(Online) ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 13, No. 3, August 2025, pp.629-670

Large language models in medicine: A systematic review of applications in medical, healthcare, and educational contexts

Humberto J. Navarro[1,2]*, Camilo L. Sandoval[1], Ixent Galpin[2]
[1] Faculty of Natural Sciences and Engineering, Unidades Tecnológicas de Santander, Bucaramanga Colombia
[2] Faculty of Natural Sciences and Engineering, Universidad de Bogotá-Jorge Tadeo Lozano, Bogotá Colombia

*Corresponding author E-mail: hnavarro@correo.uts.edu.co

ABSTRACT

Large language models have emerged as transformative tools in medicine and medical education, offering applications in aided diagnosis, automation of clinical assessments, and optimization of healthcare workflows. This article critically reviews 112 relevant publications analyzing the use of LLMs in these fields. It explores their applications in specific tasks such as biomedical classification, automated clinical assessment, medical question answering, medical report generation, and enhancement in medical education through exam simulation and personalized tutoring. Despite their advances, LLMs continue to face significant challenges, including data privacy issues, clinical validation, and algorithmic biases. However, their integration into clinical and educational settings demonstrates considerable potential to improve efficiency, accuracy, and accessibility in health care, provided these models adhere to technical and ethical rigor. This article offers a comprehensive overview for healthcare professionals and researchers who aim to adopt these models responsibly.

Keywords: large language models, biomedical natural language processing, clinical applications, medical education, automated clinical assessment, artificial intelligence, healthcare

1. Introduction

Large language models (LLMs) have emerged as one of the most disruptive technologies in artificial intelligence (AI) due to their ability to understand, generate, and interact with natural language in a highly contextual and accurate manner [1], [2]. Trained on billions of words, these models demonstrate outstanding performance in natural language processing (NLP) tasks, particularly in the biomedical domain, outperforming traditional, smaller, and more specialized models [3].

In the clinical context, LLMs applied to medicine demonstrate their usefulness in generating medical reports [4], formulating preliminary diagnoses [5], providing surgical assistance [6], answering clinical questions [7], and automating hospital workflows [8]. In addition, their ability to integrate with multi-modal data, such as images, video, physiological signals, or spoken language, extends their application to specialties such as radiology [9], ophthalmology [10], cardiology, and oncology [11], [12], as well as to specific AI-assisted clinical tasks such as extubating of ventilated patients [13].

In medical education, LLMs transform medical training through intelligent tools for exam generation and assessment [14], clinical simulation [15], personalized tutoring [16], and student performance analysis [17]. LLMs also support scientific writing and automated literature synthesis [4] and contribute to manuscript development [18], evidence review, and knowledge dissemination [8].

However, their adoption in real-world settings faces considerable challenges. These include the need to protect patient privacy, manage algorithmic biases, clinically validate their results, and address the ethical and legal implications of their use in medical decisions [19], [20], [21], [22], [23], [24]. These issues require a rigorous and systematic approach to ensure safe, effective, and equitable implementation.

This article presents a comprehensive and critical review of the current state of LLMs in medicine, medical education, and healthcare. The analysis of 112 publications selected between August 2017 and May 2025 addresses (i) their main applications in clinical, healthcare, and educational tasks; (ii) the datasets employed for

their development and evaluation; (iii) the performance metrics applied; (iv) prevailing challenges; and (v) possible technical and ethical solutions. The aim is to provide a comprehensive and accessible guide for researchers, healthcare professionals, and developers who seek to deploy LLMs in real-world settings responsibly. This review presents existing knowledge and supports the development and implementation of LLM-based applications in medicine by providing a foundation for others to build upon, thereby contributing to the design of more efficient solutions to real-world healthcare challenges.

Previous surveys in the area include [25], [26], [27], [28], [29], [30], [31]. While Singhal *et al.* [25] focus on the clinical knowledge embedded within LLMs, and Thirunavukarasu *et al.* [26] provide a broad overview of medical applications, other works, such as those by Lucas et al. [31] and Sallam [27], emphasize educational uses and pedagogical concerns. Meng *et al.* [28] provide a scoping review primarily focused on mapping applications, whereas Tian *et al.* [30] discuss broader opportunities and challenges in biomedicine. Ethical aspects are addressed in detail by Haltaufderheide and Ranish [29]. Our work complements these surveys by providing a unified, systematic analysis across medical, healthcare, and educational contexts, covering 112 studies and offering a comparative perspective across tasks, model types, and evaluation strategies, with an emphasis on practical deployment and domain-specific adaptation techniques.

1.1. Methodology for the compilation of relevant publications

This study has drawn on a systematic search of relevant literature in academic databases, including Springer, Elsevier, arXiv, and medRxiv, as well as other multidisciplinary sources. The review focuses particularly on developments that have emerged since the launch of ChatGPT [32] in November 2022, as well as other general-purpose AI models adapted to the medical environment. A combination of terms associated with application domains and large language models guides the search strategy. The general search string used is: `("medical" OR "clinical" OR "health care" OR "medical education") AND ("large language model")`. The initial review identifies many general review articles, but few include specific models directly applicable to the medical setting. Therefore, the search strategy is adjusted by replacing the generic term "large language model" with the particular names of LLMs, such as ChatGPT, BERT, LLaMA, and PaLM, among others, and combining these with terms from the medical domain. The selection of specific model names relies on references such as [1] and [3], as well as other recognized sources. This approach is based on the fact that many models designed for clinical tasks are derived from, tuned, or trained on general LLMs, which justifies their inclusion as part of the analysis. A detailed review of titles, abstracts, and keywords excludes papers that do not address specific medical, healthcare, and educational models. In addition, to provide a comprehensive view of the state-of-the-art, the analysis seeks to cover as many medical areas as possible, including clinical specialties such as ophthalmology, oncology, psychiatry, neurology, radiology, cardiology, and pediatrics. This thematic breadth enables the identification of general applications and highly specialized approaches to LLMs in different medical care and training scenarios. Finally, the review incorporates cross-sectional articles analyzing LLMs in healthcare to ensure comprehensive coverage. As a result, the study selects 112 articles closely related to its objectives.

1.2. Scope, contributions, and structure of the document

This review systematically explores the applications of LLMs in medical, healthcare, and medical education settings, considering their use in various clinical scenarios, specialized datasets, and the evaluation methodologies employed. It assesses their performance in different tasks and the main challenges faced in real-world healthcare contexts. The main objective is to provide a structured, critical, and action-oriented guide for researchers, healthcare professionals, and developers who aim to apply LLMs in clinical and educational settings.

The main contributions of this work are summarized as follows:

- A comprehensive and updated review of state-of-the-art language models, highlighting their applications in different medical, healthcare, and educational scenarios.
- The available scientific literature is categorized and analyzed, integrating relevant tasks, domains, and evaluation metrics to assess the performance of LLMs in biomedical settings.
- Current challenges associated with using LLMs in medicine and healthcare are identified and classified, and possible solutions are proposed to address these open problems critically and prospectively.

The rest of the paper is structured as follows: Section 2 presents the technical and conceptual background of LLM, covering its core architectures, the emergence of multi-modal models, and strategies for adapting them

to the biomedical domain. In Section 3, applications of LLMs in the medical, healthcare, and educational domains are discussed, covering tasks such as entity recognition, assisted diagnosis, biomedical prediction, automated clinical assessment, medical question answering, and clinical report generation. Section 4 describes the most commonly used metrics for evaluating the performance of these models in biomedical tasks. Section 5 outlines the primary limitations and challenges associated with their implementation, including concerns over data privacy, the generation of misinformation, algorithmic biases, and ethical and legal issues. Finally, Section 6 presents the study's conclusions and raises future perspectives for the responsible development of LLMs in healthcare.

2. Background

This section explores the fundamental concepts that underpin LLMs. It begins by describing the artificial neural network architectures that underpin these models, emphasizing the role of learned embeddings and the Transformer architecture. Furthermore, it explains how self-attention and multi-head attention enable LLMs to model complex linguistic structures. This background serves as a basis for understanding the capabilities and limitations of current LLMs, including phenomena such as hallucinations. It also explores how LLMs process multi-modal inputs, such as text and images, and how LLMs adapt to specific domains, like biomedicine, through prompt engineering, fine-tuning, instruction tuning, and retrieval-augmented generation (RAG).

2.1. Large language models

LLMs are machine learning models based on *artificial neural networks* (ANNs). These networks mimic how the human body organizes and processes information. ANNs comprise numerous "neurons" interconnected through several layers [33]. The first of these layers is the input layer, which receives data and transmits it to subsequent layers, where the model processes it according to its structure and parameters. One of the first approaches in this field was fully connected multi-layer networks, which served for years in deep learning models that do not process language but lay key technological foundations for developing LLMs. A fundamental challenge of these models lies in getting machines to learn from text when, in essence, machines only process numbers. To address this challenge, learned *embeddings* transform words into numeric vectors that capture their meaning and context. This allows terms with similar meanings to be placed close to each other within a mathematical space [34], [35]. However, an advanced architecture is required for a model to process these representations and capture more complex relationships in language. As such, *Transformers* function as efficient tools because Transformer architectures analyze several connections between words and generate contextual representations [36]. Additionally, it enables the capture of the syntactic and semantic structure of sentences. To achieve this, an embedding process is first performed, followed by positional coding, which helps identify word order within the input text. This allows the model to better understand the context and meaning of the processed information. One of the key mechanisms in Transformers is self-attention, which enables it to calculate attention scores for all possible word combinations [37]. To achieve this, it employs three learned matrices: Query (Q), Key (K), and Value (V), which are derived from linear transformations of the input. This mechanism can assign different weights to each word according to its level of importance and identify relationships between words, even when many others separate them within a text. The relationship between these elements appears in Equation 1.

$$Attention\ (Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where d_k represents the dimension of the key vectors, the softmax function normalizes these values, allowing the model to generate a relevance score for each word relative to the others. The scores are weighted and combined for each word. Multi-head attention extends this process, allowing several attention sets to compute in parallel, which the model then concatenates and linearly transforms to obtain the final result. This multi-head processing enables the model to identify contextual relationships from different perspectives, improving the quality of the generated representations. The data then goes to an *encoder*, which transforms the words into vectors that capture the utterance's grammatical structure, meaning, and context. Each encoder layer applies multiple self-attention mechanisms and nonlinear transformations, allowing the information to be progressively refined until more abstract representations of the content emerge. Subsequently, a *decoder* consisting of multiple layers takes the processed information and converts it into a sequence of words. The model predicts the next word in a text and generates coherent responses.

However, despite the fluency and contextual coherence of the generated outputs, LLMs exhibit a phenomenon known as *hallucination* [38]. Hallucinations occur when a model produces content that is linguistically plausible yet factually incorrect or entirely fabricated. The occurrence of hallucinations in LLMs raises significant concerns for real-world applications, particularly in information retrieval (IR), medicine, and law, where factual accuracy and consistency are crucial. Unlike earlier task-specific NLP systems, the open-ended nature of LLMs introduces distinct challenges for controlling and detecting hallucinations, necessitating new evaluation methods, mitigation strategies, and model grounding [39].

Current LLMs present variations on this structure and differ in their training methods. Generally speaking, the BERT (Bidirectional Encoder Representation from Transformers) model does not follow a unidirectional approach to predict the next word in a sentence [40]. Instead, BERT trains the model with randomly hidden words and learns to anticipate them based on the visible context surrounding them. Some notable implementations of LLMs include LLaMA and LLaMA-2 from Meta [41], GPT-4 from OpenAI [42], PaLM and PaLM-2 from Google Research [43], and DeepSeek-R1 from DeepSeek [44]. Additionally, models developed by Google Research include T5 [45] and BERT [40]. These models adapt to both general tasks, such as conversational assistants, and more specialized uses, for example, in generating answers for medical queries [5], [46], [47], [48], [49].

2.2. Multi-modal LLMs

Multi-modal LLMs represent a groundbreaking advance in the development of AI and NLP. Unlike traditional models, which are limited to text analysis and generation, multi-modal LLMs process and combine content from multiple formats, such as images, audio, and video. This enables them to better adapt to real-world situations and provides more accurate and contextually relevant responses. Integrating this type of content facilitates interaction between people and technology, leading to the application of these tools in various fields, such as medical education, research, and healthcare [3].

