LLM-Based Applications for Healthcare: A Systematic Mapping Study

Ramon Santos Malaquias(✉), Itamir de Morais Barroca Filho, Jean Mário Moreira de Lima, André Morais Gurgel, and Bruna A. O. de Brito

Digital Metropolis Institute, Federal University of Rio Grande do Norte, Natal 59078970, Brazil
{malaquias,itamir.filho,jean.lima}@imd.ufrn.br, andre.gurgel@ufrn.br, bruna.oliveira.124@ufrn.edu.br

Abstract. Large Language Models (LLMs) have been increasingly incorporated into real-world applications, including systems in the healthcare domain. In this context, these models have been applied to tasks such as clinical documentation, information extraction, medical record summarization, and clinical decision support. Despite the rapid growth of the literature on this topic, existing studies often focus on specific applications or techniques, with limited consolidation regarding the characteristics of these systems, the architectural approaches adopted, and the challenges associated with their development. This paper presents a systematic mapping study aimed at identifying the main characteristics of LLM-based healthcare applications, the architectures, patterns, and architectural styles employed in their development, as well as the challenges reported in the literature. The analysis considers recent studies describing LLM applications in healthcare contexts. The results reveal architectural approaches, including multi-stage processing pipelines, knowledge retrieval mechanisms, and integrations with external data sources, as well as common challenges related to hallucinations, data quality, privacy, and computational costs.

Keywords: Large Language Models · LLM · Software Architecture · Literature Review · Healthcare

1 Introduction

Large Language Models (LLMs) have been increasingly incorporated into real-world applications across multiple domains due to their ability to process and generate natural language at scale. Recent surveys report a rapid growth of systems integrating LLMs as core software components, highlighting their adoption in a wide range of application contexts [15,26,48].

In the healthcare domain, this trend has intensified in recent years. Studies report the use of LLMs in tasks such as clinical documentation, information

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 3–19, 2027.
https://doi.org/10.1007/978-3-032-30488-9_1

extraction from electronic health records, medical record summarization, and clinical decision support [23,44]. However, deploying these technologies in clinical environments introduces strict requirements related to patient safety, data privacy, reliability, and regulatory compliance.

In this context, the way LLMs are integrated into system architectures becomes a critical concern. Unlike traditional components, LLMs introduce probabilistic behavior, external dependencies, and challenges related to hallucination mitigation, cost control, and lifecycle governance. Recent studies show that LLM-based applications frequently rely on multi-component architectures that combine retrieval mechanisms, processing pipelines, and validation layers [11,32].

Despite the growing body of research on LLM applications in healthcare, architectural knowledge in this area remains fragmented. Most studies focus on specific applications or techniques, while systematic analyses that consolidate application characteristics, architectural approaches, and development challenges remain limited.

To address this gap, this paper presents a systematic mapping study on LLM-based healthcare applications. Based on the analysis of 43 primary studies from recent literature, the study identifies key characteristics of these applications, the architectures and architectural patterns adopted in their development, and the challenges reported in the literature.

The remainder of this paper is organized as follows. Section 2 describes the research methodology adopted in this study, including the research questions, search strategy, inclusion and exclusion criteria, and the study selection process. Section 3 presents the results of the systematic mapping, discussing the main characteristics of LLM-based healthcare applications, the architectural approaches identified in the literature, and the challenges reported in the analyzed studies. Section 4 discusses the research gaps identified from the analysis of the literature. Finally, Sect. 5 concludes the paper and outlines directions for future research.

2 Methodology

This study was conducted as a Systematic Mapping Study (SMS), following the guidelines proposed by Kitchenham et al. [21]. The goal is to identify and organize evidence on LLM-based applications in healthcare, focusing on their characteristics, architectural approaches, and development challenges.

2.1 Research Questions

The study is guided by the following research questions:

- **RQ1:** What are the main characteristics of LLM-based applications in healthcare?
- **RQ2:** What architectures, patterns, and architectural styles are used in healthcare LLM-based applications?
- **RQ3:** What are the main challenges related to the development of LLM-based applications in healthcare?

2.2 Search Process

The search was conducted in the Scopus database [5], considering titles, abstracts, and keywords. Scopus indexes the main scientific literature databases in the field of Computer Science, such as the ACM Digital Library [28], IEEE Xplore [10], ScienceDirect [38], and SpringerLink [39]. For this reason, it was selected as the primary database to execute the search string defined in the review protocol and to collect the relevant studies.

The following search string was used:

("large language model" OR LLM) AND ("application" OR "case study") AND ("healthcare" OR "health")

2.3 Inclusion and Exclusion Criteria

Studies written in English describing LLM applications in healthcare were considered. The following exclusion criteria were applied:

- **EC1:** Duplicate studies;
- **EC2:** Studies without full-text access;
- **EC3:** Studies without an abstract;
- **EC4:** Studies not addressing any research question;
- **EC5:** Studies not describing LLM applications in healthcare;

2.4 Quality Assessment

The selected studies were evaluated using five quality assessment questions, as follows:

- **QA1:** Is the paper based on research (or is it merely a "lessons learned" report based on expert opinion)?
- **QA2:** Does the paper present a clear statement of the research objectives?
- **QA3:** Does the paper clearly describe the context in which the research was conducted?
- **QA4:** Are the development challenges clearly described?
- **QA5:** Is there a clear statement of the study limitations and conclusions?

