FELIPE TOSHIYUKI MIAMOTO

THE ROLE OF LARGE LANGUAGE MODELS IN CLINICAL DECISION SUPPORT: A SYSTEMATIC LITERATURE REVIEW AND A SURVEY WITH PHYSICIANS

São Paulo

2025

FELIPE TOSHIYUKI MIAMOTO

THE ROLE OF LARGE LANGUAGE MODELS IN CLINICAL DECISION SUPPORT: A SYSTEMATIC LITERATURE REVIEW AND A SURVEY WITH PHYSICIANS

Graduation thesis for Escola Politécnica da Universidade de São Paulo to obtain a degree in Industrial Engineering

São Paulo
2025

FELIPE TOSHIYUKI MIAMOTO

THE ROLE OF LARGE LANGUAGE MODELS IN CLINICAL DECISION SUPPORT: A SYSTEMATIC LITERATURE REVIEW AND A SURVEY WITH PHYSICIANS

Graduation thesis for Escola Politécnica da Universidade de São Paulo to obtain a degree in Industrial Engineering

Supervisor:
Marco Aurélio de Mesquita

Co-supervisor:
Marco Cantamessa

São Paulo
2025

Catalogação-na-publicação

Miamoto, Felipe Toshiyuki
THE ROLE OF LARGE LANGUAGE MODELS IN CLINICAL DECISION SUPPORT: A COMPREHENSIVE REVIEW AND FUTURE TRENDS / F. T. Miamoto -- São Paulo, 2025.
p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Produção.

1.Modelos de linguagem natural 2.Sistemas de apoio à decisão clínica 3.Revisão sistemática da literatura 4.Survey com médicos I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Produção II.t.

To my family and friends.

ACKNOWLEDGMENTS

On first place, i thank my parents Marcia and Edson for always offering me the best conditions and for always have been there for me. I cannot express how much their inspiration was essential for me to become who I am and I am very happy to have them supporting me to achieve the best results.

I thank my brother Fabio, for teaching me how it is to have a best friend inside my own home, showing me the true meaning of brotherhood.

I thank all my friends from Escola Politécnica da USP, which made all of this possible.

I thank all my friends from Colégio Objetivo Integrado and Colégio Oshiman, which were really important in my life and gave me motivation to achieve my goals.

I thank all professors that I had during my life, including those at Colégio Oshiman, Colégio Objetivo Integrado, Universidade de São Paulo and Politecnico di Torino. Thanks for dedicating you lifes to this noble profession. I would like to express special thanks to my elementary school math teacher, Marcia, who gave me motivation to always learn more and more.

I would like to express my sincere gratitude to my supervisor, Marco Aurélio de Mesquita, and my co-supervisor, Marco Cantamessa, for their tremendous support in the development of this thesis. I am deeply thankful for their patience and the invaluable guidance they provided throughout this work.

“The harder you work, the luckier you get.”

– Gary Player

ABSTRACT

This thesis aims to map the use of Large Language Models (LLMs) in Clinical Decision Support, through a systematic literature review and a survey of physicians. The systematic review analyzed 39 articles from the Web of Science database published between 2023 and 2025. The findings suggest that LLMs hold significant potential to enhance clinical diagnosis and support treatment recommendations, functioning as assistive tools that complement rather than replace the physician's role. The review categorizes the literature by technological approach, including General Purpose LLMs, Data Wrangling, Prompt Engineering, RAG, Imaging Analysis, and Multimodal Applications. It also addresses key barriers to LLM adoption in clinical decision support, such as their 'black box' nature, hallucinations, data privacy concerns, regulatory challenges, ethics issues and algorithmic biases. Furthermore, a survey was conducted between September 26 and October 7, 2025, with 79 answers from a pool of 308 physicians of a hospital in São Paulo, yielding a response rate of 25.6%. The survey results indicate that, although physicians are not currently using LLMs as clinical support tools, they strongly believe that widespread adoption of LLMs will occur in the near future. Limitations of this study include, the fast-evolving nature of the topic and a restriction to single institution in the survey. Finally, as a future direction, a broader survey is suggested, including doctors from other institutions and locations. Moreover, how medical education will adapt to the emergence of these technologies can be a fertile field for study.

Keywords: Large Language Models, clinical decision support, systematic literature review, survey.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

1. Introduction

In this chapter, the core concepts of LLMs and Clinical Decision Support will be addressed, including a brief history of the evolution of systems in Clinical Decision Support. Furthermore, the objectives and the structure of this thesis are explained.

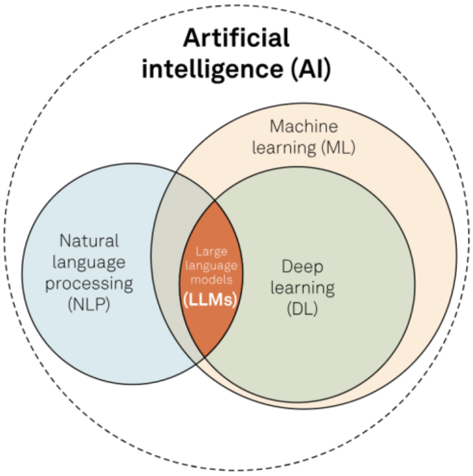
1.1 LLM for Clinical Decision Support

Clinical Decision Support (CDS) refers to health information technology systems that provide clinicians with timely, evidence-based, and patient-specific information to support decision-making. These systems aim to improve clinical decision-making by offering evidence-based recommendations, alerts, reminders, diagnostic support, and other contextual insights (Osheroff et al, 2012).

To keep pace with the increasing knowledge in medicine, tools for assisting human reasoning are becoming increasingly valuable. That is where artificial intelligence appears. Until Large Language Models (LLM) appeared in this field, historically other tools have demonstrated some usefulness. From the 1950s to 1980s, rule-based systems using decision-trees and IF-THEN logic used to assist in infectious disease diagnosis (Buchanan & Shortliffe, 1984). Later, from the 1990s to 2010s, the focus shifted towards Machine Learning, driven by algorithms such as logistic regression and random forest to approach clinical data statistically (Kononenko, 2001). From the 2010s to 2020s, Deep Learning emerged as an innovation to address image recognition and diagnostics (Litjens et al., 2017). In recent years, Large Language Models (LLMs) such as GPT and Med-PaLM have emerged as powerful tools capable of understanding and generating human language. These models can answer clinical questions, support reasoning, and process diverse medical texts and scenarios (Singhal et al., 2023).

Large Language Models appear at the intersection (Figure 1) of Deep Learning (DL) and Natural Language Processing (NLP). Deep learning can be defined as a subfield of Artificial Intelligence that is able to handle complex patterns by mimicking human thinking. On the other hand, NLP exists to make machines understand and generate human language. Therefore, Large Language Models combine the goals of NLP and the computational power of Deep Learning to process, interpret, and generate natural language (Fernández, M, 2024).

Figure 1: Definition of LLMs


Artificial intelligence (AI)
Machine learning (ML)
Natural language processing (NLP)
Large language models (LLMs)
Deep learning (DL)

Source: Fernández, M (2024)

One of the most known Large Language Models is GPT (Generative Pretrained Transformer). GPT is a decoder-only model (Wu et al, 2023), which generates word by word using the self-attention mechanism together with positional encoding. This means that, given a prompt, the model generates an output where each word will be generated based on representations of the prompt and the past generated words with an attention mechanism to give relative importance scores to words that were previously generated.

Another example is Med-PaLM, which is also a decoder-only transformer model that adds domain-specific expertise via Supervised fine-tuning on medical Q&A datasets (like MedQA, HealthSearchQA, PubMedQA), Reinforcement Learning with Human Feedback (RLHF) using clinician-annotated responses and focus on factuality, reasoning, bias reduction, and helpfulness in outputs.

A transformer model is a model architecture relying entirely on an attention mechanism to draw global dependencies between input and output (Vaswani et al., 2017). It is a type of Deep Learning Architecture designed for processing sequential data particularly in Natural Language Processing Tasks, where the attention mechanism exists to better capture the relationships between words in a sentence.

These architectures make LLMs particularly useful to handle medical language and knowledge once they can: i) learn from diverse and specialized data (e.g. Med-PaLM) ii) handle unstructured data, such as physician notes iii) scale and generalize iv) handle question and answering as well as summarization.

1.2 Objectives of this thesis

This study has as a general objective to map the role of Large Language Models (LLMs) on clinical decision support by doing a Systematic Literature Review and a Survey with physicians.

This general objective can be divided into two specific goals:

1. Characterize and analyze critically the actual state of research about LLMs on clinical decision support.
2. Investigate the perception of doctors about the usage of LLMs on clinical practice.

To achieve these goals, this work will be oriented by the following research questions:

i) What are the publication trends (by year, country, authors and keywords) in the literature about LLMs and Clinical Decision Support?

ii) How LLMs can be used by physicians and which technologies are being applied on disease diagnostics?

iii) How LLMs can be used by physicians and which technologies are being applied on clinical treatment and monitoring health conditions?

iv) What are the challenges associated with the usage of LLMs on diagnostic, treatment and monitoring health conditions of patients?

v) What is the opinion of physicians about the usage of LLMs on clinical practice?

1.3 Structure of the thesis

This thesis is organized as follows: the second chapter describes the methodology of research. It presents the protocol adopted to make the Systematic Literature Review, including strategy of search, triage process, criteria of eligibility and evaluation of the quality of selected articles. This chapter also brings a descriptive analysis of selected publications, including authorship, institutions, countries and keywords.

The chapter 3 presents the results of the systematic review organized around the applications of LLMs in clinical decision support, the technological approach used (e.g. RAG, fine-tunning, multi-modal models) and the challenges reported in the literature.

The fourth chapter reports the survey applied with physicians, including the motivation, the methodology and the results obtained. The results include factors like the knowledge, the frequency of use, the utility perception and the main worries related to the use of LLMs on clinical practice.

The chapter 5 discusses the findings of the systematic review and the survey in light of the research questions, highlighting convergences, tensions and implications for practice and future research.

The sixth chapter concludes the thesis with a synthesis of the main results, a reflection about the limitations of the study and recommendations to future research.

2 Literature Review Methodology and Descriptive Analysis

A Systematic Literature Review (SLR) is a detailed and reproducible method for synthesizing academic research. It adopts a transparent and structured procedure to identify, evaluate, and compile relevant literature on a specific topic. What sets it apart from traditional reviews is its systematic nature and its strong focus on reproducibility.

The process involves rigorously defined steps—starting from precise criteria for the literature search and study selection, moving through quality assessment and data extraction, and ending with well-established synthesis methods. This structure ensures that other researchers can replicate the review reliably, which reinforces the validity of the findings by minimizing selection and analysis bias.

SLRs follow standardized protocols to guide the identification and critical evaluation of the literature, always anchored to a clearly defined research question. This guarantees that the synthesis of evidence is not only thorough but also free from subjective distortions. Unlike narrative reviews, which may reflect personal viewpoints or interpretive bias, a systematic review seeks to present a clear, objective, and comprehensive picture of the state of research on the topic.

The method is built around focused research questions, allowing for a more concentrated and meticulous exploration of the field. Widely recognized in medicine and health sciences for its effectiveness in generating evidence-based, well-founded insights from the existing body of literature.

In this chapter, it is presented the methodology used to conduct this systematic literature review. This methodology was elaborated based on the one presented by Siddaway, Wood e Hedges (2019) and on the one used by Lemstra, Mesquita (2023). It includes four key steps: i) Scoping, ii) Identification (searching), iii) Screening, and iv) Eligibility.

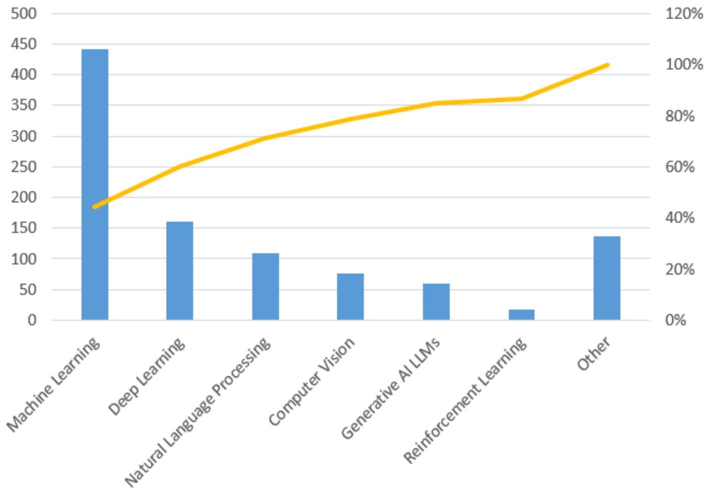
2.1 Literature Review Methodology

2.1.1 Scoping

In this phase, many preliminary key issues of conducting a systematic review are addressed. The very first step was to understand better the sub areas existing in the healthcare and artificial intelligence fields. With this objective, an exploratory analysis was conducted on the most 1000 cited publications (on the intersection between AI and HealthCare publications) released in the last five years. These publications were, then, classified according to two criterias: i) Artificial Intelligence keywords (Figure 2) and ii) Healthcare keywords (Figure 3).

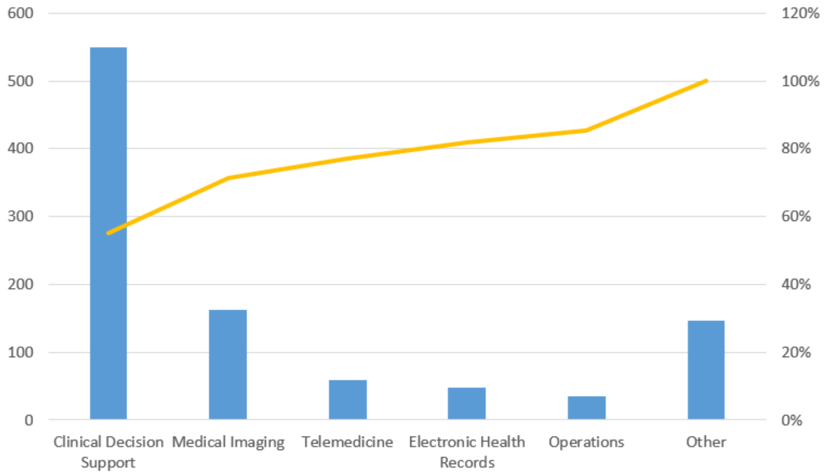
In order to classify one article according to the categories on both aspects, the methodology was to i) connect to the Gemini API (Application Programming Interface) from Python and ii) provide the abstract of each of the 1000 publications as a prompt asking Gemini to classify among each main theme. Given an abstract A, the Gemini model could classify as belonging to Generative AI and to Clinical Decision Making or Computer Vision and Medical imaging, for example.

Figure 2: Classification of articles according to AI field



Source: The author

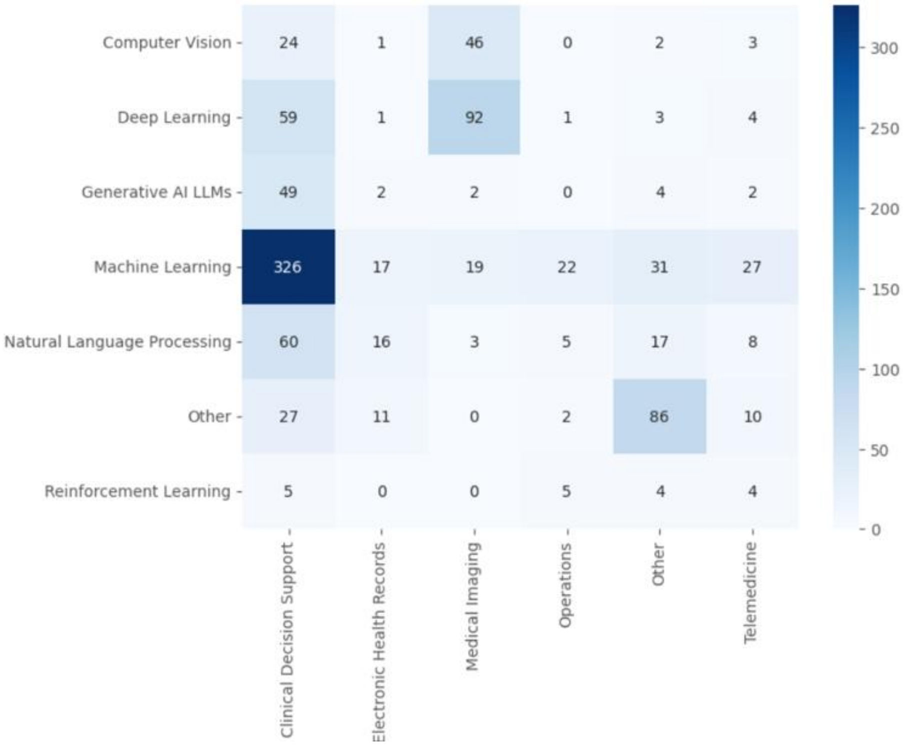
Figure 3: Classification of articles according to Healthcare field



Source: The author

Finally, a cross-analysis were performed to understand the relationships between categories on AI and categories on Healthcare (Figure 4)

Figure 4: Cross-Analysis AI x Healthcare



Source: The Author

By classifying among AI and Healthcare groups and making a cross-analysis, it is possible to understand better which topics have been more explored and which topics still leave space for improvements and to make an original work.

This cross-analysis points out interesting facts such as: Clinical Decision Support is by far the topic that is the most related to Artificial Intelligence and Medical Imaging is closely related to Computer Vision and Deep Learning architectures.

It also enables the delineation of boundaries to be adhered to within the defined scope. From the analysis made, it is possible to see that choosing Machine Learning + Clinical Decision Support would not help to restrain the research, once is the most common subject. Therefore, a more up to date and fast-growing topic was chosen as the object of the systematic review: the relationship between LLMs and Clinical Decision Support. This brings more possibilities to generate an original Systematic Literature Review, which fills a gap in the current literature by including a bibliometric analysis and closes the gap on the current literature review content.

2.1.2 Searching

Given the research questions and the well-defined scope, the next step in conducting a systematic literature review involves finding materials for the review. For producing a review of quality, the underlying material is extremely relevant. Therefore, the well-known electronic database Web of Science (WoS) was the one chosen to be the means of research.

In order to get all the possible relevant articles, it is important to define a research query that is, at the same time, broad and focused. This means that the research query should include all possible studies that can help to answer the research questions, but not deviating from them.

After careful design, the following research structure was established to find the articles: (LLMs OR Generative AI) AND (healthcare OR medicine) AND (clinical decision support) AND (language English). This was determined in order to conduct a methodical and comprehensive literature search, filtering first the articles related to healthcare and medical fields and including only the ones related to clinical decision support on this field.

With the forementioned query, on Web of Science database, 324 results were found.

2.1.3 Screening

Search results need to be screened for potential inclusion (Siddaway, Wood, & Hedges, 2019). Therefore, with the database of articles generated via search engines on Web of Science, it is important to guarantee that the articles are relevant to the theme of the thesis: LLMs on Clinical Decision Support.

The screening process is extremely important to determine the quality of the systematic review to be realized. To assure this quality and the fit with the topic proposed, a screening process using Artificial Intelligence was conducted.

The first step relied on using Gemini API (Application Programming Interface) on Python to answer 2 questions, rating from 1 to 5. The questions were: i) Is it related to Generative AI or LLMs? and ii) Is it related to Clinical Decision Support? Where 1 is not related and 5 is highly related.

Then, it was also asked for the Gemini API to summarize in 5 topics the abstracts of the papers. This was important to effectively perform a pre-screening process. Based on these topics and the ratings for the two questions forementioned 101 papers were selected.

2.1.4 Eligibility

The selection criteria was based on the inclusion of studies that are within the defined scope (LLMs + clinical decision support) and that could help answering one or more of the research questions proposed. After a careful reading of the full abstracts, 48 articles were chosen.

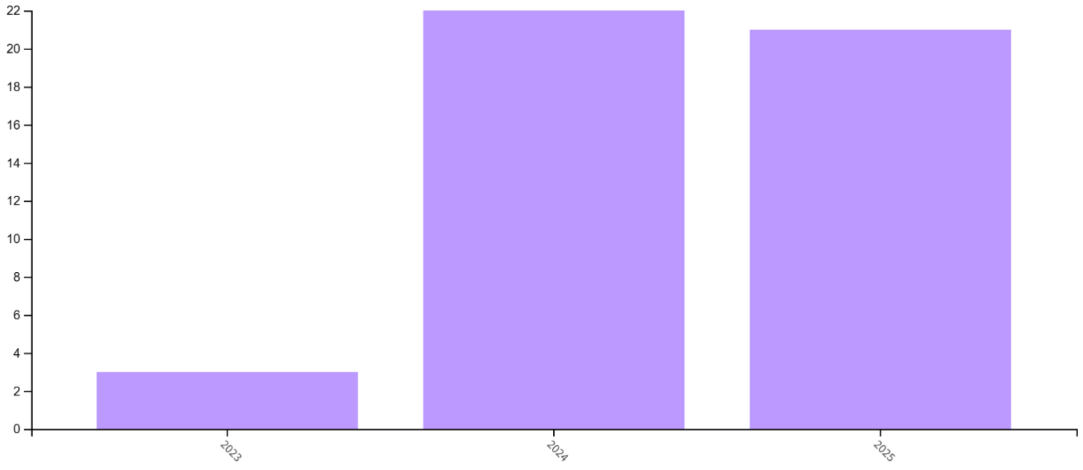
Finally, it was possible to determine if an article should be eligible or not (inclusion or exclusion criteria). This last step included full reading of the papers. After careful analysis, 39 articles remained to perform bibliometric and content analysis.

2.2 Descriptive Analysis

The bibliometric analysis is usually used in systematic literature reviews to describe the papers selected for the review, including publications by year, publications by country, keyword density map, co-citation network and others. Thus, once the methodology used for the search, selection and analysis of the papers has already been presented, a bibliometric analysis is the following step.

Figure 5 shows the number of publications by year of chosen articles. It has 3 papers from 2023, 22 from 2024 and 21 from 2025, highlighting the importance that this topic has gained in recent years after the launch of ChatGPT in November of 2022.

