SLJoAS Sri Lankan Journal of Applied Sciences

Review Article

https://sljoas.sljol.info/

DOI: https://doi.org/10.4038/sljas2.v5i1.21

Large Language Models and Medical AI Systems for Healthcare Diagnosis: A Systematic Review

M.U.K. Gunawardhna [a*], Pirunthavi Wijikumar [a], D.M.O.K. Weerasinghe[b]

[a]Department of Information and Communication Technology, University of Vavuniya, Vavuniya, Sri Lanka, 43000

[b] University of Peradeniya,Kandy,Sri Lanka,20000

*** Corresponding author email address**: udikagunawardhna@gmail.com

Received: 07 March 2026 Accepted: 30 April 2026 Published: 30 August 2026

Abstract

The rapid growth of artificial intelligence systems (AI systems) has increased interest in the use of patient care and clinical decision-making processes. There is some uncertainty regarding their reliability and safety in clinical practice. A more detailed systematic review of literature examining LLMs applied to healthcare diagnosis was conducted. A PRISMA-based systematic review has been carried out of relevant literature published in the major databases for the years 2022–2025. Key findings include a growing trend to develop multimodal models based on diverse input modalities, combining LLM models with other models as part of clinical workflows. The Usage of complementary methodologies such as retrieval-augmented generation, knowledge graphs, and federated learning is highly expanding, particularly in enhancing the efficiency and accuracy of clinical decision-making processes. Significant challenges such as hallucinations, bias, prompt sensitivity, limited explainability, and inadequate clinical validation continue to pose major obstacles. Although promising, LLM-based systems are not yet reliable enough for autonomous medical diagnosis. Overall, this review contains multiple recommendations for future research in many areas (e.g., LLMs) to ensure a high level of safety, transparency, and clinical applicability for LLMs and other AI/ML-related technologies and devices.

Keywords: Large Language Models, Healthcare Diagnosis, Clinical Decision Support, Multimodal AI, Medical Ethics

1. Introduction

This paper provides a comprehensive review of large language models (LLMs) and medical AI systems for healthcare diagnosis. Analysing clinical text, images, and physiological signals have resulted in AI taking increasing precedence as an integral component of modern healthcare. However, these newer technologies have left many questions surrounding the accuracy, reliability, and clinical safety of AI models unresolved. The available body of information on the use of LLMs for diagnosis is fairly large, yet fragmented. Most of the studies are based on small datasets that consist of heterogeneous, differ considerably in the evaluations performed (i.e., make comparative assessments impossible), and make it difficult to determine whether or not these kinds of models will continue to yield comparable predictive abilities in the commercial sector. Across the existing literature, six major themes emerge:

- The issue of generalizability in LLMs and multimodal systems used as clinical decision support tools (CDSTs) in various clinical environments represents a significant knowledge gap.
- Some LLMs perform very well when evaluated in a laboratory setting; however, as seen in several studies of LLMs in real-world conditions, when confronted with (a) distribution shifts, (b) background variability, and (c) real-world noise, their performance drops significantly.
- With all these issues considered, continued questions about the hallucination, bias, uncalibrated predictions, and the need for additional external validation remain concerning LLMs' readiness for commercialization and clinical use.
- A further challenge in evaluating the state of artificial intelligence in the clinical environment is that there is not yet a standard evaluation framework or a benchmarking protocol to evaluate the ability of AI in the health care area.
- The lack of uniform evaluation criteria related to accuracy, safety, robustness, fairness, and explainability leaves the evaluation of AI systems incomplete and inconsistent across jurisdictions.
- In addition, the ethical, legal, and social implications of AI (ELSI) have not received adequate attention, particularly with respect to accountability, transparency, and patient safety.

In response to these issues, this review synthesizes the literature published from 2022 to 2025. A total of 200 studies

were identified from databases such as Google Scholar, ScienceDirect, ResearchGate, and IEEE Xplore; 39 of these studies met our inclusion criteria. We present a PRISMA based selection process and a taxonomy of medical AI models, as well as a proposed evaluation framework and In response to these issues, this review synthesizes the literature published from 2022 to 2025. A total of 200 studies were identified from databases such as Google Scholar, ScienceDirect, ResearchGate, and IEEE Xplore; 39 of these studies met our inclusion criteria. We present a PRISMA-based selection process and a taxonomy of medical AI models, as well as a proposed evaluation framework and practical examples of how to facilitate the safe and responsible development of clinical AI systems.

2. Background

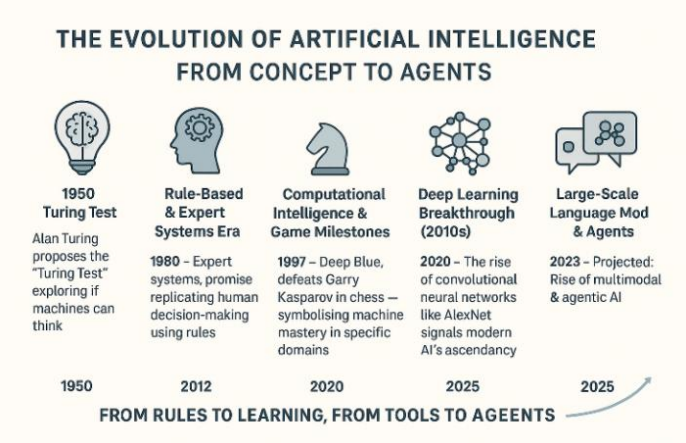

Fig. 1. History and Evolution of Artificial Intelligence

The history of AI charts the progression from the Turing Test in the 1950s to the Dartmouth Conference in 1956, which gave AI proper academia and the coining of the term. The 1980s brought about the popularity of expert systems and in 1997, the computer chess player Deep Blue from IBM broke new ground by defeating world champion Garry Kasparov.