Each criterion was scored as 0 (not satisfied), 0.5 (partially satisfied), or 1 (fully satisfied).

2.5 Extraction, Analysis and Study Selection

The selection process consisted of three phases:

- (i) execution of the search string in Scopus, resulting in 3,226 studies;
- (ii) title and abstract screening using the defined criteria, resulting in 206 studies;
- and (iii) quality assessment and full-text analysis, resulting in 43 primary studies.

Figure 1 illustrates the study selection process and the number of studies retained at each stage.

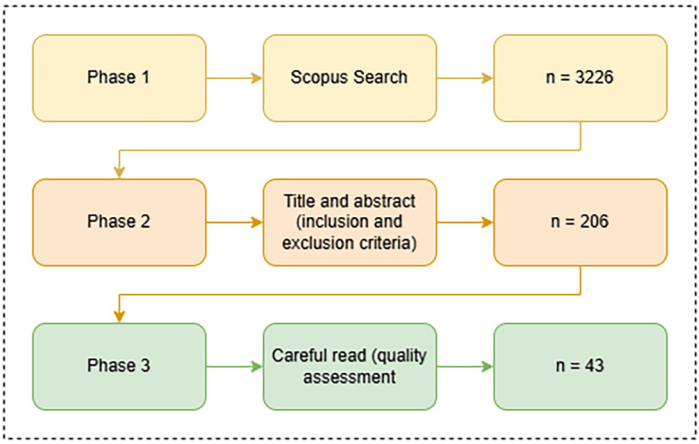

Fig. 1. Study extraction, analysis and selection process.

3 Review Results

Phase 3 of the selection process resulted in 43 primary studies, listed in Table 1. These studies were analyzed in detail to identify evidence related to the research questions defined in the protocol. To support analysis and evidence traceability, an online spreadsheet was used and is available at this link.

Table 1. Studies included in the review and their publication year.

ID	Title	Year
A01	A Multiagent Summarization and Auto-Evaluation Framework for Medical Text: Development and Evaluation Study [7]	2025
A02	A personal health large language model for sleep and fitness coaching [20]	2025
A03	A Smart System for Continuous Sitting Posture Monitoring, Assessment, and Personalized Feedback [35]	2025
A04	Advancing entity recognition in biomedicine via instruction tuning of large language models [19]	2024
A05	AI driven cardiovascular risk prediction using NLP and Large Language Models for personalized medicine in athletes [24]	2025
A06	An innovative X-RAG technique combined with GPT-4o for summarizing medical information from EHR and EMR to assist doctors in clinical decision-making effectively and efficiently [46]	2025
A07	Application of Generative Artificial Intelligence Models for Accurate Prescription Label Identification and Information Retrieval for the Elderly in Northern East of Thailand [43]	2025
A08	Application of generative language models to orthopaedic practice [6]	2024
A09	Benchmarking Large Language Models on Communicative Medical Coaching: A Dataset and a Novel System [16]	2024
A10	CARE-AD: a multi-agent large language model framework for Alzheimer's disease prediction using longitudinal clinical notes [27]	2025

(*continued*)

Table 1. (*continued*)

ID	Title	Year
A11	ChatDiet: Empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework [50]	2024
A12	ChatENT: Augmented Large Language Model for Expert Knowledge Retrieval in Otolaryngology–Head and Neck Surgery [30]	2024
A13	Chat-ePRO: Development and pilot study of an electronic patient-reported outcomes system based on ChatGPT [8]	2024
A14	ChatGPT provides inconsistent risk-stratification of patients with atraumatic chest pain [14]	2024
A15	Clinical risk prediction using language models: Benefits and considerations [1]	2024
A16	ClinicalAgent: Clinical Trial Multi-Agent System with Large Language Model-based Reasoning [53]	2024
A17	Critical care studies using large language models based on electronic healthcare records: A technical note [55]	2025
A18	DALK: Dynamic Co-Augmentation of LLMs and KG to answer Alzheimer's Disease Questions with Scientific Literature [25]	2024
A19	Development and evaluation of a large language model of ophthalmology in Chinese [56]	2024
A20	Development and Evaluation of a Retrieval-Augmented Large Language Model Framework for Ophthalmology [33]	2024
A21	Effectiveness of the Medical Chatbot PROSCA to Inform Patients About Prostate Cancer: Results of a Randomized Controlled Trial [3]	2024
A22	Efficient and Personalized Mobile Health Event Prediction via Small Language Models [47]	2024
A23	Efficient Inference Offloading for Mixture-of-Experts Large Language Models in Internet of Medical Things [52]	2024
A24	Empowering large language models for automated clinical assessment with generation-augmented retrieval and hierarchical chain-of-thought [13]	2025
A25	Enhancing Large Language Model Reliability: Minimizing Hallucinations with Dual Retrieval-Augmented Generation Based on the Latest Diabetes Guidelines [22]	2024
A26	Genomics-Enhanced Cancer Risk Prediction for Personalized LLM-Driven Healthcare Recommender Systems [31]	2025
A27	Identifying signs and symptoms of urinary tract infection from emergency department clinical notes using large language models [17]	2024
A28	Large language models forecast patient health trajectories enabling digital twins [37]	2025
A29	Leveraging long context in retrieval augmented language models for medical question answering [54]	2025
A30	Leveraging Medical Knowledge Graphs Into Large Language Models for Diagnosis Prediction: Design and Application Study [12]	2025
A31	LlamaMTS: Optimizing Metastasis Detection with Llama Instruction Tuning and BERT-Based Ensemble in Italian Clinical Reports [29]	2024
A32	Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data [49]	2024
A33	Multimodal AI and Large Language Models for Orthopantomography Radiology Report Generation and Q&A [9]	2025
A34	PharmaLLM: A Medicine Prescriber Chatbot Exploiting Open-Source Large Language Models [2]	2024
A35	Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation [45]	2025
A36	Scaffolding Empathy: Training Counselors with Simulated Patients and Utterance-level Performance Visualizations [41]	2025
A37	Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications [4]	2024
A38	Systematic evaluation of common natural language processing techniques to codify clinical notes [42]	2024
A39	Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults [51]	2024
A40	Trustworthy AI Psychotherapy: Multi-Agent LLM Workflow for Counseling and Explainable Mental Disorder Diagnosis [36]	2025
A41	Using the Retrieval-Augmented Generation to Improve the Question-Answering System in Human Health Risk Assessment: The Development and Application [34]	2025
A42	Virtual Agents for Alcohol Use Counseling: Exploring LLM-Powered Motivational Interviewing [40]	2024
A43	ZebraMap: A Multimodal Rare Disease Knowledge Map with Automated Data Aggregation & LLM-Enriched Information Extraction Pipeline [18]	2026