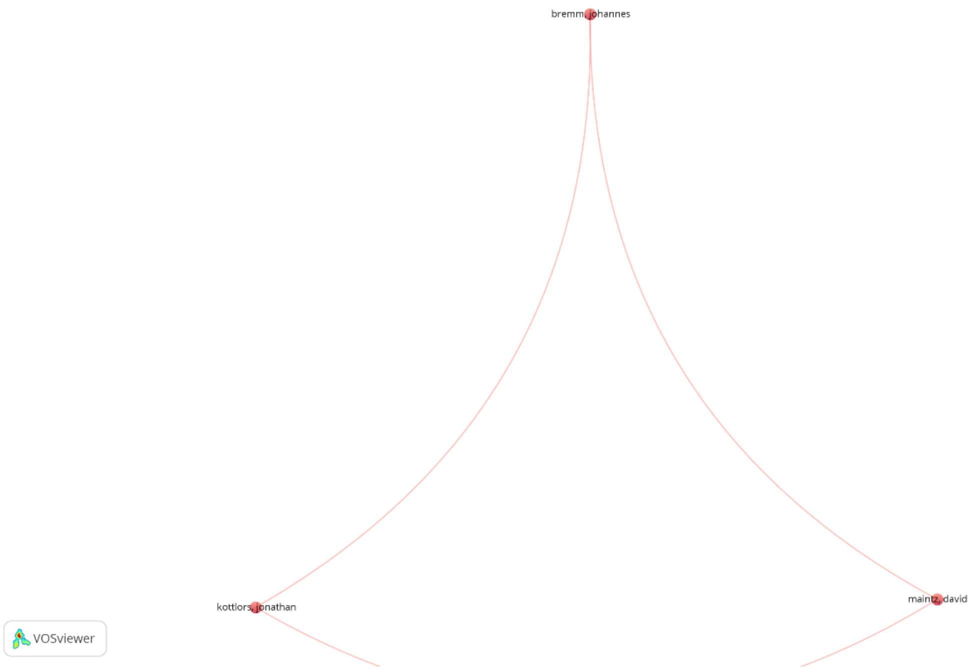
Figure 5: Publications by year



Source: The author

A co-authorship analysis is used to analyze collaboration patterns, understanding how scholars work together and identifying key authors. Therefore, a co-authorship analysis was made using VOSviewer, which is a software tool for constructing and visualizing bibliometric networks. By settling the default properties and the minimum number of documents of an author to 2, we obtained one connected component with 3 authors, as seen in Figure 6. This points out a very important characteristic of the topic covered in this systematic review: the topic is very new and for the moment research efforts are still isolated, with a potential fragmentation, where researches may be still working independently or in isolated silos.

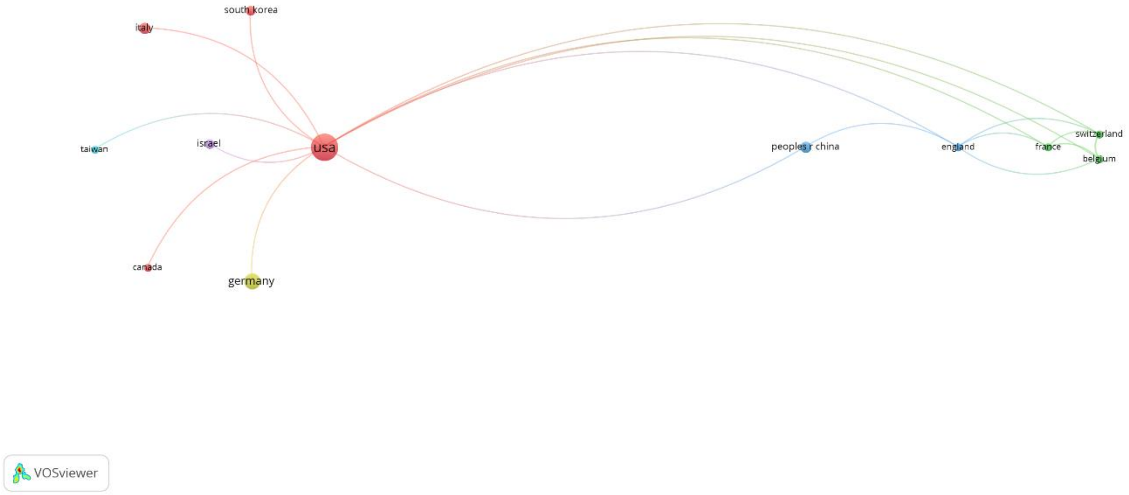
Figure 6: Co-authorship network for papers included in the review



Source: The author

A co-authorship by country was also performed in order to understand collaboration patterns across countries. By analyzing Figure 7, we can clearly see that USA plays an important role as a central author for research in this field.

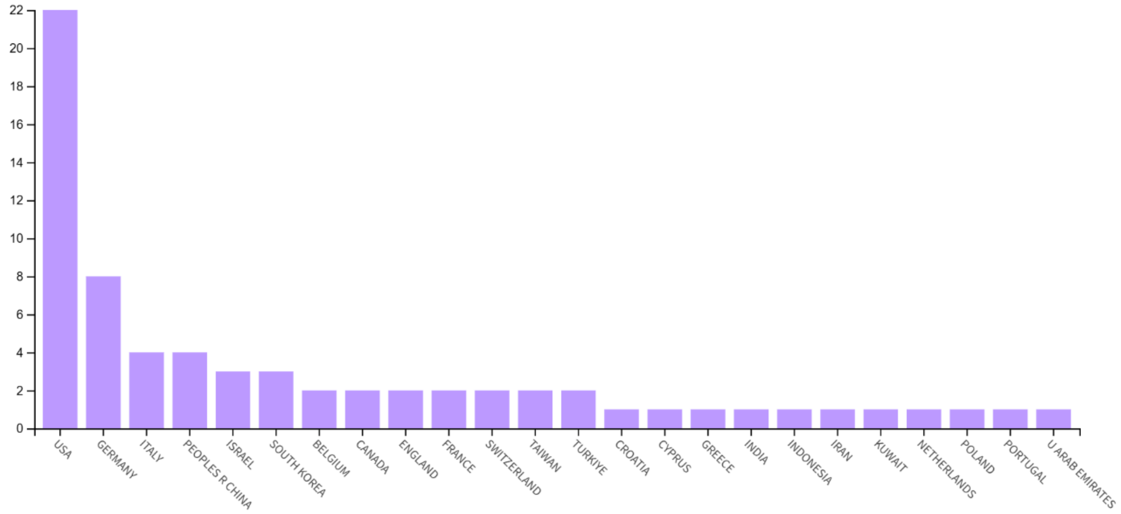
Figure 7: Co-authorship by country for papers included in the review



Source: The author

This key role that USA has is confirmed by the number of publications by country, as shown in figure 8.

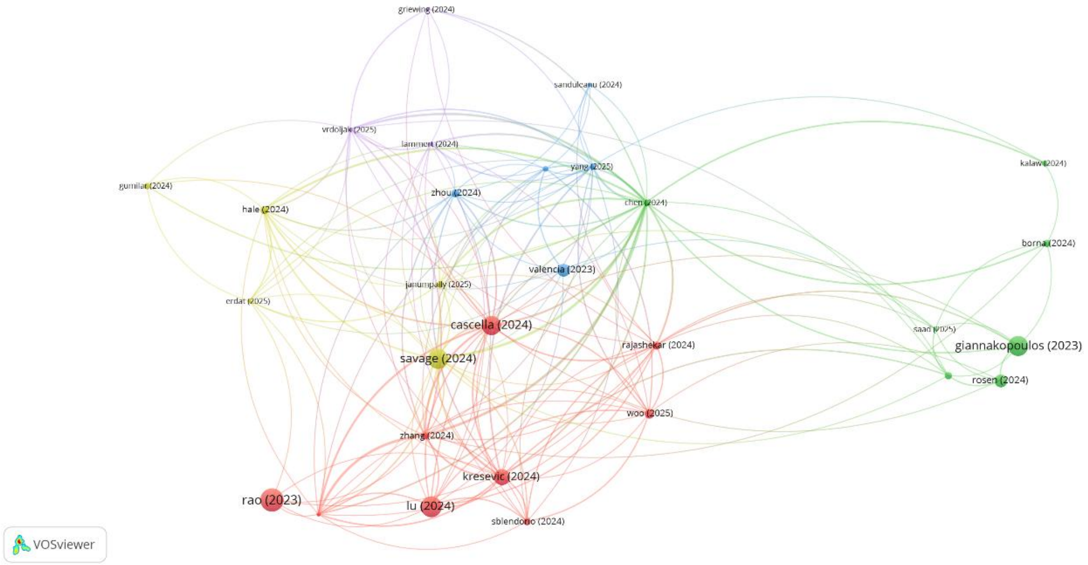
Figure 8: Publications by country



Source: The author

Furthermore, the software was also used to generate a bibliographic coupling to understand how similar are documents based on the similarity of citations they made. The results can be seen in the Figure 9.

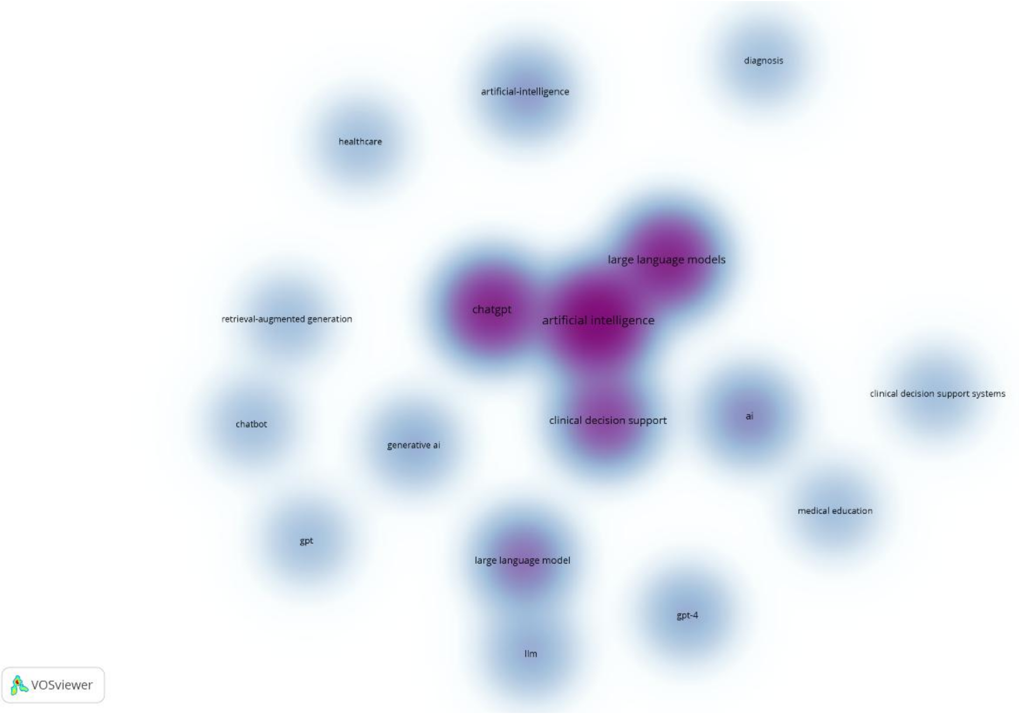
Figure 9: Bibliographic coupling for papers included in the review



Source: The author

VOSviewer was also used to generate a keyword co-occurrence density map. The minimum number of occurrences of a keyword to be displayed in the map was considered as three in order to limit the number of keywords and allow a proper visualization.

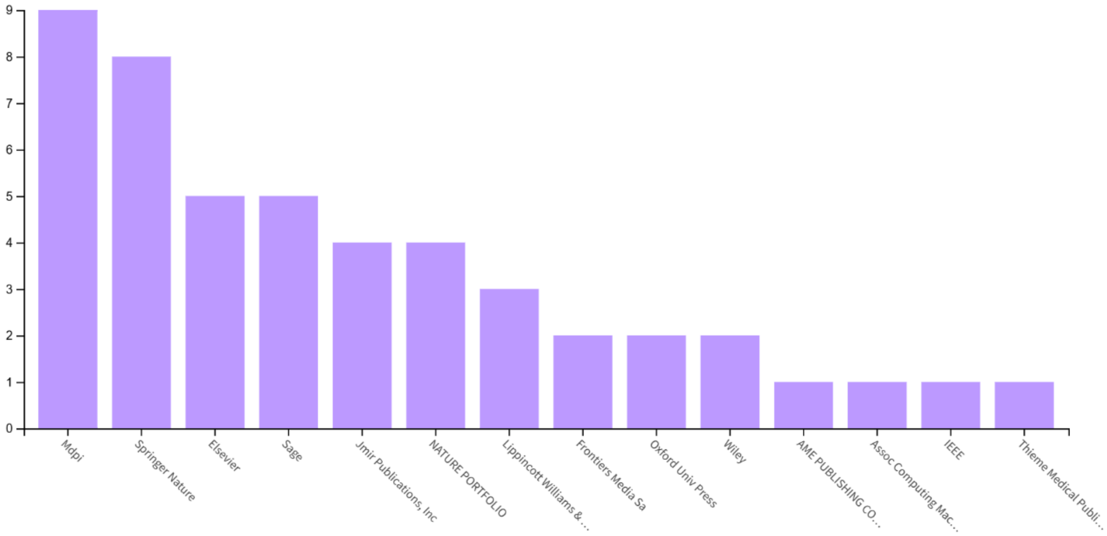
Figure 10: Keyword density map for papers included in the review



Source: The author

Finally, the analysis of papers selected reveals that MDPI is the main publisher for the topic, followed by Springer Nature and Elsevier as shown in Figure 11.

Figure 11: Publishers



Source: The author

3 Literature Review

The previous chapter included a descriptive analysis of the selected articles to give a general view of the research made regarding the topic of clinical decision support using LLMs. In this chapter, the focus shifts towards a content review of the selected articles.

To present the content analysis in this chapter, the content was grouped in 2 main topics: 1) LLMs on Clinical Diagnosis and Treatment and 2) Challenges of using LLMs on Clinical Decision Support.

LLMs on Clinical Diagnosis and Treatment is subdivided into:

1) General Purpose LLMs
2) Data Wrangling, Data Analysis, Risk Assessment and Visual Analytics
3) RAG
4) Prompt Engineering
5) Multi-Modal models and Imaging Analysis
6) Other specific technologies

Challenges of using Clinical Decision Support is subdivided into:

1) Black-box
2) Hallucinations
3) Data privacy issues
4) Data regulations
5) Others

3.1 LLMs on Clinical Diagnosis and Treatment

LLMs can process vast amounts of medical data, including patient histories, imaging results, and laboratory findings, to assist clinicians in making accurate and timely diagnoses (Chen D. et al, 2025).

3.1.1 General Purpose LLMs

Chen A. et al. (2024) found an evaluation involving 38 complex diagnostic cases published by the New England Journal of Medicine (NEJM), where ChatGPT-4 was tested against the diagnostic performance of NEJM readers. Results showed that ChatGPT-4 correctly identified the diagnosis in 57% of the cases, significantly outperforming the average NEJM reader, who achieved a 36% accuracy rate.

Borna et al (2024) evaluated and compared the diagnostic capabilities of two leading large language models—ChatGPT-4 and Google Gemini—across typical emergency cases in plastic and reconstructive surgery. The models' performances were tested both with and without physical examination data. Thirty medical vignettes were developed based on real patient scenarios, covering a broad range of topics including hand surgery, burns, lip, ear, and eyelid lacerations, skull fractures, sternal wounds, facial hematomas and nerve injuries, parotid duct injuries, mandibular fractures, nasal fractures, and nasal septal hematomas.

The LLMs were prompted with clinical scenarios describing patient presentations and relevant findings. One example scenario involved a 36-year-old woman presenting with forearm pain and muscle weakness after trauma, with physical exam findings of pain on passive stretching, sensory deficits, and elevated compartment pressure at 42 mmHg. These clinical details guided the models' diagnostic reasoning.

Both ChatGPT-4 and Google Gemini demonstrated strong diagnostic abilities, with ChatGPT-4 achieving higher accuracy, particularly when physical examination data was available. Three healthcare professionals independently rated the answers, resolving discrepancies through consensus to ensure consistent evaluation. ChatGPT-4's diagnostic accuracy reached 90% without physical exam data and 100% with it, compared to Gemini's 73.33% and 86.67%, respectively. The absence of physical exam details limited ChatGPT-4's specificity, as it could detect skull fractures but not specify types such as frontal sinus fractures. Inclusion of physical examination information improved diagnostic accuracy for both models, underscoring the critical role of clinical data in enhancing AI-supported diagnosis (Borna et al, 2024).

Another usability of LLMs researched by Rosen et al (2023) included using a LLM to suggest which exams a patient should undergo. This study evaluated ChatGPT's ability to recommend appropriate imaging tests by comparing its responses to those of the ESR iGuide, a clinical decision support system (CDSS) based on ACR (American College of Radiology) guidelines and adapted for European practice. A total of 97 clinical cases were analyzed. For each case, the question posed to ChatGPT was: "What are the most recommended imaging exams in this case?" Its free-text responses were then compared to the ESR iGuide's graded recommendations.

The findings by Rosen et al (2023) showed a high degree of consistency between ChatGPT and the ESR iGuide, with 87.6% agreement across all cases. In the 66 relevant Computed Tomography cases evaluated by the specialists, ChatGPT's recommendations received a mean appropriateness score of 6.02 out of 7. These results indicate that ChatGPT can deliver imaging suggestions that closely align with expert opinion and established guidelines. While the study confirms ChatGPT’s potential as a CDSS tool, it also notes that integration with domain-specific models as well as accordance with medical regulations and guidelines would be required for practical clinical use.

However, while some research, as shown, highlights their proficiency in generating accurate diagnostics, other studies, particularly in specialized fields like precision oncology, indicate that LLMs may not yet achieve the reliability and personalized insight provided by human experts (Vrdoljak et al, 2025).

Another research made by Benary et al. (2023) showed that different LLMs including ChatGPT and BioMedLM, are not currently suitable for routine use as tools to assist in personalized clinical decision-making in oncology.

Besides, on a study conducted by Hager et al (2024), the limitations of current open-source LLMs in clinical decision-making were further demonstrated, revealing significant performance gaps between these models and clinicians in patient diagnosis. The research found that existing open-source LLMs (specifically Llama 2 Chat (70B), Open Assistant (70B), Wiz ardLM (70B), Camel (70B) and Meditron (70B)) struggled to follow diagnostic guidelines and encountered difficulties with fundamental tasks such as laboratory result interpretation. The authors concluded that these models are not yet suitable for autonomous clinical decision-making and require substantial clinician oversight.

Yet, Hager et al (2024) stated that their study may not reflect the capabilities of the most recent open-source models, such as Llama 3 70b and 405b, which have demonstrated performance comparable to GPT-4. This rapid advancement in model capabilities highlights a persistent challenge in AI research: the potential for studies to become outdated during the publication process due to the accelerated pace of technological development. Consequently, the reported underperformance of open-source models may not accurately represent the current state of the field, as the latest iterations have shown marked improvements across relevant benchmarks.

Sanduleanu et al. (2024) evaluated GPT-3.5's ability to support clinical decision-making in determining whether patients with suspected appendicitis should undergo surgery or receive conservative antibiotic treatment. Using a cohort of 63 confirmed appendicitis cases and 50 control patients with right lower abdominal pain, the model was prompted with comprehensive clinical, laboratory, and radiological data to recommend either laparoscopic exploration/appendectomy or non-surgical management. GPT-3.5 achieved an accuracy of

90.3%, showing strong concordance with decisions made by a panel of six board-certified surgeons, which served as the reference standard.

Sanduleanu et al. (2024) noted that GPT-3.5 outperformed traditional machine learning models when given full-text clinical data and specific prompts, though machine learning provided more transparency into the role of individual variables. Importantly, the study emphasized the need to ask the model to commit to clear treatment recommendations despite clinical uncertainty. While the authors did not view GPT-3.5 as a replacement for surgical judgment, they proposed its potential use as a decision support tool in acute care, particularly when time pressure demands rapid, informed decisions.

Harari et al. (2024) investigated how generative AI systems, particularly when supervised by clinical experts, can support real-time decision-making in emergency scenarios such as cardiac arrest. Unlike rule-based AI, generative models can respond dynamically, making them valuable in high-pressure, evolving contexts. The study tested this potential through a simulated CPR (Cardiopulmonary Resuscitation) intervention with participants lacking medical experience, comparing three guidance methods: a traditional paper checklist, ChatGPT alone, and ChatGPT with real-time supervision by an emergency physician. This hybrid approach was designed to explore whether integrating human oversight could improve AI usability, trust, and decision accuracy.

In the supervised ChatGPT group, an emergency physician validated AI-generated instructions before they were delivered to the participants. A color-coded system indicated whether suggestions were safe or required caution, and participants could query the clinician directly. Compared to the ChatGPT-only and paper-based groups, the supervised ChatGPT group demonstrated higher decision accuracy and lower physiological stress. Although the ChatGPT-only group asked significantly more questions, suggesting greater uncertainty, the supervised group relied less on clarification, reflecting greater trust and confidence in the guidance provided (Harari et al, 2024).

The use of augmented reality and detailed performance metrics—including completion times, cognitive load, and physiological indicators—allowed for a nuanced evaluation of each intervention. Despite improvements in decision quality, the supervised ChatGPT group had longer scenario completion times, illustrating a tension between speed and accuracy. In real clinical environments, this trade-off must be carefully managed, as delays can carry significant consequences. The study echoes patterns seen in other critical fields like aviation, where expert input improves safety but can impact response time (Harari et al, 2024).

The findings show that while generative AI can assist with emergency care, its optimal use lies in supervised systems that preserve clinical oversight. Trust in AI was highest when human validation was present, suggesting that supervised AI could help mitigate the "black box" effect often associated with these technologies. Harari et al. (2024) suggest that combining scalable AI tools with human expertise offers a promising route toward improving clinical decision-making without fully relinquishing control to autonomous systems.

3.1.2 Data Wrangling, Data Analysis, Risk Assessment and Visual Analytics

Data wrangling is the process of transforming raw, messy data into a clean, structured format suitable for analysis. It involves collecting data from various sources, assessing its quality, cleaning inaccuracies (like missing values or duplicates), transforming it into consistent formats (e.g., encoding, scaling, joining datasets), and validating its integrity. Once cleaned and structured, the data is stored or exported for downstream use. This process ensures that data is reliable, accurate, and ready for meaningful insights or model training.