In 2012, deep learning revolutionized computer vision and pattern recognition methods. In the 2020s, large language models such as GPT-3 and GPT-4 pushed fellows in conversational AI, and by 2025 [8], it is speculated that we will see multimodal, agentic AI (see Fig.1). AI systems that utilize machine learning, deep learning, and natural language processing (NLP) can study and analyse extensive amounts of clinical and biomedical data and identify patterns that may not be detected by human experts [9].There are multiple theoretical foundations that provide context for how AI, specifically larger language model medical AI, learns and reasons in order to assist clinicians with clinical decision-making ([1] - [5]). Artificial Intelligence (AI) applications in the medical industry have had a profound impact on how patients are managed, processed for diagnosis, and treated. Machine learning-related advancements along with leading edge tools for analytics and automation created an enormous growth rate in AI applications in U.S. healthcare during the time period spanning from 2014 through 2025. In 2014, the estimated value of AI software solutions was about $544 million (see Fig. 2) [10].

The primary technology underpins machine learning and deep learning, which allows models to learn automatically by finding and utilizing patterns when presented with large amounts of medical data. Common algorithms include transformer based LLMs as one example of using natural language processing (NLP); convolutional neural network (CNN) algorithms as one example of using deep neural networks to process/interpret images; and multimodal fusion models that can process, understand, and integrate data types (textual, visual, and sensor) ([6], [7]).

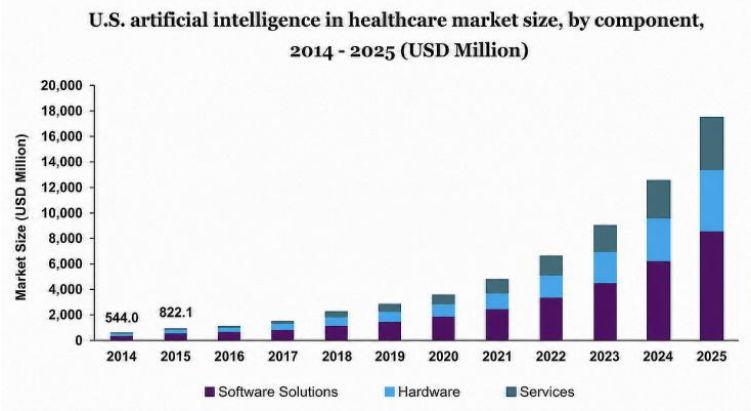

Fig. 2. U.S. Artificial Intelligence in Healthcare Market Size by Component, 2014–2025 (USD Million)

In addition to foundational algorithms, many domain-specific terms, such as clinical ontologies and structured data from electronic health records, as well as diagnostic coding standards, help AI models identify, understand, and interpret medical contexts accurately. Together these theoretical foundations will provide an essential component in developing diagnostic support tools that can provide sophisticated reasoning, prediction, and clinical decision support. The basic, fundamental characteristics that describe how medical AI systems function also provide the structure for assessing their capabilities, weaknesses, and roles within the context of clinical practice.

3. Materials and Methods

This study adopted a structured review approach based on PRISMA guidelines to promote transparency, repeatability, and methodological integrity in the analysis of changes in Medical Artificial Intelligence (MAI)(Fig.3). Using PRISMA and combined elements of a scoping review with a Systematic Review (SR) methodology, this article assesses the rapid advancement and application of Large Language Models (LLMs) and multimodal artificial intelligence systems (AI) in the healthcare industry.

Databases used to perform the literature review include Google Scholar, IEEE Xplore, ScienceDirect, PubMed, Scopus, and ACM Digital Library; all publications were

published in English from 2022 until 2025. Databases were designed with specific Boolean queries that corresponded to the specific indexing options provided by each platform to produce an explicit and reproducible search strategy. All of the Boolean expressions and exact keywords utilized to perform a search on each database are shown in Table 1. These search queries were created to find articles discussing the topics listed above. In the interest of methodological transparency and consistency while determining appropriate information sources, the same search queries can be utilized to perform searches in Google Scholar, IEEE Xplore, PubMed, ScienceDirect, Scopus, ACM Digital Library, and ResearchGate.

Table 1
Databases and Boolean Search Queries Used in the Review.

	Materials and Methods
Database	**Keywords / Boolean Search Queries**
Google Scholar	("medical artificial intelligence" AND "large language models") OR ("clinical diagnosis" AND "AI")
IEEE Xplore	("LLM" AND "healthcare") OR ("transformer model" AND "medical applications")
PubMed	("clinical diagnosis" AND "AI") OR ("NLP" AND "medicine") OR ("multimodal model" AND "healthcare")
ScienceDirect	("AI diagnosis" AND "patient safety") OR ("deep learning" AND "medical imaging")
Scopus	("artificial intelligence" AND "clinical decision support") OR ("healthcare LLM" AND "evaluation")
ACM Digital Library	("machine learning" AND "healthcare") OR ("multimodal AI" AND "biomedical data")
Research Gate	("medical AI systems") AND ("robustness" OR "hallucination" OR "bias")

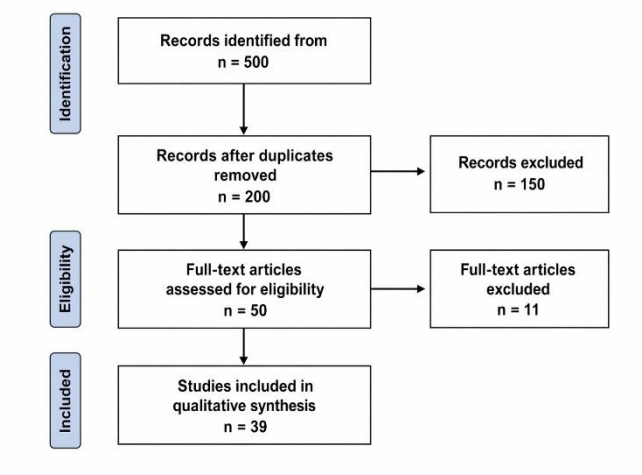

Fig. 3 .PRISMA-based study selection methodology

A structured quality-assessment methodology was developed to evaluate whether each article met the selection criteria based on four critical components: The articles were evaluated for risk of bias, relevance to the objectives of the systematic review, methodological rigor, quality of data, and reproducibility. Each article was rated, and borderline articles were evaluated again by a second evaluator for consistency. A flowchart was created using PRISMA to summarize the entire article selection process, from identification of retrieved articles to initial screening and removing duplicates, then to full-text review for determining eligibility, ending with the final selection of 39 articles. This flowchart depicts the sequential order in which the article-review process occurs (refer to Fig.2).