2.2.1. Applications and operation of multi-modal LLMs

Advances in multi-modal LLMs drive the development of applications in multiple fields. In this context, Visual ChatGPT functions as a tool that integrates images and text to answer complex questions more accurately [50]. Similarly, models such as BLIP-2 employ the Qformer mechanism, designed to integrate visual and textual information, allowing for improved interaction between the two formats [51]. In tasks such as Visual Question Answering (VQA), multi-modal LLMs interpret images and generate answers based on their content. In the biomedical domain, some models are specifically designed for processing medical language, such as MedPaLM-2, which is based on PaLM-2, a model trained on diverse data that has the potential to integrate textual information with medical images [49]. Similarly, architectures such as BioGPT and PubMedGPT, developed for analyzing clinical and scientific texts, can be complemented with multi-modal techniques to improve the interpretation and generation of medical reports [10], [52], [53].

Multi-modal LLMs feature an architecture composed of modules that work together to process information from various input sources. Figure 1 shows a general framework for handling multiple input modalities, including images, audio, video, and text. Text is already in the LLMs space and can be processed directly. In contrast, other modalities are first passed through a Modality Encoder, which transforms them into modality-specific embeddings, such as numerical representations of images, audio, or video vectors. These embeddings are then passed through an Aligner, which projects them into the same space as the language model, allowing unified processing. Once aligned, all inputs, regardless of modality, exist as vectors within a shared representation space. This enables the LLMs to perform reasoning and understanding in a unified manner. From this point, the model can generate output text. Optionally, this output can be passed through an Output Decoder module, which converts the text into other formats such as images (Text2Img), audio (Text-to-Speech), or video (Text-to-Video), depending on the application. In specific implementations, such as BLIP-2, additional modules, like Q-Formers, are used to generate learnable queries that enhance information extraction from visual inputs. Likewise, some models incorporate Multi-Layer Perceptrons (MLPs) for intermediate processing or Multi-Head Attention (MH-Attn) mechanisms to improve the model's ability to identify patterns and relationships within multi-modal data.

One of the most recent innovations in developing multi-modal LLMs is the incorporation of the Mixture of Experts (MoE) architecture, which demonstrates significant improvements in performance and efficiency [54]. This approach relies on multiple specialized submodels (experts) trained to address specific tasks or modalities, such as visual recognition or NLP. Only the experts most relevant to the task are activated during inference,

enabling an efficient allocation of computational resources and reduced processing costs. Representative models, such as MoVA [55] and MoE-LLaVA [56], demonstrate that this strategy enhances accuracy in complex multi-modal scenarios and improves system scalability. Significantly, the benefits of the MoE architecture extend beyond multi-modal LLMs. Models such as DeepSeek-R1, which do not inherently operate in a multi-modal manner, also adopt this architecture to enhance specialization and computational efficiency. This demonstrates that MoE functions with a versatile design applicable across many LLM configurations [57]. Overall, the integration of MoE into multi-modal LLMs strengthens their adaptability and positions them as a robust solution for a wide range of applications in real-world environments, from human-machine interaction to dynamic multi-modal content generation [58].

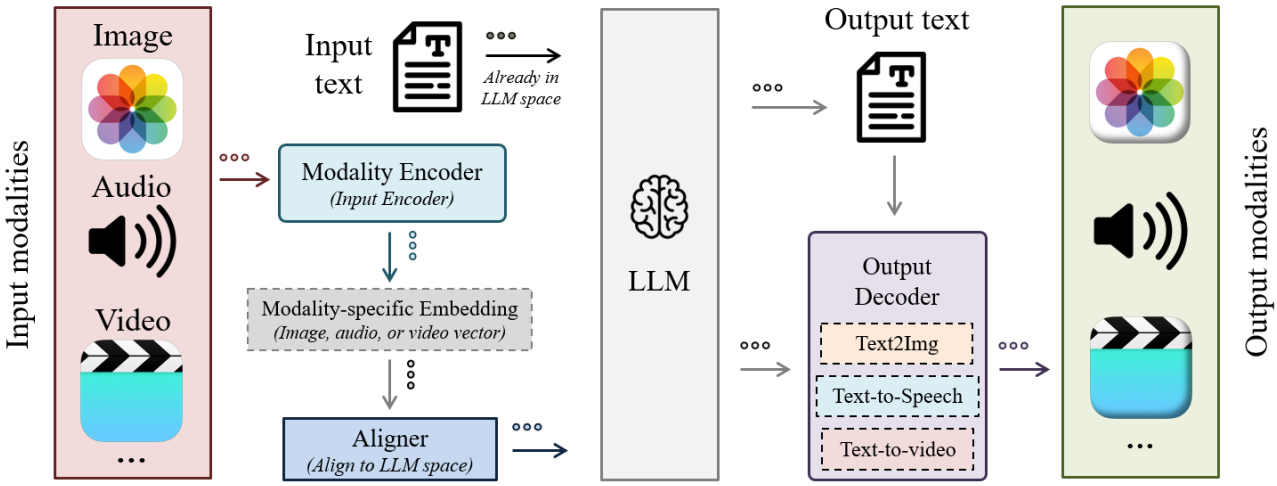

Figure 1. The typical multi-modal LLM architecture, based on [3].

In summary, advances in multi-modal LLMs enable more seamless integration between different data types, greatly expanding their applicability in complex and specialized contexts such as biomedical. The evolution of architectures such as MoE represents an essential step toward more efficient, scalable, and accurate models. These innovations not only enhance the ability of systems to interpret heterogeneous information but also pave the way for new forms of human-machine interaction, solidifying the role of multi-modal LLMs as fundamental tools in developing intelligent and adaptive applications.

2.3. Adaptation strategies for LLMs in the medical field

The development of LLMs represents a key advance in artificial intelligence, especially in their generative branch. These models revolutionize NLP, unlocking new possibilities across multiple domains. In general terms, AI architectures that learn from large volumes of unlabeled data, enabling them to perform various tasks such as language understanding, text generation, and natural language interaction, are referred to as *base models*. Models such as GPT-3 and GPT-4, PaLM and PaLM2, and LLaMA and LLaMA 2 demonstrate the potential to perform complex tasks without domain-specific training [59]. These models also significantly impact the biomedical field, where AI plays a key role in analyzing clinical and scientific data. Models such as BioBERT, BioMedLM, PubMedGPT, and MedPaLM-2 are specifically designed to process medical information, facilitating tasks including the classification of biomedical texts, extraction of relevant information, and generation of clinical summaries [49], [53], [60], [61]. Similarly, BioBERT and PubMedBERT are trained on specialized biomedical corpora, whereas ClinicalT5 demonstrates its ability to generate summaries from real clinical data [62], [63].

One of the main features of these models is their self-supervised training capability, which allows them to identify patterns in large datasets without the direct intervention of specialists. This approach proves effective in models such as GatorTron, which trains on more than 90 billion words extracted from de-identified clinical records, allowing it to analyze medical information accurately [64]. Despite their potential, implementing these models in clinical settings poses significant challenges. While tools such as BioGPT and PubMedGPT facilitate automation in biomedical research, their integration into healthcare systems requires adjustments to ensure their reliability and applicability in real-world settings [52].

Depending on the available data, goals, and technical resources, various strategies serve to adapt LLMs in the medical domain. These adaptation methods include:

- Prompt engineering, where carefully crafted inputs guide the model's behavior without changing parameters. This is particularly useful when fine-tuning is not feasible [65].
- Fine-tuning updates the model's weights using domain-specific datasets, enabling deep customization. This approach is also featured in models such as BioMedLM and MedPaLM-2 [66].
- Instruction tuning, where the model trains on examples of instruction-following tasks tailored to clinical language or workflows [7].
- Adapter layers and parameter-efficient methods, such as LoRA or PEFT, introduce small, trainable components into the frozen base model, which proves particularly helpful in resource-constrained environments [67].
- RAG augments the model with real-time access to external biomedical documents, improving accuracy without modifying the model [68].

These techniques prove essential in customizing LLMs for biomedical tasks, where precision, contextual sensitivity, and safety remain critical. While tools like BioGPT and PubMedGPT already enhance workflows, deploying such systems in clinical practice remains challenging, especially regarding validation, privacy, and trust.

2.4. Fine-tuning LLMs for the medical field

Developing and customizing LLMs enable significant advances in clinical and biomedical tasks. These models, refined through various fine-tuning strategies, address specific functions in the medical domain, including medical entity recognition, clinical question answering, and scientific document analysis. Table 1 presents some specifically adapted models to the medical field, including their tasks, base model, adaptation method, training data, size, and release date.

Table 1. LLMs in the medical field

Model	Tasks	Base Model	Adaptation Method	Training Data	Size	Release Date
BioELMo [69]	Named Entity Recognition (NER) and Biomedical Text Classification	ELMo	Pretraining from scratch	10M recent abstracts (2.46B tokens) from PubMed	~93M	2019
BioBERT [60]	NER, Biomedical Questions, and Semantic Analysis	BERT-Base	Fine-tuning	BooksCorpus + Wikipedia + PubMed	110M	Jan-19
PubMedBERT [62]	Relation Extraction and Medical Document Analysis	BERT-Base	Pretraining from scratch	PubMed abstracts only (3.1B words, 21GB)	110M	Jul-20
BioMegatron [70]	Predictions in Medical Texts	Megatron	Pretraining from scratch	PubMed abstract + PMC full-text (6.1B words)	345M	Nov-20
KeBioLM [71]	Medical Terminology Analysis	BERT-Base	Pretraining from scratch	PubMed abstracts + UMLS-linked entities (3.5M docs)	110M	Apr-21
ELECTRAMed [72]	Biomedical Information Classification and Extraction	ELECTRA	Pretraining from scratch	28.7M PubMed abstracts (~4B words, 26GB)	-	Apr-21
GatorTron [64]	Medical Records Processing	BERT	Pretraining from scratch	>90B words of de-identified clinical texts (PubMed + Wikipedia + MIMIC-III)	Up to 8.9B	Feb-22

BioGPT [52]	Specialized Medical Conversation	GPT-2	Pretraining from scratch	15M PubMed abstracts (~4B words)	347M	Nov-22
BioMedLM [61]	Biomedical Text Generation	GPT-2	Pretraining from scratch	PubMed abstracts and full articles (34.6B tokens)	2.7B	Dec-22
PubMedGPT [53]	Biomedical Text Generation	GPT-2	Pretraining from scratch	PubMed abstracts and full articles	2.7B	Dec-22
ClinicalT5 [63]	Automatic Clinical Summarization	T5-Base	Fine-tuning	MIMIC-III clinical notes (~2M documents)	220M	Dec-22
MedPaLM [49]	AI-Based Medical Assistance	PaLM-540B	Instruction tuning	MultiMedQA (Diverse medical data)	540B	Dec-22
MedPaLM-2 [49]	AI-Based Medical Assistance	PaLM 2	Instruction tuning + ensemble prompting	MultiMedQA (Diverse medical data)	-	May-23
EriBERTa [73]	Clinical Text Processing	RoBERTa	Pretraining from scratch	Spanish & English corpora (PubMed + MIMIC-III + EMEA)	-	Jun-23
BioMistral [74]	Advanced Biomedical Text Processing	Mistral-7B Instruct	Fine-tuning	PubMed Central (3B tokens)	7B	Feb-24
Medical mT5 [75]	Multilingual Medical Text Generation and Comprehension	mT5-Base	Fine-tuning	3B words from PubMed + ClinicalTrials + EMEA	738M	Apr-24

2.5. IA agents based on LLMs

AI agents represent a significant evolution in the use of LLMs. Unlike models that generate responses, agents operate autonomously, interact with multiple sources, execute complex tasks, and maintain long-running dialogues. These agents often combine LLMs with external tools, sequential reasoning capabilities, and context persistence, making them ideal for dynamic medical environments such as continuous clinical assistance [76], hospital system navigation [5], or personalized medical tutoring [16]. Despite their potential, the development and implementation of these approaches still present technical and ethical challenges that warrant further exploration.

Recent studies, such as Almanac [77], ArgMed-Agents [78], and MedChatZH [7], have begun to explore this paradigm in medicine, demonstrating practical applications in automated clinical responses, therapeutic simulations, and specialized expert dialogues.