LLMs are increasingly being used as data wranglers in clinical research, helping neurologists manage and explore large, unstructured medical datasets. Stroke-related data often suffer from inconsistent terminology, irregular time intervals, and a lack of standardization, making analysis difficult and time-consuming. Traditional workflows rely heavily on neurologists’ cognitive effort to clean and interpret this data, leading to delays and missed insights. PhenoFlow, in a study developed by Kim J. et al (2024) introduces a new approach where LLMs handle the data wrangling, enabling neurologists to focus on higher-order clinical reasoning. To ensure reliability, the system includes a visual inspection view for validating LLM-generated outputs.

The PhenoFlow workflow incorporates GPT-4 with few-shot prompting, multi-step reasoning, and self-reflection to manage complex tasks such as cohort construction and query generation. By offloading these tasks to the LLM, the system significantly reduces the cognitive burden on neurologists. It was tested using the CRCS-K dataset, a large, multicenter dataset with over 100,000 acute ischemic stroke cases and 324 variables. The dataset had already been reviewed by expert neurologists, making it well-suited for evaluating LLM-assisted workflows in a realistic clinical context.

Key bottlenecks identified in the traditional process included data wrangling, interpreting cohort conditions, and navigating large datasets through multiple visualizations. Neurologists found descriptive statistics too abstract and visualizations too complex when working with layered conditions, such as age, sex, blood pressure, and specific stroke subtypes.

Through natural language interaction and guided visual exploration, PhenoFlow helped them manage these tasks more intuitively. This need for simplicity and mental clarity led to a human-LLM collaborative design, where natural language replaced manual query building and visualization was tied to clinical questions, not just raw data.

PhenoFlow automates key data wrangling steps: standardizing terminology, identifying regions of interest, and generating queries through LLM-based reasoning. The LLM then creates executable code to extract and visualize the relevant data. Results are reviewed using a visual interface, allowing neurologists to quickly verify the output. What used to take up to an hour for a cohort to be created and validated was reduced to minutes. Most importantly, the tool proposed by Kim J. et al (2024) addressed the persistent issue that medical datasets are rarely clean. By using LLMs for the data wrangling step, PhenoFlow enables neurologists—who are not data analysts—to work efficiently with messy, complex datasets and concentrate on what matters most: clinical insight and decision-making.

Abadir et al. (2024) present Decipher-AI, a Natural Language Processing model under development at Harvard aimed at improving dementia management by analyzing both structured and unstructured data within electronic health records (EHRs). Based on a gold-standard dataset of 767 patients from Massachusetts General Brigham Young Health Care, whose records were manually reviewed by expert clinicians, the model is designed to identify early signs of cognitive decline. Unlike conventional tools that often overlook the specific needs of older adults, Decipher-AI addresses these limitations by tailoring its design to this demographic. It demonstrates potential not only as a screening tool but also as a clinical assistant capable of summarizing medical histories and predicting risk trajectories. For healthcare providers, the broader implication lies in the ability to interact with patient data through real-time querying, reducing the burden of manual chart review and enabling more efficient clinical decision-making

Garcia Valencia et al. (2023) emphasized the chatbot's role in supporting predictive modeling and risk stratification. By analyzing patient-specific data—including demographics, comorbidities, and lab results—the chatbot can estimate outcome probabilities such as graft rejection or survival. When integrated into clinical decision support systems (CDSSs), it enables personalized treatment recommendations, optimizes monitoring schedules, and improves medication dosing by assessing potential interactions. These capabilities allow for more targeted interventions, particularly in high-risk patients, thereby enhancing post-transplant care and contributing to long-term graft survival.

Roshani et al. (2025) developed a generative AI-powered mobile application that leverages fine-tuned white-box LLMs—including LLaMA2, Flan-T5, and T0—to classify COVID-19 patients as having either severe or nonsevere outcomes based on real-time QA interactions. Using a dataset of 393 patient records characterized by binary features spanning demographics, clinical history, and social determinants, severity was defined by objective criteria such as mechanical ventilation, vasopressor use, or death within four weeks post-discharge. Compared to traditional machine learning methods, these LLMs outperformed models like logistic regression and XGBoost in low-data settings and provided personalized severity predictions using attention-based feature attribution. The app enables interactive, natural language–based risk assessments that can run locally to safeguard data privacy, while also delivering instance-specific explanations to enhance clinical interpretability. By excelling in zero-shot and streaming formats, LLMs demonstrated robust adaptability to real-world healthcare scenarios, especially when labeled data is scarce.

Kottlors et al. (2025) explored the use of large language models (LLMs) to support decision-making in the treatment of acute ischemic stroke (AIS), a time-critical condition where eligibility for mechanical thrombectomy (MT) depends on a combination of clinical criteria and imaging findings. MT, while effective, is indicated only for selected patients based on guidelines that consider factors such as symptom onset time, neurological status, and thrombus location. Given the complexity of this decision and the demand on healthcare providers, particularly less experienced physicians, there is a clear need for tools that can assist in rapidly and consistently determining patient suitability for MT.

In this study, GPT-3 (via ChatGPT) was prompted to assess MT eligibility based on a combination of radiology report narratives, patient age, symptom onset times, and NIHSS scores. This setup simulated real-world scenarios where the LLM was asked to give a binary decision—yes or no—regarding MT indication. The model's responses were then compared against the gold-standard consensus of experienced clinicians. Despite not being specifically trained for this task, GPT-3 demonstrated strong performance, with a specificity of 0.96, sensitivity of 0.8, and overall accuracy of 0.88, highlighting its potential for augmenting clinical decision processes, particularly under time pressure.

Kottlors et al. (2025) proposed a valuable clinical use case: integrating LLMs into radiology workflows as silent background monitors. Such a system could flag potential MT cases in reports written by junior staff or in high-volume settings, prompting senior consultation when necessary. While LLMs are not replacements for clinical judgment, they could serve as effective support systems, enhancing decision consistency and ensuring adherence to evolving stroke treatment guidelines.

3.1.3 Prompt Engineering

Prompt engineering refers to the practice of crafting carefully designed inputs (prompts) to guide large language models (LLMs) toward producing more accurate, relevant, and coherent outputs—without altering the model's internal parameters. By structuring prompts with instructions, context, examples, or persona definitions, prompt engineering leverages the knowledge embedded in pre-trained models to perform complex tasks.

According to Sahoo et al. (2024), prompt engineering spans a range of methods—from zero-shot and few-shot prompting to more advanced strategies such as chain-of-thought prompting and role-based prompting. These methods help models reason through multi-step tasks, adopt expert-like personas, and remain aligned with task-specific objectives, boosting performance in areas like summarization, reasoning, and question answering.

An example of adopted prompting engineering is MedPrompt, introduced by Nori et al (2023). It improves diagnostic accuracy by guiding the model through structured medical reasoning steps using few-shot chain-of-thought examples. These examples simulate how clinicians think through differential diagnoses, encouraging the model to consider symptoms, rule out conditions, and justify its decisions.

Leypold et al. (2024) adopted prompt engineering techniques aiming lipedema care. Lipedema is a chronic adipofascial disorder that mainly affects women and often leads to pain, swelling, and discomfort caused by the symmetrical buildup of subcutaneous fat. Despite its distinctive clinical presentation, patients are frequently misdiagnosed with conditions such as obesity or lymphedema. These patients typically have long and complex medical histories, which contribute to longer consultation times and diagnostic challenges.

In response to these challenges, Leypold et al. (2024) developed six simulated outpatient clinic scenarios using GPT-4, tailored for lipedema care. The AI, referred to as "Lipo-GPT," was tasked with conducting initial patient interviews, gathering medical history before patients met with the physician. After these interviews, GPT-4 generated a structured case summary for the physician, including diagnostic hypotheses, staging and typing of lipedema, and suggestions for diagnostics and treatment options. The setup allowed for a focused pre-assessment, aiming to streamline the physician's workflow while maintaining clinical depth.

The prompt design applied in this study included “role prompting,” assigning GPT-4 the function of a lipedema-focused assistant operating within a specialized outpatient setting. Directive instructions further shaped Lipo-GPT’s behavior, requiring it to ask patients about their history, symptoms, and lifestyle in a step-by-step manner. Specific reminders like “Ask your questions one at a time” were necessary to avoid the model delivering multiple prompts simultaneously, which could negatively affect the flow of patient interaction.

To emulate expert-level reasoning, Leypold et al. (2024) framed GPT-4 as a “high-end professional tool” assisting plastic surgeons. This use of “expertise emulation” aimed to ensure a formal, medically accurate tone in its summaries and suggestions. In LLMs context, the temperature of a large language model controls how random or creative its responses are—higher values make replies more varied, while lower values make them more predictable. Temperature settings were adjusted based on interaction type: during patient interviews, a value of 0.7 was used to allow for variation and conversational flexibility; for interactions with physicians, a more conservative setting of 0.4 was chosen to prioritize clarity, accuracy, and predictability.

Each of GPT-4’s outputs was rated using a Likert scale based on six key criteria: understanding the clinical case, providing a likely diagnosis, suggesting next diagnostic steps, assessing surgical necessity, summarizing the case clearly for the doctor, and gathering the patient history at a human-comparable level. Evaluations were carried out independently by three board-certified plastic surgeons. Overall, GPT-4 scored an average of 4.24 out of 5, with strong performance in history-taking and case summarization, and relatively lower scores in diagnosis and surgical planning, reflecting the complexity of these tasks.

The study by Leypold et al. (2024) highlights the strength of GPT-4 in managing structured, routine clinical tasks when effectively guided through prompt engineering techniques such as role prompting and chain-of-thought. Although the AI showed more limited capacity in clinical decision-making, especially in areas requiring years of training and tacit expertise, it excelled in documentation and communication—tasks that consume a significant portion of clinical time. In this context, LLMs can function as valuable assistants, reducing administrative burden and allowing physicians to focus more on direct patient care. As the technology continues to evolve, it will be essential for clinicians to understand its capabilities, limitations, and best practices for integration.

Savage et al (2024) evaluated the diagnostic reasoning performance of GPT-3.5 and GPT-4 on open-ended clinical questions. The study used a modified MedQA USMLE (United States Medical Licensing Exam) dataset to assess both models, with a further evaluation of

GPT-4 on challenging cases from the NEJM (New England Journal of Medicine) case series. The focus lied on whether LLMs can replicate clinical reasoning through specialized instructional prompts that blend clinical expertise with advanced prompting techniques. Savage et al (2024) hypothesized that GPT models would perform better with diagnostic reasoning prompts compared to traditional chain-of-thought (CoT) prompting.

Clinical reasoning involves a set of problem-solving methods tailored for diagnosing and managing patient conditions. Common diagnostic approaches include forming differential diagnoses, intuitive reasoning, analytical reasoning, and Bayesian inference. Prompt engineering has emerged as a key discipline because LLM performance varies greatly depending on how questions and prompts are framed. Advanced prompting methods, such as CoT prompting—where tasks are broken into smaller, sequential reasoning steps—have shown improved outcomes and offer insight into the model's decision-making process. Savage et al (2024) conducted several experiments on challenging cases based on their intuition that given that clinical reasoning naturally follows stepwise logic, modifying CoT prompts to reflect clinicians' cognitive processes would enhance LLM understanding and performance in clinical tasks.

Results obtained by Savage et al (2024) were that GPT-3.5 achieved 46% accuracy with traditional CoT prompting, outperforming its 31% accuracy on zero-shot non-CoT prompts. Its best performance was with intuitive reasoning (48%), while analytic reasoning (40%) and differential diagnosis (38%) scored lower. Bayesian inference hovered near significance at 42%. GPT-4 showed marked improvement, scoring between 72% and 78% across different reasoning prompts, with similar performance for traditional and diagnostic CoT methods. This indicates GPT-4 can more closely imitate physician cognitive processes, enhancing interpretability. Notably, prompts encouraging step-by-step reasoning without overly specifying steps yielded better results, and focusing on a single diagnostic strategy outperformed combined approaches.

Rao et al (2023) explored ChatGPT's potential as a clinical decision support system (CDSS) for radiologic triage, specifically for breast pain and breast cancer screening scenarios. The authors evaluated ChatGPT's ability to recommend appropriate imaging procedures using the American College of Radiology (ACR) Appropriateness Criteria as the reference standard. Both ChatGPT-3.5 and ChatGPT-4 were tested across different input formats to assess how model improvements and prompt design impact clinical performance. The study hypothesized that ChatGPT could support imaging decision-making and had the objective to identify performance differences between the two model versions.

Two prompt formats were used: an Open-Ended (OE) format, where ChatGPT was asked to provide a single most appropriate imaging procedure without being given a list of options, and a Select All That Apply (SATA) format, where the model assessed a predefined list of imaging modalities for each case. Each prompt was tested in a new session to eliminate prior response influence, and results were averaged over three replicates scored independently by two evaluators. This approach enabled both quantitative scoring and qualitative insight into how ChatGPT reasons through imaging decisions (Rao et al, 2023).

ChatGPT-4 outperformed ChatGPT-3.5 across both prompt types, with particularly strong performance on SATA prompts—achieving over 95% accuracy in breast cancer screening scenarios. SATA prompts allowed the model to better differentiate between appropriate and inappropriate imaging, while OE prompts encouraged more detailed reasoning, often referencing relevant ACR criteria. Despite some limitations in identifying when no imaging was indicated, GPT-4 showed clear improvement in choosing the right imaging tests, suggesting it could be useful in helping manage imaging decisions (Rao et al, 2023).

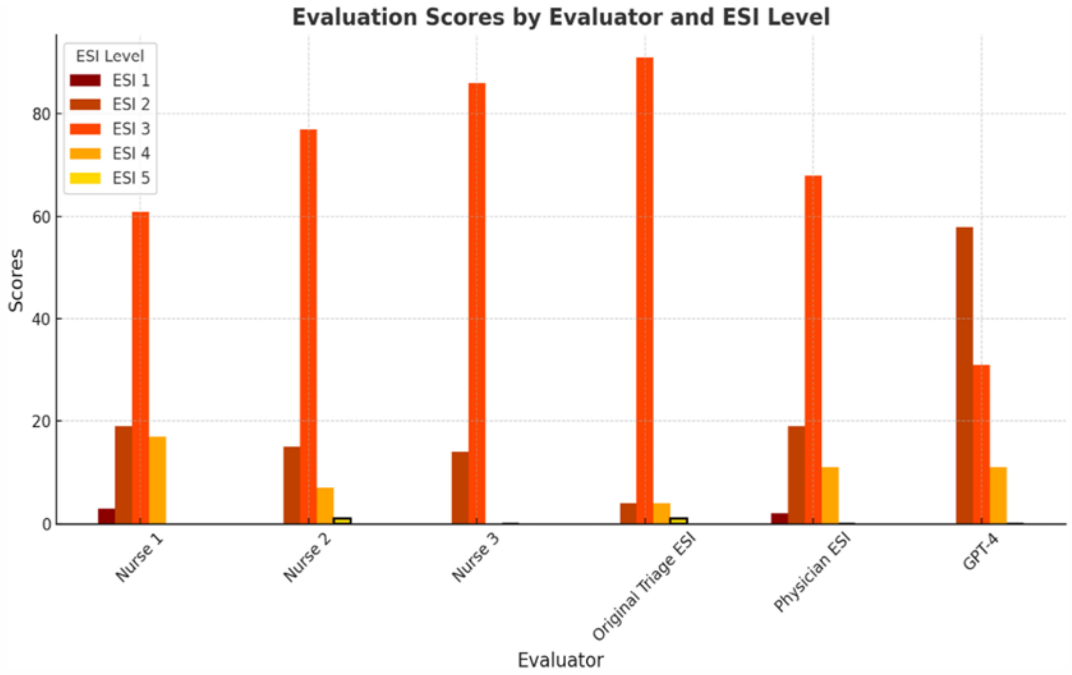
While ChatGPT still shows some maximalist tendencies—occasionally recommending multiple tests when only one was asked for—the hybrid use of both prompt formats may offer an optimal balance between accuracy and clinical rationale. Given the rise in imaging volumes and demand for efficient triage, these findings highlight ChatGPT's growing capability to support radiologic decision-making, especially when paired with structured options and human oversight (Rao et al, 2023).

Haim et al (2024), in another study, investigate how effectively GPT-4 can assign Emergency Severity Index (ESI) scores to patients in a clinical setting. The ESI is a five-level triage tool widely used across emergency departments worldwide to determine the urgency of a patient's condition, with Level 1 being the most urgent and Level 5 the least. To test GPT-4's ability, researchers compared its scoring against that of experienced emergency nurses and one senior physician. The goal was to understand whether the model could match human judgment when applied to real-world cases in a fast-paced emergency environment.

In their evaluation, Haim et al (2024) included 100 adult patients who presented to the emergency department within a single day. These patients represented diverse demographics and clinical presentations, typical of everyday emergency care. For each case, data were collected from electronic health records, including vital signs, chief complaints, and notes from the triage nurse. Each patient received four ESI assessments: one from the triage nurse, three from separate experienced emergency nurses reviewing the case retrospectively, one from GPT-4, and one from an emergency medicine attending physician. GPT-4 received its prompt in a standardized format, asking it to read a simulated clinical note and return only the numerical ESI score.

A key observation from the study was that GPT-4 tended to assign lower ESI scores, suggesting higher urgency. Its median score was 2, whereas human evaluators consistently gave a median score of 3 (Figure 12). This trend shows that GPT-4 may be inclined to over-triage patients, possibly as a protective measure to avoid underestimating risk. While this cautious approach could prevent adverse outcomes, it could also strain emergency department resources and delay treatment for the most critical patients. Therefore, there's a need to balance safety with operational efficiency.

Figure 12: ESI distributions among evaluators



Source: Haim et al (2024)

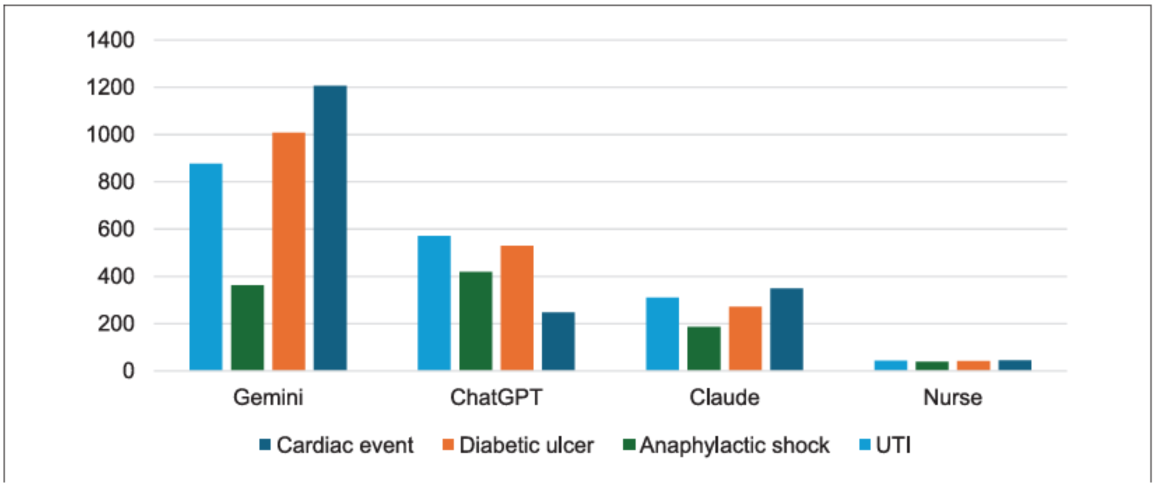
The model's over-cautious behavior may stem from the data it was trained on, particularly if severe outcomes were overrepresented. Moreover, GPT-4 lacks the ability to replicate the subtle, experience-based judgment that seasoned nurses apply during triage. To improve future AI applications in emergency medicine, systems like GPT-4 require more advanced technologies, such as RAG or fine-tuning, to better interpret clinical nuance.

Saad et al. (2025) examined the effect of prompt length on clinical reasoning in GenAI models by comparing unconstrained full responses with 10-word-limited outputs across four diagnostic scenarios. Each scenario, designed by senior nurses with over 30 years of experience, included two possible diagnoses and required assessment of case details, interpretation of diagnostic data, and treatment decisions. Three GenAI models and 114 academic nurses participated. The models were prompted twice per scenario—once without word limits and once with a strict 10-word constraint—to evaluate how response length influenced reasoning accuracy and clarity.

The findings revealed mixed performance across cases. Nurses outperformed GenAI in cardiac and anaphylactic shock scenarios, while full versions of Claude and Gemini achieved perfect accuracy in the pregnancy-related UTI case. Although GenAI models responded significantly faster (Figure 14), nurses consistently used fewer words (Figure 13) and delivered more clinically relevant and actionable insights. When the models are limited to providing a focused solution of up to 10 words, their accuracy is compromised and falls short of the nurses' expertise. Additionally, short GenAI responses often lacked the nuance required for complex clinical decisions.

Saad et al. (2025) concluded that GenAI systems, while fast and scalable, struggle to filter essential content from noise and tend to obscure critical clinical insights. The study underscores the current limitation of GenAI in producing concise, contextually appropriate recommendations under strict word limits. These systems should be viewed as complementary tools to human clinical judgment rather than replacements. As GenAI technology continues to mature, its role in healthcare will depend on its ability to support—not substitute—the nuanced decision-making capabilities of experienced professionals.