4. Results and Discussion

4.1 Study Selection Results

The procedure for selecting the studies was based on the guidelines established by PRISMA. There were 500 unique records found across various databases to begin with. After duplicates were removed and articles deemed irrelevant by the authors had been deleted, 200 records were still available for further evaluation. All remaining records underwent a full-text review to assess whether or not each was eligible for inclusion; 39 studies met the criteria necessary for acceptance into the final qualitative evaluation. Fig. 3 represents the results of the entire review process from start to finish.

4.2 Diagnostic Performance of LLMs

Disease diagnosis is heavily influenced by AI technology, as many different types of machine learning and deep learning algorithms have been created that can be used for a variety of different types of diseases. For example, Table 2 contains the various examples of using AI-based approaches that have been developed for diagnosing a variety of diseases.

Table 2
Diagnosis of disease using artificial intelligence techniques.

	Diagnostic Performance of LLMs		
Disease Type	**AI Technique**	**Best Model**	**Accuracy**
Heart Disease	Machine Learning	Support Vector Machine (SVM)	95%
Breast Cancer	Deep Learning	CNN + SVM Hybrid	99.5%
Lung Cancer	Deep CNN	Not identified one mode	84%
Alzheimer's Disease	Deep Learning	EfficientNetB0	92.98%
Diabetes	Random Forest	Not identified one model	91%

The research that has been published recently concerning the capabilities of Large Language Models (LLMs) to assist clinicians with diagnosing illnesses has shown great strides made both in how well these tools are performing for clinicians but also highlights the obstacles which will arise as LLMs start to receive clinical uptake in a real-life setting. [3] found that although LLMs outperformed trained physicians at accurately diagnosing illness using an objective assessment, the results do not suggest that there is

any benefit to physicians because utilizing LLMs for diagnosis does not lead to a statistically significant increase in human reasoning. [4] have been conducted demonstrating limitations related to how LLMs integrate information, perform multi-modal reasoning, and highlight ethical implications of LLM use along with providing future direction for LLMs, including continued improvements in the ability of LLMs to learn continuously as well as developing generalist medical models. LLMs also frequently lack sensitivity to clinically relevant information which brings into question the reliability of LLMs when used to diagnose patients [5].

The ongoing evaluation of GPT-4 is demonstrating that it is necessary to evaluate the diagnostic ability of LLMs versus physicians on a blind comparison basis. [6] published a technical review which showed that recent advancements in innovation related to multi-query attention (when more than one query (question) is asked) and mixture-of-experts scaling and alignment tuning has resulted in significant improvements in the ability of LLMs to perform as an efficient diagnostic tool. Additionally, domain-specific AI-CAD platforms, as described by [7], further demonstrate that AI-enhanced tools will allow non-specialists to diagnose effectively across all aspects of ophthalmology.

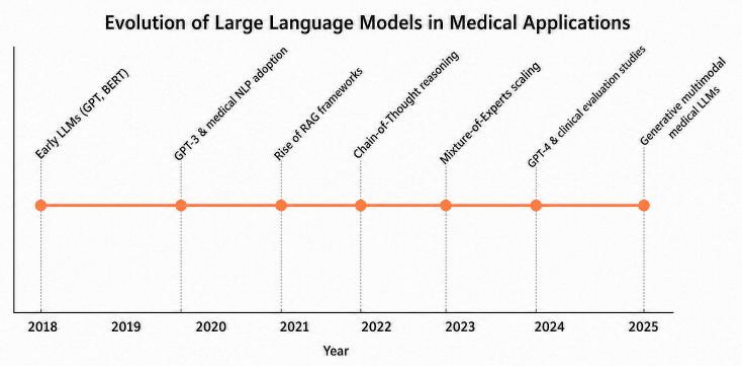

Fig. 4. Evolution of Large Language Models in Medical Applications

Further, [5] demonstrate the importance of ethical and regulatory implications associated with LLM-enhanced diagnostic processes. Collectively, the evolution of LLMs demonstrates movement towards developing a multimodal, reliable, clinically integrated AI-enhanced diagnostic system. The introduction of domain-oriented multimodal medical LLMs could increase the roles of AI in medicine by integrating imaging, text, and physiological information, thus adding another layer of versatility to the diagnostic process for 2025 (refer to Fig. 4).