3.1 Publication Year Distribution

Figure 2 shows the distribution of the 3,226 studies identified in the initial search. Between 2019 and 2022, the number of publications is minimal, with only one study per year. From 2023 onward, there is a significant increase, with 173 studies, followed by sharp growth in 2024 (886) and 2025 (1,845). Although 2026 was still incomplete at the time of data collection, 318 studies had already been identified.

This increase may be associated with technological and ecosystem shifts in the LLM landscape. In particular, the public release of conversational foundation models such as ChatGPT in late 2022 significantly increased awareness and accessibility of LLM capabilities among researchers and practitioners. At the same time, the availability of commercial LLM APIs (e.g., GPT-based services) enabled rapid prototyping and integration of language models into software systems without requiring extensive machine learning infrastructure.

Another relevant factor is the emergence of open-source foundation models, such as the LLaMA family and other community-driven models, which lowered the barrier for experimentation and allowed researchers to fine-tune models for domain-specific tasks, including biomedical and clinical applications. In parallel, the research community has introduced several medical and biomedical benchmarks designed to evaluate LLM performance in clinical reasoning, question answering, and medical knowledge tasks.

Together, these developments have accelerated the exploration of LLM-based systems in healthcare, contributing to the rapid growth of publications observed from 2023 onward.

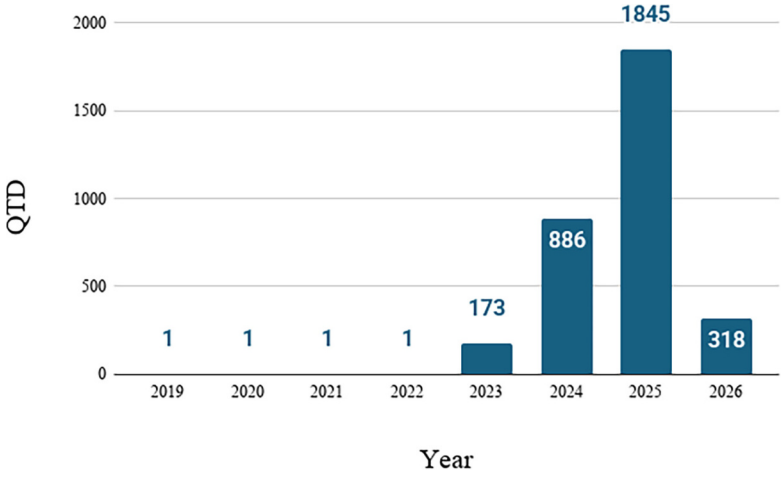

Fig. 2. Distribution of publications per year.

3.2 Application Types

Figure 3 presents the distribution of the selected studies by application type. Question Answering is the most frequent category (19 studies), followed by Information Extraction (13). Other categories include Prediction (8), Conversational Assistant (6), Summary and Recommendation (4 each), as well as Educational Training (3), NER and Clinical Assessment (2 each).

3.3 Model Family Distribution

Figure 4 presents the distribution of model families used in the analyzed studies. The GPT family (OpenAI) is the most frequent (26 studies), followed by LLaMA (16). Other relevant families include Mistral/Mixtral (8), Qwen (6), and both Gemini and biomedical variants of BERT (4 each). Several other model families appear less frequently, indicating the diversity of models used in the analyzed applications.

3.4 Research Question Analysis

This section presents the analysis of the selected studies based on the research questions defined in this systematic mapping. For each question, the extracted data were organized and analyzed to identify patterns, recurring characteristics, and trends in healthcare applications based on LLMs.