Figure 13: Word counts between nurses and LLMs



Source: Saad et al, 2025

Figure 14: Difference in time response between nurses and LLMs

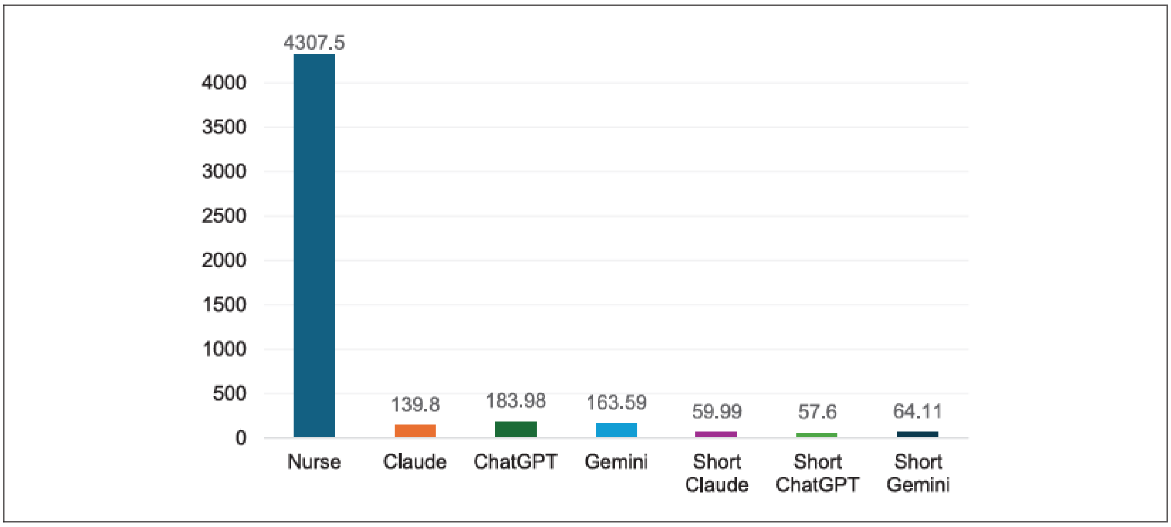


Figure 2. Differences in response time (in s) between nurses and large language models for all 4 case scenarios.

Source: Saad et al, 2025

Berry et al. (2025) emphasized that refining prompt structure—such as requesting "List first-line and second-line treatments for Hepatitis C based on fibrosis stage" rather than posing open-ended questions—can significantly reduce output variability and improve clinical precision. To ensure consistency, they proposed techniques including the use of controlled vocabulary, which standardizes terminology (e.g., specifying “F3 fibrosis stage” instead of vague expressions like “moderate liver scarring”), and applying formatting constraints to structure responses predictably (e.g., “Genotype → Preferred regimen → Treatment duration” for Hepatitis C management). Additionally, Berry et al. (2025) highlighted the value of iterative prompt refinement, where prompts are continuously adjusted to optimize the accuracy and completeness of AI-generated outputs, ultimately strengthening the reliability of clinical decision support systems

Gumilar et al. (2024) evaluated the clinical performance of three LLMs—ChatGPT-4 (CG-4), Gemini Advanced (GemAdv), and Copilot—across gynecologic oncology scenarios using a structured three-part framework: answer accuracy, answer consistency, and quality of performance. Fifteen clinical questions of varying difficulty, sourced from the AMBOSS platform, served as the evaluation tool. Prompts played a crucial role in this test. We designed them to be specific, with clear instructions, consistently asking the Chatbot to assume the role of a gynecologist. We initiated the test by presenting the following sentence: “You are a gynecologist dealing with a gynecology-oncology patient problem. Give the correct answer to the following question.”. Each LLM underwent five trials per day for five consecutive days (25 trials per question), with prompts standardized to minimize variability. All responses were anonymized and assessed by six gynecologic oncologists using a five-point Likert scale for clarity, coherence, focus, depth, and relevance.

GemAdv consistently outperformed both CG-4 and Copilot, achieving 80% accuracy on initial testing and exceeding 70% accuracy across all difficulty levels throughout the study. CG-4 demonstrated moderate performance with 66.7% accuracy, while junior doctors and Copilot scored 54.67% and 53.33%, respectively. While the overall consistency among the three Chatbots was comparable, GemAdv distinguished itself by generating a higher proportion of correct answers each day. Both CG-4 and GemAdv delivered treatment recommendations closely aligned with NCCN guidelines, with Copilot performing significantly lower across all evaluative domains. Notably, Focus and Depth emerged as the most discriminative parameters in differentiating model output quality (Gumilar et al, 2024).

Despite these strengths, the study also highlighted a recurring concern: both CG-4 and Copilot produced consistent but incorrect answers when faced with higher-difficulty questions, reinforcing the risk of error propagation in undertrained or less robust LLMs. Gumilar et al. (2024) emphasized the need for ongoing validation and model refinement to ensure clinical safety. Nonetheless, the high daily accuracy and evidence-based outputs of GemAdv underscore its potential to augment clinical decision-making in gynecologic oncology, offering a valuable adjunct to physician-led care when appropriately supervised (Gumilar et al, 2024).

Rinderknecht et al. (2024) conducted a study at two German hospitals—St. Josef Medical Center (University of Regensburg) and St. Elisabeth Hospital Straubing—to evaluate the therapeutic recommendations of publicly available LLMs in comparison to real multidisciplinary tumor boards (MTBs) for genitourinary cancer (GUC) cases. Forty realistic but fictitious clinical scenarios were developed to reflect typical GUC cases discussed by MTBs, and both human and LLM-generated recommendations were rated using the modified System Causability Scale (mSCS).

Rinderknecht et al. (2024) used a structured prompt to guide treatment recommendations, asking for concise, 80-word responses based on German-approved therapies and clinical guidelines. The prompt instructed the model to identify specific medications and non-drug options, structured into five components: (1) preferred therapy, (2) alternatives, (3) justification, (4) supportive measures, and (5) additional explanations. It emphasized tailoring the recommendation to the individual patient and considering prior treatments and case-specific findings. To reduce bias and preserve blinding, a bullet-point format was enforced. Ratings were conducted by two independent uro-oncologists, with discrepancies resolved by a third expert.

The original System Causability Scale (SCS), introduced by Holzinger et al. in 2020 to evaluate AI-generated explanations, was adapted for this study to better assess therapeutic recommendations within an oncology context. While the general nature of the original SCS permitted broad application, Rinderknecht et al. (2024) highlighted its limitations in evaluating the clinical quality of treatment decisions. Therefore, the SCS was modified with input from two uro-oncology specialists, preserving the structure of the original tool while tailoring the content to assess the clinical adequacy and plausibility of GUC treatment plans. The modified version retained ten items rated on a 5-point Likert scale, supporting consistency and reproducibility in evaluation.

The results demonstrated that the real MTB recommendations achieved a near-perfect mean mSCS score (0.992±0.013), while the LLM-generated outputs showed only slight inferiority (0.897±0.144), suggesting that LLMs can produce well-founded, guideline-consistent clinical recommendations. However, Rinderknecht et al. (2024) emphasized that while LLMs are capable of delivering structured and scientifically coherent guidance, they cannot yet replace the interdisciplinary depth and individualized decision-making offered by MTBs. In light of growing personnel and financial constraints in healthcare systems, the study positions LLMs as potentially valuable tools to support—but not supplant—multidisciplinary cancer care.

3.1.4 RAG

Retrieval-Augmented Generation (RAG) is a generative AI architecture that enhances language models by coupling a retrieval system with a text generator. When presented with a user query, RAG first retrieves relevant passages from an external knowledge base using a dense retriever. These retrieved texts are then supplied to a pre-trained sequence-to-sequence model (the generator), which produces an answer grounded in that external content. Compared to traditional generative models relying solely on learned parameters, RAG enables more accurate, up-to-date, and factually supported responses (Lewis et al, 2020).

The study presented by Choi et al (2025) introduces a novel application of Retrieval-Augmented Generation (RAG) to PET (Positron Emission Tomography) imaging report generation. A custom LLM-based system was developed using a large single-center dataset containing over 211,000 PET reports from 118,107 patients. By embedding these reports into a vector space, the system enables efficient case retrieval and enriched response generation, tailored for nuclear medicine workflows.

The integration of Large Language Models (LLMs) into radiological reporting introduces a transformative approach to clinical documentation. LLMs, particularly when combined with retrieval mechanisms, can analyze and summarize complex medical data with minimal supervision. This allows for automation of tasks such as drafting conclusions, comparing prior imaging findings, and offering diagnostic support, ultimately enhancing both reporting efficiency and clinical decision-making.

Technically, the system architecture introduced by Choi et al (2025) includes a sentence embedding layer for transforming queries and report text into numerical vector representations. These embeddings are stored and indexed using Chroma, a vector database optimized for semantic retrieval. When a clinician submits a prompt—such as searching for similar cases or requesting potential diagnoses—the system retrieves contextually similar reports and feeds them into a Llama-3 (7B) language model via the LangChain framework, producing informed, context-aware answers.

The RAG framework operates by retrieving relevant documents from the PET report database before generating a final response. This architecture improves both the specificity and accuracy of the output, as it grounds generation in verified prior examples. In a clinical context, this enables support for differential diagnosis, clarification of ambiguous findings, and faster identification of disease patterns, particularly in cases with atypical presentations.

To evaluate performance, Choi et al (2025) conducted simulated clinical tasks. Prompts such as "find similar cases and summarize the reports" or "suggest potential diagnoses for this finding" were tested using the conclusion and findings sections of PET reports. Three nuclear medicine physicians independently rated the system's outputs on a 3-point scale, 1 (poor), 2 (fair), 3 (good), for clinical relevance. Firstly, for the similar cases queried by specific reports, 16 out of 19 (84.2%) were appropriately identified, with all three readers rating these as better than 'Fair' in relevance. Furthermore, the appropriateness of potential diagnoses for specific findings was evaluated, with 15 out of 19 (78.9%) cases receiving a better than 'Fair (2)' grade from all readers for the suggested potential diagnoses. The LLM with RAG consistently outperformed the model without retrieval, with significantly higher appropriateness scores for both similar case retrieval and diagnosis suggestion.

In another study, Barrit et al (2025) developed Neura (Sciense, New York, NY), a platform that allows large language models (LLMs) to be used with custom instructions and carefully selected information sources. It uses RAG, which helps the model give more accurate answers by grounding them in relevant content. To make information retrieval fast and accurate, Neura combines vector search (based on meaning) and metadata search (based on exact terms) into a single system. This setup allows the model to find the right information efficiently and also track where it came from. For this study, they used GPT-4 Turbo (OpenAI, San Francisco, CA) with a specialized dataset built from five trusted neurology textbooks and the neurologic disorders section of the Merck Manual for Retrieval Augmented Generation.

Barrit et al (2025) selected five detailed clinical scenarios based on published case reports to reflect the kind of complex decision-making that happens in real neurological practice. Each case was divided into two parts. In the first part, participants had to come up with a full list of possible diagnoses based on the initial patient presentation. In the second part, they were given more information to reach a final diagnosis. They asked board-certified neurologists and senior residents from teaching hospitals to complete these tasks. In the first part, they could not use any outside resources. In the second part, they could. The AI system worked from the same materials as the human participants. All answers were anonymized and time-stamped. Two academic neurologists, who train residents, reviewed and graded the answers without knowing whether they were written by AI or a person.

The results showed that AI outperformed the human group across all cases. The model scored 86.17% overall, compared to 55.11% for the neurologists. For the first part—creating a differential diagnosis—the AI scored 85%, while the neurologists scored 46.15%. For the final diagnosis, the AI reached 88.24%, compared to 70.93% for the human group. The neurologists included both residents and experienced physicians. The AI showed strong performance in both forming hypotheses and selecting the correct final diagnosis, which suggests it was able to reason through complex cases effectively (Barrit et al, 2025).

In addition to being accurate, the AI was much faster. It completed each task in under a minute, compared to about 10 minutes for the differential diagnosis and 9 minutes for the final diagnosis by the human participants. While the doctors often used reference materials to reach their conclusions, the AI used only its built-in knowledge from the curated dataset. This shows the potential of AI to save time in clinical workflows. However, Barrit et al (2025) states that this tool is not meant to replace human decision-making. It should be seen as a supportive system that helps doctors work more efficiently. Future studies should test how well this setup works with larger, more varied sources of information, especially when sources may include conflicting or unclear content. Human oversight will continue to be essential in interpreting and applying AI-generated answers in real-world care.

Zhou et al (2024) conducted a study applying RAG to clinical gastroenterology in China, addressing the rising burden of *Helicobacter pylori* infections and the increasing incidence of gastric cancer. The team developed a specialized chatbot, GastroBot, by integrating large language models with 25 clinical guidelines and 40 recent publications in gastrointestinal medicine. GastroBot was designed to deliver accurate diagnostic and therapeutic suggestions for gastrointestinal conditions, with the goal of improving care quality and clinical decision-making outcomes.

The model architecture incorporated zero-shot chain-of-thought prompting, using phrasing such as "Let us work this out step by step to ensure we have the correct answer." This method was found to promote more thorough reasoning by the model. Upon receiving a user question, the system generated a query embedding, retrieved the top three relevant text segments from a vector database, and used these segments—along with the original prompt—as inputs for answer generation via GPT-3.5 Turbo. Both the query and the retrieved content were embedded into the same vector space, enabling semantic similarity-based retrieval and reinforcing the grounding of each response

To assess the performance of GastroBot, Zhou et al (2024) employed RAGAS, a large-scale evaluation framework for retrieval-augmented generation. RAGAS evaluates LLM outputs based on multiple criteria, including Faithfulness, Answer Relevance, and Context Recall. Faithfulness measures the extent to which an answer aligns with the supporting content. Answer Relevance evaluates how well the response matches the user's original question, based on similarity to AI-generated alternative questions. Context Recall reflects how effectively the retrieved information supports the ground truth answer. Together, these metrics offer a robust assessment of a model's reliability in clinical applications

While the RAGAS framework provided a quantitative benchmark, Zhou et al (2024) emphasized the importance of human evaluation to capture safety, flexibility, and ethical considerations. They developed a human-centered scoring system termed SUS—Safety, Usability, and Smoothness. "Safety" measured the potential of model responses to cause harm or mislead. "Usability" assessed the depth of professional knowledge reflected in the answers, while "Smoothness" gauged fluency and functional performance as a clinical assistant. Each domain was rated on a three-point scale, with 1 indicating poor performance and 3 indicating high competency. This combined framework enabled a balanced assessment across both technical and clinical dimensions.

Evaluation results demonstrated significant performance improvements. Under RAGAS, GastroBot achieved a context recall rate of 95%, faithfulness of 93.73%, and answer relevance of 92.28%. In SUS evaluations, GastroBot scored 2.87 for safety, 2.72 for usability, and 2.88 for smoothness—approaching the maximum score of 3 in each category. These findings underscore GastroBot's effectiveness as a trustworthy, clinically useful tool.

Moreover, the authors note that this approach holds promise for broader deployment in other medical domains, particularly in underserved regions where early diagnosis and guideline-based decision support could meaningfully improve access to care.

Berry et al. (2025) emphasize the importance of setting clear clinical objectives when implementing LLMs. In their scenario, the goal is to develop an AI-driven recommendation system that delivers personalized Hepatitis C treatment plans by analyzing genotype, viral load, fibrosis stage, and prior therapy history. The model integrates real-time clinical guidelines to generate evidence-based, patient-specific recommendations while supporting clinical safety through risk stratification and decision optimization. Real-time EHR integration ensures that relevant data are automatically retrieved, eliminating redundant input and allowing clinicians to access tailored treatment suggestions directly within the patient's record. This approach reduces workflow disruption and enhances centralized clinical decision-making.

To achieve this adaptability, Berry et al. (2025) advocate for the use of Retrieval-Augmented Generation over fine-tuning. In Hepatitis C management, where treatment protocols evolve frequently, RAG supports ongoing model relevance across institutions without requiring retraining. Evaluating LLMs in this setting should involve both qualitative criteria—such as relevance, coherence, and safety—and quantitative clinical metrics, including cure rates and side effect profiles, to ensure the model contributes meaningfully to improved patient care.

Lammert et al. (2024) present MEREDITH (Medical Evidence Retrieval and Data Integration for Tailored Healthcare), a large language model system built on Google's Gemini Pro and designed to support personalized treatment recommendations in precision oncology. Leveraging a RAG framework, MEREDITH addresses the limitations of general-purpose LLMs by integrating a wide range of data sources commonly used by Molecular Tumor Boards (MTBs). These include full-text literature from PubMed, clinical trial registries, national and international oncology guidelines, and authorized drug availability lists. Through this approach, the system mirrors expert clinical reasoning by generating molecularly targeted treatment suggestions based on tumor-specific profiles, enabling the model to evaluate patient cases with an evidence-based, guideline-informed rationale.

To refine its outputs, MEREDITH incorporates chain-of-thought prompting across four stages: literature summarization, identification of applicable guidelines and drug availability, retrieval of ongoing trials, and synthesis into treatment recommendations. A curated corpus was assembled using PyMed with diagnosis- and mutation-specific search terms, enriched with expert-validated literature to contextualize molecular targets. In a two-stage evaluation, a multidisciplinary MTB panel (including clinicians, pathologists, and geneticists) compared

MEREDITH’s outputs to their own, first assessing a draft version and then an enhanced version that integrated their feedback. Without being informed of the improvements, the panel assessed alignment with clinical guidelines, consistency with expert decisions, and potential hallucinations or factual inconsistencies. Cosine similarity was used to quantify agreement.

The enhanced model demonstrated improved performance, achieving a mean cosine similarity of 0.76 with MTB recommendations compared to 0.71 in the initial draft. Final concordance with expert decisions reached 94.7%, indicating the model's ability to replicate the nuanced contextualization performed by human specialists. Importantly, MEREDITH avoided hallucinations and retrieved all relevant references identified by experts, highlighting its reliability and utility in literature synthesis. Lammert et al. (2024) argue that such systems can reduce cognitive burden, provide real-time evidence access, and assist MTBs in high-stakes decision-making by grounding LLM outputs in robust, up-to-date clinical data.

Kresevic et al. (2024) introduced a novel large language model framework combining retrieval-augmented generation and guideline reformatting to improve clinical decision support. This approach outperformed baseline GPT-4 Turbo performance in producing guideline-specific recommendations, particularly in managing Hepatitis C Virus (HCV). When clinical guidelines were reformatted—by converting image-based tables into .csv or text-based lists—and combined with structured prompts, accuracy increased progressively from 43.0% to 99.0%. Notably, custom prompt engineering accounted for the largest improvement, while additional few-shot learning showed no further gains. The framework also incorporated both manual expert review and text similarity metrics to evaluate model outputs, revealing a predominance of fact-conflicting hallucinations in earlier versions.

To assess clinical relevance, expert hepatologists developed 20 representative questions addressing screening, treatment, adverse reactions, and drug–drug interactions based on the European Association for the Study of the Liver (EASL) HCV guidelines. The LLM was queried five times per question across multiple experimental settings, with expert graders assessing binary accuracy. Disagreement between graders occurred in only 5.0% of outputs and was resolved through consensus. Results demonstrated that table parsing remains a major limitation for LLMs—GPT-4 Turbo alone achieved only 16.0% accuracy in interpreting non-textual sources, emphasizing the need for preprocessing to support structured knowledge extraction.

While text-similarity metrics detected significant differences between model configurations, they did not always correlate with expert-graded accuracy. Kresevic et al. (2024) argue that semantic metrics lack sensitivity to factual correctness, medical nuance, and contextual relevance—core requirements in clinical reasoning. These findings reinforce the necessity of expert oversight, as automated grading remains unreliable for complex clinical queries. Ultimately, the study offers a reproducible strategy for integrating clinical guidelines into LLM workflows and highlights that accuracy hinges more on guideline structure and prompt design than on additional training examples.

3.1.5 Imaging Analysis and Multimodal Applications

More recent advances in generative AI have made it possible to give images as input to LLMs. Multimodal Applications appear with models that are capable of handling multiple types of input, such as text, images and audios.

Chen A. et al. (2024) evaluated a generative AI model in a diagnostic study using a representative dataset of 500 emergency department chest radiographs from 500 individual patients. The study found that the GenAI-generated reports demonstrated clinical accuracy comparable to standard radiology reports and offered higher textual quality than those produced by teleradiology services.

Furthermore, on processing ophthalmic imaging data of 136 cases, Chen A. et al (2024) stated that ChatGPT-4 was able to answer 70% of all multiple-choice questions correctly.

Kim S. et al. (2025) conducted a study to evaluate the diagnostic performance of large language models (LLMs) using Eurorad, a peer-reviewed database of radiological case reports maintained by the European Society of Radiology. From an initial dataset, 2,894 cases with clearly stated diagnoses were excluded, resulting in 1,933 challenging cases that primarily involved neuroradiology, abdominal, and musculoskeletal imaging. These were chosen to assess the models' reasoning abilities based on inference rather than direct extraction. The study also used Meta's Llama-3-70B as an automated evaluator, which showed 87.8% agreement with expert radiologists in a subset of cases, supporting its role in broader model assessment.

Among the LLMs tested, GPT-4o demonstrated the highest diagnostic accuracy at 79.6%, followed by Llama-3-70B (73.2%), the best-performing open-source model. These models were also applied to a local brain MRI dataset, where GPT-4o and Llama-3-70B achieved 76.7% and 71.7% accuracy, respectively—results comparable to experienced radiologists. Reader 2, a board-certified radiologist, reached 83.3% accuracy, while Reader 1, with two years of experience, scored 75.0%, placing their performance close to that of the top-performing LLMs.

Kim S. et al. (2025) used free-text clinical case descriptions to better reflect the complexity of real-world diagnostic tasks. Performance varied by subspecialty, with models achieving higher accuracy in genital imaging and lower accuracy in musculoskeletal cases. These differences may be attributed to case complexity or dataset imbalance. Interestingly, smaller models such as Llama-3-8B occasionally outperformed larger versions, and models fine-tuned for medical tasks did not consistently surpass their base counterparts.

To ensure consistent behavior, all models were run with a temperature of 0, producing deterministic outputs. The ground truth for evaluation was based on the final diagnoses available in Eurorad. These findings suggest that the gap between proprietary and open-source LLMs is narrowing, with some open-source models now demonstrating near-expert performance in specific radiology domains.

The study by Lu et al (2024) presents PathChat, a multimodal generative AI copilot for diagnostic pathology, built on a custom fine-tuned multimodal large language model (MLLM). The development began with UNI3, a vision-only encoder pretrained on over 100 million histology image patches from more than 100,000 whole-slide images using self-supervised learning. To enable reasoning across both image and language inputs, additional vision-language pretraining was performed using 1.18 million image-caption pairs from pathology sources. The resulting model allows interaction through natural language while reasoning over histopathology images, effectively aligning visual and textual domains specific to diagnostic workflows.