4.3 Risks and Limitations

Recent research from 2023 through 2025 indicates that Large Language Models (LLMs) are currently not safe to support autonomous medical diagnoses based on key limitations (refer to Table 3). The most important limitation of the LLMs is their lack of true clinical reasoning ability, as they rely only on the identification of statistical correlations and not on the understanding or appreciation of physiological and contextual relationships that govern medical practice. This leads to incorrect diagnoses, incomplete clinical information, and inconsistent clinical outputs, all of which become even more erratic with any rephrasing of the prompt to the model or a change of systems to an updated model of the LLM. Diagnostic accuracy as assessed in empirical studies is less than that of any trained clinician. Most LLMs do not accurately follow evidence-based medical practice guidelines, frequently misinterpret laboratory test results, and frequently reverse their performance when provided with clearly defined clinical history or additional context. This demonstrates a major flaw of the LLM where they lack robustness and consistency. In addition, LLMs are particularly susceptible to distortion by manipulation and non-relevant context, which can lead to a change of diagnosis of 30-40% within a limited amount of time. Furthermore, LLMs regularly hallucinate or make up information and re-enforce the bias of the data from which they are trained. Moreover, there are ethical, legal, and regulatory considerations that prevent the clinical employment of LLMs, including privacy concerns, disclosure of accountability, and the absence of regulation and surveillance. Therefore, without transparency in reasoning, the use of LLMs to independently diagnose patients is currently considered an unsafe and unproven application of LLMs.

Overall, the evidence currently available to support the use of LLMs to assist clinicians in their clinical practice should only be done so with strict supervision by the clinician, continuing reassessment of the bias of the data used to train the model, and ongoing validation through appropriate regulatory governance. Table 3 below highlights the key limitations of LLMs based on the most recent published research.

4.4 Comparative Analysis

Today's best medical AI systems are categorized into several different subclasses: general-purpose LLMs (large language models), medically aligned LLMs (MLLMs), and biomedical text model(s) + real-world clinical decision support system(s) + task/data diagnostic system(s). General LLMs (such as GPT-4) have demonstrated very strong reasoning skills, and they perform on average comparably to effectively within clinical practice settings[3], [9], [11], [15].

MLLMs (medically aligned LLMs that have been trained specifically with medical content as input training data) are able to perform better than general LLM models when tested against specific medical question-answering tasks, although they also struggle with issues related to providing accurate responses when there is ambiguity, insufficient context (incomplete), and also the presence of bias in their training datasets[4], [10], [16].Other LLMs have been specifically trained to read and extract information from medical literature, in addition to providing high-level summaries of the data within this literature, and are being utilized to assist with locating relevant articles through advanced search

techniques. While these LLMs do excel in completing these tasks, they still lack sufficient evidence to support the use of these systems to aid with the diagnosis of individual patients[4], [6]. While advanced models are able to

demonstrate some level of performance drop upon testing in a real-world healthcare workflow setting, this performance

Table 3
Current limitations and risks of large language models in medical diagnosis (2023–2025 research).

	Risks and Limitations		
Limitation	**Description**	**Key Findings**	**Recent Citations**
Lack of Clinical Reasoning	Models rely on statistical patterns instead of causal medical reasoning.	Produces plausible but incorrect diagnoses, especially for atypical cases.	[8], [9]
Bias in Medical Data	Models inherit demographic and systemic biases from training data.	Unequal diagnostic performance across population groups.	[8], [10]
Hallucinations	Generation of confident but incorrect or unsafe clinical statements.	Hallucinated treatments and diagnostic steps observed in trials.	[11]
Prompt Sensitivity	Output changes drastically with small prompt modifications	30–40% diagnostic changes with irrelevant context.	[9], [12]
Limited Multimodal Integration	Difficulty combining text, images, and laboratory data into coherent reasoning.	Multimodal fusion remains inconsistent and unstable.	[8]
Poor Rare Disease Performance	Weak generalization to underrepresented conditions.	Lower accuracy outside common diseases; misclassification of rare disorders.	[13]
Ethical & Legal Issues	Privacy risks, unclear accountability, and regulatory gaps.	Safety, liability, and transparency concerns remain unresolved.	[10], [14]
Lack of Explainability	Models operate as "black boxes."	No transparent reasoning chain for clinical verification	[13]
Over Reliance Risk	Users may trust LLM outputs excessively.	Potential for harmful decisions without human oversight.	[11]
Context Manipulation	Models change predictions with irrelevant input.	Demonstrates fragile, noise-sensitive reasoning	[12]

was demonstrated to be inconsistent with established clinical guidelines and was also shown to be poorly reliable when using Electronic Health Record (EHR) data [11], [13]. Studies focused upon reliability have shown that these models are very sensitive to changes in prompt and can have varying levels of reliability across a number of similar clinical cases [5], [12]. On the other hand, specialized imaging-based AI models, like CAD models for retinal disease, do show excellent performance in their limited diagnostic scope but are not able to generalize their results to a multimodal clinical reasoning approach [7].Digital pathology tools have similar task-specific pros and cons, as these tools enhance workflows with regard to efficiency but still require the oversight of a qualified expert [8]. LLMs that are used in the field of medical education lend to providing support with respect to the learning process and explanations of concepts for learners; however, they may provide misleading information to learners due to a lack of accuracy and/or an overconfidence level that has been demonstrated to be inaccurate[18].

In summary, the overall themes across the body of literature are that all model categories have their own set of pros and cons, and none of the models currently have been shown to contain the level of accuracy, reliability, or contextual knowledge required to allow for independent clinical diagnosis. As such, the need for qualified persons/supervision and domain-specific validation will continue to be paramount. (see Table.4).

4.5 Emerging Hybrid Approaches

4.5.1 Applications and Advancements

AI is having an excellent impact on almost everything involved in healthcare, from original research to the writing of our patients' clinical documentation and how patients engage with/receive cancer treatment, etc. The outbreak of COVID-19 resulted in many practitioners' utilizing AI wearable technology in addition to utilizing remote patient monitoring devices for a virtual delivery of service.

The metaverse has recently offered new ways of delivering care through the implementation of telehealth applications, allowing for a continued improvement on standard methods of video conferencing [1]. Historically, AI was seen through its earlier stages as being dominated by rule-based symbolic systems, and all logical rules set out had to be followed until the system was able to develop intelligence. The majority of these early-stage systems did not have the capacity for additional development and were not able to expand their implementation capabilities due to standard variables of uncertainty created by the lack of scalability in these systems. The introduction of machine learning during the 1980s created the turning point in the industry whereby systems were able to use real-world experiences to improve their performance by learning without the constraints of how learning occurred and what would be learned [2].