3. Applications of LLMs in medical, healthcare, and educational contexts

LLMs' advanced linguistic understanding and text-generation capabilities have a significant impact on various contexts, including medical, healthcare, and educational. In the medical care field, understood as clinical activities directly linked to disease diagnosis, treatment, and follow-up [79], LLMs serve as key tools for classifying and extracting biomedical relationships, supporting computer-aided diagnosis, automating clinical

evaluation, and generating medical reports [4]. LLMs can also summarize medical records, answer specialized questions, and recognize relevant entities in clinical texts. These applications enhance diagnostic accuracy and facilitate medical staff's clinical decision-making.

In contrast, healthcare refers to the management and operation of healthcare services at the institutional and population levels [5]. In this context, LLMs demonstrate their usefulness by integrating into large-scale platforms to optimize workflows in hospitals and healthcare centers, reduce the administrative operational burden, and facilitate more agile, efficient, and patient-centered care. Their ability to automate bureaucratic, organizational, and logistical processes directly improves service quality, enabling healthcare professionals to devote more time to high-value clinical tasks.

Finally, LLMs are revolutionizing formal medical training and patient-focused education. In the academic context, LLMs simulate professional assessments, explain complex clinical concepts, and offer personalized pedagogical support tailored to each student's level and needs. In parallel, these models also facilitate health education in the general population, enabling patients to understand their condition better and adhere to treatment. This versatility positions LLMs as a valuable tool for improving training and care processes.

Table 2 summarizes some of the primary applications of LLMs in the medical, healthcare, and educational fields, as described in the functions above.

Table 2. Applications of LLMs

Medical Domain	Healthcare Domain	Education Domain
▪ Biomedical relationship classification ▪ Computer-aided diagnosis ▪ Automated clinical evaluation ▪ Medical report generation* ▪ Clinical summary generation ▪ Specialized question answering* ▪ Medical entity recognition	▪ Workflow optimization in hospitals ▪ Automation of complex processes ▪ Operational workload reduction ▪ Improved service quality ▪ Patient-centered care ▪ Medical report generation*	▪ Simulation of medical evaluations ▪ Explanation of clinical concepts ▪ Personalized educational support ▪ Patient-targeted education ▪ Improved understanding of treatments ▪ Specialized question answering*

Note. Items marked with * represent applications that are also relevant in other domains.

3.1. LLMs for medical applications

The development of LLMs enables their integration into various specialized medical tasks, where advanced linguistic processing and a deep contextual understanding of the clinical domain are required. These applications range from the automated extraction of relevant information from electronic health records to the generation of assisted diagnoses, from the classification of mental disorders in social networks to the multi-modal analysis of medical images and natural language [67], [79], [80].

Given the highly sensitive, technical, and contextual nature of medical language, many recent works opt for specific approaches that prioritize semantic accuracy and clinical relevance. This section presents some of these studies selected for their direct applicability in real medical settings, exemplifying how current models support clinical work, optimize diagnostic processes, or contribute to medical decision-making. Although the sample remains limited, it responds to a deliberate criterion of focusing on uses linked to the medical field without losing sight of the fact that the ecosystem of health applications is considerably broader and more diverse. Table 3 presents specific characteristics of the approaches used in each work, including the type of model, year of publication, and medical area of application. This table provides a complementary summary of the works reviewed, enabling us to identify methodological trends and the clinical domains covered by the analyzed studies.

Table 3. Summary of studies applying LLMs to clinical tasks in medical specialties

Ref.	Model	Year	Medical Area
[81]	ClinicalBERT	2019	Internal medicine and hospital management
[82]	ChatGPT + Isabela Pro	2023	Ophthalmology
[83]	GPT-4	2023	Ophthalmology
[84]	CHiLL	2023	Multispecialty
[85]	Trap-VQA	2023	Pathology
[86]	GPT-4 + InstructGPT-3 + LLaMA	2023	Psychiatry

McInerney *et al.* [84] present CHiLL, a method based on the Flan-T5 language model to automatically extract interpretable clinical features from medical notes and chest radiographs. This approach allows the unsupervised identification of relevant clinical features, facilitating more transparent and reliable medical predictions, especially in estimating hospital readmissions within 30 days. On the other hand, in mental health contexts, Yang *et al.* [86] use advanced models such as GPT-4, InstructGPT-3, and LLaMA to analyze social network posts, demonstrating that specific prompt engineering strategies significantly improve the detection of emotions, depression, anxiety, and suicide risk. This work highlights ChatGPT's ability to provide explanations and its near-human performance in clinical interpretations. Furthermore, Balas and Ing [82] compare ChatGPT with Isabel Pro in ophthalmological diagnoses, reporting that ChatGPT correctly identifies specific diagnoses in 9 out of 10 cases, significantly outperforming Isabel Pro's accuracy. Likewise, Ćirković and Katz [83] evaluate GPT-4 in clinical classification to determine patient eligibility for refractive surgery, obtaining promising results. This work highlights the potential of GPT-4 to aid in complex medical decisions informed by clinical data. For their part, Huang *et al.* [81] propose ClinicalBERT, a model specifically trained to predict hospital readmissions from structured clinical notes in internal medicine and hospital management settings. Finally, Naseem *et al.* [85] developed Trap-VQA, a visual-linguistic system based on Transformers that answers medical questions about pathological images. This approach increases predictive accuracy and provides interpretable explanations, which is crucial for clinical adoption in complex visual diagnoses. Although this selection represents only a fraction of the reviewed literature, it follows a specific criterion focused on applications directly related to the medical field. This delimitation enables a more nuanced understanding of how specific technological solutions are designed to intervene in clinical, diagnostic, and hospital processes, distinguishing them from other, more general, or cross-cutting works in digital health.

3.2. LLMs for healthcare applications

One of the most widespread applications of LLMs in the biomedical field is structuring information from unstructured clinical language. This task involves processes such as the automatic recognition of medical entities, the classification of clinical concepts, and the extraction of relationships between them. Together, these techniques enable the transformation of free text in medical records, scientific literature, or social forums into organized representations that feed clinical decision-support systems [76], specialized search engines [87], or predictive models [88].

As these models become more sophisticated, their applications in healthcare grow significantly, enabling complex clinical tasks to be addressed with greater accuracy, reasoning power, and contextualization. This section examines five key areas where LLMs demonstrate relevant impact in the healthcare environment. First, clinical entity recognition, biomedical relation extraction, and classification are introduced and developed in Section 3.2.1. Subsequently, in Section 3.2.2, computer-aided diagnosis receives attention, highlighting the use of models to simulate clinical reasoning and generate diagnostic explanations. Section 3.2.3 addresses automated clinical evaluation, including screening, therapeutic recommendation, and postoperative follow-up tasks. Section 3.2.4 presents advances in biomedical prediction to anticipate risks, disease progression, and clinical needs from textual and multi-modal data. Section 3.2.5 examines the application of models for addressing medical questions, with a focus on facilitating automated interaction between professionals and patients through natural language. Finally, Section 3.2.6 addresses report generation in the healthcare domain, analyzing how LLMs generate automated clinical documentation from both structured and unstructured data, with applications in specialties such as radiology, oncology, and internal medicine.

3.2.1. Recognition of entities, extraction, and classification of biomedical relations in healthcare

The automatic recognition of entities and subsequent classification and extraction of biomedical relationships represents a fundamental step in structuring clinical information dispersed in electronic records, scientific literature, or social networks. These tasks enable the identification of relevant entities, such as diseases, symptoms, treatments, or social factors, and the establishment of semantic relationships, thereby optimizing analysis processes, clinical decisions, and epidemiological surveillance. In recent years, these areas have undergone significant transformation through the advancement of LLMs such as BERT, GPT, and T5, as well as their clinical variants. These models are applicable in multiple clinical settings, utilizing various architectures and methodologies. Table 4 presents specific characteristics of the approaches used in each work, including the type of model, year of publication, and medical area of application. This table provides a complementary summary of the reviewed works, allowing the identification of methodological trends and clinical domains covered by the analyzed studies.

Table 4. Summary of studies applying LLMs to biomedical entity recognition and relation extraction tasks

Ref.	Model	Year	Medical Area
[89]	MentalBERT + MentalRoBERTa	2021	Psychiatry
[90]	LFBERT	2024	Psychiatry and mental health
[91]	medBERT.de	2024	Multispecialty
[92]	DepGPT	2024	Psychiatry and mental health
[93]	Pediatric Stroke GPT-3.5	2024	Pediatric Neurology
[94]	Flan-T5 XXL + Flan-T5 XL + GPT-3.5 + GPT-4	2024	Oncology, Internal medicine, Radiotherapy,
[95]	GPT-3.5 Turbo	2024	Oncology, Pathology
[96]	GPT-4 API	2024	Oncology, Internal medicine, Translational pharmacology
[15]	KLUE-RoBERTa + KLUE-BERT + KoBERT + KorBERT	2024	Emergency medicine
[97]	BiomedRAG	2024	Oncology, Internal medicine, Clinical pharmacology
[98]	AssistMED	2024	Cardiology
[99]	EHR-BERT	2024	Cardiology, Internal medicine, Clinical pathology
[100]	CaseGPT	2024	Internal medicine, Legal medicine, Clinical pharmacology
[16]	GPT-3.5	2024	Endocrinology and otolaryngology
[68]	RECTIFIER	2024	Cardiology
[101]	MentaLLaMA	2024	Psychiatry
[77]	Almanac	2024	Cardiology, Cardiac surgery, Neurology, Gastroenterology, Nephrology, Infectious diseases, Pediatrics

In the context of systematic reviews, GPT-4 integrates with the RAG architecture through LangChain to classify articles and extract biomedical data, including medical conditions and model architectures. This automation allows the structuring of key information in patient-clinical trial matching studies, facilitating the extraction of complex biomedical relationships in areas such as oncology, neurology, and rare diseases [96]. Similarly, Flan-T5 XL and XXL models are trained by Guevara *et al.* [94] to extract social determinants of health (SDoH) from narrative clinical notes, encompassing categories such as employment, housing, transportation, parental status, relationships, and social support, derived from free text. These models facilitate the structuring of critical information that is poorly documented in electronic records, revealing semantic relationships between social factors and clinical profiles.

In narrative clinical settings, ChatGPT demonstrates its ability to transform unstructured clinical notes into organized data [95], while AssistMED automates the extraction of diagnoses and medications from electronic records [98]. BiomedRAG structures multisource biomedical knowledge by integrating relevant entities in the contexts of oncology and internal medicine [97]. Models such as EHR-BERT are developed to identify anomalous clinical events within medical records [99], or CaseGPT extracts information from EHRs to generate clinical recommendations [100]. Fiedler *et al.* [93] developed Pediatric Stroke GPT (PS-GPT), a GPT-3.5-based model designed to extract clinical information from medical notes of pediatric stroke patients and automatically complete records for the International Pediatric Stroke Study (IPSS).

In mental health, the LFBERT model applies to detect anxiety or suicidal ideation from text [90], revealing relationships between emotional language and psychological states. Ji *et al.* [89] developed MentalBERT, a model that classifies mental disorders based on social networks, while Yang *et al.* [101] proposed MentaLLaMA to categorize anxiety and depression. Models such as DepGPT identify depressive publications, connecting linguistic patterns with clinical variables [92]. Regarding medical entity recognition, models such as KoBERT and KLUE-BERT are employed by Lee *et al.* [14] to identify symptoms and clinical histories in simulated doctor-patient conversations during emergencies. The system enables automatic recognition of key clinical entities from natural language, showing its usefulness in supporting automated registration in triage settings. In German clinical texts, models such as medBERT.de are also used to identify diseases, medical devices, and clinical procedures. This adaptation to specialized medical language shows substantial improvements over general models, facilitating the automatic structuring of medical documents in hospital environments [91].

RECTIFIER was developed by Unlu *et al.* [68] to automate the extraction of inclusion and exclusion criteria from unstructured clinical notes, supporting participation in a study focused on patients with heart failure. The system analyzes clinical records using directed questions and semantic retrieval, determines patient eligibility with high accuracy and efficiency, and reduces the burden on staff in the screening process for clinical trials. Almanac, a GPT-4-based system for answering open-ended clinical questions, was also developed by Zakka *et al.* [77]. The model identifies key entities, such as diseases, interventions, and outcomes, and generates clinical answers by organizing explicit relationships between diagnosis, treatment, and prognosis. This structure enables the model to function as a reliable clinical assistant, providing accurate, verifiable, and error-resistant responses. ChatGPT is used by Sievert *et al.* [16] to extract relevant information about thyroid nodules from clinical reports and support diagnostic decisions. The model identifies textual patterns associated with the risk of malignancy and links clinical descriptions with therapeutic decisions, demonstrating its support for text-based risk stratification.