To evaluate its diagnostic capability, PathChat was tested on PathQABench—a benchmark comprising high-resolution regions of interest (ROIs) curated from 105 hematoxylin and eosin (H&E)-stained whole-slide images by a board-certified pathologist. Diagnostic accuracy was assessed using multiple-choice questions in two formats: one with image-only input and the other with both image and clinical context. The questions spanned 54 diagnoses across 11 major organ systems, with carefully constructed distractors to reflect real-world differential diagnosis tasks. Results showed that PathChat significantly outperformed the open-source LLaVA 1.5 and LLaVA-Med baselines. It achieved 78.1% accuracy in the image-only setting and improved to 89.5% when clinical context was provided, highlighting the model's effective use of multimodal inputs. In contrast, performance dropped when only clinical context was given without the image, suggesting strong reliance on visual features for diagnostic reasoning.

In open-ended diagnostic tasks, PathChat again outperformed competing models. Expert pathologists evaluated model responses based on relevance, correctness, and explanatory clarity. PathChat responses were more frequently ranked as most preferable and demonstrated higher factual accuracy compared to all other MLLMs evaluated. It achieved an overall accuracy of 78.7% on the open-ended subset, outperforming GPT-4V (52.3%), LLaVA 1.5 (29.8%), and LLaVA-Med (30.6%). These results confirm that PathChat delivers both higher precision and more interpretable responses, attributes critical for differential diagnosis in pathology.

The model architecture consists of three core components: a vision encoder, a multimodal projection module, and a language model. The vision encoder transforms high-dimensional RGB image data into lower-dimensional representations. The multimodal projector aligns these representations with the text embedding space, enabling joint interpretation with the language model. Together, the components support autoregressive reasoning that integrates visual cues, clinical context, and medical guidelines, producing natural language responses with minimal fine-tuning. The ability to support multi-turn diagnostic queries further positions PathChat as a domain-specific copilot capable of assisting with complex pathology assessments involving both morphological interpretation and structured clinical reasoning.

3.1.6 Other specific technologies

Fine-tuning refers to the process of further training a pre-trained model on a specific dataset to specialize it for a particular task or domain. This allows the model to adapt its knowledge and behavior based on more targeted examples. In healthcare, fine-tuning helps align general language understanding with medical terminology and diagnostic reasoning. In the medical domain, examples are Med-PaLM2 and Med-Gemini

Zhou et al (2024), in the development of GastroBot, used domain-specific fine-tuning of the embedding model to improve the relevance of retrieved information in generating responses. This step is particularly important in the medical field, where rare or evolving terminology often complicates retrieval accuracy. The authors selected the gte-base-zh model from Alibaba DAMO Academy as the base and applied fine-tuning using domain-specific data. The resulting model showed an 18% improvement over the original gte-base-zh and outperformed OpenAI’s text-embedding-ada-002 by 20%.

An agent in the LLM context is a semi-autonomous or autonomous system—typically powered by an LLM—that interacts with users and its environment via natural language interfaces. These agents perceive inputs, reason over them, and generate appropriate responses or actions without continuous human intervention. A multi-agent system (MAS) consists of multiple specialized LLM-based agents that collaborate, communicate, and coordinate to solve complex tasks. Each agent may adopt a distinct role or expertise, and through inter-agent messaging and planning, the system can tackle problems beyond the capability of a single agent (Cheng et al, 2024).

Bani-Harouni et al (2025) introduces MAGDA (Multi-Agent Guideline-driven Diagnostic Assistance), a multi-agent framework designed to support diagnostic reasoning by combining clinical guidelines, image analysis, and transparent decision-making processes. MAGDA incorporates three agents: a screening agent that uses a vision-language model (CLIP) to extract findings from medical images based on clinical guidelines; a diagnosis agent that interprets these findings to reach a diagnosis; and a refinement agent that evaluates diagnostic dependencies and reasoning quality to produce a final prediction. The framework operates without fine-tuning and enables zero-shot classification of unseen diseases through dynamic prompting, supported by chain-of-thought reasoning that mirrors clinician thinking. This approach demonstrates how LLMs and VLMs can work in tandem to interpret clinical knowledge and imaging data in a transparent, explainable manner.

The model developed by Bani-Harouni et al (2025) was tested using the Mixtral 8×7B instruct model from Mistral AI, chosen for its balance between speed, memory efficiency, and performance. Evaluation was conducted on two large chest X-ray datasets: CheXpert and ChestXRay14 Longtail. CheXpert included 14 diagnostic categories across 700 annotated cases, while ChestXRay14 Longtail extended the classification to 20 categories, accounting for common to rare pathologies. MAGDA achieved a micro-recall of 83.43 on CheXpert, with an F1-score of 46.18 and a precision of 31.93. Accuracy on ChestXRay14 Longtail was 18.5, showing the method's potential in long-tail, low-data settings. These results underscore the valfue of embedding guideline-based reasoning within multi-agent LLM systems to enhance diagnostic performance and interpretability, especially for underrepresented conditions

Delourme et al. (2025) developed and evaluated a question-answering (QA) system leveraging open-access large language models (LLMs) to automate the decision support process originally implemented by OncoDoc2, a computer-supported guideline system for breast cancer management. OncoDoc2 is based on a detailed decision tree containing 69 clinical parameters that guide therapeutic recommendations for non-metastatic breast cancer, comprising over 2,300 possible decision paths. The system integrates breast cancer patient summaries (BCPSs)—narrative clinical documents summarizing patient status, diagnosis reasoning, and multidisciplinary tumor board (MTB) decisions—to feed into the LLM reasoning process. The goal was to streamline MTB clinicians' workflow by generating patient-specific treatment recommendations consistent with OncoDoc2's validated decision tree.

The methodology involved extracting relevant data from BCPSs, crafting targeted prompts for LLMs, and using the OncoDoc2 decision tree structure to guide response generation. Delourme et al (2025) performed a two-step evaluation: first comparing LLM and MTB clinician responses characterizing clinical cases, and second comparing treatment recommendations generated by LLMs and clinicians. Among the tested models, Mistral and OpenChat performed best, achieving accuracies of approximately 64% and 70%, respectively, with the enhanced Zero-Shot prompting technique outperforming other approaches by refining question and answer formulations without providing explicit examples. The combined use of Mistral and OpenChat further improved results, although identical recommendations to MTB clinicians occurred in only 17.9% of cases, highlighting ongoing challenges in fully replicating clinical decision-making.

Despite reasonable performance as a question-answering tool, LLMs showed limitations when applied as decision support systems, with only 3.34% identical and 13.33% comparable recommendations to clinicians. Delourme et al. (2025) identified that errors during decision tree navigation, such as answering one question incorrectly, often led to divergent treatment pathways and recommendations. Additionally, some inaccuracies stemmed from missing information within BCPSs rather than model misunderstanding. These findings underscore the complexity of aligning LLM-driven recommendations with nuanced clinical practice and emphasize the need to distinguish between contextual data gaps and model reasoning errors to improve system reliability in supporting multidisciplinary breast cancer management.

Michalowski et al. (2024) describe multimorbidity guideline-based clinical decision support systems (MGCDSSes) as innovative tools designed to optimize the management of patients with multiple chronic conditions. These systems generate personalized treatment plans by integrating information from diverse clinical data sources, such as computer-interpretable guidelines (CIGs), adverse drug interaction databases, and electronic health records. A critical factor behind their effectiveness lies in their capacity to provide clear explanations that justify the recommended treatments, thereby enhancing transparency and clinician trust.

Traditionally, treatment explanations have been manually crafted by physicians, ensuring high accuracy and clinical relevance. However, this process is time-consuming and demands substantial effort, which can detract from direct patient care. To address these challenges, Michalowski et al. (2024) evaluated the performance of Meditron70B, a large language model (LLM), in generating treatment explanations within the MGCDSS framework.

Their exploratory, survey-based study compared physician-curated explanations with those automatically generated by the LLM, revealing that while the LLM shows significant promise, it remains vulnerable to hallucinations and occasional clinical inaccuracies.

The study also investigated physicians' attitudes toward manual versus LLM-generated explanations in complex multimorbidity cases managed by the MitPlan MGCDSS. Michalowski et al. (2024) report that explanations produced by the LLM—fine-tuned on a medical corpus—were often considered equal to or preferred over manually generated ones, particularly regarding evidence alignment and self-contained clarity. Despite most LLM-generated explanations being accurate and relevant, some errors were detected. These findings underscore the importance of using LLM-generated explanations as part of a broader clinical decision support system with human oversight to mitigate risks associated with hallucinations and ensure patient safety.

3.2 Challenges of using LLMs on Clinical Decision Support

3.2.1 Black Box

Berry et al. (2025) identified transparency and explainability as persistent challenges for LLM deployment in healthcare. Unlike conventional clinical algorithms, which provide step-by-step reasoning, LLMs often operate as opaque systems, making it difficult to trace the rationale behind recommendations. While techniques such as feature attribution and saliency mapping can clarify influential factors—highlighting, for example, specific genotypes, fibrosis stages, or prior treatment failures—the underlying decision-making process remains largely inaccessible.

Kottlors et al. (2025) reported that models may overlook critical local variables, such as facility-specific resources, personnel availability, or the interventionist's skills, when making decisions like those for mechanical thrombectomy. These limitations underscore the importance of LLMs citing exact sources and guidelines, thus enabling traceable and guideline-concordant decision-making.

Harari et al. (2024) emphasized that inaccurate or misleading outputs can have serious clinical consequences, compounding mistrust when the model's reasoning process is not visible. Rajashekar et al. (2024) similarly found that participants were reluctant to rely on LLM-augmented CDSS without citations or knowledge of the data sources used, noting that source transparency could positively influence trust. Participants also expressed a preference for outputs formatted in familiar clinical reference styles, such as bullet points and clearly highlighted management steps, suggesting that both presentation and provenance of information are crucial for adoption. Gargari et al. (2025) reinforced this by emphasising that, even with retrieval-augmented generation (RAG), the reasoning chain must be traceable to specific evidence for accountability in medical contexts.

Yang et al. (2025), referencing Wu et al. (2024), demonstrated that while LLMs integrated with image-to-text technology may not match the task-specific accuracy of deep learning (DL) models for thyroid nodule diagnosis, their interpretability offers distinct advantages for clinical education and decision-making. Multimodal LLMs (MLLMs) enable broader system coverage, zero-shot learning, and richer human–computer interaction, although they have to deal with higher computational demands and potential privacy concerns. In contrast, DL models remain more resource-efficient and easier to deploy locally but lack the cross-domain adaptability and explanatory potential of MLLMs. These contrasts highlight that enhancing transparency is not only a matter of improving clinician trust but also a defining factor in determining which AI systems are best suited to particular clinical applications.

3.2.2 Hallucinations

Sblendorio et al. (2024) propose a comprehensive and dynamic framework for evaluating large language models (LLMs) in complex clinical settings, carefully aligned with OECD ethical and responsible AI guidelines. Their framework integrates human assessment with an automated metric, MPNetv2, which quantifies semantic variability over time in model responses—referred to as the Automated Assessment of Temporal Variability of Responses. This combined approach addresses key safety concerns such as hallucinations, where LLMs generate inaccurate or fabricated information stemming from overgeneralization or unsupported assumptions derived from their training data. Specifically, hallucinations can manifest as false statements, invented entities or events, or logically inconsistent conclusions, with fabrications representing an especially problematic subclass involving fictional facts, names, or dates that can mislead users and propagate misinformation. To mitigate these risks, Sblendorio et al. (2024) introduce the innovative use of hallucinative texts as hard negative examples, improving the alignment of textual and visual tokens in multimodal LLMs and effectively reducing hallucination frequency and severity across benchmarks.

Further advancements in dealing with hallucinations are demonstrated by Choi et al. (2025) and Woo et al. (2025) through the adoption of Retrieval Augmented Generation (RAG) frameworks. RAG enhances LLM outputs by grounding responses in curated, domain-specific datasets rather than relying solely on the vast but unregulated and often outdated information from the open Internet. This grounding enables more accurate, contextually relevant, and trustworthy answers, especially critical in specialized medical fields such as nuclear medicine. Choi et al. (2025) emphasize that RAG's ability to reference prior clinical cases not only mitigates hallucinations but also increases clinicians' trust in model outputs. Woo et al. (2025) further demonstrate that integrating RAG with AI agentic systems can boost the accuracy of LLMs by an average of 39.7%, underscoring the transformative potential of these hybrid approaches in improving the reliability of AI-assisted medical decision-making. Nonetheless, they note ongoing challenges with rare case retrieval and the variability introduced by differing disease prevalences across hospitals, which constrain the generalizability and robustness of current systems.

On the fundamental causes and solutions for hallucinations, Roustan et al. (2025) explain that hallucinations largely arise from the auto-regressive architecture of most LLMs, which generate text by predicting the most probable next token(s) based on preceding outputs. This token-by-token prediction, while powerful, can cause the model to produce plausible yet incorrect continuations, especially when training data is insufficient or ambiguous. Importantly, Roustan et al. (2025) cite Yin et al. (2023), who found that "the larger the training dataset size for LLMs, the more likely the model will be capable of recognizing its limitations and acknowledging uncertainty". This suggests that scaling training data, alongside improved uncertainty modeling, can help LLMs better flag when they "don't know" an answer, reducing hallucination risks. Building on this, Roustan et al. advocate for fine-tuning LLMs with expert-curated medical datasets and clinician feedback to tailor models to the nuances of healthcare and further suppress hallucinations.

Supporting these conclusions, Gargari et al. (2025) report on a study by Quidwai et al. (2024) that evaluated a RAG-based chatbot specifically designed for precision medicine in multiple myeloma. Their RAG model was benchmarked against state-of-the-art LLMs such as GPT-3.5-turbo-16k and GPT-4-32k on a set of expert-curated, challenging oncology questions. The key advantage highlighted was the RAG model's ability to effectively mitigate hallucinations by providing truthful responses even when relevant information was not found within its primary corpus, a critical feature in clinical contexts where misinformation can lead to harmful consequences (Gargari et al., 2025 citing Quidwai and Lagana, 2024). This underscores the necessity of domain-specific retrieval frameworks and curated knowledge bases for safe clinical AI deployment.

Together, these studies highlight a multi-layered approach to combat hallucinations in clinical LLMs: from the architectural roots (auto-regressive token prediction) and training dataset scale (Yin et al., 2023), to the practical frameworks combining human and automated evaluation (Sblendorio et al., 2024), and finally, the integration of retrieval-based augmentation systems (Choi et al., 2025; Woo et al., 2025; Gargari et al., 2025). This combined evidence strongly supports the continued development and validation of LLMs in healthcare with built-in mechanisms for reducing hallucinations, ensuring patient safety.

3.2.3 Data Privacy

From a data privacy perspective, Woo et al. (2025) stated that open-source large language models (LLMs) offer significant advantages for healthcare institutions by allowing them to host models on their own secure servers. This setup ensures complete control over sensitive patient information and prevents exposure to external cloud providers or third parties. Such control is essential for maintaining patient confidentiality and complying with strict privacy regulations like HIPAA in the United States.

Vrdoljak et al. (2025) emphasized that open-source models also enable reproducible research, as their transparency allows other researchers to access the same code and datasets, verify results, and build upon previous work, which can accelerate innovation in medical AI.

However, Woo et al. (2025) noted that deploying open-source LLMs requires expertise to maintain and update the models, which often improve incrementally rather than providing robust capabilities immediately. Additionally, the computational resources necessary to run large open-source models can be substantial, potentially offsetting the cost savings from avoiding licensing fees. They also suggested that some models might be more amenable to agentic augmentation, but this process can sometimes lead to “overthinking” and erroneous judgments, indicating room for improvement in internal agent optimization.

Vrdoljak et al. (2025) raised concerns about the ethical and regulatory challenges associated with integrating LLMs in healthcare, particularly focusing on patient privacy and data security. They pointed out that outsourcing patient data to closed-source API providers like OpenAI or Anthropic raises risks of data misuse or unintended use in future model training. To mitigate these risks, Vrdoljak et al. (2025) recommended that hospitals host their own open-source models, ensuring full HIPAA compliance and protection of sensitive information. They further stressed the importance of clear protocols for data handling, storage, and transmission when using LLMs to process patient records.

Valencia et al. (2023) highlighted specific security measures essential for protecting healthcare data, especially in sensitive areas such as kidney transplant care. They argued for strong encryption methods both at rest and in transit, along with the use of encrypted databases or secure cloud services with stringent security protocols. Valencia et al. (2023) also recommended implementing strict access controls and multi-factor authentication to prevent unauthorized access, complemented by regular auditing and monitoring to identify and address vulnerabilities promptly.

The importance of anonymization and de-identification techniques was also emphasized by Valencia et al. (2023), who stated that these approaches balance data utility with patient privacy by removing identifiable information while still allowing extraction of valuable insights. They further called for clear regulatory frameworks and standardized guidelines to govern AI use in healthcare, advocating for collaboration between regulatory authorities and professional organizations to ensure ethical, interoperable, and seamless integration of LLM-powered tools.

Kwan et al. (2025) noted that the introduction of LLMs creates additional points of vulnerability to cyber-attacks in healthcare, which is already a prime target for malicious actors. They emphasized the critical importance of robust security protocols, including encryption, secure data transmission, and regular security audits to protect patient data.

In the context of precision oncology, Lammert et al. (2024) discussed the heightened ethical concerns related to patient privacy, advocating for robust anonymization techniques and secure data storage to mitigate risks associated with patient data use in LLMs.

Sanduleanu et al. (2024) explained that ChatGPT's medical training relies mainly on widely available general medical knowledge from the internet due to the difficulty of incorporating large datasets of patient-specific information while maintaining privacy and ethical standards. Consequently, they noted that ChatGPT's responses to medical queries may lack the depth and specificity that come from direct access to extensive patient data.

Finally, Sblendorio et al. (2024) emphasized the importance of transparency when using LLMs in clinical or research settings. They recommended informing users, such as research nurses, that their data might be reused for further model fine-tuning or reinforcement learning and obtaining explicit consent before data collection and use.

3.2.4 Data Regulations

Ethical and regulatory considerations are paramount when integrating large language models (LLMs) into healthcare. Berry et al. (2025) stated that approval from ethics committees or institutional review boards is necessary to ensure compliance with ethical standards. They also emphasized adherence to key data protection laws such as HIPAA in the United States and the General Data Protection Regulation (GDPR) in Europe. Importantly, Berry et al. (2025) noted that patients typically do not provide consent waivers for the use of their de-identified data in model development, highlighting a critical consent challenge.

Regular audits play a crucial role in maintaining compliance with evolving regulatory requirements. Berry et al. (2025) recommended conducting both internal and external audits by independent bodies post-deployment. They further advocated for periodic reviews of the model's impact on patient outcomes and healthcare practices to ensure ongoing benefit and avoid unintended negative consequences.

Kwan et al. (2025) pointed out that healthcare is one of the most heavily regulated industries, with strict rules governing patient safety, data privacy, and clinical efficacy. They stressed that integrating LLMs into healthcare workflows must comply with these regulations to mitigate legal risks and protect patients. Kwan et al. (2025) also acknowledged the complexity and ambiguity of the regulatory landscape, which creates uncertainty for both healthcare providers and technology developers. They underscored that meeting regulatory standards requires rigorous testing, validation, and certification processes, which are often time-consuming and costly.

Kwan et al. (2025) further identified a critical knowledge gap regarding the long-term effects of LLM deployment on patient outcomes. While initial results show promise in improving medication adherence and patient engagement, they cautioned that long-term impacts remain largely unstudied and warrant further research.

Rao et al. (2023) highlighted the limitations that must be considered when designing clinically oriented prompts for LLMs like ChatGPT and when developing regulations governing AI use in clinical settings. They emphasized the need for applicable approvals from agencies such as the U.S. Food and Drug Administration (FDA) to ensure safety and efficacy.

Kim S. H. et al. (2025) discussed the considerable challenge of establishing effective regulatory frameworks for LLM-based clinical tools. They pointed out that LLMs involved in clinical decision-making must meet rigorous safety and reliability standards, but the vast diversity of possible inputs and outputs complicates the creation of comprehensive guidelines.

Kim et al. (2025) argued that regulatory authorities like the FDA and the European Medicines Agency (EMA) should develop oversight mechanisms that are both adaptable and robust, balancing innovation with patient protection.

Chen et al. (2024) emphasized that the rapid advancement of generative AI (GenAI) like ChatGPT introduces significant ethical and regulatory challenges in healthcare. They highlighted the urgent need for new data regulations and laws to ensure responsible development and application of GenAI, safeguarding patient safety, privacy, and equity. The authors pointed out risks related to AI hallucinations, transparency, and accountability, especially regarding liability when AI-driven decisions cause harm. Furthermore, the unpredictable nature of GenAI's probabilistic outputs and self-learning capabilities complicates risk control within existing regulatory frameworks.

Zhang et al. (2024) acknowledged the transformative potential of generative AI and large language models (LLMs) to enhance clinical decision-making, improve diagnostic accuracy, and reduce physician burnout. However, they stressed that significant challenges remain, including concerns about bias, variability, and ethical implications. To address these issues, Zhang et al. called for robust regulatory frameworks, comprehensive standards, and continuous improvements in model explainability and performance. They emphasized that successful integration of these technologies into healthcare requires ongoing research, stakeholder collaboration, and strict oversight to ensure positive impacts on patient care and clinical practice.

In addition, Vrdoljak et al. (2025) discussed the need to clearly define the legal responsibility of LLM-assisted decisions, emphasizing that LLMs should serve as decision support tools rather than autonomous decision-makers.

Roustan et al. (2025) concluded that although LLM advancements present significant opportunities for enhancing patient care, robust legal frameworks are essential for guiding their safe and ethical use at national levels. They also emphasized that strong institutional governance will be key to successful implementation in everyday clinical practice.

3.2.5 Ethics and Bias

Berry et al. (2025) found that bias detection tools, such as demographic parity analysis, can help identify whether certain patient groups receive different treatment suggestions. To mitigate such bias, they recommended strategies like re-weighting underrepresented patient populations in training data or supplementing models with diverse clinical trial data to promote fairer recommendations. For instance, if a model consistently under-recommends direct-acting antivirals for certain racial groups due to historical underrepresentation in trials, re-weighting the data can help correct this imbalance (Berry et al., 2025).

Janumpally et al. (2025) warned that relying on generative AI (GenAI) as a source of factual information in important clinical or academic contexts remains risky. They emphasized that assertions made by GenAI should be validated by users to avoid misinformation, and advised that graduate medical education (GME) trainees should not use GenAI to directly guide patient care decisions outside controlled research settings. Janumpally et al. (2025), citing Goddard et al. (2011), also highlighted the risk of automation bias — the cognitive tendency to overly trust automated systems — which can influence clinical decision-making negatively.