Table 4
Comparative analysis of state-of-the-art medical AI models based on strengths, limitations, and supporting evidence.

	Comparative Analysis			
Model Type	**Example Models / Key Papers**	**Strengths**	**Limitations**	**Citations**
General Large Language Models (LLMs) Used for Clinical Diagnosis	GPT-4 clinical diagnostic studies	Strong reasoning ability; competitive with clinicians in structured vignettes; useful for differential diagnosis generation	Hallucinations; misleading prompts; poor guideline adherence; limited real-world validation	[3], [9], [15], [17]
Medically Aligned LLMs (Domain Specialized)	Medically tuned GPT-4 variants; models covered in medical LLM reviews	Higher accuracy in medical QA; better domain grounding; improved interpretability through chain-of-thought	Vulnerable to ambiguity; dataset bias; cannot achieve fully reliable autonomous diagnosis	[4], [10], [16]
Biomedical Text Specialized LLMs	BioGPT,GatorTron like models (as discussed in surveys)	Strong biomedical literature summarization; extraction; named-entity recognition; supports research workflows	Not validated for direct clinical diagnosis; limited patient level evaluation	[4], [6]
Real-World Diagnostic Decision Support LLMs	GPT-4 evaluated on real EHR and guideline workflows	Provides helpful guideline-based suggestions under structured prompts	Performance drop on real clinical data; inconsistent outputs; poor generalization to complex EHR cases	[9], [11], [13]
LLM Reliability and Safety Evaluations	Studies on consistency, manipulation, and sensitivity testing	Reveal failure patterns; frameworks for safety evaluation; highlight bias sources	Show instability, prompt sensitivity, inconsistent reasoning	[5], [12]
AI Medical Imaging Diagnostic Models (Task Specific)	Retinal disease CAD systems	High accuracy in single-modality imaging tasks; strong clinical validation	Cannot generalize beyond narrow diagnostic domains	[7]
Digital Pathology & Domain Specific AI	LLMs in pathology workflows	Helpful for annotation, classification, and workflow efficiency	Requires expert oversight; risk of hallucinating morphological features	[8]
AI for Medical Education and Training	LLMs used as teaching assistants	Supports case-based learning; good explanation capability	Overconfidence; factual errors common; may mislead learners	[18]
AI for Healthcare System Support	General AI for triage, documentation, and summarization	Reduces workload; useful for administrative tasks	Not suitable for unsupervised medical decision-making	[14]

4.5.2 LLMs + Robotics / Embodied AI

The multimodal planning system based on GPT-4V creates plans for robots using visual and text information. The robot capabilities include combining the two forms of input into a single process for producing plans and ustilizing an established collection of symbolic actions for the implementation of the plans (see Fig. 5). The framework functions without adjustment as a zero-shot learning type of planner by leveraging the multimodal knowledge to gain a grasp of the scenes and task context (the purpose of this study is zero-shot task planning without robot-specific fine-tuning [19]).

The results of testing across nine robotic manipulation data sets demonstrated that the action sequences produced by the framework were in close agreement with the trajectories determined by ground truth; self-judged matching average scores were assessed at 8.7/10 (refer to Table 5).

Table 5
Evaluation results across datasets.

LLMs + Robotics / Embodied AI			
Dataset	**Task Type**	**Performance (/10)**	**Notes**
NYU VINN	Structured tabletop	9-10	Clear spatial layout
USC Jaco Play	Manipulation	9-10	High accuracy
Freiburg Franka Play	Tabletop manipulation	9-10	Minimal ambiguity
BC-Z	Multi-step tasks	7-9	Visual complexity
QT-Opt	Grasping	7-9	Depth-dependent tasks
Average		8.7	Strong overall alignment

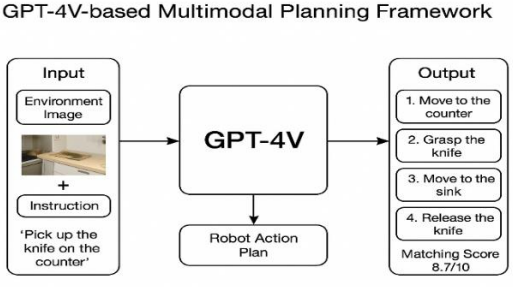

Fig. 5. Experimental framework

The effectiveness of the framework was shown to be especially strong in well-defined scenarios; areas of limitation in the capabilities of GPT-4V include that it can only use a single RGB image as input, which can result in incorrect assumptions about the nature of factor actions in complicated scenarios, and that the specific terms used to prompt GPT-4V were very sensitive to the writer of the prompt. Although the framework produces coordinated action sequences, it does not possess direct execution parameters regarding specific embodiments, which means that the output action sequence must be applied through other means by motion planners. Overall, the testing indicates that there is a strong degree of generalization across multiple dynamic manipulations involving robot actions.

4.5.3 Multimodal LLMs (vision + text + other signals), especially in medicine

This research uses an innovative method for producing US images. Using a cGAN model, this method produces more realistic ultrasound images while maintaining key anatomical details. Additionally, the objective of this study is to improve how ultrasound images are created to enhance medical training and enhance the accuracy of ultrasound image segmentation, as well as provide better diagnostics through improved ultrasound image authenticity [20] (see Fig. 6).