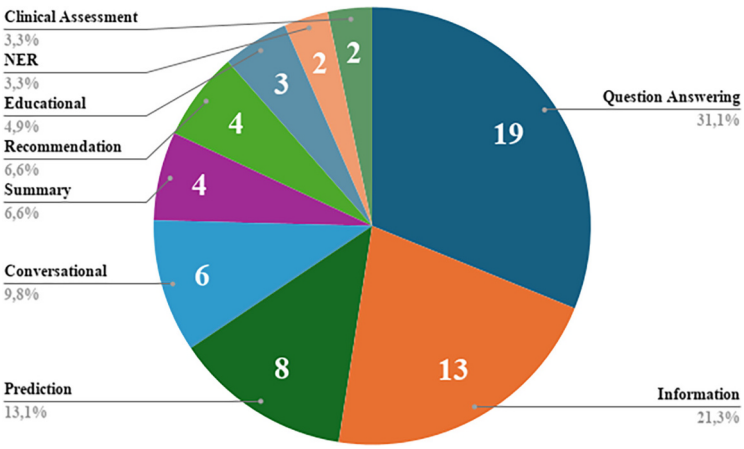

Fig. 3. Distribution of studies by application type.

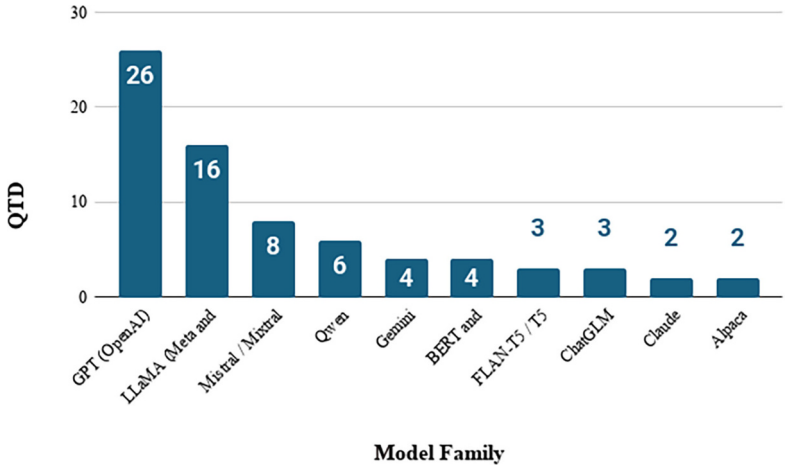

Fig. 4. Distribution of Top 10 model families used in the analyzed studies.

RQ1: What are the main characteristics of LLM-based applications in healthcare?

The analysis of the selected studies suggests a recurring set of functional and non-functional requirements among the analyzed healthcare LLM applications. In general, functional requirements are mainly associated with the processing and interpretation of clinical data, while non-functional requirements focus on response quality, data security, and integration with clinical systems.

One of the most common functional requirements is the ability to process and summarize clinical information. For example, Chen et al. [7] propose a multi-agent framework for automatic medical text summarization, where different agents collaborate to generate and evaluate clinical summaries. Similarly, Wang et al. [46] use a Retrieval-Augmented Generation approach combined with GPT-4o to synthesize information from electronic health records and support clinical decision-making.

Another frequently identified requirement is the automatic extraction of clinical information from unstructured texts. Keloth et al. [19] investigate the use of instruction tuning to improve biomedical entity recognition, while Iscoe et al. [17] use LLMs to identify signs and symptoms of urinary tract infections in emergency department clinical notes.

Applications related to health risk and condition prediction also appear frequently. Acharya et al. [1] investigate clinical risk prediction using medical textual data, while Xu et al. [49] propose Mental-LLM, a model designed to predict mental health conditions from online text data. Similarly, Li et al. [27] introduce the CARE-AD multi-agent framework for predicting Alzheimer's disease using longitudinal clinical notes.

Another group of applications involves conversational interaction with users. Baumgärtner et al. [3] present the PROSCA medical chatbot to provide infor-

mation about prostate cancer, while Yang et al. [51] propose Talk2Care, an LLM-based voice assistant to support communication between healthcare professionals and older adults. Azam et al. [2] introduce PharmaLLM, a clinical chatbot designed to support medication prescription.

Some studies also explore personalized recommendation and health coaching. Khasentino et al. [20] present a language model designed for personalized sleep and fitness coaching, while Yang et al. [50] propose ChatDiet, a system that uses LLMs to provide personalized nutritional recommendations.

Regarding non-functional requirements, the studies emphasize aspects such as clinical accuracy, response reliability, and data security, which are critical in healthcare applications. Other frequently mentioned requirements include scalability, interpretability, hallucination mitigation, and integration with clinical systems, such as electronic health records.

Table 2 summarizes the main non-functional requirements identified and the studies that mention them and Fig. 5 presents a word cloud highlighting the most frequently cited non-functional requirements in the analyzed applications.



Fig. 5. Word cloud of non-functional requirements in healthcare LLM applications.

3.5 RQ2: What Architectures, Patterns, and Architectural Styles Are Used in Healthcare LLM-Based Applications?

The analysis of the selected studies shows that many of the analyzed healthcare LLM applications adopt modular and composite architectures in which language models are integrated with components responsible for knowledge retrieval, clinical data processing, and user interaction. Rather than relying on standalone models, most systems embed LLMs within broader software architectures designed to support clinical workflows and ensure response reliability.

