model capability is only one component; clinical validation, human oversight, privacy, safety monitoring, and governance determine whether that capability can translate into responsible healthcare value. Future research should therefore prioritize longitudinal clinical evaluation, standardized safety benchmarks, transparent reasoning assessment, real-world patient interaction testing, and adaptive monitoring after deployment.

6. Conclusion

Large language models are becoming an important component of emerging intelligent healthcare systems, offering capabilities that extend from clinical summarization and patient interaction to diagnostic reasoning and longitudinal health forecasting. The provided literature demonstrates that these applications have considerable potential, but it also establishes that linguistic fluency and benchmark performance cannot by themselves demonstrate clinical reliability.

The central contribution of this review is the identification of a lifecycle-oriented perspective for healthcare LLM deployment. Effective systems should integrate contextual data management, language processing, clinical functionality, human oversight, safety mechanisms, and governance. Such an architecture recognizes that healthcare AI is fundamentally a high-stakes socio-technical system rather than a standalone language-generation technology.

The literature further indicates that evaluation must become more clinically grounded. Real-world patient queries, reasoning fidelity, hallucination risk, confidentiality, and downstream clinical consequences should be incorporated into assessment frameworks. The distinction between task correctness and clinically trustworthy reasoning is particularly important for diagnostic applications.

Future research should focus on standardized evaluation protocols, longitudinal validation, uncertainty-aware prediction, robust hallucination mitigation, privacy-preserving architectures, and clinically meaningful human-AI collaboration. Research into digital twins and health trajectory forecasting should receive particular attention because these applications could substantially extend the role of LLMs from information processing toward predictive healthcare. At the same time, their deployment should remain proportional to demonstrated evidence and risk.

Ultimately, LLM-powered healthcare systems are most appropriately understood as intelligent augmentation technologies. Their long-term value will depend less on replacing healthcare professionals than on reliably extending human capacity while preserving clinical accountability, patient safety, confidentiality, and regulatory responsibility.

References

1. Agrawal M, Chen IY, Gulamali F, Joshi S. The evaluation illusion of large language models in medicine. npj Digit Med. 2025;8:600.
2. Bedi S, Jiang Y, Chung P, Koyejo S, Shah N. Fidelity of medical reasoning in large language models. JAMA Netw Open. 2025;8(8):e2526021.
3. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med. 2023;3:141.
4. Communications Medicine Editorial/Review corpus on patient-facing LLMs and limitations—see Truhn et al. 2025 systematic review for taxonomy of design/output risks. Commun Med. 2025;5:26.
5. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969.
6. Hatem R, Simmons B, Thornton JE. A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. Cureus. 2023;15(9):e44720.
7. Jo E, Zhang C, Thomas MR, et al. Assessing GPT-4's performance in delivering medical advice using realworld user-generated queries. JMIR Med Educ. 2024;10(1):e51282.
8. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med. 2025;31:77-86.
9. Maity S, Bhattacharjee A, Sharma A. Large language models in healthcare and medical applications: opportunities, challenges, and future directions. Bioengineering. 2025;12(6):631.
10. Makarov N, Bordukova M, Quengdaeng P, et al. Large language models forecast patient health trajectories enabling digital twins. npj Digit Med. 2025;8:588.
11. Perlis RH, Fihn SD. Editorial context on evaluation standards and confidentiality when using LLMs with clinical data. JAMA Netw Open. 2023;6(10):e2335924.
12. Truhn D, Cuocolo R, Adams LC, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med. 2025;5:26.

13. US Food and Drug Administration (FDA). Artificial Intelligence and Machine Learning Software as a Medical Device. Updated Mar 25, 2025. fda.gov.