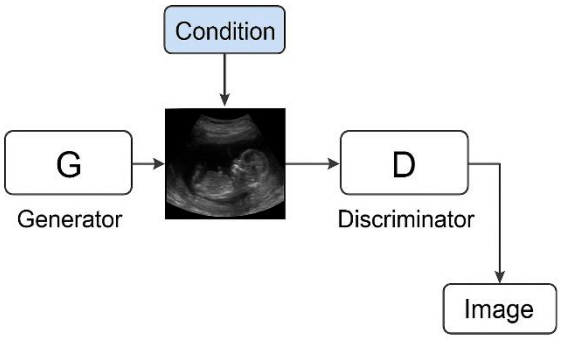

Fig. 6. Multimodal LLMs (vision + text + other signals), especially in medicine

4.5.4 Federated Learning (FL) + LLMs (privacy-preserving LLMs)

Large language models (LLMs) combined with federated learning (FL) can lead to the development of privacy-preserving, scalable, and high-performance AI systems and are a very promising area of research. In the paper written by [30], it discusses how LLMs perform numerous tasks well; however, they require large amounts of good-quality data for training. In comparison, FL allows for collaborative training of models from many different data sources located in multiple places, without having to put sensitive data in one central location. The authors of this work have shown that sub-technologies of LLMs, such as pre-training and prompt engineering, can play a critical role in improving FL by increasing speed of convergence, alleviating problems associated with non-iid datasets, and increasing the ability to create personalized responses (see Fig. 7).

Conversely, FL provides LLMs with capabilities that are not available to them when working with single databases, such as protection of privacy through distributed computing and secure prompt generation, as well as more reliable or robust chain-of-thought selection. If both LLMs and FL were successfully integrated, then they would provide significant advantages in enhancing generalization, privacy protection, and lifelong learning. The authors also identified several areas of concern that would be encountered as a result of integrating these technologies, including the potential transmission of data bias, synchronization challenges of heterogeneous devices, communication bandwidth required for transmitting large amounts of data, and new security risks that would be created. Overall, the paper written by Chen and the authors [30] highlights the complementary nature of both LLMs and FL working together collectively under the term" FedLLMs" and their ability to change the future of data-sensitive industries such as healthcare, finance, and education, in that they could create equal and secure access to state-of-the-art models.

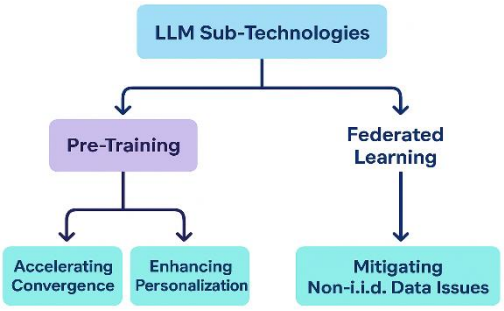

Fig. 7. LLM-Sub Technologies.

4.5.5 Knowledge Graphs (KGs) + LLMs (symbolic grounding and explainability)

[31] explores ways of integrating large language models (LLMs) with knowledge graphs (KGs) to develop artificial general intelligence (AGI). It identifies the synergy between LLMs and KGs (i.e., complementing strengths and weaknesses) and provides three different integration frameworks (see Fig. 8). The first framework, the KG-enabled LLM, helps improve factual accuracy, contextual reasoning, and knowledge grounding by incorporating KG data during the pre-training, fine-tuning, and inference phases of LLM training. The second framework, the LLM-enhanced KG, includes that LLMs help enhance how KGs are completed, constructed, and used for types of QA, which boosts their scalability, adaptability, and efficiency in the KGs. The third framework, the Synergistic LLM-KG System, utilizes LLMs and KGs to enable real-time, bidirectional reasoning between them, creating dynamic and AGI-enabled frameworks from traditionally static knowledge repositories. This integration technique draws strength from addressing the LLMs' tendency to "hallucinate" (provide false or vague information) and KG's rigidity in adopting new knowledge.

Examples of how all three frameworks can enhance systems in industries like health care, recommended systems, and other types of web services by creating smart and reliable end-user-focused solutions. The ultimate goal is to better support the design and creation of next-generation AGI architectures.

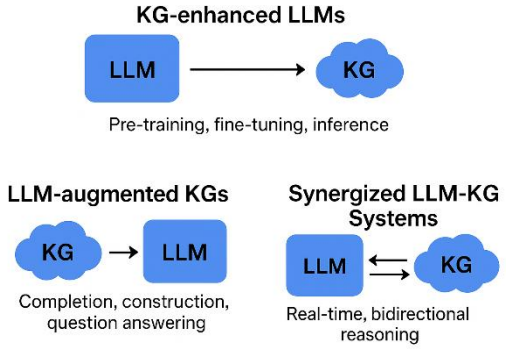

Fig. 8. Large language models (LLMs) with knowledge graphs (KGs)

4.5.6 Retrieval-Augmented Generation (RAG) Retrieval + LLM pipelines)

Grounding model outputs in verified clinical knowledge using the Retrieval Augmented Generation (RAG) framework improves the accuracy and reliability of medical chatbots. In a study from Guan et al., it is shown that RAG-based systems outperform traditional LLMs in many aspects of medical education and patient support[32]. However, this study also reveals the introduction of privacy risks for users through retrieval mechanisms. The study described by Bora and Cuayáhuitl demonstrates that using RAG leads to an increase in performance for medical question answering (e.g., Mistral-7B had a 57% correct answer rate when used with fine-tuning and RAG), and they also found that direct answer selection methods performed better than generative scoring [33]. Both studies emphasize that RAG pipelines can enhance reliability, factual grounding, and safety for most medical AI systems. They suggest that factors such as model architecture and tuning approaches can affect the performance improvement obtained from the successful utilization of RAG. Another important finding from these studies was that the degree of privacy risks associated with RAG pipelines increases with depth of retrieval, highlighting the importance of privacy-by-design principles as a foundation for building RAGenabled systems. Therefore, future studies need to integrate coherent policies with strong technical infrastructures to develop trusted and scalable RAG-based solutions for healthcare settings.