These proposals demonstrate that the automatic recognition of clinical entities, followed by their classification and the extraction of biomedical relationships, establishes itself as a key sequence in applying LLMs in healthcare. The reviewed approaches reveal a broad spectrum of practical uses in multiple medical specialties, from the automated structuring of clinical notes to identifying social, psychological, and pathophysiological factors. The integration of these models enables the organization of large volumes of dispersed information and the inference of clinically relevant links, ultimately contributing to more accurate, personalized, and data-driven medicine. This convergence of language, clinical information, and intelligent systems marks a significant advance toward more efficient care environments and better-informed research processes.

3.2.2. Computer-aided diagnosis in the healthcare field

Computer-aided diagnosis represents a significant advance in the clinical setting, enabling improved accuracy and efficiency in disease identification by analyzing various data types. In recent years, LLMs such as GPT, BERT, and LLaMA have revolutionized this task, enabling automated clinical reasoning, generating structured diagnostic explanations, and integrating text with visual data. Table 5 presents specific characteristics of the

approaches used in each work, including the type of model, year of publication, and medical area of application. This table provides a complementary summary of the reviewed works, allowing the identification of methodological trends and clinical domains covered by the analyzed studies.

Table 5. Summary of research on the use of LLMs in computer-assisted diagnosis tasks

Ref.	Model	Year	Medical Area
[9]	GPT-3.5	2023	Radiology
[102]	DR. KNOWS	2023	Multispecialty
[67]	PneumoLLM (LLaMA-7B + ViT-L/14 from CLIP)	2024	Nephrology
[103]	GPT-3.5	2024	Ophthalmology
[78]	ArgMed-Agents	2024	Multispecialty
[104]	SkinGEN	2024	Dermatology
[105]	GPT-3.0 + GPT-4, with clinical-specific CoT prompts	2024	Multispecialty

A tool based on GPT-3.5 and GPT-4 was developed by Savage *et al.* [105] that simulates medical reasoning using prompts. The model represents processes such as differential diagnosis, intuitive analysis, and Bayesian inference, i.e., the ability to update diagnostic hypotheses based on new patient information and progressively evaluate disease probabilities. This generates a step-by-step explanation that improves interpretability in complex clinical settings. This focus on diagnostic transparency is complemented by [102], where DR.KNOWS, a system that incorporates medical graphs into LLMs to improve diagnostic accuracy from clinical notes, is proposed. The system predicts conditions such as sepsis or pneumonia by analyzing semantic paths between medical concepts extracted from text, thereby structuring more coherent diagnostic hypotheses.

In the same vein, ArgMed-Agents, a multi-agent system that simulates clinical discussions using argumentation schemes, is presented in [78]. This tool evaluates treatments, generates counterarguments, and selects justified therapeutic options, contributing to the transparency and reliability of the medical decision-making process. This multi-modal approach is complemented by [103], where ChatGPT-3.5 is used to diagnose different types of glaucoma from clinical cases written in natural language. The model identifies relevant patterns in the text and generates consistent differential diagnoses, demonstrating a capability comparable to that of resident physicians, particularly in settings with limited access to specialists.

Song *et al.* [67] propose PneumoLLM, a model that integrates computer vision with LLaMA to directly diagnose pneumoconiosis from chest radiographs, eliminating the need for textual input. This approach enables accurate diagnoses with limited data, making it useful in resource-poor clinical settings. In turn, SkinGEN, a dermatological diagnostic system that combines visual-linguistic models to analyze skin images, generate diagnoses, and produce explanatory visualizations, was developed by Lin *et al.* [104]. This strategy is purported to improve patient understanding and strengthen confidence in automated diagnosis.

On the other hand, Rao *et al.* [9] evaluate ChatGPT's ability to select appropriate imaging studies in cases of breast pain and breast cancer screening. The tool demonstrates high accuracy when analyzing structured clinical questions, and its recommendations align with the American College of Radiology criteria, underscoring its usefulness in radiological clinical decision-making. These proposals reflect remarkable methodological progress: today's LLMs classify, explain, and reason. This capability extends successfully to multiple medical specialties, including pulmonology [67], dermatology [104], radiology [9], and general clinical care scenarios [78], [102], [103], [105]. The convergence between text, image, and structured medical knowledge points toward more accurate, explainable, and data-driven medicine.

3.2.3. Automated clinical assessment in healthcare

Integrating LLM models in the clinical setting enables the automation of traditionally complex tasks such as differential diagnosis, treatment selection, and complication assessment. These tools, trained with large volumes of biomedical and clinical text, interpret symptoms, generate reasoned explanations, and propose interventions

tailored to the patient's context. Several studies illustrate the scope and evolution of these technologies in real medical settings. Table 6 presents specific characteristics of the approaches used in each work, including the type of model, year of publication, and medical area of application. This table provides a complementary summary of the reviewed works, allowing the identification of methodological trends, medical specialties addressed, and the degree of integration between text, image, and structured clinical knowledge.

Table 6. Summary of research on the use of LLMs in clinical assessment tasks

Ref.	Model	Year	Medical Area
[18]	GPT-3.5	2023	Psychiatry
[106]	Chatbot based on Facebook Messenger	2023	Oncology
[8]	GPT-4	2024	Orthopedics and sports medicine
[107]	GPT-4 + Google Gemini	2024	Ophthalmology
[108]	GPT-3.5 Turbo	2024	Radiation oncology
[109]	GPT-4 + LLaMA Chat-2.0 (Meta AI)	2024	Otorhinolaryngology
[17]	GPT-3.5 Turbo + Google Bard (PaLM 2)	2024	Emergency medicine
[110]	GPT-3.5 Turbo + GPT-4 + GPT-3.5 Clinical Assistant + Aya-101 + Nemotron Clinical Assistant	2024	Psychiatry
[111]	GPT-4 + Google Gemini Pro	2024	Plastic surgery
[112]	GPT-3.5 + GPT-4 + Bing AI (Microsoft)	2024	Ophthalmology
[113]	GPT-4	2024	Urological oncology
[114]	GPT-4 + Mixtral 8x7b	2024	Infectious diseases

In mental health, Gargari *et al.* [110] compare multiple models, including ChatGPT-3.5, GPT-4, and specialized clinical models, for diagnosing mental disorders according to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) cases, highlighting the accuracy of GPT-3.5 and the argumentative robustness of GPT-4. Franco D'Souza *et al.* [18] evaluate ChatGPT-3.5 using 100 psychiatric clinical vignettes, where 61% of the responses receive an "A" rating from experts. These proposals reflect the potential of LLMs as support tools in mental health screening, diagnosis, and follow-up.

In the surgical and postoperative setting, Gomez-Cabello *et al.* [111] analyze the usefulness of several LLMs in generating recommendations after cosmetic surgeries, highlighting GPT-4 for its clinical depth. In addition, Hsueh *et al.* [113] studied the ability of GPT-4 to detect complications after renal surgery, achieving an 86.7% accuracy rate in identifying postoperative problems. For their part, Huang *et al.* [106] developed a chatbot for symptomatic follow-up during chemotherapy, reducing unplanned hospitalizations. These initiatives consolidate the role of LLMs in improving clinical monitoring and continuity of care in surgical and oncological scenarios.

In emergency settings, Garg *et al.* [17] evaluate Bard and ChatGPT applying the START (Simple Triage and Rapid Treatment) protocol, where Bard shows greater accuracy under time pressure. An analogous approach is presented in [8], which utilizes GPT-4 to prioritize causes of knee pain in outpatient triage. Both studies highlight the use of LLMs to accelerate clinical classification processes and support decision-making under high-demand conditions.

On the other hand, Carlà *et al.* [107] evaluate ChatGPT-4 and Gemini in the context of surgical recommendations for glaucoma, observing greater consistency and accuracy in GPT-4, which reinforces its value as a support tool in ophthalmology. In the context of prevention and automated surveillance, Gopalakrishnan *et al.* [112] propose a personalized screening system for diabetic retinopathy tailored to individual clinical risk factors. Complementarily, Wiemken and Carrico [114] use GPT-4 and Mixtral in an RAG system to verify compliance with National Healthcare Safety Network (NHSN) clinical definitions, such as Central Line-Associated Bloodstream Infection (CLABSI) and Catheter-Associated Urinary Tract Infection (CAUTI). Both works demonstrate how language-based models support preventive and automated monitoring strategies in clinical settings with high operational burdens.

In radiation therapy, Dennstädt *et al.* [108] studied the ability of GPT-3.5 to answer clinical questions. It performed well on closed questions, those with defined answers, such as dose indications or standard schedules. Still, it identified limitations when facing open questions, which require broader clinical reasoning, such as justifying the choice of treatment or adapting the decision to the patient's context. This contrast underlines the importance of integrating explicit medical knowledge in tasks that require contextual interpretation, clinical judgment, or individualized adaptation.

Finally, Dronkers *et al.* [109] analyze therapeutic decision-making in bilateral vocal cord paralysis. ChatGPT and LLaMA 2 show partially correct but risky strategies, revealing the need for models trained specifically for highly specialized domains.

Together, these works solidify the utility of LLMs in automating clinical assessments with applications in multiple medical specialties. The ability of these models to assess, reason, and personalize clinical recommendations suggests a paradigm shift toward more innovative, more accessible, and patient-centered medicine. However, clinical validation, bias management, contextual interpretation, and ongoing medical monitoring remain significant challenges. The future of these systems depends on their responsible integration, emphasizing explainability, equity, and patient safety.

3.2.4. Biomedical prediction in the healthcare field

Biomedical prediction seeks to anticipate risks, disease progression, and clinical needs based on textual and multi-modal information, playing a vital role in developing personalized, proactive, and data-driven medicine. In this context, models such as BERT, GPT-4, and specialized variants like Health-LLM or AD-BERT facilitate the integration of clinical language, enabling accurate and contextually informed predictions about a patient's condition and progression. Table 7 presents the characteristics of the approaches used in each work, including the type of model, year of publication, and medical area of application. This table complements the clinical prediction analysis, facilitating the identification of domains addressed and models adapted to the specific context of each task.

Table 7. Summary of research on the use of LLMs in clinical prediction tasks

Ref.	Model	Year	Medical Area
[5]	KM-BERT	2023	Multispecialty
[79]	BERT + GPT-2.0	2023	Oncology
[115]	AD-BERT	2023	Neurology
[88]	GPT-4	2024	Cardiology
[116]	Health-LLM	2024	Gastroenterology, Pulmonology, Endocrinology

One of the most relevant approaches focuses on automated triage from the first point of contact. In this regard, KM-BERT is a pre-trained model developed by Kim *et al.* [5] for Korean medical texts, capable of identifying the most relevant medical specialty for patient care based on patients' self-reported descriptions of symptoms. The system, which encompasses 27 clinical areas, demonstrates how linguistic adaptation to the biomedical domain and everyday language substantially enhances efficiency in automated clinical referrals, particularly in high-demand or resource-limited settings.

Complementing this early classification approach, Mao *et al.* [115] developed AD-BERT, a model trained on unstructured clinical notes to predict conversion from mild cognitive impairment to Alzheimer's disease. This model stands out for its ability to anticipate longitudinal data contained in electronic clinical records. It provides a valuable tool for early detection in neurological settings, where time to intervention proves critical.

Beyond traditional diagnosis, some studies address more subjective dimensions of care, such as patient quality of life. Mao *et al.* [79] explore this aspect through the combined use of BERT and synthetic data generated by GPT-2, applied to analyzing transcribed interviews of patients with thyroid cancer. This approach allows the identification of physical and emotional trajectories without relying on structured questionnaires, opening the door to non-intrusive monitoring systems focused on patient well-being.

In internal medicine, Jin *et al.* [116] present Health-LLM, a model trained to predict disease in gastroenterology and endocrinology. Its strength lies in combining clinical feature extraction and contextual retrieval of medical knowledge, which allows it to surpass previous models in accuracy and diagnostic personalization. This integration between clinical reasoning and semantic text representation represents a key advance in automated medical decision-making.