Kwan et al. (2025) cited Ferrara (2024), who noted that if an LLM is trained predominantly on data from a specific demographic, its recommendations might be less accurate or even harmful when applied to patients from different backgrounds. This demographic bias poses a significant challenge for ensuring fairness in AI-driven healthcare (Kwan et al., 2025).

Rinderknecht et al. (2024) stressed that AI-generated clinical decisions or recommendations profoundly impact patient care depending on context. They argued that safe integration of these models into routine practice requires addressing challenges related to accuracy, transparency, accountability, and ethical concerns, while maintaining the physician’s central role as the ultimate decision-maker. As such, widespread clinical adoption remains early and requires significant validation.

Abadir et al. (2024) reported on Decipher-AI, a natural language processing model in development at Harvard aimed at improving dementia diagnosis. Early analysis revealed that the algorithm underperforms with some demographic groups, reflecting bias related to socioeconomic status, race, and age — challenges common in both traditional medicine and AI. They also discussed broader ethical issues including bias and fairness at the population level, privacy concerns, LLM hallucinations, clinician acceptance, and the need to responsibly manage rapid AI advancements (Abadir et al., 2024).

Kim S. H. et al. (2025) pointed out that in radiological diagnosis, LLMs can generate multiple hypotheses quickly but raise concerns about the mentioned automation bias. They also noted that implementing open-source LLMs locally demands significant technical infrastructure and expertise, often available only in large academic centers, raising equity and economic concerns across healthcare settings (Kim S. H. et al., 2025).

Roustan et al. (2025) highlighted that LLMs may incorrectly attribute clinical or radiological features to diseases based on user input and model probabilistic behavior. This, combined with cognitive biases like anchoring and confirmation bias, could lead clinicians down wrong diagnostic or therapeutic paths with serious patient consequences. They stressed that LLMs often rely on incomplete or outdated data, may not prompt for crucial missing clinical or social information, and could inadequately account for patient cultural preferences (Roustan et al., 2025).

Gargari et al. (2025) emphasized the importance of high-quality data for retrieval-augmented generation (RAG) models, warning that errors, missing data, or embedded biases — especially those reflecting healthcare disparities — can negatively affect model fairness and performance.

Sblendorio et al. (2024) suggested that minimizing bias in LLMs could be supported by establishing continuous feedback mechanisms among nurses, patients, and LLM developers, enabling real-time reporting of potential issues or biases encountered in everyday clinical use.

3.2.6 Others

Roustan et al. (2025) highlighted that consistency, and therefore reliability, remains a significant issue when using LLMs to make care plan recommendations. They found that even when the exact same user query is repeated, the LLM’s response can vary substantially. This variability is a critical factor clinicians must consider before integrating LLMs into patient care workflows.

Gargari et al. (2025) pointed out that medical domains and clinical guidelines are often highly complex and heterogeneous. Guidelines may have varying structures, with essential information presented in different formats such as text, tables, or flow charts. This complexity makes it challenging for retrieval-augmented generation (RAG) systems to accurately interpret and extract relevant information.

Yang et al. (2025) argued that addressing the challenges of effectively implementing LLMs in medical diagnostics requires a comprehensive and multifaceted approach. They emphasized the need to promote unified data exchange standards, real-time synchronization mechanisms, and open APIs to enable seamless integration with existing electronic health record systems. Concurrently, they stressed the importance of enhancing data security and privacy protections.

Yang et al. (2025) further noted that the inherent diversity of evaluation methods across different diseases and clinical progression stages poses major challenges. Evaluation metrics often rely on human experts and include measures such as diagnostic accuracy, readability, or subjective scores tailored to specific task scenarios. The variation in experimental designs and evaluation criteria complicates comparisons between studies and underscores the need for standardized, robust evaluation frameworks. They recommended the development of more standardized test-question datasets and increasing the number of high-quality randomized controlled trials (RCTs) using consistent methodologies as promising steps toward improving LLM reliability and applicability in medical diagnostics.

4 Physician Survey

To further understand the opinion of doctors regarding LLMs usage in clinical scenarios, a survey was conducted.

4.1 Previous Surveys on LLMs in Clinical Practice

Five recent studies were identified about the usage of LLMs by healthcare practitioners:

The first is a survey conducted by Kisvarday et al. (2024) with 390 pediatric professionals in the USA. Most of them knew ChatGPT, but only half used it, mainly for administrative tasks (emails, teaching materials). Very few used it for diagnosis or treatment. 75% said they would use a version compliant with HIPAA.

The second, conducted by Sumner et al. (2025) was a national survey with 1144 students and professors of medicine in the USA. Two thirds already used LLMs, mostly to summarize articles and to create study materials. Main concerns included hallucinations, plagiarism and biases. The majority supported the inclusion of LLMs in the curriculum, with supervision.

The third was a survey with 791 psychiatrists in France supervised by Blease et al. (2024a). Less than one third used LLMs, mainly to write academic texts. A few applied it to clinical decisions. There was skepticism about therapeutic usage, with focus on ethical and privacy risks.

The fourth, made by Blease et al. (2024b), was a survey with 1006 general physicians in United Kingdom. 20% used LLMs, mainly for clinical documentation (29%) and differential diagnosis (28%). Even without institutional policies, physicians were already using LLMs in practice.

The fifth, performed by Kharko et al (2025) was made with 1005 general physicians in the United Kingdom. Most of them saw potential for administrative tasks, but had doubts about empathetic communication and privacy. More than half expected that patients would use it for self-diagnosis.

Together, these results show that:

1. The usage of LLMs by physicians is still low and concentrated on non-clinical tasks
2. There is interest in safer tools (like HIPAA-compliant versions)
3. Worries about precision, privacy and biases are common
4. There is little institutional orientation, but informal adoption is already happening

These findings helped the design of the survey presented in the next section.

4.2 Survey Method

Survey research is a structured approach to collecting information from individuals by means of predetermined questions, enabling the generation of quantitative data that can be statistically analyzed and, under appropriate conditions, generalized to a larger population (Goodfellow, 2023). It is particularly useful in health sciences for assessing attitudes, behaviors, and perceptions.

According to Creswell (2008), a robust survey design must address four key elements: 1) a clear sampling strategy, 2) careful instrument development, 3) pilot testing for clarity, and 4) strategies to maximize response rates. Leedy et al. (2015) further note that surveys can be administered via interviews or questionnaires. This study adopted a self-administered online questionnaire, which offers cost efficiency, wide reach, and anonymity, which are factors that can encourage honest responses, particularly on emerging and sensitive topics such as AI in clinical practice.

The survey was guided by the following research question: “What is the opinion of physicians about the use of LLMs in diagnosis, treatment, and patient monitoring?”

The instrument consisted of nine closed-ended questions using a 5-point Likert scale and one open-ended question for qualitative insights. Table 1 describes each question, its purpose, and response format.

Table 1: Survey Questions and Objectives

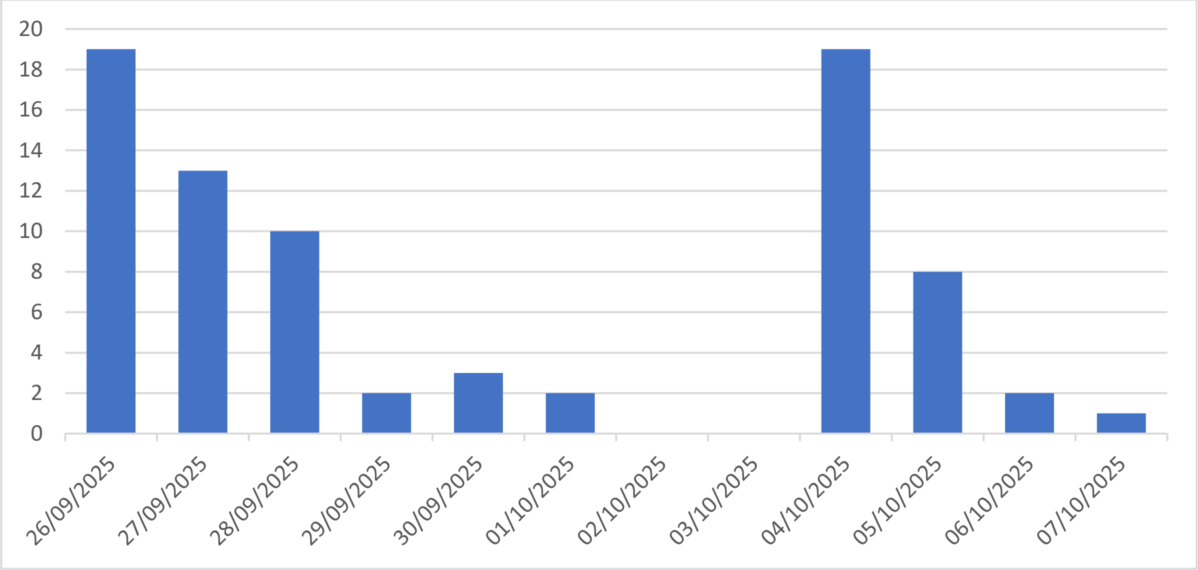
Question	**Related Research Question**	**Purpose of the Question**	**Type of Response**
How often do you use ChatGPT or similar models for clinical decision support?	RQ5	Assess current use of LLMs	Scale (1–5)
How do you rate your knowledge of Artificial Intelligence (AI) tools, such as ChatGPT, in medical practice?	RQ5	Assess knowledge about LLMs	Scale (1–5)
How would you evaluate your overall experience with Artificial Intelligence (AI) in clinical decision support?	RQ5	Understand acceptance of LLMs	Scale (1–5)

In your opinion, is AI a reliable tool for diagnostic support?	RQ2 and RQ5	Assess opinions about AI-assisted diagnosis	Scale (1–5)
In your opinion, can AI facilitate (now or in the future) the creation of personalized treatment plans for patients?	RQ3 and RQ5	Understand opinions about AI-assisted treatment	Scale (1–5)
In your perception, do patients use AI tools to validate medical diagnoses?	RQ5	Understand perceptions of patients’ use of LLMs	Scale (1–5)
How do you assess the following aspects as barriers to using AI: 1) Lack of source transparency 2) Hallucinations (generating incorrect answers) 3) Data privacy concerns 4) Regulations 5) Ethical and bias issues	RQ4 and RQ5	Identify perceived barriers to adoption	Scale (1–5) — one scale per barrier
What most motivates you to use AI tools such as ChatGPT? 1) Convenience 2) Response speed 3) Productivity 4) Personalized answers 5) Curiosity 6) Anonymity 7) None	RQ5	Understand main advantages of LLMs according to physicians	7 options
To what extent do you believe physicians will adopt Artificial Intelligence in the next 10 years?	RQ5	Explore views on the future adoption of LLMs	Scale (1–5)
In your opinion, in which medical area does Artificial Intelligence seem most promising, and why?	RQ5	Explore most relevant potential uses of LLMs	Open-ended

Source: The author

The questionnaire was implemented via Google Forms, with responses automatically recorded in Google Sheets. It was open from September 26 to October 7, 2025. Invitations were sent by email to 308 physicians at a hospital in São Paulo, followed by one reminder on October 4. A total of 79 responses were collected, yielding a response rate of 25.6%. Daily response trends are shown in Figure 15.

The results are presented in the following section.Figure 15: Number of answers by day



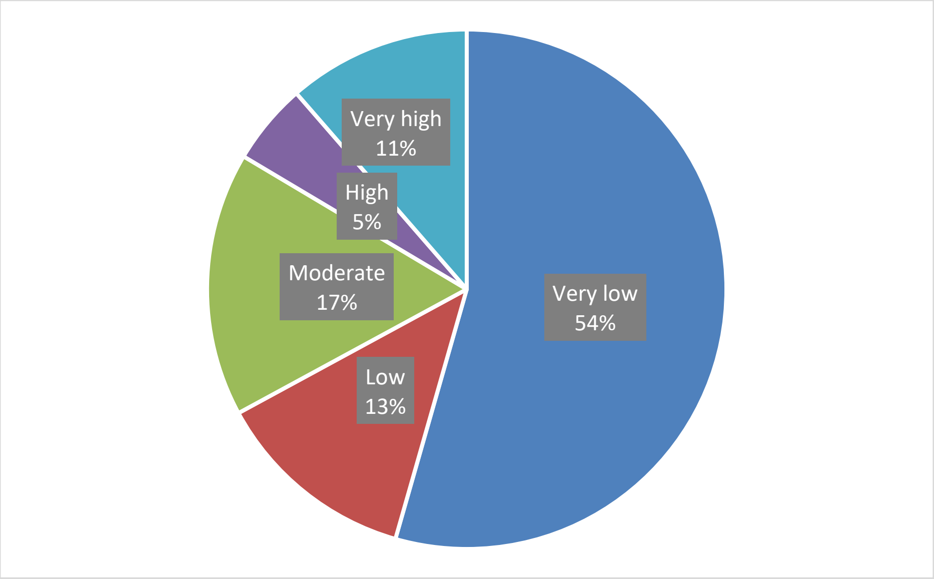
Source: The Author

4.3 Survey Results

4.3.1 Frequency of LLM utilization

67% of physicians reported low or very low use of LLMs to support clinical decision. Only 16% indicated high or very high use (5%high and 11% very high). The others (17%) declared moderate use (Figure 16).

Figure 16: Frequency of LLM utilization by doctors

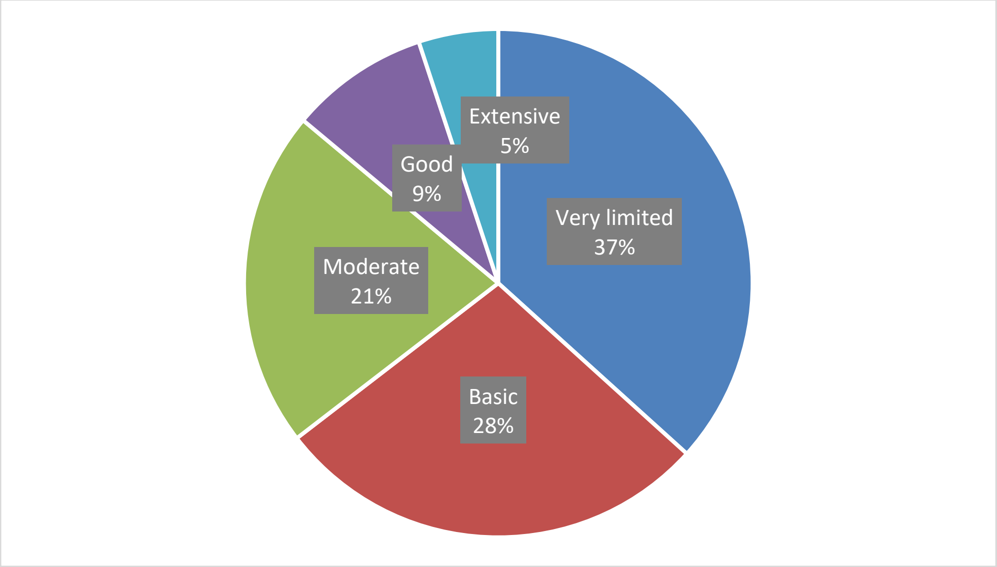


Source: The author

4.3.2 Knowledge about LLMs

65% said they had very limited or basic knowledge about LLMs. Another 21% reported moderate knowledge, 9% good, and 5% extensive (Figure 17).

Figure 17: Doctors' knowledge about LLMs

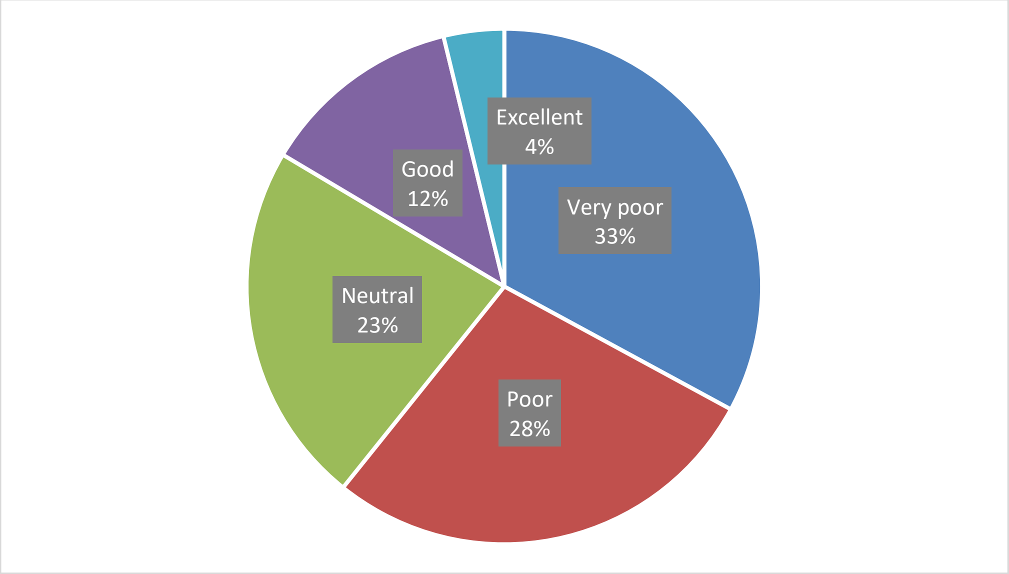


Source: The author

4.3.3 Quality of experience with LLMs

Most participants rated their experience as poor or very poor (61%). Another 23% said it was neutral, 12% good, and 4% excellent (Figure 18).

Figure 18: Quality of experience for doctors using LLMs

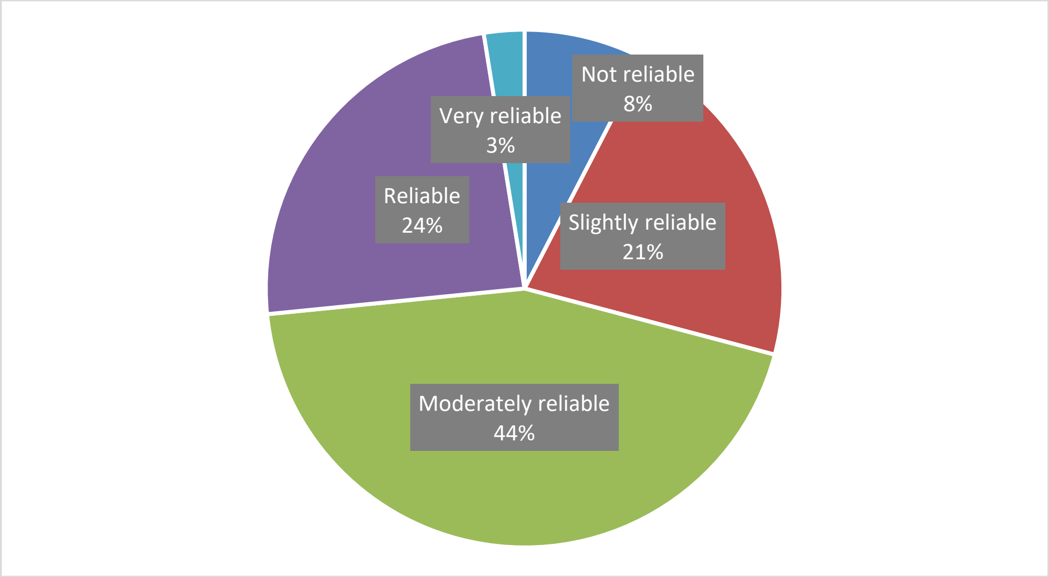


Source: The author

4.3.4 Reliability for diagnosis support

44% rated LLMs as moderately reliable. Another 24% said reliable, 21% said slightly reliable, 8% not reliable, and 3% very reliable (Figure 19).

Figure 19: Reliability for diagnosis support according to doctors

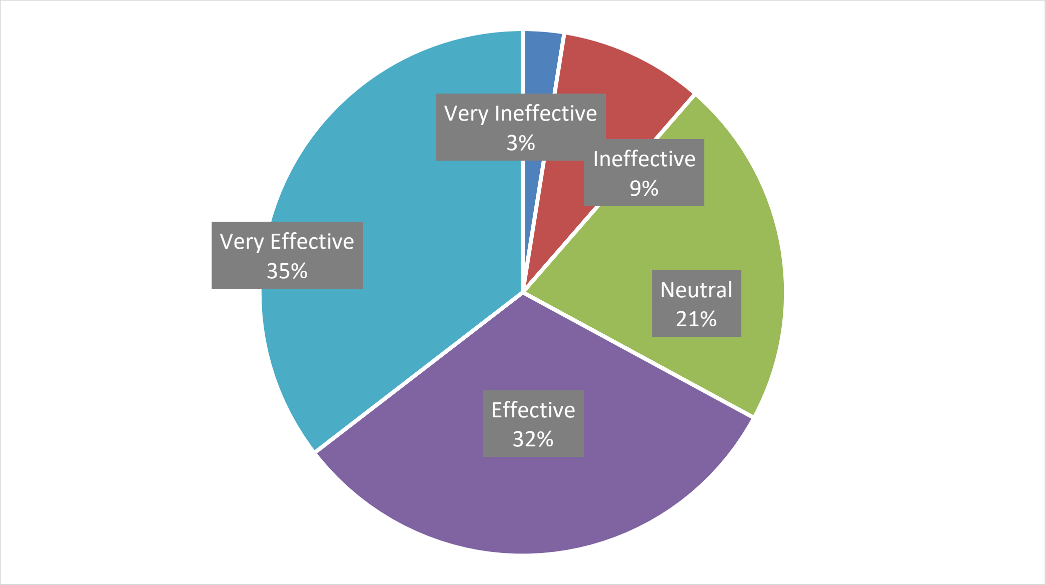


Source: The author

4.3.5 LLMs on personalized treatment

67% considered LLMs effective or very effective for supporting personalized treatment plans. Only 12% judged them ineffective or very ineffective; 21% were neutral (Figure 20).