4.5.7 LLM Agents, Reinforcement Learning, and Decision Systems)

This research examines if Reinforcement Learning from Human Feedback (RLHF) could be used to enhance LLM's capabilities for deployment as a therapeutic chatbot designed specifically for use within the field of psychology. The authors did not detect substantial improvements in performance based on their findings when comparing these same models (pre-trained and tuned via RLHF) as assessed by psychologists licensed by the state. They believe that limitations regarding dataset size, hardware limitations, and the complexity associated with providing emotional support through the process of counselling had caused there to be no tangible evidence of increased value following the application of RLHF methods [21].

4.5.8 Evaluation Frameworks and Benchmarking in Medical AI

Medical artificial intelligence (AI) evaluation frameworks utilize dedicated benchmark datasets such as MedQA, PubMedQA, and MIMIC-III to evaluate clinical reasoning, biomedical question answering, and performance on Electronic Health Record (EHR) tasks using real-world data. However, benchmark datasets do not encompass all the Medical artificial intelligence (AI) evaluation frameworks utilize dedicated benchmark datasets such as MedQA, PubMedQA, and MIMIC-III to evaluate clinical reasoning, biomedical question answering, and performance on Electronic Health Record (EHR) tasks using real-world data. However, benchmark datasets do not encompass all the complexity of working within an actual clinical setting. Two newer types of research studies have now emerged: Human-AI Collaboration Research, testing how AI tools enhance clinical decision-making, and Clinical Validation and Deployment, which investigates how AI models are used in actual care settings, the generalizability of the model to various patient populations, and performance decline or new error types that do not occur in a benchmark setting. Clinical validation and deployment investigations are therefore necessary to identify the effectiveness of the AI in real-world settings.

4.6 Ethical, Legal, and Social Implications (ELSI)

AI-based diagnosis raises ethical, legal, and social implications (ELSI). These implications involve accountability, traceability, and transparency in clinical decision-making. Maintaining patient trust is also important; patients must know how their data are used and the role of AI in their care. Strong human oversight and appropriate validation of AI will ensure that errors are minimized, reduce bias, and allow AI to supplement, not supplant, clinicians' professional judgment.

5. Future Directions

Research on medical LLMs will advance over time by establishing increasingly more reliable and trustworthy systems. Medical professionals and patients will have increased confidence in their use of these systems. Some of the possibilities that research in this field will explore include the integration of multimodal reasoning (text, images, genomic data, and other clinical indicators) into medical LLMs to provide greater diagnostic accuracy and more accurate decision support for clinicians. A major opportunity within this research area is to create models that allow for continuous learning from the acquisition of real-time data in he. However, Hile economically utilizing the protections afforded by established frameworks for

maintaining patient confidentiality, protecting the integrity of the patient care delivery systems, and continuing to provide safe and effective patient care. The establishment of clear and sound policies and regulations will allow for the safe and ethical development and use of medical LLMs while ensuring compliance with the highest standards of medical practice.

6. Conclusion

This review explored the newly developed LLMs along with AI technology for healthcare diagnostics, identifying both their future potential as well as their current limitations. Based on performance tests conducted in controlled environments, LLMs show considerable promise as decision support tools when used together with a clinician. However, many challenges remain (reliability, robustness, ability to explain decisions made by AI and clinicians) that will slow or restrict the use of existing models for autonomous clinical diagnosis without appropriate human oversight. Future research will need to concentrate on integrating multiple modalities (i.e., using photos/images and data), implementing standardized evaluation methodologies to measure the effectiveness of using LLMs/AI, documenting transparent and simplified explanations of their models, and developing rigorous clinical validation processes to verify the safe and appropriate use of LLMs/AI in the clinical setting.

Acknowledgments

The authors thank the Department of ICT,University of Vavuniya for research support.

Competing Interests

The authors have no competing interests to declare

References

[1] A. and N. K. and A.-R. A. and A.-S. S. and A.-M. A. and S. A. V. and A. M. D. and A.-M. F. A. Al Kuwaiti, “A Review of the Role of Artificial Intelligence in Healthcare,” *J. Pers. Med.*, vol. 13, p. 951, 2023.

[2] N. and K. E. and N. A. Ghaffar Nia, “Evaluation of Artificial Intelligence Techniques in Disease Diagnosis and Prediction,” *Discover Artificial Intelligence*, vol. 3, no. 5, 2023.

[3] S. and U. G. Gencturk, “Rodent Tests of Depression and Anxiety: Construct Validity and Translational Relevance,” *Cogn. Affect. Behav. Neurosci.*, vol. 24, pp. 191–224, 2024.

[4] Y. and Z. Y. and C. Z. and G. B. and A. M. and F. R. and L. Y.-W. D. and X. Z. and D. A. D. and L. Y. and C. Q. and D. X. Zhai, “Machine Learning Predictive Models to Guide Prevention and Intervention Allocation for Anxiety and Depressive Disorders Among College Students,” *Journal of Counseling and Development*, 2024.

[5] W. and A. D. and others Olabiyi, “The Evolution of AI: From Rule-Based Systems to Data-Driven Intelligence,” 2025.

[6] Y. and S. Q. and K. D. R. Liu, “Integrating Large Language Models into Robotic Autonomy: A Review of Motion, Voice, and Training Pipelines,” *AI*, vol. 6, p. 158, 2025.

[7] J. and S. E. and H. H. and M. C. and L. Y. and W. X. and Y. Y. and L. X. and G. B. and Z. S. Wang, “Large Language Models for Robotics: Opportunities, Challenges, and Perspectives,” *Journal of Automation and Intelligence*, vol. 4, pp. 52–64, 2025.

[8] R. Manuel, “The history of artificial intelligence: Key milestones in AI”.

[9] M. and K. E. and N. H. Nia, “Artificial intelligence in clinical data analysis and diagnosis,” *J. Pers. Med.*, vol. 13, no. 9, p. 951, 2023.