For its part, Han *et al.* [88] evaluate the performance of GPT-4 in predicting 10-year cardiovascular risk using data from multicenter population-based cohorts. The model demonstrates comparable performance to established tools, such as the Framingham scales or ACC/AHA guidelines, with the added advantage of maintaining accuracy in scenarios with incomplete clinical data, a critical feature in real-world healthcare settings.

Together, these proposals demonstrate the growing sophistication of LLMs for clinical prediction tasks in various domains, including neurology, oncology, ophthalmology, cardiology, and internal medicine. These models no longer limit themselves to interpreting clinical language; LLMs now learn from it, anticipate disease trajectories, and propose personalized treatment plans. This convergence between artificial intelligence, language, and medical data opens new possibilities for predictive, accessible, and adaptive medicine.

3.2.5. Answers to medical questions in the healthcare environment

Generating automated clinical responses represents a strategic advance in patient-model interaction, with direct applications in counseling, health education, and initial screening. Answering medical questions accurately and with an evidence-based approach is essential to supporting patients and professionals. LLMs such as ChatGPT, Bard, and Med-PaLM 2 revolutionize this task by processing complex clinical text and generating accessible explanations. Table 8 summarizes the characteristics of work that applies LLMs to medical response tasks, organized by model, year, and specialty addressed. This table complements the analysis of responses to medical questions, facilitating the identification of domains addressed and models adapted to the specific context of each task.

Table 8. Summary of research on the use of LLMs to answer medical questions

Ref.	Model	Year	Medical Area
[89]	MentalBERT + MentalRoBERTa	2021	Psychiatry
[117]	GPT-3.5	2023	Gastroenterology and Hepatology
[118]	GPT-3.5	2023	Urology
[119]	GPT-3.5 + GPT-4 + Google Bard	2023	Ophthalmology
[120]	Bard AI + GPT-3.5	2023	Ophthalmology
[121]	GPT-3.5	2023	Multispecialty
[49]	Med-PaLM 2	2023	Multispecialty

Ref.	Model	Year	Medical Area
[122]	EyeGPT	2024	Ophthalmology
[66]	GPT-3.5 + LLaMA2-7b + LLaMA2-7b-Chat + LLaMA2-13b + LLaMA2-13b-Chat	2024	Ophthalmology
[123]	GPT-3.5 + GPT-4 + Google Gemini (formerly Bard)	2024	Plastic Surgery
[124]	GPT-3.5 + Google Bard (now Gemini)	2024	Cardiology
[125]	Psy-LLM	2024	Psychiatry and Mental Health
[126]	GPT-3.5 + Google Bard (now Gemini)	2024	Spine Surgery and Traumatology
[127]	GPT-3.5 + GPT-4	2024	Radiology and Cardiology
[128]	GPT-4 + Gemini (Google AI, formerly Bard) + Microsoft Copilot (formerly Bing Chat) + ChatSpot (HubSpot) + PiAI (Inflection AI)	2024	Cardiology, Oncology, Dermatology
[129]	GPT-4	2024	Endocrinology
[130]	GPT-3.5 Turbo + GPT-4	2024	Geriatrics and Cognitive Neuroscience
[48]	GPT-3.5 + GPT-4 + LLaMA-2-chat (7B and 70B)	2024	Internal Medicine, Surgery, Neurology, Gynecology, Pediatrics, Pediatrics
[131]	Aeyeconsult	2024	Ophthalmology
[7]	MedChatZH	2024	Traditional Chinese Medicine
[68]	RECTIFIER	2024	Cardiology
[101]	MentaLLaMA	2024	Psychiatry
[77]	Almanac	2024	Multispecialty

One of the most consistent findings across studies is the ability of LLMs to generate clinically relevant responses with high levels of empathy. Ayers *et al.* [121] compare ChatGPT-3.5 responses with those of specialist physicians in response to real questions posed by patients in social forums. In an analysis of 195 exchanges, healthcare professionals rate the reactions and, surprisingly, prefer ChatGPT responses in 78.6% of cases. The model is 3.6 times more likely to produce "good or excellent" responses, which were nearly 10 times more empathetic. These results suggest a complementary role in asynchronous settings, such as clinical messaging or digital community care. This same pattern also appears in postoperative settings [123] and in chronic conditions, such as heart failure [124] or liver diseases [117], where the models exhibit high accuracy rates, albeit with limitations in actionability or in delivering clinical thresholds.

In ophthalmology, specialized models are developed with outstanding results. EyeGPT, trained by Chen *et al.* [122] on a Chinese ophthalmic clinical corpus, outperforms generalist models such as ChatGPT-3.5 and HuatuoGPT in report generation tasks, technical explanations, and clinical support. Complementarily, Tan *et*

al. [66] evaluate five LLMs tuned with 400 real ophthalmological questions and observe that GPT-3.5 achieves an accuracy of 87.1%, although all models exhibit clinical errors and hallucinations. To improve traceability, Aeyeconsult by Singer *et al.* [131] uses GPT-4 and RAG architecture. It achieves an accuracy of 83.4% against OKAP-type questions, demonstrating higher stability and lower variability than ChatGPT-4.

In mental health, Ji *et al.* [89] introduce MentalBERT and MentalRoBERTa, models specifically adapted for forum language, such as Reddit. MentalRoBERTa outperforms models such as BERT and BioBERT in detecting depression, anxiety, and suicidal risk. More recently, MentaLLaMA by Yang *et al.* [101], trained with the IMHI dataset, combines reinforcement learning and interpretable explanation generation, outperforming even GPT-4 in reasoning quality. In addition, Psy-LLM by Lai *et al.* [125] proposes a conversational approach in Chinese to provide psychological support with ethical and patient-centered responses.

The use of LLMs extends to areas such as endocrinology and pregnancy [129], where ChatGPT-4 demonstrates high reliability but low readability, and urology [118], where limitations in consistency and source citation are identified. On the other hand, models such as Almanac [77] and RECTIFIER [68], based on GPT-4 with RAG architectures, demonstrate higher accuracy, efficiency, and traceability in open clinical responses and patient screening for trials, respectively.

Comparative studies reflect apparent differences between models. ChatGPT-4 outperforms GPT-3.5 in accuracy, completeness, and consistency in several clinical tasks [119], [127], particularly in cardiac imaging and myopia. Sandmann *et al.* [48] confirm these advantages in diagnosis and clinical examination, with no statistically significant differences in therapeutic suggestions. However, even GPT-4 exhibits significant errors in rare diseases or with ambiguous prompts, necessitating ongoing clinical validation.

From a technical point of view, models that integrate augmented retrieval, such as Aeyeconsult, RECTIFIER, and Almanac, prove to be more accurate and traceable. These approaches enable the citation of verified sources, such as PubMed or clinical textbooks [68], [77], [131], thereby reducing the risk of hallucination. Additionally, the proposal to utilize GPT-4 as an automated medical content evaluator [66] represents a significant step toward large-scale clinical validation.

Despite progress, readability remains low for patients with low health literacy [129]. Clinical errors, hallucination, or information overload risks persist, even in the best-performing models [127], [128]. Adherence to official guidelines is still limited [118], which imposes barriers to their autonomous implementation.

These proposals reflect a growing maturity of LLMs in medicine but also emphasize the need for regulated frameworks, human review, and more specialized developments. Models such as EyeGPT [122], AeyeConsult [131], MentaLLaMA [101], or MedChatZH [7] demonstrate that integrating structured sources and the multilingual approach enhances clinical utility. However, their adoption must consider their performance, ethics, traceability, and safety in genuine care settings.

3.2.6. Clinical report generation in the healthcare environment

The automated generation of clinical reports is another key area where LLMs demonstrate a relevant impact. These models enable the creation of medical reports from both structured and unstructured clinical data, thereby enhancing the efficiency, consistency, and quality of healthcare documentation. Table 9 summarizes the most representative works in this line, organized by model, year, and medical specialty. A variety of architectures are used, ranging from generalist models such as GPT-3.5 and GPT-4 to specialized variants like Radiology-LLAMA2 or ChatCAD+, which are tailored to different clinical contexts, including radiology, oncology, internal medicine, and geriatrics. This compilation compares approaches and highlights the steady growth in using LLMs to automate specialized medical reports.

Table 9. Summary of research on the use of LLMs in clinical report generation tasks

Ref.	Model	Year	Medical Area
[63]	ClinicalT5	2022	General Medicine
[64]	GatorTron	2022	General Medicine
[12]	GPT-3.5	2023	Radiation Oncology

Ref.	Model	Year	Medical Area
[132]	ChatCAD+	2023	Radiology
[133]	Claude-instant-v1.0 + GPT-3.5-Turbo + Command-xlarge-nightly + Bloomz	2023	Multispecialty
[134]	Radiology-LLaMA2	2023	Radiology
[46]	GPT-4	2024	Radiology
[47]	ChatGPT + BART + T5	2024	Multispecialty
[65]	LLaMA 2	2024	Internal Medicine, Surgery, Intensive Care
[76]	T5 + BART + RNN+LSTM	2024	Radiology
[135]	LLaMA 2 (13B)	2024	Geriatrics and Clinical Nutrition

One of the most relevant approaches in the literature focuses on utilizing LLMs for clinical report generation, with applications spanning general medicine to specialties such as radiology, oncology, and clinical nutrition. ClinicalT5, as introduced by Lu *et al.* [63], is one of the first models explicitly tailored to the medical domain, designed to generate clinical text, summarize medical notes, and classify clinical documents and records. This line of work evolves into more complex systems, such as the one developed by Wilhelm *et al.* [133], which combines Claude-instant-v1.0, GPT-3.5-Turbo, Command-xlarge-nightly, and Bloomz to generate automated therapeutic recommendations for clinical questions.

In the hospital setting, LLaMA 2 by Goswami *et al.* [65] summarizes discharge reports, prioritizing clarity to facilitate understanding by other professionals. Similarly, Latif and Kim [47] propose a hybrid architecture integrating ChatGPT with BART and T5 to reformulate complex clinical texts using large-scale language models. For its part, Guckenberger *et al.* [12] use ChatGPT (GPT-3.5) to write scientific texts and generate research hypotheses in medical environments.

Radiology remains one of the most explored areas. Liu *et al.* [134] developed Radiology-LLaMA2, a model designed for generating radiology reports and analyzing medical findings. In a similar vein, Gulati *et al.* [46] use GPT-4 to transform technical language into patient-friendly versions, suggesting a potential use in accessible clinical communication. Zhao *et al.* [132] present the ChatCAD+ proposal, which moves in this direction by applying models such as ChatGPT and LLaMA to generate reports directly from diagnostic images.

A noteworthy technical approach is presented by López-Úbeda *et al.* [76], who evaluate a multi-modal architecture combining T5, BART, RNN, and LSTM for writing medical reports on knee MRI. Furthermore, Alkhalafin *et al.* [135] applied LLaMA 2 to generate reports and de-identify sensitive clinical data in geriatric patient records, striking a balance between accuracy and privacy protection. Finally, Yang *et al.* [64] introduce GatorTron, which is oriented toward processing electronic health records, with the ability to extract clinical concepts and map semantic relationships within texts.

These studies reflect a consolidated trend toward integrating LLMs in clinical tasks beyond practitioner support, including generating understandable text for patients, report automation, image analysis, and ethical data management. Despite advances, challenges such as ensuring traceability, conducting clinical validation, and adapting to the local context remain. However, the progress made so far suggests a promising path toward more efficient, accessible, and personalized medicine through the strategic use of LLMs.

3.3. Applications of LLMs in medical education contexts

LLMs emerge as innovative tools across multiple fields, including health sciences education. Their ability to process large volumes of information, generate coherent content, and adapt to different contexts positions them as key resources in transforming teaching, learning, and clinical assessment [136]. This section examines how these models are applied in education through three primary applications.

Section 3.3.1 examines the application of LLMs in entity recognition, extraction, and classification of biomedical relationships, which is crucial for structuring medical knowledge and facilitating education. This technology enables the automated identification of clinical concepts and their interrelationships, supporting academic training by analyzing specialized texts. Subsequently, Section 3.3.2 focuses on automated clinical assessment and the generation of reports. In this context, LLMs aid in creating clinical simulations and questionnaires and generate educational reports that enhance feedback and strengthen students' diagnostic competencies. Finally, Section 3.3.3 explores the application of these models in addressing medical questions, highlighting their potential as virtual tutors that can resolve doubts, reinforce knowledge, and foster self-directed learning. From exam preparation to clinical case simulation, LLMs redefine how knowledge is accessed and transmitted in medical education. These three areas represent a significant evolution in utilizing artificial intelligence to enhance educational processes, providing scalable, personalized, and evidence-based solutions.