Few recurring architectural patterns were identified. One of the most common is *Retrieval-Augmented Generation* (RAG), where LLMs are combined with

retrieval mechanisms that access external medical sources such as scientific literature, clinical guidelines, or electronic health records. For example, Wang et al. [46] apply this approach to summarize information extracted from EHR and EMR systems to support clinical decision-making.

Another pattern is the use of *multi-agent architectures*, in which multiple LLM-based agents perform specialized roles within the system. Chen et al. [7] propose a multi-agent framework for medical text summarization, while Li et al. [27] apply a similar approach to predict Alzheimer's disease using longitudinal clinical notes. Additional architectural approaches include multimodal architectures, NLP processing pipelines, knowledge graph integration, and domain-specific fine-tuned models. Table 3 summarizes the architectural patterns identified in the analyzed studies.

The analysis also revealed a relationship between application types and the architectural approaches adopted. Table 4 summarizes the dominant architectures observed across different healthcare application categories.

These findings suggest that question answering systems predominantly rely on RAG architectures to improve factual grounding through external medical knowledge retrieval. In contrast, information extraction tasks frequently adopt structured NLP pipelines to process unstructured clinical texts. Conversational assistants typically employ multi-agent or chatbot architectures to manage interaction and dialogue, while prediction-oriented systems tend to rely on domain-specialized or fine-tuned models capable of integrating textual reasoning with structured clinical data.

Table 2. Non-functional requirements identified in healthcare LLM applications and the studies that mention them.

Non-functional requirement	Studies
Clinical accuracy	A01, A05, A06, A14, A15, A24, A28, A33, A36
Response reliability	A01, A14, A24, A25, A28, A36
Scalability	A03, A22, A23, A31
Data privacy and security	A08, A13, A21, A34
Interpretability/explainability	A15, A24, A30, A38
Hallucination mitigation	A24, A25, A28
Integration with clinical systems (EHR/EMR)	A06, A27, A32
Response personalization	A02, A11, A26
Model robustness	A24, A25, A28
Computational efficiency	A22, A23
Multimodal support	A03, A33
Interoperability	A06, A32

Table 3. Architectural patterns identified in healthcare LLM applications and the studies that adopt them.

Architectural Pattern	Studies
Retrieval-Augmented Generation (RAG)	A06, A07, A12, A20, A25, A35, A37, A41, A43
Multi-agent architecture	A01, A09, A10, A16, A36, A40, A42
Multimodal architecture	A02, A07, A33, A34, A43
NLP pipeline/text processing	A04, A14, A21, A27, A38, A43
Conversational/chatbot architecture	A13, A21, A42
Specialized/fine-tuned models	A02, A15, A19
LLM + Knowledge Graph integration	A18, A30
Mixture-of-Experts (MoE)	A23, A26

Across several analyzed studies, similar architectural compositions can be observed in which user requests are processed through input processing or prompt engineering layers, followed by retrieval mechanisms that access clinical data sources. The retrieved context is then provided to the LLM reasoning component, whose output is typically validated through guardrails or post-processing layers before producing the final clinical response.

Such architectural composition is particularly important in healthcare systems due to the critical nature of their non-functional requirements. Applications in this domain must satisfy strict requirements related to clinical accuracy, reliability, explainability, privacy, and regulatory compliance, while ensuring safe interaction with sensitive patient data and clinical workflows. Consequently, domain-specialized architectures that incorporate mechanisms such as knowledge grounding, validation layers, auditability, and human oversight are essential to mitigate risks such as hallucinations and unsafe clinical recommendations.

Table 4. Relationship between healthcare application types and dominant architectural approaches.

Application Type	Dominant Architecture
Question Answering	Retrieval-Augmented Generation (RAG) architectures combined with external medical knowledge retrieval mechanisms
Information Extraction/NER	NLP processing pipelines integrating LLM reasoning with text preprocessing, entity extraction, and classification modules
Conversational Assistants	Multi-agent or chatbot-oriented architectures combining dialogue management and LLM-based interaction modules
Prediction/Risk Assessment	Fine-tuned domain-specific models or hybrid architectures combining LLM reasoning with structured clinical data
Recommendation and Coaching	Personalized architectures integrating LLM reasoning with contextual patient data and knowledge bases
Summarization	RAG-based pipelines and multi-agent architectures responsible for generating and validating medical summaries

RQ3: What are the main challenges related to the development of LLM-based applications in healthcare?
The analysis of the selected studies revealed a recurring set of challenges associated with the development of applications based on LLMs in the healthcare domain and Fig. 6 presents this challenges. In general, these challenges are concentrated in five main dimensions: reliability of generated responses, availability and quality of clinical data, privacy and regulatory concerns, computational limitations, and difficulties related to integrating heterogeneous data and systems. Figure 6 presents a word cloud summarizing the challenges discussed in the analyzed studies.

The most frequently reported challenge concerns the reliability of model-generated outputs. Several studies report issues such as hallucinations, inconsistencies, and low accuracy in clinical tasks. Chen et al. [7], for example, highlight this issue in medical text summarization systems, while Heston and Lewis [14] observe inconsistencies in risk stratification tasks performed by ChatGPT.