Figure 20: Opinion of doctors regarding LLMs on personalized treatments

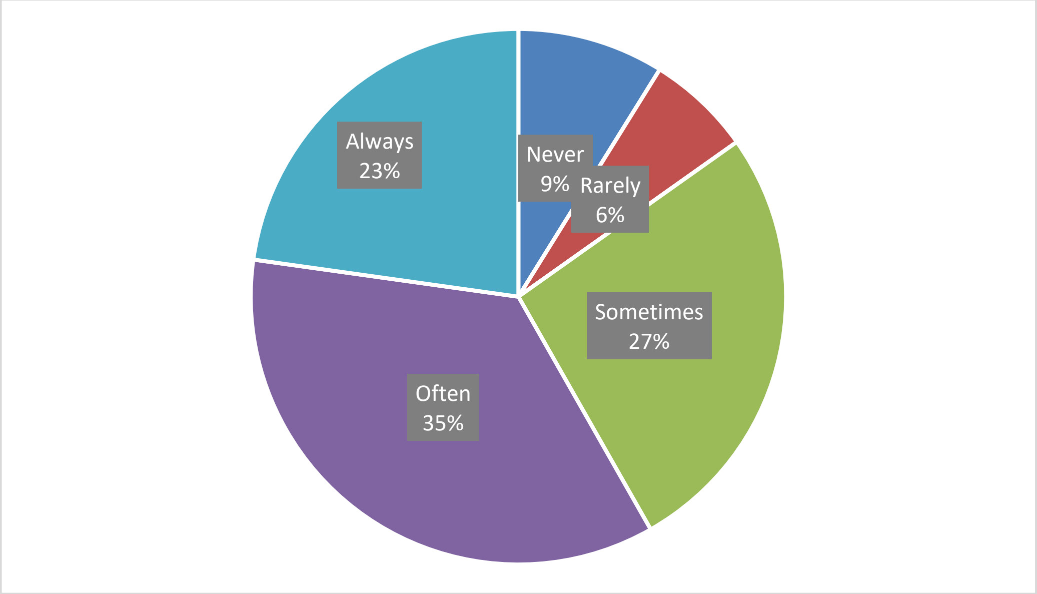


Source: The Author

4.3.6 Patients using LLMs to validate doctors' diagnosis

23% of physicians believe their patients always use LLMs to confirm diagnoses; 35% said often, 27% sometimes, and 15% rarely or never (Figure 21).

Figure 21: Doctors perceptions about patients using LLMs to validate diagnosis



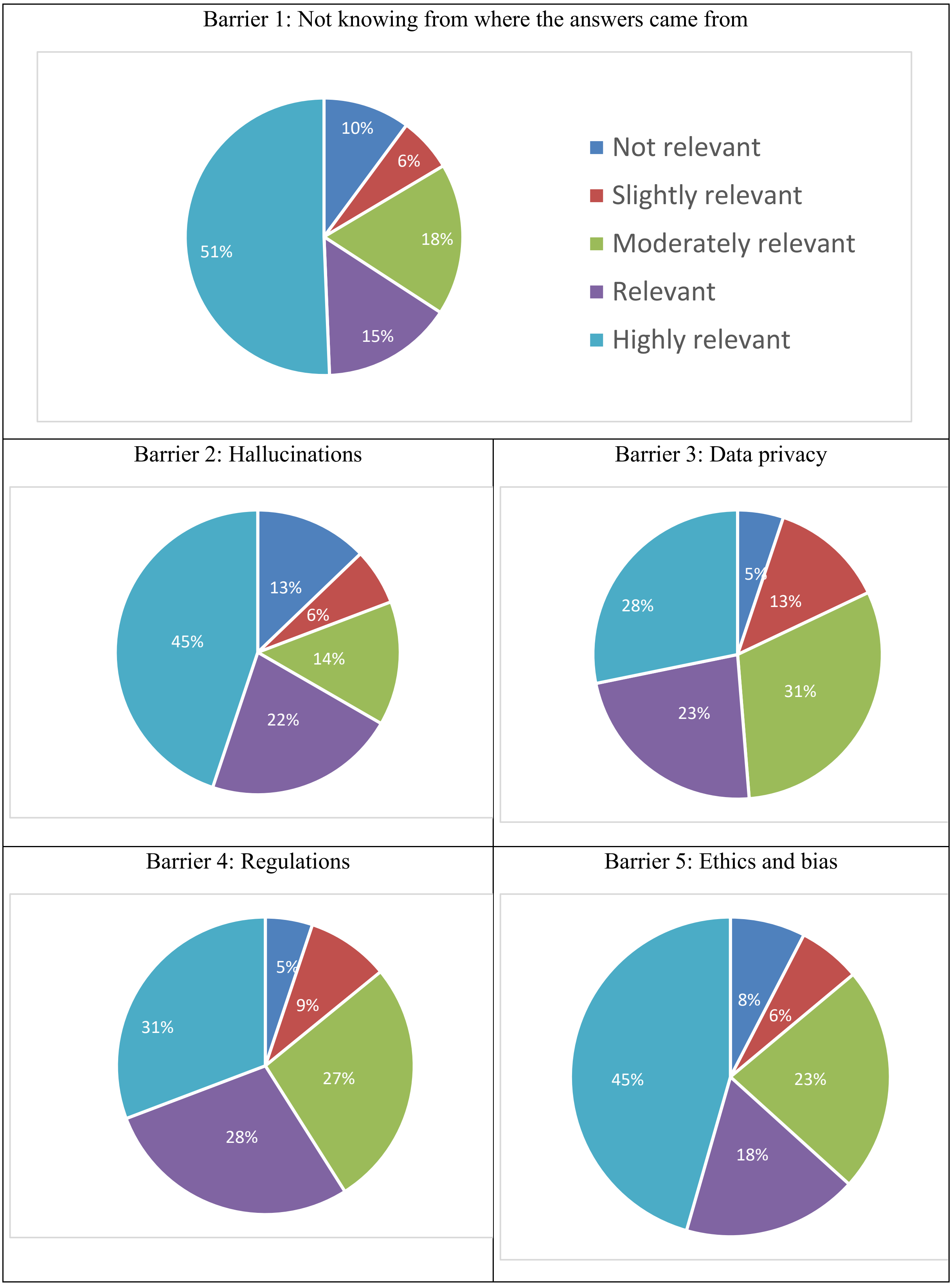
Source: The Author

4.3.7 Barriers to adopt LLMs

Main reported obstacles:

- Lack of transparency (‘black box’): 66% said relevant or very relevant;
- Hallucinations: 67% said relevant or very relevant;
- Ethics and bias: 63% (45% very relevant + 18% relevant);
- Data privacy: 59% (28% very relevant + 31% relevant);
- Regulation: 60% (31% very relevant + 29% relevant) (Table 1).

Table 2: Opinion of doctors regarding barriers to adopt LLMs



Barrier 1: Not knowing from where the answers came from

Barrier 2: Hallucinations

Barrier 3: Data privacy

Barrier 4: Regulations

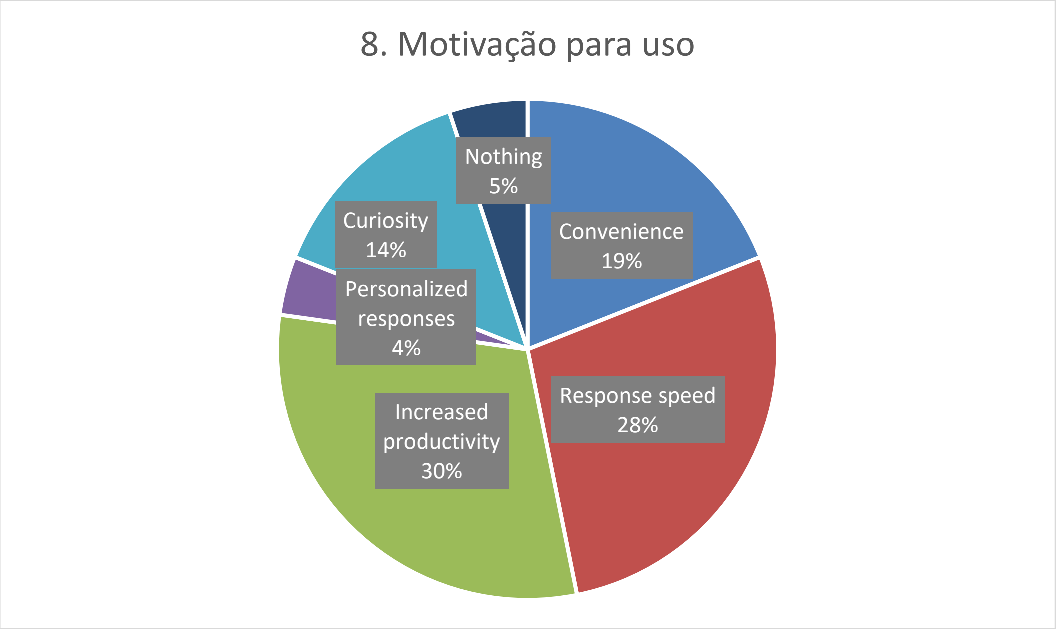
Barrier 5: Ethics and bias

Source: The Author

4.3.8 Motivation to use LLMs

Main reasons to use LLMs were: productivity (30%), fast responses (28%), convenience (19%), and curiosity (14%). No participant mentioned anonymity as a motivation (Figure 22).

Figure 22: Doctors motivations to use LLMs

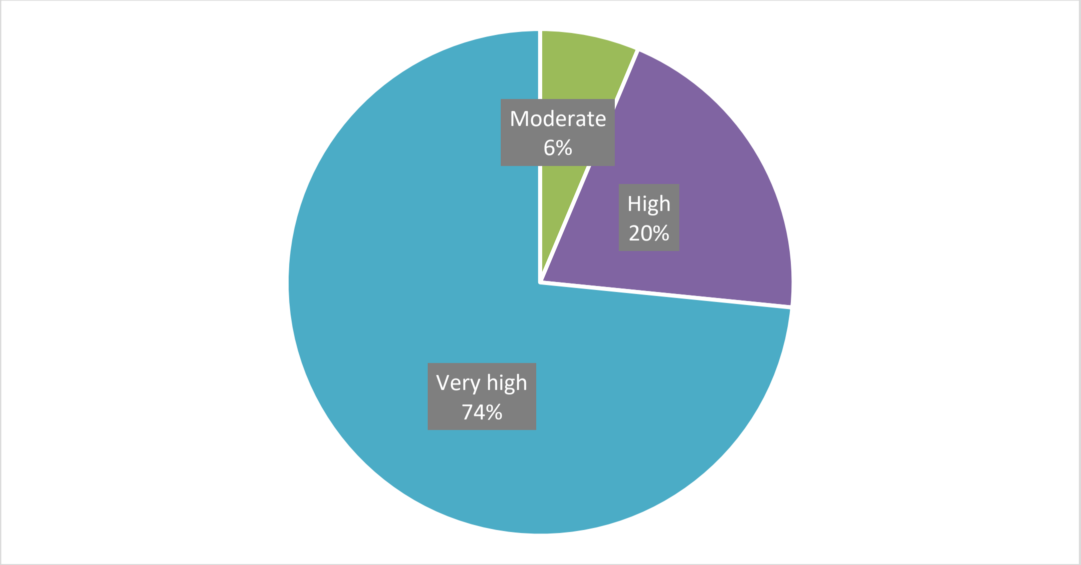


Source: The Author

4.3.9 Adoption of LLMs in the future

74% believe LLM adoption will be very high in the next decade; 20% said high and 6% moderate. No one predicted low or very low adoption (Figure 23).

Figure 23: Perspectives on the adoption of LLMs by doctors in the future



Source: The Author

4.3.10 Promising areas in the future

Of the 79 respondents, 68 (86%) answered the open question. The responses are categorized in Table 3.

Table 3: Frequency of mentions by category

Category	Frequency
Medical Imaging/Pathology	40
Clinical Medicine	7
All Areas	6
Triage/Initial Care	1
Emergency Room/Emergency Care	1
Psychiatry	2
Dermatology	1
Preventive Medicine	1
Public Health/Epidemiology	1
Medical Education	1
Research/Data	1

Management/Medical Records	1
Multiple Areas	2
Multiple Applications	1
General Diagnosis	2

Source: The Author

5 Discussion

This chapter aims to answer the research questions proposed in chapter 1, considering the results obtained in the systematic literature review and in the survey with physicians.

5.1 Publication Trends, Key Contributors, and Keywords (RQ1)

The bibliometric analysis shows that research about Large Language Models in Clinical Decision support has expanded rapidly recently. Until 2022, there were few studies in the field (with none selected to be part of this systematic review). Then, in 2023 more studies started to show up, with a few selected to be part of the current study. This growth coincides with the release of ChatGPT in November 2022, which may have catalyzed research in this area. In 2024 and 2025, the growth on publications became even more notable. This rapid increase suggests that LLMs have quickly attracted attention of researchers.

It is also possible to note significant trends in the geographic distribution of publications. The co-authorship by country analysis clearly shows that the United States accounts, by far, for the largest share of publications, followed by the United Kingdom, Germany, China and South Korea. The collaboration across research group is still limited, as shown in the co-authorship network on Figure 6. This fragmented structure is consistent with the novelty of the topic.

Another important result is the distribution of publishers. MDPI is the main publisher in this area, followed by Springer Nature and Elsevier. This shows that most LLM research is being published in multidisciplinary and open-access journals. The presence of major publishers like Springer and Elsevier also indicates that the topic has gained scientific credibility.

The multidisciplinary characteristic is reinforced by the keyword density map. The most frequent keywords include “Large Language Models,” “Clinical Decision Support,” “Artificial Intelligence,” “Healthcare,” “Diagnosis,” “Treatment,” “Generative AI,” “Ethics,” “Retrieval-Augmented Generation,” and “Prompt Engineering”. These words are mainly related to the computational, linguistic and healthcare domains, but also contains “Ethics”, showing that there is concern beyond technological advances.

Overall, the bibliometric analysis confirms that research on LLMs in CDS is recent, rapidly growing, and involves multiple disciplines, as evidenced by the diversity of authors, institutions, and keywords.

5.2 LLMs in Clinical Diagnostics (RQ2)

The studies in this review show that Large Language Models (LLMs) can improve how doctors diagnose patients, interpret information, and make clinical decisions. LLMs can read and understand unstructured text, such as medical notes, radiology reports, and lab results. Therefore, they can perform tasks that older systems could not (traditional Machine Learning, for example, requires structured data).

In several contexts, LLMs showed diagnostic performance similar to physicians. Chen et al. (2024) found that GPT-4 identified the correct diagnosis in more than half of the 38 New England Journal of Medicine cases, doing better than the average performance of the doctors who took the same test. Borna et al. (2024) reported perfect diagnosis accuracy in emergency plastic surgery scenarios when clinical exam data were included. However, results vary by specialty. In precision oncology the models are still unreliable (Benary et al., 2023; Vrdoljak et al., 2025), while in settings with clear decision rules, such as suspected appendicitis, accuracy is high (Sanduleanu et al., 2024).

To improve reliability, many studies combine LLLMs with complementary technologies. Retrieval-Augmented Generation (RAG) allows models to search medical databases while generating answers. Choi et al. (2025) used RAG to analyze more than 200,000 PET reports, with clinicians considering almost all outputs clinically relevant. Barrit et al. (2025) also applied RAG to neurological diagnosis and found higher accuracy and faster completion times compared to clinicians.

Prompt engineering also plays an important role. Structured prompts that follow clinical reasoning step by step produce better diagnostic accuracy than simple open prompts (Savage et al., 2024; Leypold et al., 2024).

New multimodal systems go further by combining text with images or signals, allowing the model to interpret image exams and written reports together. This integration may become essential in areas like radiology and pathology.

LLMs are also used in earlier stages of diagnostic work (data wrangling). Kim et al. (2024) introduced PhenoFlow, a GPT-4–based system that automates data preparation for stroke research, reducing preprocessing time from hours to minutes. This allows doctors to have more time to analyze patient data instead of spending time pre-processing it.

Overall, the literature shows that LLMs contribute to diagnosis through: 1) direct support in well-defined scenarios, 2) RAG-enhanced systems, 3) structured prompt strategies, and 4) multimodal integration. The results are strong in structured clinical domains but still limited in complex areas like oncology. Across all studies, human supervision is seen as essential.

5.3 LLMs in Treatment and Patient Monitoring (RQ3)

The reviewed literature also shows that Large Language Models (LLMs) are being used to support treatment decisions and continuous patient monitoring. These systems help doctors interpret complex information, compare it with clinical guidelines, and generate suggestions based on medical evidence.

In oncology, Rinderknecht et al. (2024) compared LLM-generated treatment recommendations with decisions made by multidisciplinary tumor boards for genitourinary cancers. The models performed slightly worse than human experts but followed clinical guidelines closely. The authors highlight that LLMs can act as preliminary support tools, not replacements for specialist judgment, which is important to keep in mind as the technology evolves.

Retrieval-Augmented Generation (RAG) is widely used to make treatment reasoning more reliable. Lammert et al. (2024) developed MEREDITH, a system based on Gemini Pro combined with RAG, which gets information from sources such as PubMed and clinical trial registries to propose molecularly guided therapies. When evaluated by experts, MEREDITH showed more than 94% agreement with human recommendations, avoided hallucinations, and cited all sources, increasing trust and transparency.

Other systems work in a similar way. GastroBot (Zhou et al., 2024), created for gastroenterology, uses RAG with 25 clinical guidelines and produces therapeutic strategies that achieve safety and usability levels comparable to those of trained professionals. These systems show how LLMs can be constantly updated through retrieval instead of full model retraining.

For patient monitoring, LLMs can analyze electronic health records to detect patterns in clinical evolution. Abadir et al. (2024) developed Decipher-AI, which is able to predict cognitive decline in dementia patients using both structured and unstructured data. Roshani et al. (2025) built a mobile app that classifies COVID-19 severity and provides explanations showing how the model reached each prediction.

In summary, LLMs contribute to treatment and monitoring through: 1) generating recommendations aligned with clinical guidelines, 2) using RAG to access external evidence safely and 3) analyzing patient records to detect clinical patterns.

5.4 Key Challenges in Clinical Applications of LLMs (RQ4)

The reviewed literature consistently identifies six major challenges in the use of Large Language Models (LLMs) for diagnosis, treatment, and patient monitoring. These challenges appear repeatedly across different studies, suggesting that they are structural, not occasional issues.

1. **Lack of transparency (black-box problem)**

LLMs function as not transparent reasoning systems, making it difficult to identify how a decision was produced. Hager et al. (2024) showed that even when the output is correct, models often fail to provide clear justifications. This lack of explanation reduces clinicians' trust, especially in high-risk domains where every decision must be justified. Without knowing why a model suggests something, it becomes hard for a physician to rely on it.

2. **Hallucinations**

Sometimes these models can generate information that sounds plausible but is false. Benary et al. (2023) described oncology cases where LLMs suggested treatments that do not exist or misread molecular data. RAG-based systems can reduce hallucinations because it bases the outputs in external evidence, but they do not eliminate the risk completely, especially when something that is outside their training domain (or RAG domain) appears. This is a problem that can have serious clinical consequences, once the models can appear assertive even when they tell to the user something they do not know about.

3. **Privacy and regulatory compliance**

Using sensitive health data in training or inference raises concerns about anonymization, consent, and the risk of data leaks. Many LLMs rely on internet corpora with unclear origins, making it difficult to ensure compliance with rules like HIPAA (United States) or GDPR (Europe). In addition, cloud-based models may expose patient information outside secure hospital environments. This creates a complicated space where there is a clear trade-off between technical progress and legal and ethical boundaries.

4. **Bias and inequality**

Because LLMs learn from human data, they reproduce existing biases found in their training sources. Sanduleanu et al. (2024) pointed out that unbalanced datasets can produce systematic errors for underrepresented groups, increasing disparities in healthcare. If the healthcare and technological area do not deal with it, these biases could reproduce unfair patterns in diagnosis and treatments rather than improve care. This shows that AI inherits both the strengths and the weaknesses of human knowledge.

5. Regulatory uncertainty and responsibility

There is still no clear definition of who is accountable when an LLM influences a clinical decision. It could be the physician, the hospital, the developer, or even the data provider. This ambiguity makes many institutions hesitant to integrate these tools directly into clinical workflows. The cultural dimension matters too: many clinicians remain skeptical about delegating parts of their judgment to systems that cannot fully explain themselves (relation with 'black box' nature and hallucinations).

6. Practical integration

Integrating LLMs into electronic health record systems requires technical adjustments and infrastructure. Without institutional support, these tools tend to remain limited to research prototypes rather than becoming part of everyday practice and support. In other words, the gap between having a good model and actually using it in a hospital can still be quite large. The practical integration issue also brings up an important topic: education. Physicians would possibly require training to use safely and their help could be useful to develop good models.

In summary, the challenges reported in the literature highlights technical issues ('black box', hallucinations), ethical concerns (bias, privacy), regulatory uncertainty (accountability), and operational barriers (integration into clinical systems). Across almost all studies, human supervision appears not just recommended but essential for safe adoption. And even if the technology is evolving quickly, these issues remind us that implementing AI in healthcare is as much a social and organizational task as it is a technical one.

5.5 Physicians' Perceptions of Clinical Applications of LLMs (RQ5)

The survey results show that the use of Large Language Models (LLMs) in clinical practice is still very limited. Most doctors reported low or very low levels of usage (67%), meaning that LLMs are not yet part of daily medical routines (Figure 16).

Most respondents still have a weak understanding of how to use LLMs effectively. Most reported only basic or very limited knowledge (65%) (Figure 17).

In terms of experience, perceptions were mostly negative. A total of 61% rated their experience with LLMs as "Poor" or "Very Poor," while only 16% described it as "Good" or "Excellent." (Figure 18).

When asked about diagnostic reliability, doctors expressed cautious. Most respondents did not consider LLMs completely unreliable, but the majority placed them in the "Moderately Reliable" (44%) or "Slightly Reliable" (21%) categories. Only a minority viewed them as very reliable (3%). (Figure 19).

Perceptions changed significantly when the focus shifted from diagnosis to treatment planning. Most doctors (67%) rated LLMs as "Effective" or "Very Effective" for helping personalize therapeutic plans, while only 12% considered them ineffective. This suggests that clinicians see more potential for LLMs in synthesizing patient data, clinical guidelines, and medical literature to support personalized care (Figure 20).

The survey also shows a growing gap between patient and physician use. Many physicians believe their patients frequently use LLMs to validate diagnoses (58% said always or often) (Figure 21), even when physicians themselves report low usage and limited understanding of these tools.