3.3.1. Recognition of entities, extraction, and classification of biomedical relationships in the field of education in medical contexts

Several research efforts focus on entity recognition, information extraction, and classification in biomedical texts, which are essential for enhancing information quality and supporting both educational and research processes. Table 10 summarizes the characteristics of each model, including year and specialty. This table complements the analysis of entity recognition, extraction, and classification of biomedical relations, facilitating the identification of domains addressed and models adapted to the specific context of each task.

Table 10. Summary of research on using LLMs for applications in the educational domain, entity recognition, extraction, and classification of biomedical relationships

Ref.	Model	Year	Medical Area
[71]	KeBioLM	2021	Multispecialty
[137]	Chat-ePRO (based on ChatGPT-3.5 Turbo)	2024	Oncology
[138]	GPT-3.5	2024	Gynecology and Urogynecology

Progress in biomedical language processing advances through the development of models that combine precision, automation, and contextual understanding, generating direct implications for health sciences education. A conversational system that interacts directly with users is introduced in the context of data collection in educational and clinical settings. Based on generative language techniques, it improves participation, enriches the patient-student learning experience, and maintains content accuracy, demonstrating its usefulness as a pedagogical tool in patient-centered education [137]. Complementarily, there is a need to identify texts generated by AI, especially in academic contexts. Automatic systems outperform human evaluators in this task. This reinforces the need for reliable tools to preserve academic integrity, particularly in the development of papers, essays, and scientific publications within educational environments [138]. Finally, biomedical LLMs integrate structured knowledge from ontological systems [71]. By incorporating this type of information, the model achieves superior performance in concept recognition and complex relation extraction tasks, facilitating its application in specialized educational platforms and continuing education environments where semantic precision and contextual understanding are essential. This work reflects a steady evolution toward more autonomous, accurate, and interpretable models, which improve clinical and scientific systems and positively transform teaching, assessment, and learning strategies in the biomedical domain.

3.3.2. Automated clinical assessment and reporting in education settings

With the rapid advancement of language models, the methods by which medical training processes are evaluated and receive support are evolving. A significant body of work explores their applicability in automated clinical assessment, demonstrating their potential to assist in preparing medical and surgical examinations. Table 11 summarizes the characteristics of the works, organized according to model, year, and specialty addressed. This table complements the analysis of automated clinical evaluation and report generation in education, facilitating the identification of the domains addressed and models adapted to the specific context of each task.

Table 11. Summary of research on the use of LLMs for automated clinical assessment and report generation in the educational domain

Ref.	Model	Year	Medical Area
[139]	BioBART	2022	Multispecialty
[52]	BioGPT	2023	Multispecialty
[140]	ChatGPT	2023	General Medicine
[141]	GPT-3.5	2023	Neurosurgery
[142]	Med-HALT	2023	Multispecialty
[14]	Bing Chat (Microsoft) + GPT-3.5 + Bard (Google) + LLaMA 2 (Meta)	2024	Multispecialty
[143]	GPT-4	2024	General Surgery
[144]	GPT-3.5 + GPT-4 + Gemini (Google AI) + LLaMA 2 (Meta AI) + Copilot (Microsoft, GPT-4 Turbo-based)	2024	Oral and Maxillofacial Surgery
[145]	GPT-3.5 + GPT-4 + Bard (Google) + Bing Chat (Microsoft, GPT-4-based)	2024	Dentistry
[146]	GPT-4, Bing (Microsoft), and Bard (Google)	2024	Bariatric Surgery

Developing biomedical LLMs promotes new ways to support health sciences education. These models automate clinical assessment and generate helpful content for training, facilitating learning across different levels and medical specialties.

Quah *et al.* [144] evaluate ChatGPT, Gemini, and Copilot for answering certification exam questions in oral and maxillofacial surgery, obtaining promising results in terms of accuracy and usefulness as a study tool. Similarly, Lee *et al.* [146] discuss their application in the training of bariatric surgeons, highlighting how these models facilitate both autonomous learning and preparation for high-level assessments. For its part, Tsoutsanis and Tsoutsanis [14] address the use of multilingual models to support the preparation of medical exams across multiple specialties, including internal medicine, psychiatry, and gynecology. This study reveals that, in many cases, the performance of the models equals or exceeds that of doctors in training, which reinforces their usefulness as an educational resource.

In the dental field, Yamaguchi *et al.* [145] explore their use in preparing national exams for dental hygienists in Japan, confirming their effectiveness as a complementary tool in technical and professional training. In addition, Beaulieu-Jones *et al.* [143] investigate the performance of ChatGPT-4 in general surgical training, noting its ability to replicate and even surpass human clinical reasoning on standardized tests. Further work [141] combines automated assessment with report generation, focusing on neurosurgical clinical questions. It highlights the clarity of the answers generated and their consistency against established medical sources, positioning it as a didactic and feedback resource for students. Other studies focus on the use of models for developing educational and clinical reports. Pal *et al.* [142] propose using models to reduce hallucinations in medical texts, thus improving the reliability of the content generated for academic contexts. Along the same lines, [139] and [52] utilize specialized models, such as BioBART and BioGPT, to produce summaries of medical documents, simulate clinical dialogues, and synthesize information, which is practical in case-based training contexts. Finally, Eysenbach [140] proposes educational applications, such as generating clinical simulations, questionnaires, and assisted article writing, that integrate AI as an active resource in the teaching and learning process. This work evidences a growing trend towards incorporating LLMs in medical educational environments, not only as assistants in clinical assessment but also as facilitators of autonomous learning, critical thinking, and academic production.

3.3.3. Answers to medical questions in the field of education

Utilizing LLMs in education unlocks new opportunities for enhancing comprehension, self-learning, and personalized education in the health sciences. Several proposals show that these models can effectively assist in answering medical questions, conducting clinical training, and developing professional competencies. Table 12 summarizes the characteristics of the works, organized according to model, year, and specialty addressed. This table complements the analysis of medical question-answering in the educational setting, facilitating the identification of domains addressed and models adapted to the specific context of each task.

Table 12. Summary of research on using LLMs for answering medical questions in the educational domain

Ref.	Model	Year	Medical Area
[61]	BioMedLM	2022	Multispecialty
[19]	GPT-3.5	2023	Ophthalmology
[147]	ChatGPT	2023	Urology and Oncology
[148]	HuaTuo, based on LLaMA-7B	2023	General Medicine
[149]	GPT-4	2023	Ophthalmology
[150]	GPT-3.5	2023	Multispecialty
[151]	GPT-3.5 + GPT-4	2023	Ophthalmology
[11]	Bing (Microsoft)	2024	Cardiology and Emergency Medicine
[74]	BioMistal 7B	2024	Multispecialty
[87]	GPT-4	2024	Oncology
[152]	ChatGPT	2024	Gynecologic Oncology
[153]	AI-guide bot (based on ChatGPT-3.5)	2024	Oncology
[154]	GPT-3.5 + GPT-4	2024	Cardiology
[155]	GPT-3.5 + GPT-4 + Google Bard + Bing Chat (Microsoft) + Claude + Sage	2024	Dentistry
[156]	GPT-4 + Bard (Google) + Bing Chat (Microsoft)	2024	Anesthesiology
[157]	GPT-4	2024	Neurology
[158]	Xiaoqing (based on ChatGLM-6B)	2024	Ophthalmology
[159]	ChatDiet	2024	Nutrition

Utilizing LLMs in education unlocks new opportunities for enhancing comprehension, self-learning, and personalized education in the health sciences. Various studies show that these models can effectively assist in answering medical questions, conducting clinical training, and developing professional competency. BioMedLM, designed by Lee *et al.* [61], answers biomedical questions in either free-form or multiple-choice format, serving as a versatile resource to reinforce knowledge across multiple clinical areas. Similarly, BioMistal 7B proves helpful in clinical text analysis, with educational applications focused on interpreting unstructured medical information and training medical students' critical reading and synthesis skills [74].

On the other hand, ChatGPT has been extensively evaluated in various studies. For example, [150] and [151] explore its ability to solve clinical questions in ophthalmology. Furthermore, [19] tests its use with the Ophthalmology Clinical Competency Examination (OKAP), highlighting its usefulness as a structured student review tool. Coskun *et al.* [147] expand the focus to combined specialties such as urology and oncology, evaluating their performance in simulated clinical scenarios. Alternative models, such as HuaTuo [148] and Xiaoqing [158], show potential as multilingual educational assistants, providing general medical information or focusing on specific conditions, such as glaucoma or diabetes, while adapting to different cultural and linguistic contexts. In general medical education, [149] and [157] discuss ChatGPT's ability to provide accurate and helpful answers in diverse clinical contexts, including neurology.

In nutrition, Yang *et al.* [159] introduce ChatDiet, a model focused on dietary habits and personalized recommendations aimed at preventive education. Other work, focusing on preparing students and professionals to interact with patients [154], examines the use of ChatGPT for clinical report writing and improving communication skills in cardiology. Likewise, Birkun and Gautam [11] evaluate Bing Chat in emergency medicine settings, highlighting its usefulness for clinical reasoning under pressure. In surgical and specialty areas, works such as those of [152], [156], and [155] analyze combined models (including ChatGPT, Bard, Bing, Claude, and Sage) in anesthesia, gynecology, and dental education. These works underscore the value of multilingual and multi-modal models as comprehensive tools in advanced clinical and educational settings. For its part, Lee *et al.* [153] present the AI-guide bot, a tool based on ChatGPT-3.5 oriented to oncology education. This model generates patient simulations, questionnaires, and automatic feedback on clinical understanding, integrating it as a pedagogical resource in active learning environments. Finally, Wei *et al.* [87] apply ChatGPT-4 in similar tasks, showing its ability to accurately answer clinical questions related to oncology and serve as simulated clinical decision support for advanced learners. These works reflect a clear trend: LLMs not only answer medical questions with remarkable accuracy but also position themselves as valuable tools in medical education by facilitating autonomous learning, exam preparation, clinical case simulation, and improved evidence-based decision-making.

4. Evaluation metrics

Rigorous performance evaluation of LLMs is crucial for their safe and practical application in medical settings. Due to the inherent complexity and sensitivity of clinical settings, evaluation metrics must capture both technical accuracy and the semantic validity and contextual relevance of the responses generated. The most commonly employed metrics include ROUGE, BERTScore, and BLEU, which aim to compare textual responses generated by the model and reference responses for classification or clinical decision tasks. Accuracy, precision, recall, and F1-score are used to quantify the model's ability to identify relevant medical entities, diagnoses, or symptoms correctly. In addition, some researchers propose more specialized measures, such as medical concept coverage (MCC) [160], which evaluates whether the model adequately includes critical clinical concepts such as diseases, treatments, organs, or denied conditions. This metric is particularly relevant in scenarios where omission of information may affect patient safety or quality of service.

On the other hand, given the risks associated with generating erroneous or misleading clinical content, there is a need for metrics designed to assess potential harms. Roy *et al.* [161] present the average number of unsafe hits (AUM), a measure designed to estimate the probability that a model produces clinically dangerous, inappropriate, or serious health consequences for the patient. This metric allows a quantitative assessment of the potential negative impact of the model when confronted with medically sensitive questions.

Given the potential for bias, hallucinations, and unsafe outcomes in LLMs applied to the medical domain, any evaluation process must consider overall performance, clinical feasibility, ethical risk, and the model's ability to operate within acceptable margins of error in healthcare. In this context, Table 13 synthesizes the evaluation metrics used in various works, providing a comparative view of the methodological approaches employed to assess the performance of LLMs in healthcare-related tasks.