Another relevant group of challenges relates to the availability and quality of clinical data. Many studies report scarcity of annotated datasets, the high cost of expert labeling, and the presence of noise or bias in clinical records. Keloth et al. [19] and Li et al. [24] show that the limited availability of high-quality biomedical datasets restricts model training and generalization.



Fig. 6. Word cloud of challenges in the development of LLM-based healthcare applications.

The literature also emphasizes challenges related to privacy, ethics, and regulatory compliance, particularly in applications that use electronic health records or other sensitive patient data [8,27,33]. In addition, issues such as computational cost, latency, and infrastructure requirements are frequently reported, along with difficulties in the integration of heterogeneous and multimodal data sources [2,22,31,47]. Finally, many studies point to limitations associated with clinical reasoning, domain adaptation, and rigorous evaluation of solutions [25,50]. Overall, these findings indicate that progress in this area depends not

only on improvements in foundation models, but also on more robust architectures, better data ecosystems, and systematic mechanisms for evaluation and human oversight.

Taken together, the results of RQs 1, 2, and 3 show that LLM-based healthcare applications already cover multiple tasks and usage contexts, while adopting composite architectures that combine LLMs with retrieval mechanisms, external knowledge bases, specialized agents, and integration with clinical systems. At the same time, important challenges remain regarding response reliability, data quality and availability, privacy, regulation, computational costs, and interoperability. These findings suggest that advances in this area may depend not only on improvements in foundation models but also on the design of robust software architectures.

4 Identified Research Gaps

The analysis of the selected studies revealed several gaps in the literature on healthcare applications based on LLMs. Overall, these gaps relate to architectural aspects, engineering and evaluation practices, and challenges associated with real-world adoption.

- **Lack of reference architectures.** Most applications described in the literature adopt ad hoc architectures designed for specific tasks. Reference architectures that systematically organize recurring components such as LLMs, retrieval mechanisms, multimodal pipelines, and clinical knowledge bases are still scarce.
- **Limited discussion of architectural trade-offs.** Although quality attributes such as clinical accuracy, latency, computational cost, privacy, and explainability are frequently mentioned, few studies explicitly discuss how architectural decisions balance these competing requirements.
- **Limited architectural governance mechanisms.** Techniques such as RAG, specialized agents, and knowledge base integration are commonly used, but rarely accompanied by comprehensive architectural strategies for monitoring, auditing, fallback mechanisms, and lifecycle management of LLM components.
- **Limited integration with existing clinical systems.** Although the need for integration with electronic health records and hospital systems is frequently acknowledged, few studies propose robust architectural approaches to ensure interoperability at scale.

5 Conclusions and Future Work

This study presented a systematic mapping of the literature on healthcare applications based on LLMs, aiming to identify their main characteristics, architectural approaches, and development challenges.

The results show that LLM-based applications are being explored in several contexts, including clinical decision support, medical data analysis, interaction with patients and healthcare professionals, and processing of unstructured clinical data. These applications emphasize requirements such as clinical accuracy, reliability, personalization, and integration with multiple medical data sources.

From an architectural perspective, the analyzed studies frequently adopt retrieval mechanisms, natural language processing pipelines, and modular architectures that combine LLM reasoning with external knowledge sources and specialized processing components.

Despite recent progress, several challenges remain, including hallucinations, limited availability and quality of clinical data, privacy and security concerns, computational cost, and the need for human supervision. These challenges highlight that the adoption of LLM-based systems in healthcare requires architectures capable of ensuring reliability, explainability, and safe interaction with clinical workflows.

The results also provide relevant implications for research and practice. From a research perspective, the findings indicate the need for reference architectures for healthcare LLM systems, more standardized evaluation protocols, and architectural strategies to mitigate risks such as hallucinations, bias, and privacy concerns. From a practical perspective, the development of these systems should consider early integration with clinical infrastructures, explicit treatment of non-functional requirements, and validation in real-world healthcare environments.

As future work, we intend to investigate reference architectures for healthcare LLM systems and explore architectural strategies to mitigate issues such as hallucinations, bias, privacy, and integration with clinical systems. In addition, further studies involving validation in real clinical environments are essential to better understand the practical impact and reliability of these applications.

Acknowledgments. This study was funded by its authors.

Disclosure of Interests. The authors have no competing interests to declare that they are relevant to the content of this article.

References

1. Acharya, A., et al.: Clinical risk prediction using language models: benefits and considerations. J. Am. Med. Inf. Assoc. **31**(9), 1856–1864 (2024)
2. Azam, A., Naz, Z., Khan, M.U.G.: Pharmallm: a medicine prescriber chatbot exploiting open-source large language models. Human-Centric Intell. Syst. **4**(4), 527–544 (2024)
3. Baumgärtner, K., et al.: Effectiveness of the medical chatbot prosca to inform patients about prostate cancer: results of a randomized controlled trial. Eur. Urol. Open Sci. **69**, 80–88 (2024)
4. Bora, A., Cuayáhuitl, H.: Systematic analysis of retrieval-augmented generation-based LLMs for medical chatbot applications. Mach. Learn. Knowl. Extract. **6**(4), 2355–2374 (2024)