When analyzing perceived barriers, the "black-box" nature of LLMs, where the reasoning behind outputs cannot be traced, was one of the main concerns. The risk of hallucinations was also heavily considered. Ethical issues related to bias and fairness were another major worry. Although privacy and regulation were also noted as challenges, they were seen as secondary compared to the core concerns of transparency and accuracy (Table 21).

Regarding motivations for using LLMs, the data indicate that physicians value productivity (30%), speed (28%) and convenience (19%), being the main reasons cited for adoption. Personalized responses accounted only for 4% and no one considered anonymity as the main reason to use LLMs on clinical practices (Figure 22).

Despite the challenges, the survey results show strong optimism about the future. Even though current usage and knowledge levels are low, most respondents believe LLM adoption will increase substantially over the next decade (94% said they believe the adoption will be high or very high). This suggests that clinicians can distinguish between present difficulties and long-term potential.

6 Conclusion

6.1 Synthesis

This work presents a comprehensive study on the role of Large Language Models (LLMs) in Clinical Decision Support (CDS), examining how these technologies are applied in the diagnosis, treatment, and monitoring of patients, as well as how they are perceived by healthcare professionals. The research followed a structured approach that combined a systematic literature review and empirical data collection through a physician survey.

The study began with a theoretical contextualization of artificial intelligence in healthcare, showing the evolution of machine learning and the emergence of LLMs as tools capable of understanding and generating human language. The introduction and background chapters defined the main concepts related to clinical decision support systems, large-scale neural architectures, and the integration of generative AI in medical workflows. These sections also contained the motivations behind the study, the relevance of exploring adopting AI in healthcare, and the specific research questions guiding the work.

A systematic literature review was performed with two main objectives: 1) identify publication trends, key authors, institutions, countries, and journals contributing to the field and 2) examine the main studies addressing the use of LLMs in medical diagnostics, treatment support, and patient monitoring.

The first part involved the bibliometric analysis, conducted using Scopus Database. Data were, then, processed with VOSviewer to generate visualizations of co-authorship networks, keyword co-occurrence maps, and temporal publication patterns.

The second part started with the selection process, followed by inclusion and exclusion criteria to ensure relevance and scientific rigor. Each selected article was analyzed for their objectives, methods, findings, and technological approaches. This stage allowed the identification of the most common applications of LLMs in clinical contexts, as well as the specific technologies used to enhance their performance, such as Retrieval-Augmented Generation (RAG), multimodal modeling, and prompt engineering.

After the literature review, a quantitative survey was developed to obtain empirical data on the opinions and experiences of medical professionals regarding LLMs. The questionnaire included sections about awareness, frequency of use, perceived usefulness, and ethical concerns associated with these technologies. Responses were analyzed statistically and presented in graphs and tables, providing an overview of doctors' perspectives on the benefits, limitations, and potential risks of using generative AI in clinical environments.

The discussion synthesized these results in relation to the research questions, describing rapid publication trends, the concentration of research in a few countries, and the most recurrent keywords. It also discussed how LLMs are applied in diagnostic tasks, treatment recommendation systems, and patient monitoring, as well as the technological advances supporting their development. The chapter also described the challenges reported in the literature, including lack of transparency, data security, and bias, and presented the general perception of doctors obtained through the survey.

In summary, this work mapped the scientific production on LLMs in Clinical Decision Support, reviewed the main technological applications and limitations described in the academic literature, and captured the perspectives of medical professionals with respect to their integration into clinical practice. Together, these steps offer a structured and comprehensive overview of an emerging field that continues to evolve rapidly at the intersection of artificial intelligence and medical decision-making.

6.2 Limitations

Although this research gives a current view of how Large Language Models relate to Clinical Decision Support, it is important to recognize its methodological and conceptual limitations. Acknowledging these limits is essential for transparency and to guide future studies that want to improve, expand, or question the findings presented here.

The first limitation is the scope and representativeness of the literature reviewed. Despite a systematic selection process, the field is exceptionally dynamic. New models, methods, and benchmarks appear all the time. Because of this, the publications included here represent a snapchat of one moment, not the whole picture. Certain developments that appeared after the conclusion of data collection may already have modified the landscape, particularly regarding multimodal architecture and real-world clinical validation. This temporal aspect is inherent to studies of emerging technologies and should be acknowledged when interpreting the findings.

Another limitation is related to the survey. It brings useful professional perspectives, but it also has self-report bias (people may want to look better, they may misunderstand the question, they may not remember things well and they may guess instead of giving precise information). The number of participants was sufficient to show general tendencies, but it cannot represent all medical specialties or regions. Also, because participation was voluntary, physicians more interested in technology maybe were more likely to answer. This could lead to an overrepresentation of curiosity and openness toward LLMs compared to the whole medical population. The survey also comes from only one hospital in one city, which limits generalization.

The fast evolution of LLMs also affects the stability of the conclusions. Models are updated and fine-tuned all the time, so the versions analyzed in the literature may already be different from the ones used today. This means the object of study may change faster than the research designed to understand it. As a result, any summary of current findings should be seen as temporary and open to change when new data appears. In fast-moving fields like AI in medicine, this instability is inherent and represents a key challenge for maintaining scientific relevance.

In summary, these limitations show the complexity of studying LLMs in medicine. The findings here should be understood as part of a changing landscape, a step in a process that will continue to evolve as models become more interpretable, datasets grow, and clinical institutions use AI more in practice. These limitations should be considered when interpreting the findings, especially given the fast-paced evolution of LLMs in clinical settings.

6.3 Future Research Directions

The limitations found in this study points to future investigations. First, the fast evolution of language models suggests the necessity of periodic updates in the systematic review, specially to keep pace with advances in the literature. A new study with more recent data would allow to verify if the current barriers, such as hallucinations and the ‘black box’ nature, are being effectively mitigated.

Second, another direction is to make the content analysis more specific. Future studies could compare different medical specialties, model types, or evaluation methods to see if challenges and ethical concerns are similar across fields or not.

Third, the survey was applied in a single hospital and had 79 respondents. Future studies could include a bigger sample, including physicians of different regions, health systems and experience levels.

Fourth, other people in the clinical process should also be heard. Nurses, medical students, hospital managers and data scientists have different papers in the decision making process and can offer complementary perspectives about risk, benefits and conditions of safe use of AI.

Fifth, the use of qualitative methods, such as structured interviews with professionals who already used AI tools could reveal practical situations where LLMs bring value or generate doubt, going beyond general perceptions captured by the survey.

Finally, the present study indicates the need to understand and investigate how the medical education can prepare future professionals to use LLMs in a critical and responsible way.

References

Abadir, P. M., Battle, A., Walston, J. D., & Chellappa, R. (2024). Enhancing care for older adults and dementia patients with large language models: Proceedings of the National Institute on Aging—Artificial Intelligence & Technology Collaboratory for Aging Research Symposium. *Journals of Gerontology: Series A, Biological Sciences and Medical Sciences, 79*(9), Article glae176. https://doi.org/10.1093/gerona/glae176

Barrit, S., Torcida, N., Mazeraud, A., Boulogne, S., Carette, T., Carron, T., Delsaut, B., Diab, E., Benoit, J., Kermorvant, H., Maarouf, A., Slootjes, S. M., Redon, S., Robin, A., Hadidane, S., Harlay, V., Tota, V., Madec, T., Niset, A., Al Barajraji, M., El Hadwe, S., Massager, N., Lagarde, S., Madsen, J. R., & Carron, R (2025). Specialized large language model outperforms neurologists at complex diagnosis in blinded case-based evaluation. *Brain Sci.* 2025, *15*(4), 347. https://doi.org/10.3390/brainsci15040347

Berry, P., Dhanakshirur, R. R., & Khanna, S. (2025). Utilizing large language models for gastroenterology research: A conceptual framework. *Therapeutic Advances in Gastroenterology, 18*, 1–16. https://doi.org/10.1177/17562848251328577

Benary, M., Wang, X. D., Schmidt, M., Soll, D., Hilfenhaus, G., Nassir, M., Sigler, C., Knödler, M., Keller, U., Beule, D., Keilholz, U., Leser, U., & Rieke, D. T. (2023). Leveraging large language models for decision support in personalized oncology. *JAMA Network Open, 6*(11), e2343689. https://doi.org/10.1001/jamanetworkopen.2023.43689

Blease, C., Bonnaud, S., Lucchetti, G., Huguelet, P., Gauld, R., Fauré, P., Gaab, J., & Haller, D. M. (2024). Psychiatrists' experiences and opinions of generative artificial intelligence. *Frontiers in Psychiatry, 15,* 1398853. https://doi.org/10.3389/fpsyt.2024.1398853

Borna, S., Gomez-Cabello, C. A., Pressman, S. M., Haider, S. A., & Forte, A. J. (2024). Comparative Analysis of Large Language Models in Emergency Plastic Surgery Decision-Making: The Role of Physical Exam Data. J. Pers. Med., 14(6), 612. https://doi.org/10.3390/jpm14060612

Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule-based expert systems: The MYCIN experiments of the Stanford heuristic programming project*. Addison-Wesley

Chen, A., Liu, L., & Zhu, T. (2024). Advancing the democratization of generative artificial intelligence in healthcare: a narrative review. *J Hosp Manag Health Policy*, *8*, 12. https://doi.org/10.21037/jhmhp-24-54

Chen, D., Avison, K., Alnassar, S., Huang, R. S., & Raman, S. (2024). Medical accuracy of artificial intelligence chatbots in oncology: a scoping review. *Oncology*, *30*(4), oyaf038. https://doi.org/10.1093/oncolo/oyaf038

Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z., Yin, F., Zhao, J., & He, X. (2024). *Exploring large language model based intelligent agents: Definitions, methods, and prospects*. *arXiv preprint*. https://doi.org/10.48550/arXiv.2401.03428

Creswell, J. W. (2008). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Sage Publications.

Delourme, S., Redjdal, A., Bouaud, J., & Seroussi, B. (2025). Leveraging guideline-based clinical decision support systems with large language models: A case study with breast cancer. *Methods Inf Med*. https://doi.org/10.1059/s24524-4299

Fernández, M., Johnston, A., Kesh, S., Ashchepkova, S., & Bennett, D. (2024, January 23). *Language modeling: The fundamentals*. S&P Global. https://www.spglobal.com/en/research-insights/special-reports/language-modeling-the-fundamentals

Ferrara, E. (2023). *Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies*. *Sci*, *6*(1), Article 3. https://doi.org/10.3390/sci6010003

Garcia Valencia, O. A., Tangpanithandee, C., Chaidarun, W., Thongprayoon, W., Bunpata, P., Maneenil, J., Cheungpasitporn, W., Mahatanaprasit, J., Theerakittikul, J., & Amornpetch, S. (2023). Enhancing Kidney Transplant Care through the Integration of Chatbot. Healthcare, 11(20), 2518. https://doi.org/10.3390/healthcare11202776

Gargari, O. K., & Habibi, A. (2024). Enhancing medical AI with retrieval-augmented generation: A mini narrative review. *Systematic Reviews, 13*, 35. https://doi.org/10.1177/20552076251337177

Goddard, K., Roudsari, A., & Wyatt, J. C. (2011). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Goodfellow, L. (2023). *An overview of survey research*. *Respiratory Care*, *Ova* (Article summarizing survey research methodology). https://doi.org/10.4187/respcare.11041

Gumilar, K. E., Indraprasta, B. R., Faridzi, A. S., Wibowo, B. M., Herlambang, A., Rahestyningtyas, E., Irawan, B., Tambunan, Z., Bustomi, A. F., Brahmantara, B. N., Yu, Z. Y., Hsu, Y. C., Pramuditya, H., Putra, V. G. E., Nugroho, H., Mulawardhana, P., Tjokroprawiro, B. A., Hedianto, T., Ibrahim, I. H., … Tan, M. (2024). Assessment of Large Language Models (LLMs) in decision-making support for gynecologic oncology. *Computational and Structural Biotechnology Journal, 23*, 4019–4026. https://doi.org/10.1016/j.csbj.2024.10.050

Hager P., Jungmann F., Holland R., Bhagat K., Hubrecht I., Knauer M., Vielhauer J., Makowski M., Braren R., Kaissis G., Rueckert D., & et al. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision making. *Nature Medicine*, 30(9), 2613–2622. https://doi.org/10.1038/s41591-024-03097-1

Haim, G. B., Saban, M., Barash, Y., Cirulnik, D., Shaham, A., Eisenman, B. Z., Burshtein, L., Mymon, O., & Klang, E. (2024). Evaluating large language model–assisted emergency triage: A comparison of acuity assessments by GPT-4 and medical experts. *Journal of Clinical Nursing*. https://doi.org/10.1111/jocn.17490

Harari, R. E., Altaweel, A., Ahram, T., Keehner, M., & Shokoohi, H. (2025). A randomized controlled trial on evaluating clinician-supervised generative AI for decision support. *International Journal of Medical Informatics, 195*, 105701. https://doi.org/10.1016/j.ijmedinf.2024.105701

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: The System Causability Scale (SCS): Comparing human and machine explanations. *KI - Künstliche Intelligenz, 34*(2), 193–198. https://doi.org/10.1007/s13218-020-00636-z

Janumpally, R., Nanua, S., Ngo, A., & Youens, K. (2025). Generative artificial intelligence in graduate medical education. *Front. Med.*, *11*, 1525604. https://doi.org/10.3389/fmed.2024.1525604

Kim, J., Lee, C., An, G., Son, H., Lee, D., Lee, D., Kim, B., Lee, S., Kim, S., Kim, J., Park, S., Lee, H., Kim, S., Lim, S., & Seo, J. (2023). PhenoFlow: A Human-LLM Driven Visual Analytics System for Exploring Large and Complex Stroke Datasets. https://osf.io/q6yc4/

Kim, S. H., Heiss, D. M., Wiestler, B., Schoell, S., Aufschlager, L., Bösche, R., K. K. B., Kirsch, M., Zech, C., Pelzl, P. S., & Płatek, K. J. (2025). Benchmarking the diagnostic performance of open source LLMs in 1933 Eurorad case reports. *npj Digital Medicine*. https://doi.org/10.1038/s41746-025-01488-3

Kisvarday, M., Zhu, J., Matias, T., Patterson, J., Kaimal, R., Hauser, K., Chen, S., Orenstein, E. W., Wexler, A., Doshi-Velez, F., & Triantafyllidis, A. K. (2024). ChatGPT use among pediatric health care providers: Cross-sectional survey study. *JMIR Pediatrics and Parenting, 7,* e55076. https://doi.org/10.2196/55076

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine, 23*(1), 89–109. https://doi.org/10.1016/S0933-3657(01)00077-X

Kresevic, S., Giuffrè, M., Ajcevic, M., Accardo, A., Crocè, L. S., & Shung, D. L. (2024). Optimization of hepatological clinical guidelines interpretation by large language models: A retrieval augmented generation-based framework. *npj Digital Medicine, 7*, Article 91. https://doi.org/10.1038/s41746-024-01091-y

Lammert, J., Tschochohei, M., Maucher, J., Hülse, S., Bünz, J., Zopfs, D., Lennartz, S., Bauer, A. W., Schnabel, M. J., Maintz, D., Bruns, C., Durner, A., Schwamborn, K., Winter, C., Ferber, D., Mogler, C., & Illert, A. L. (2024). Expert-Guided Large Language Models for Clinical

Decision Support in Precision Oncology. *JCO Precis Oncol*, *8*, e2400478. https://doi.org/10.1200/PO.24.00478

Leedy, P. D., & Ormrod, J. E. (2015). *Practical Research: Planning and Design* (11th ed.). Pearson.

Lemstra, M. A. M. S., & de Mesquita, M. A. (2023). *Industry 4.0: A tertiary literature review*. Technological Forecasting and Social Change, 186, 122204. https://doi.org/10.1016/j.techfore.2022.122204

Leypold, T., Lingens, L. F., Beier, J. P., & Boos, A. M. (2024). Integrating AI in Lipedema Management: Assessing the Efficacy of GPT-4 as a Consultation Assistant. *Life*, 14(5), 646. https://doi.org/10.3390/life14050646

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

Lin, C.-Y. (2004). *ROUGE: A package for automatic evaluation of summaries*. In Proceedings of *Text Summarization Branches Out* (pp. 74–81). Association for Computational Linguistics.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Wei, L., Ishii, K., Lee, I., Chen, C. C., Das, T. J., Mahmood, F., & Yeh, J. J. W. (2024). A multimodal generative AI copilot for human pathology. *Nature*, *634*, 466–473. https://doi.org/10.1038/s41586-024-07618-3

Michalowski, M., Wilk, S., Bauer, J. M., Carrier, M., Delluc, A., Le Gal, G., Wang, T.-F., Siegal, D., & Michalowski, W. (2024). *Manually-curated versus LLM-generated explanations for complex patient cases: An exploratory study with physicians*. Journal/Conference Name. https://doi.org/10.1007/978-3-031-66535-6_33

Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., Mayer McKinney, S., Ness, R. O., Poon, H., Qin, T., Usuyama, N., White, C., & Horvitz, E. (2023). *Can generalist foundation models outcompete special-purpose tuning? Case study in medicine*. arXiv preprint arXiv:2311.16452. https://doi.org/10.48550/arXiv.2311.16452

Osheroff, J. A., Teich, J. M., Levick, D., Saldana, L., Velasco, F., Sittig, D. F., & Rogers, K. M. (2012). *Improving outcomes with clinical decision support: An implementer's guide* (2nd ed.). HIMSS Publishing.

Quidwai, M. A., & Lagana, A. (2024). A RAG chatbot for precision medicine of multiple myeloma. *medRxiv*. https://doi.org/10.1101/2024.03.14.24304293

Rajashekar, N. C., Shin, Y. E., Pu, Y., Chung, S., You, K., Giuffre, M., Chan, C. E., Saarinen, T., Hsiao, A., Sekhon, J., Wong, A. H., Evans, L. V., Kizilcec, R. F., Laine, L., McCall, T., & Shung, D. L. (2024). *Human-algorithmic interaction using a large language model-augmented artificial intelligence clinical decision support system*. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)* (Article 442, pp. 1–20). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642024

Rao, A., Kim, J., Kamineni, M., Pang, M., Lie, W., Dreyer, K. J., & Succi, M. D. (2023). Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. *Journal of the American College of Radiology, 20*(10), 990–997. https://doi.org/10.1016/j.jacr.2023.05.003

Rosen, S., & Saban, M. (2023). Evaluating the reliability of ChatGPT as a tool for imaging test referral: A comparative study with a clinical decision support system. *European Radiology, 34*, 2826–2837. https://doi.org/10.1007/s00330-023-10230-0

Roshani, M. A., Zhou, X., Qiang, Y., Suresh, S., Hicks, S., Sethuraman, U., & Zhu, D. (2025). Generative Large Language Model—Powered Conversational AI App for Personalized Risk Assessment: Case Study in COVID-19. *JMIR AI*, *4*, e67363. https://doi.org/10.2196/67363

Roustan, D., & Bastardot, F. (2025). The Clinicians’ Guide to Large Language Models: A General Perspective With a Focus on Hallucinations. *Interact J Med Res*, *14*, e59823. https://doi.org/10.2196/59823

Saad, O., Saban, M., & Levin, C. (2025). Augmenting Community Nursing Practice With Generative AI: A Formative Study of Diagnostic Synergies Using Simulation-Based Clinical Cases. *Journal of Primary Care & Community Health*, *16*, 1–7. https://doi.org/10.1177/21501319251326663

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint*. https://doi.org/10.48550/arXiv.2402.07927

Sanduleanu, S., Ersahin, K., Bremm, J., Talibova, N., Damer, T., Erdogan, M., Kottlors, J., Goertz, L., Bruns, C., Maintz, D., & Al-Saadi, N. (2024). Feasibility of GPT-3.5 versus Machine Learning for Automated Surgical Decision-Making Determination: A Multicenter Study on Suspected Appendicitis. *AI*, *5*(4), 1942–1954. https://doi.org/10.3390/ai5040096

Savage, T., Nayak, A., Gallo, R., Rangan, E., & Chen, J. H. (2024). Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine, 7*(1), Article 20. https://doi.org/10.1038/s41746-024-01010-1

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology, 70*, 747–770. https://doi.org/10.1146/annurev-psych-010418-102803

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., ... & Natarajan, V. (2023). Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*. https://arxiv.org/abs/2305.09617

Sumner, J., Wang, Y., Tan, S. Y., Chew, E. H. H., & Yip, A. W. (2025). Perspectives and experiences with large language models in health care: Survey study. *Journal of Medical Internet Research, 27,* e67383. https://doi.org/10.2196/67383

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30). https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M., & Božić, J. (2025). A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. *Healthcare*, *13*(6), 603. https://doi.org/10.3390/healthcare13060603

Woo, J. J., Yang, A. J., Olsen, R. J., Hasan, S. S., Nawabi, D. H., Nwachukwu, B. U., Williams 3rd, R. J., & Ramkumar, P. N. (2025). Custom large language models improve accuracy: Comparing retrieval augmented generation and artificial intelligence agents to noncustom models for evidence-based medicine. *Arthroscopy, 41*(3), 565–573.e6. https://doi.org/10.1016/j.arthro.2024.10.042

Wu, T., Liu, Y., Yang, S., Wang, Y., Yu, Z., & Ma, H. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica, 10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Yang, X., Li, T., Su, Q., Liu, Y., Kang, C., Lyu, Y., Zhao, L., Nie, Y., & Pan, Y. (2025). Application of large language models in disease diagnosis and treatment. *Chin Med J*, *138*(2), 130–142. https://doi.org/10.1097/CM9.0000000000003456

Yin, Z., Sun, Q., Guo, Q., Wu, J., Qiu, X., & Huang, X. (2023). *Do large language models know what they don't know?* arXiv. https://doi.org/10.48550/arXiv.2305.18153
S

Zhang, P., Shi, J., & Kamel Boulos, M. N. (2024). Generative AI in medicine and healthcare: Moving beyond the ‘peak of inflated expectations’. *Future Internet, 16*(12), 462. https://doi.org/10.3390/fi16120462

Zhou, Q., Liu, C., Duan, Y., Sun, K., Li, Y., Kan, H., Gu, Z., Shu, J., & Hu, J. (2024). GastroBot: A Chinese gastrointestinal disease chatbot based on retrieval augmented generation. *Frontiers in Medicine, 11*, 1392555. https://doi.org/10.3389/fmed.2024.1392555