Table 13. Summary of evaluation metrics used in different LLM research studies

Metrics	Description	Research papers
ROUGE	Evaluates the quality of generated text by comparing the overlap of words or phrases with reference texts. It is commonly used to assess the output of LLMs in summarization and text generation tasks.	[7], [47], [65], [76], [125], [160], [162], [163], [164], [165], [166]
BERTScore	Measures semantic similarity between texts using representations from models such as BERT. It helps evaluate whether LLMs capture the intended meaning beyond surface-level matching.	[47], [76], [112], [162], [165], [166], [167]
BLEU Scores	Calculates the n-gram precision between the generated text and a reference, commonly used in machine translation. It helps measure how accurately LLMs reproduce expected sequences in generation tasks.	[7], [65], [76], [168], [169]
Accuracy	The proportion of correct predictions over the total number of predictions made and applied to evaluate classification performance in tasks handled by LLMs.	[5], [6], [8], [13], [58], [71], [82], [93], [102], [107], [110], [119], [121], [122], [126], [128], [131], [136], [140], [142], [144], [146], [162]
Precision	Measures the proportion of correct optimistic predictions among all instances predicted as positive. It is essential when LLMs are used for information extraction or question answering.	[16], [17], [19], [68], [79], [92], [96], [109], [113], [117], [122], [123], [157], [173]
Recall	Measures the proportion of actual positive cases correctly identified by the model. Used to assess how well LLMs detect relevant information or correct responses.	[80], [89], [94], [95], [98], [99], [174]
F1-score	The harmonic mean of precision and recall summarizes the balance between both metrics in a single score. It provides a balanced view of LLM performance, especially in uneven class distributions.	[15], [88], [90], [91], [97], [100], [115], [137], [147]
Medical concept coverage	Measures how well a generated medical text captures the relevant clinical concepts (e.g., symptoms, medications, diagnoses) found in the reference and is commonly used in medical summarization tasks to evaluate clinical accuracy and completeness beyond surface-level text overlap.	[4], [160], [175]
The average number of unsafe matches	Quantifies the average number of instances in which generated text includes medically unsafe or incorrect content compared to the reference and is used to assess the safety and reliability of LLM outputs in clinical or health-related applications.	[161], [176]

As shown in Table 13, these metrics provide a robust quantitative framework for analyzing the performance of LLMs in different clinical scenarios. However, their concrete application varies depending on the nature of the task, the type of model, and the medical context. The following review encompasses several recent studies that illustrate how these metrics evaluate models in real-world tasks, enabling us to identify their strengths and current limitations.

One of the most advanced fields is computer-aided diagnosis. For example, in the field of mental health, the MentalBERT and MentalRoBERTa models are specifically trained to detect symptoms of depression and

anxiety from social media posts. MentalBERT achieves an F1-score of 94.23% and a Recall of 94.33% when analyzing user posts on the Reddit platform [89]. In comparison, MentalRoBERTa obtains an F1-score of 93.38% on eRisk, a dataset specifically designed to assess the early detection of mental health problems through texts posted on social networks, demonstrating its ability to identify symptoms of anxiety and depression in informal and unstructured contexts. In parallel, computer vision applied to clinical diagnosis also shows remarkable results. A model based on deep neural networks detects cataracts, jaundice, and strabismus with an accuracy of 99.31%, 99.77%, and 97.82%, respectively, demonstrating the potential of artificial intelligence to assist in ocular and pediatric diagnosis in contexts with limited resources [6].

In more complex surgical scenarios, the performance of GPT-4 is evaluated in detecting postoperative renal complications, where it achieves an F1-score of 0.87 and an accuracy of 86.7% in recognizing primary tumors. However, its accuracy is lower, according to the Clavien-Dindo scale, which classifies postoperative complications by severity and type of intervention required, with only 37.4% accuracy, revealing its capacity and current limitations for more structured analysis [113]. Complementarily, in bariatric surgery, ChatGPT-4 demonstrates an accuracy of 83% in correctly classifying clinical scenarios, particularly in recognizing surgical complications, with a rate of 91.7%, surpassing models such as Bard and Bing [15].

In ophthalmology, the performance of ChatGPT is evaluated in the context of clinical question banks, achieving an accuracy of 55.8% in the BCSC set and 42.7% in OphthoQuestions, with improvements in its Plus version to 59.4%. These results reflect a still incipient but evolving knowledge in highly specialized domains [19]. In parallel, the structured extraction of information from clinical records is a task where LLMs demonstrate significant efficacy. ChatGPT is used to extract key information from oncology pathology reports. It achieves an F1-score of 0.91, with precision metrics of 92% in lymph node identification and 99% in histological diagnoses, validating its usefulness in automated data streams [95]. In another application, the AssistMED system is able to characterize cohorts of cardiological patients in Polish clinical registries, obtaining an accuracy of 99.5% and an F1-score of 0.988. Its performance was comparable to human work, optimizing the identification of treatments and ultrasound findings [98].

Personalized clinical prediction is another growing field. Health-LLM, a model that integrates medical scores and feature extraction from clinical text, is evaluated for predicting diseases in fields such as endocrinology, gastroenterology, and internal medicine. It achieves an accuracy of 83.3% and an F1-score of 0.762, surpassing reference models such as GPT-4 and LLaMA-2 and consolidating itself as a valuable tool for preventive diagnosis [116].

Clinical text generation models, such as BART and T5, are applied to augment data in medical records. In the generation of clinical notes, BART achieves a ROUGE-1 of 52.35, a ROUGE-2 of 41.59, and a ROUGE-L of 50.71, while T5 shows better performance in treatment content [47]. In a second study, T5 is used to generate automatic knee MRI summaries, achieving a ROUGE-L of 63.8 and a 70% match in fluency and content with reports made by radiologists, highlighting its ability to automate reports in radiology [76].

Models such as BiomedRAG, which incorporate augmented retrieval, address the extraction of complex relationships between biomedical concepts. This system achieves an F1-score of 88.83% in extracting biomedical triples in the ChemProt ensemble, reducing semantic hallucinations and improving accuracy in pharmacological and molecular interaction contexts [97].

Finally, one of the most popular applications of LLMs in healthcare is answering medical questions. Aeyeconsult, a specialized clinical chatbot based on GPT-4 and leveraging verified ophthalmology literature through RAG, achieved an accuracy of 83.4% on OKAP questions, outperforming ChatGPT-4 overall (69.2%) and showed higher consistency and lower error rate [131]. In hepatology, ChatGPT achieves 79.1% accuracy in answering questions on cirrhosis and 74% accuracy on hepatocellular carcinoma, with 76.9% correct answers on treatment, although it shows deficiencies in content completeness [117]. In broader scenarios, ChatGPT-4 demonstrates 100% accuracy and 83.2% completeness in answering questions on heart failure, while GPT-3.5 scores 98.1% and 78.5%, respectively, positioning these models as viable assistants in educational and clinical support settings [124].

In short, rigorous and systematic evaluation of the performance of LLMs in clinical contexts is essential to validate their technical accuracy and ensure their ethical feasibility, practical utility, and safety in patient care. The results reviewed evidence that, although LLMs achieve remarkable performance in tasks such as diagnosis, information extraction, and medical text generation, their applicability remains highly dependent on the type of task, the specific clinical domain, and the quality of training. As these technologies evolve, the integration of

technical metrics with contextualized clinical indicators becomes increasingly necessary, capable of capturing not only superficial accuracy but also semantic depth, relevant medical coverage, and potential health risks. Thus, the roadmap to a reliable implementation of LLMs in medicine depends on improving architectures and designing comprehensive evaluation frameworks that reflect the real complexities of the clinical environment.

5. Limitations and future challenges

Integrating LLMs in the medical, healthcare, and educational domains presents many technical, ethical, and legal constraints that require rigorous attention. In the medical domain, explainability, reliability of clinical decisions, and mitigating errors, such as hallucinations, are crucial for avoiding risks in critical contexts [49]. In the healthcare domain, protecting sensitive data, ensuring algorithmic fairness, and maintaining operational transparency are crucial for maintaining the trust of patients and healthcare professionals. Meanwhile, in the healthcare educational domain, challenges arise related to content accuracy, intellectual property, and the generation of misleading information that affects learning and professional training processes. Although solutions such as constructing new datasets and specific assessment techniques have been proposed [177], [178], risks persist. Therefore, the analysis categorizes the problems into five key areas: privacy and security of patient data, risk of misinformation, fairness and bias, explainability and reliability, and plagiarism and liability issues. This section examines these limitations across all five domains, highlighting both challenges and prospects for addressing them in the future.

5.1. Privacy and security of patient data

In the medical and healthcare domain, the handling of personal data requires robust security measures to ensure patient confidentiality. Although current systems enable the advanced analysis of large volumes of data, the accidental inclusion of identifiable information in training sets remains a concern, as LLMs infer personal attributes from seemingly neutral data [179]. Chuang *et al.* [164] mention that ensuring patient privacy is even more complex than achieving good clinical outcomes. While digitizing healthcare services makes accessing and reusing data easier, it also increases the risk of critical information leaks.

Against this backdrop, it is essential to combine technological strategies such as automated anonymization and recognition of sensitive entities with updated regulatory frameworks and governance plans that address the entire data lifecycle [180]. The challenge lies in complying with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) [181] and the General Data Protection Regulation (GDPR) [182], as well as ensuring the ethical and responsible use of technology, even in contexts where legislation does not evolve at the same pace. Creating appropriate regulatory frameworks remains a significant challenge.

5.2. Generation of misinformation and hallucinations

One of the most discussed obstacles to applying LLMs in medicine and education is the generation of answers that, although linguistically plausible, lack veracity. This phenomenon, known as hallucination, poses a hazard when such responses serve in clinical or educational decision-making, where accuracy remains critical. Documented cases show that models such as GPT-3 offer harmful suggestions in mental health contexts [183].

To mitigate this problem, several strategies emerge, such as learning in context, prompting approaches based on explicit reasoning (Chain-of-Thought), and model evaluation using benchmarks designed explicitly for medical tasks [184]. In this regard, Med-HALT, a benchmark dataset for assessing hallucination phenomena in LLMs in medical contexts, is presented by Pal *et al.* [142]. This corpus comprises two distinct categories of tests: one focuses on reasoning and the other on memory. Both aim to measure the model's ability to solve problems and retrieve specific information within the medical domain. However, neither technical solution replaces the need for professional supervision and human verification mechanisms.

5.3. Algorithmic biases and fairness

Training LLMs with data extracted from the web or scientific corpora reproduces and amplifies pre-existing biases related to gender, ethnicity, nationality, age, or other social factors. In the medical domain, this translates into diagnostic or therapeutic decisions that are not equitable for all patients [185]. In the educational context, it manifests in responses that reinforce cultural stereotypes or exclude regional or linguistic realities, thus limiting the inclusiveness and diversity of the knowledge represented. One of the main concerns associated with using LLMs is the risk of generating erroneous or biased information. When trained on large volumes of text, these models produce inaccurate or misleading content, including trusted and unverified sources. Particularly

in the scientific literature, where historical biases such as gender or racial biases persist, LLMs may replicate and inadvertently amplify these biases in their generated results.

Algorithmic fairness cannot be approached solely from a technical standpoint. It is imperative to reassess the representativeness of the data used and establish mechanisms to detect and mitigate biases during both the training phase and inference. Investigating strategies to identify, reduce, and prevent these biases represents a fundamental line in the ethical development of LLMs. Implementing systematic audits and rigorous processes for validating and verifying results is crucial. Strategies such as Counterfactually Fair Prompting (CFP) [186] represent essential advances in bias mitigation as they modify model instructions or inputs to promote more equitable responses. However, these technical solutions must be framed within an interdisciplinary approach that involves professionals from health, education, law, ethics, and computer science. As highlighted in [187], it is crucial to prioritize equity and inclusivity when applying LLMs in biomedical contexts. This requires carefully selecting and preprocessing training data to minimize inherent biases and facilitate collaboration among subject matter experts, data scientists, and ethicists. Only through collaborative governance and the establishment of clear guidelines can genuinely unbiased, accountable, and helpful AI systems be developed for the benefit of all society.

5.4. Explainability and reliability

LLMs output must be understandable and reliable for end users in contexts where decisions directly impact people's lives, such as health and education. However, many LLMs operate as opaque systems or black boxes, which makes it difficult to trace the process leading to a specific recommendation or conclusion. This lack of transparency represents a significant barrier to adoption in critical domains, where practitioners require clear and defensible evidence to support their decisions [118].

In the healthcare sector, this concern becomes even more urgent. LLM-generated recommendations influence diagnoses, prognoses, or therapeutic choices. Without a clear understanding of the reasoning behind these suggestions, there is a risk of compromising patient safety. The presence of biases in training data further exacerbates this situation, as such biases lead to incorrect inferences that disproportionately affect certain population groups [113], [118].