[10] R. Manuel, “Artificial intelligence in healthcare market is the next big thing!!,” 2021.

[11] “AI-powered wearable devices and telehealth systems in healthcare innovation,” *J. Pers. Med.*, 2023.

[12] “The Evolution of AI: From Symbolic Systems to Machine Learning Paradigms”.

[13] A. C. and others Goh, “Influence of large language models on physician diagnostic reasoning,” *JAMA Netw. Open*, vol. 7, p. 4, 2024.

[14] Y. and Z. X. and L. H. and others Zhou, “Large language models in medical diagnosis: A comprehensive scoping review,” *NPJ Digit. Med.*, vol. 8, p. 11, 2025.

[15] Z. and H. W. and L. R. and others Yan, “A sensitivity evaluation framework for large language models in clinical diagnosis,” 2024.

[16] M. and U. S. and others Naveed, “A comprehensive overview of large language models: Architectures, scaling, alignment, and applications,” 2024.

[17] J. H. and A. J. and others Yoon, “Diagnostic accuracy of an AI-based computer-aided detection system for retinal disease,” *JMIR Med. Inform.*, vol. 2, p. e48142, 2024.

[18] T. and H. M. and H. D. and R. F. and H. N. and T. T. D. and G. M. I. and U. A. and K. C. a Roeschl and J. and D. H. and O. B. and H. G. and B. F. and F. V. and M. A. Kempfert, “Assessing the Limitations of Large Language Models in Clinical Practice Guideline-Concordant Treatment

Decision-Making on Real-World Data: Retrospective Study," 2024.

[19] Ano, "GPT-4 access and diagnostic performance study," *2024*.

[20] J. T. and R. P. N. and G. T. and M. C. J. and D. D. and V. G. and C. J. H. and C. E. Reese, "On the limitations of large language models in clinical diagnosis," 2023.

[21] E. and P. A. and B. M. M. and S. R. Ullah, "Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review," 2024.

[22] J. C. L. and L. K. Chow, "Large Language Models in Medical Chatbots: Opportunities, Challenges, and the Need to Address AI Risks," 2023.

[23] K. Subedi, "The Reliability of LLMs for Medical Diagnosis: An Examination of Consistency, Manipulation and Contextual Awareness," 2023.

[24] P. and J. F. and H. R. and B. K. and H. I. and K. M. and V. J. and M. M. and B. R. and K. G. and R. D. Hager, "Evaluation and mitigation of the limitations of large language models in clinical decision-making," 2024.

[25] N. M. and A. A. and P. S. P. and B. Y. and T. N. R. and S. V. and N. G. Priya, "Revolutionizing Healthcare with Large Language Models: Advancements, Challenges, and Future Prospects in AI-Driven Diagnostics and Decision Support," 2024.

[26] X. and Z. Y. and C. Z. Yang, "Large language models in clinical diagnosis and treatment: Opportunities and challenges," *Curr. Med. Sci.*, vol. 44, pp. 345–360, 2024.

[27] J. and V. B. N. Kim, "Assessing the Current Limitations of Large Language Models in Advancing Health Care Education," 2024.

[28] Anonymous, "GPT-4 diagnostic case and tool validation (eTables 4--5)," 2024.

[29] Jiamin Liang and Xin Yang and Yuhao Huang and Haoming Li and Shuangchi He and Xindi Hu and Zejian Chen and Wufeng Xue and Jun Cheng and Dong Ni, "Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis," 2022.

[30] C. and F. X. and L. Y. and L. L. and Z. J. and Z. X. and Y. J. Chen, "Integration of large language models and federated learning," *Patterns*, vol. 5, p. 101098, 2024.

[31] Linhao Luo and Carl Yang and Shirui Pan and Evgeny Kharlamov, "Integrating Large Language Models and Knowledge Graphs for Next-level AGI," *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.

[32] S. and K. H. C. and L. N. F. and S. G. and Q. H. and H. V. Guan, "Privacy Challenges and Solutions in RAG-Enhanced LLMs for Healthcare Chatbots: A Review of Applications, Risks, and Future Directions," *Comprehensive review of RAG applications and privacy risks in healthcare*, 2024.

[33] A. and C. H. Bora, "Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications," *Machines*, vol. 6, no. 4, p. 116, 2024.

[34] D. and E. T. Bill, "Fine-tuning a LLM using Reinforcement Learning from Human Feedback for a Therapy Chatbot Application," *KTH Royal Institute of Technology*, 2023

Author Biographies

First Author M.U.K. Gunawardhana received the Honours Bachelor's degree in Information and Communication Technology from Rajarata University of Sri Lanka in 2024. She is currently pursuing the MSc degree in Data Science and Artificial Intelligence at University of Peradeniya. She is affiliated with the Department of Information and Communication Technology, University of Vavuniya. Her research interests include artificial intelligence, machine learning, natural language processing, and medical AI systems.

Second Author Pirunthavi Wijikumar received the BSc (Special) degree in Computer Science and Technology from Uva Wellassa University of Sri Lanka and the Master of Computer Science degree from University of Peradeniya. She is currently a Lecturer (Probationary) in the Department of Information and Communication Technology, Faculty of Technological Studies, University of Vavuniya. She has several years of experience in teaching and research in the fields of computer science and information technology. Her research interests include machine learning, artificial intelligence, computer vision, data mining, natural language processing, remote sensing, and medical image analysis.

Third Author D.M.O.K. Weerasinghe received the BSc Eng. degree in Electronic and Telecommunication Engineering from Sri Lanka Technological Campus. He is currently pursuing the MSc degree in Data Science and Artificial Intelligence at University of Peradeniya. He works as an Application Engineer. His research interests include wireless communication, microgrid systems, and artificial intelligence.