5. Boyle, F., Sherman, D.: Scopus^{TM}: the product and its development. Ser. Libr. **49**(3), 147–153 (2006)
6. Caterson, J., Ambler, O., Cereceda-Monteoliva, N., Horner, M., Jones, A., Poacher, A.T.: Application of generative language models to orthopaedic practice. BMJ Open **14**(3), e076484 (2024)
7. Chen, Y., Wen, B., Zulkernine, F.: A multiagent summarization and auto-evaluation framework for medical text: development and evaluation study. JMIR AI **4**, e75932 (2025)
8. Chen, Z., Wang, Q., Sun, Y., Cai, H., Lu, X.: Chat-epro: development and pilot study of an electronic patient-reported outcomes system based on chatgpt. J. Biomed. Inf. **154**, 104651 (2024)
9. Dasanayaka, C., Dandeniya, K., Dissanayake, M.B., Gunasena, C., Jayasinghe, R.: Multimodal ai and large language models for orthopantomography radiology report generation and q&a. Appl. Syst. Innov. **8**(2), 39 (2025)
10. Durniak, A.: Welcome to IEEE xplore. IEEE Power Eng. Rev. **20**(11), 12 (2000)
11. Esposito, M., et al.: Generative AI for software architecture: applications, challenges, and future directions. J. Syst. Softw. (2026). https://doi.org/10.1016/j.jss.2025.112607, https://www.sciencedirect.com/science/article/abs/pii/S0164121225002766
12. Gao, Y., et al.: Leveraging medical knowledge graphs into large language models for diagnosis prediction: design and application study. JMIR AI **4**, e58670 (2025)
13. Gu, Z., Jia, W., Piccardi, M., Yu, P.: Empowering large language models for automated clinical assessment with generation-augmented retrieval and hierarchical chain-of-thought. Artif. Intell. Med. **162**, 103078 (2025)
14. Heston, T.F., Lewis, L.M.: Chatgpt provides inconsistent risk-stratification of patients with atraumatic chest pain. PLoS ONE **19**(4), e0301854 (2024)
15. Hou, X., et al.: Large language models for software engineering: a systematic literature review. ACM Trans. Softw. Eng. Methodol. (2024). https://doi.org/10.1145/3695988. https://dl.acm.org/doi/10.1145/3695988
16. Huang, H., Wang, S., Liu, H., Wang, H., Wang, Y.: Benchmarking large language models on communicative medical coaching: a dataset and a novel system. In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 1624–1637 (2024)
17. Iscoe, M., et al.: Identifying signs and symptoms of urinary tract infection from emergency department clinical notes using large language models. Acad. Emerg. Med. **31**(6), 599–610 (2024)
18. Islam, M.S., Jamal, A., Alkhathlan, A.: Zebramap: a multimodal rare disease knowledge map with automated data aggregation & LLM-enriched information extraction pipeline. Diagnostics **16**(1), 107 (2026)
19. Keloth, V.K., et al.: Advancing entity recognition in biomedicine via instruction tuning of large language models. Bioinformatics **40**(4), btae163 (2024)
20. Khasentino, J., et al.: A personal health large language model for sleep and fitness coaching. Nat. Med. **31**(10), 3394–3403 (2025)
21. Kitchenham, B., Brereton, O.P., Budgen, D., Turner, M., Bailey, J., Linkman, S.: Systematic literature reviews in software engineering-a systematic literature review. Inf. Softw. Technol. **51**(1), 7–15 (2009)
22. Lee, J., Cha, H., Hwangbo, Y., Cheon, W.: Enhancing large language model reliability: minimizing hallucinations with dual retrieval-augmented generation based on the latest diabetes guidelines. J. Pers. Med. **14**(12), 1131 (2024)

23. Leiser, F., Guse, R., Sunyaev, A.: Large language model architectures in health care: scoping review of research perspectives. J. Med. Internet Res. (2025). https://doi.org/10.2196/70315. https://www.jmir.org/2025/1/e70315
24. Li, A., Wang, Y., Chen, H.: AI driven cardiovascular risk prediction using NLP and large language models for personalized medicine in athletes. SLAS Technol. **32**, 10028 (2025)
25. Li, D., et al.: Dalk: dynamic co-augmentation of LLMs and kg to answer Alzheimer's disease questions with scientific literature. arxiv 2024. arXiv preprint arXiv:2405.04819
26. Li, J., Gao, J., et al.: Fundamental capabilities and applications of large language models: a survey. ACM Comput. Surv. (2024). https://doi.org/10.1145/3735632. https://dl.acm.org/doi/10.1145/3735632
27. Li, R., Wang, X., Berlowitz, D., Mez, J., Lin, H., Yu, H.: Care-ad: a multi-agent large language model framework for Alzheimer's disease prediction using longitudinal clinical notes. npj Dig. Med. **8**(1), 541 (2025)
28. Library, A.D.: ACM digital library home page (2021). https://dl.acm.org/
29. Lilli, L., et al.: Llamamts: optimizing metastasis detection with llama instruction tuning and bert-based ensemble in Italian clinical reports. In: Proceedings of the 6th Clinical Natural Language Processing Workshop, pp. 162–171 (2024)
30. Long, C., et al.: Chatent: augmented large language model for expert knowledge retrieval in otolaryngology-head and neck surgery. medrxiv 2023 (2023)
31. Lu, K., et al.: Genomics-enhanced cancer risk prediction for personalized LLM-driven healthcare recommender systems. ACM Trans. Inf. Syst. **43**(6), 1–30 (2025)
32. Lu, Q., Zhu, L., Xu, X., Liu, Y., Xing, Z., Whittle, J.: A taxonomy of foundation model based systems through the lens of software architecture. In: Proceedings of the IEEE/ACM International Conference on AI Engineering – Software Engineering for AI (2024). https://doi.org/10.1145/3644815.3644956. https://dl.acm.org/doi/10.1145/3644815.3644956
33. Luo, M.J., et al.: Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. JAMA Ophthalmol. **142**(9), 798–805 (2024)
34. Meng, W., Li, Y., Chen, L., Dong, Z.: Using the retrieval-augmented generation to improve the question-answering system in human health risk assessment: the development and application. Electronics **14**(2), 386 (2025)
35. Odesola, D.F., Kulon, J., Verghese, S., Partlow, A., Gibson, C.: A smart system for continuous sitting posture monitoring, assessment, and personalized feedback. Sensors **25**(18), 5610 (2025)
36. Ozgun, M.C., Pei, J., Hindriks, K., Donatelli, L., Liu, Q., Wang, J.: Trustworthy AI psychotherapy: multi-agent LLM workflow for counseling and explainable mental disorder diagnosis. In: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, pp. 2263–2272 (2025)
37. Rodriguez-Esteban, R., Schmich, F., Menden, M.P.: Large language models forecast patient health trajectories enabling digital twins (2025)
38. ScienceDirect: Science direct home page (2021). https://www.sciencedirect.com/
39. of Springer Nature, S.N.S.A.P.: Springer link home page (2021). https://link.springer.com/
40. Steenstra, I., Nouraei, F., Arjmand, M., Bickmore, T.: Virtual agents for alcohol use counseling: exploring LLM-powered motivational interviewing. In: Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents, pp. 1–10 (2024)

41. Steenstra, I., Nouraei, F., Bickmore, T.: Scaffolding empathy: training counselors with simulated patients and utterance-level performance visualizations. In: Proceedings of the 2025 Chi Conference on Human Factors in Computing Systems, pp. 1–22 (2025)
42. Tavabi, N., Singh, M., Pruneski, J., Kiapour, A.: Systematic evaluation of common natural language processing techniques to codify clinical notes. medrxiv. Preprint posted online October **10** (2022)
43. Thetbanthad, P., Sathanarugsawait, B., Praneetpolgrang, P.: Application of generative artificial intelligence models for accurate prescription label identification and information retrieval for the elderly in northern east of thailand. J. Imaging **11**(1), 11 (2025)
44. Thirunavukarasu, A.J., Ting, D.S.J., et al.: Large language models in medicine. Nat. Med. (2023). https://doi.org/10.1038/s41591-023-02448-8. https://www.nature.com/articles/s41591-023-02448-8
45. Wada, A., et al.: Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation. NPJ Dig. Med. **8**(1), 395 (2025)
46. Wang, J.F., et al.: An innovative x-rag technique combined with gpt-4o for summarizing medical information from EHR and EMR to assist doctors in clinical decision-making effectively and efficiently. Health Inf. J. **31**(3), 14604582251381232 (2025)
47. Wang, X., Dang, T., Kostakos, V., Jia, H.: Efficient and personalized mobile health event prediction via small language models. In: Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, pp. 2353–2358 (2024)
48. Xu, M., et al.: Resource-efficient algorithms and systems of large language models: a survey. ACM Comput. Surv. (2025). https://doi.org/10.1145/3706418. https://dl.acm.org/doi/10.1145/3706418
49. Xu, X., et al.: Mental-LLM: leveraging large language models for mental health prediction via online text data. Proc. ACM Interact. Mob. Wearable Ubiq. Technol. **8**(1), 1–32 (2024)
50. Yang, Z., et al.: Chatdiet: empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework. Smart Health **32**, 100465 (2024). https://doi.org/10.1016/j.smhl.2024.100465
51. Yang, Z., et al.: Talk2care: an LLM-based voice assistant for communication between healthcare providers and older adults. Proc. ACM Interact. Mob. Wearable Ubiq. Technol. **8**(2), 1–35 (2024)
52. Yuan, X., Kong, W., Luo, Z., Xu, M.: Efficient inference offloading for mixture-of-experts large language models in internet of medical things. Electronics **13**(11), 2077 (2024)
53. Yue, L., Xing, S., Chen, J., Fu, T.: Clinicalagent: clinical trial multi-agent system with large language model-based reasoning. In: Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 1–10 (2024)
54. Zhang, G., et al.: Leveraging long context in retrieval augmented language models for medical question answering. npj Dig. Med. **8**(1), 239 (2025)
55. Zhang, Z., Ni, H.: Critical care studies using large language models based on electronic healthcare records: a technical note. J. Intensive Med. **5**(02), 137–150 (2025)
56. Zheng, C., et al.: Development and evaluation of a large language model of ophthalmology in Chinese. Br. J. Ophthalmol. **108**(10), 1390–1397 (2024)