Explainability, defined as the ability of a model to provide understandable reasons behind its decisions, becomes an indispensable condition for the ethical and practical integration of artificial intelligence in medicine [159]. This is especially important in areas such as clinical risk prediction or medical image analysis, where automated decisions must be independently validated by human experts [83], [188].

Faced with these challenges, various solutions emerge. Some initiatives aim to develop interpretable mechanisms for tracking model reasoning, while others introduce control labels to guide text generation and enhance transparency. Strategies such as visualizations, decomposition of contributions by token, or the design of hybrid architectures (combining statistical models and explicit rules) show promising results. However, absolute trust builds only if these systems are auditable and their operation aligns with defined ethical and regulatory frameworks [189]. Therefore, effective governance of LLMs encourages traceability, oversight, and continuous human intervention. The challenge lies in making the models explain what the systems themselves do and making these explanations understandable, relevant, and valuable to the various stakeholders involved in decision-making. Achieving this goal requires a multidisciplinary and collaborative approach that includes experts in health, education, informatics, ethics, and public policy.

5.5. Legal, ethical, and intellectual property aspects

The deployment of LLMs in sensitive areas, such as healthcare and education, raises new challenges regarding accountability, authorship, and regulatory compliance. As these systems acquire the capacity to generate clinical, pedagogical, or administrative content, key questions emerge: Who takes responsibility if a model provides an erroneous recommendation? How is authorship attributed in a text generated without explicit references? What mechanisms exist to avoid copyright infringement or the dissemination of unsupported information?

These concerns are compounded when LLMs operate without adequate human oversight. Their ability to generate plausible but incorrect content, reproduce text fragments without attribution, or even mimic a person's style and identity leads to plagiarism, misinformation, or impersonation [143]. In clinical applications, delivering sensitive diagnoses without adequate contextual or emotional support has significant ethical and legal consequences.

In response, various jurisdictions have formulated regulatory frameworks to mitigate these risks. For example, the European Union's Artificial Intelligence Act (AI Act) proposes classifying AI systems according to their level of risk [126]. It establishes specific obligations for providers of general-purpose models, including requirements for transparency, risk assessment, and technical documentation.

In the United States, the Health Insurance Portability and Accountability Act (HIPAA) establishes strict guidelines for the use of sensitive data in healthcare settings, directly affecting how LLMs manage confidential patient information [181]. This involves everything from implementing robust encryption to regular audits and granular access controls. Other relevant frameworks, such as the General Data Protection Regulation (GDPR) [182] and the Medical Device Regulation (MDR) [190], serve as additional regulatory barriers that must be overcome to fully adopt generative AI in biomedical contexts. These regulations aim to promote traceability, minimize risk, and foster the responsible use of algorithmic systems, thereby helping to build trust among users, professionals, and institutions.

However, the existence of regulations does not eliminate the current vacuum in terms of global standards and international coordination. Rapid technological evolution often surpasses the legislative capacity to respond with agility and scope. It is, therefore, essential to move towards models of shared responsibility in which developers, users, regulators, and service providers collaborate in designing, implementing, and monitoring these systems. This approach not only facilitates stronger governance but also ensures that technological innovation is aligned with fundamental ethical principles and respects the rights of individuals [138].

6. Conclusions

This systematic review demonstrates that LLMs have a profound impact on medicine, healthcare, and medical education. Their implementation in clinical, administrative, and training tasks evolves from experimental approaches to concrete applications in real-world settings, with promising results in multiple medical specialties.

In the clinical setting, specialized LLMs such as ClinicalBERT [81], MedPaLM-2 [49], and PubMedGPT [53] demonstrate outstanding capabilities to generate differential diagnoses, summarize medical histories, detect pathologies through imaging, and assist in complex therapeutic decisions [67], [82], [86]. These models automate clinical processes and provide interpretable explanations, which enhance diagnostic transparency [84], [104].

In the healthcare context, tools such as GatorTron [64], BiomedRAG [97], and AssistMED [98] enable the structuring of unstructured medical information, extraction of biomedical relationships, and automation of triage, screening, and clinical risk prediction tasks [68], [88], [96]. These capabilities strengthen institutional efficiency, epidemiological surveillance, and the design of data-driven health policies.

In the educational area, LLMs facilitate personalized medical training, exam simulation, feedback generation, and explanation of complex clinical concepts [143], [144], [145]. Models such as BioGPT [52], Med-HALT [142], and BioBART [139] support the generation of summaries, clinical reports, and case studies, improve pedagogical quality, and reduce the teaching load.

However, deploying these technologies faces several critical challenges: rigorous clinical validation, reduction of algorithmic bias, interpretability of results, and the need for both ethical and technical regulations for their safe implementation. In particular, the field requires avoiding information hallucination in sensitive clinical contexts, strengthening traceability for architectures such as RAG [68], [77], [131], and ensuring cultural and linguistic appropriateness in diverse populations [71], [122], [137].

For all these reasons, LLMs have the potential to redefine the future of healthcare and medical education. Their adoption must be responsible, transparent, and patient-centered, integrating robust regulatory frameworks, ongoing scientific validation, and close collaboration between developers, clinicians, and educators. This convergence between AI and biomedical sciences creates the conditions to consolidate more accurate, equitable, and accessible medicine for all.

Declaration of competing interest

The authors declare that they have no any known financial or non-financial competing interests in any material discussed in this paper.

Funding information

No funding was received from any financial organization to conduct this research.

References

[1] Z. W. Xin, Z. Kun, L. Junyi, T. Tianyi, W. Xiaolei, H. Yupeng, M. Yingqian, Z. Beichen, Z. Junjie, D. Zican, *et al.*, "A Survey of Large Language Models," *arXiv preprint arXiv:2303.18223*, 2023, [Online]. Available: http://arxiv.org/abs/2303.18223

[2] M. U. Hadi, Q. Al Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, S. Z. Hassan, *et al.*, "Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects," *TechRxiv*, 2025, doi: 10.36227/techrxiv.23589741.v8.

[3] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A Survey on Multimodal Large Language Models," *arXiv preprint arXiv:2306.13549*, 2023, [Online]. Available: https://arxiv.org/abs/2306.13549

[4] Y. Zhou, F. Ringeval, and F. Portet, "A Survey of Evaluation Methods of Generated Medical Textual Reports," in *Proceedings of the 5th Clinical Natural Language Processing Workshop*, T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, and A. Rumshisky, Eds., Toronto, Canada: Association for Computational Linguistics, 2023, pp. 447–459. doi: 10.18653/v1/2023.clinicalnlp-1.48.

[5] Y. Kim, J.-H. Kim, Y.-M. Kim, S. Song, and H. J. Joo, "Predicting medical specialty from text based on a domain-specific pre-trained BERT," *International Journal of Medical Informatics*, vol. 170, p. 104956, 2023, doi: https://doi.org/10.1016/j.ijmedinf.2022.104956.

[6] U. Garg, S. Dhyani, S. Nautiyal, A. Chand, and N. Gupta, "Medical Assistance for Detecting Cataract, Jaundice and Strabismus using Deep Learning," in *2023 IEEE Pune Section International Conference (PuneCon)*, 2023, pp. 1–6. doi: 10.1109/PuneCon58714.2023.10449988.

[7] Y. Tan, Z. Zhang, M. Li, F. Pan, H. Duan, Z. Huang, H. Deng, Z. Yu, C. Yang, G. Shen, *et al.*, "MedChatZH: A tuning LLM for traditional Chinese medicine consultations," *Computers in Biology and Medicine*, vol. 172, p. 108290, 2024, doi: https://doi.org/10.1016/j.compbiomed.2024.108290.

[8] K. N. Kunze, N. H. Varady, M. Mazzucco, A. Z. Lu, J. Chahla, R. K. Martin, A. S. Ranawat, A. D. Pearle, and R. J. Williams, "The Large Language Model ChatGPT-4 Demonstrates Excellent Triage Capabilities and Diagnostic Performance for Patients Presenting with Various Causes of Knee Pain," *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 2024, doi: https://doi.org/10.1016/j.arthro.2024.06.021.

[9] A. Rao, J. Kim, M. Kamineni, M. Pang, W. Lie, and M. D. Succi, "Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making," *medRxiv*, 2023, doi: 10.1101/2023.02.02.23285399.

[10] W. J. Jaimes, W. J. Arenas, H. J. Navarro, and M. Altuve, "Detection of retinal diseases from OCT images using a VGG16 and transfer learning," *Discover Applied Sciences*, vol. 7, no. 3, p. 160, Feb. 2025, doi: 10.1007/s42452-025-06565-6.

[11] A. A. Birkun and A. Gautam, "Large Language Model-based Chatbot as a Source of Advice on First Aid in Heart Attack," *Current Problems in Cardiology*, vol. 49, no. 1, Part A, p. 102048, 2024, doi: https://doi.org/10.1016/j.cpcardiol.2023.102048.

[12] M. Guckenberger, N. Andratschke, M. Ahmadsei, S. M. Christ, A. E. Heusel, S. Kamal, T. E. Kroese, E. L. Looman, S. Reichl, E. Vlaskou Badra, *et al.*, "Potential of ChatGPT in facilitating research in radiation oncology?," *Radiotherapy and Oncology*, vol. 188, p. 109894, 2023, doi: https://doi.org/10.1016/j.radonc.2023.109894.

[13] C. J. Arismendi Pereira, C. L. Sandoval-Rodríguez, B. F. Giraldo, and E. H. Solano, "Extubating of a patient undergoing mechanical ventilation: What is the right time? A retrospective study assisted by

artificial intelligence techniques," *Periodicals of Engineering and Natural Sciences*, vol. 12, no. 3, pp. 604–615, 2024, doi: 10.21533/pen.v12.i3.60.

[14] P. Tsoutsanis and A. Tsoutsanis, "Evaluation of Large language model performance on the Multi-Specialty Recruitment Assessment (MSRA) exam," *Computers in Biology and Medicine*, vol. 168, p. 107794, 2024, doi: https://doi.org/10.1016/j.compbiomed.2023.107794.

[15] S. Lee, J. Lee, J. Park, J. Park, D. Kim, J. Lee, and J. Oh, "Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage," *The American Journal of Emergency Medicine*, vol. 77, pp. 29–38, 2024, doi: https://doi.org/10.1016/j.ajem.2023.11.063.

[16] M. Sievert, O. Conrad, S. K. Mueller, R. Rupp, M. Balk, D. Richter, K. Mantsopoulos, H. Iro, and M. Koch, "Risk stratification of thyroid nodules: Assessing the suitability of ChatGPT for text-based analysis," *American Journal of Otolaryngology*, vol. 45, no. 2, p. 104144, 2024, doi: https://doi.org/10.1016/j.amjoto.2023.104144.

[17] R. K. Gan, J. C. Ogbodo, Y. Z. Wee, A. Z. Gan, and P. A. González, "Performance of Google bard and ChatGPT in mass casualty incidents triage," *The American Journal of Emergency Medicine*, vol. 75, pp. 72–78, 2024, doi: https://doi.org/10.1016/j.ajem.2023.10.034.

[18] R. Franco D'Souza, S. Amanullah, M. Mathew, and K. M. M. M. M. Surapaneni, "Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes," *Asian Journal of Psychiatry*, vol. 89, p. 103770, 2023, doi: https://doi.org/10.1016/j.ajp.2023.103770.

[19] F. Antaki, S. Touma, D. Milad, J. El-Khoury, and R. Duval, "Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings," *Ophthalmology Science*, vol. 3, no. 4, p. 100324, 2023, doi: https://doi.org/10.1016/j.xops.2023.100324.

[20] M. Liebrenz, R. Schleifer, A. Buadze, D. Bhugra, and A. Smith, "Generating scholarly content with ChatGPT: ethical challenges for medical publishing," *The Lancet Digital Health*, vol. 5, no. 3, p. e105, 2023, doi: 10.1016/S2589-7500(23)00019-5.

[21] H. J. Navarro, S. A. Salinas, W. J. Arenas, S. A. Sotelo, C. L. Rueda, M. E. Otero, M. A. Altuve, and W. J. Jaimes, "Gait events detection using inertial sensors, Apple Watch, and the G-WALK reference system," in *2021 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*, 2021, pp. 1–6. doi: 10.1109/GMEPE/PAHCE50215.2021.9434858.

[22] H. J. Navarro, W. J. Arenas, W. J. Jaimes, and S. A. Salinas, "Automated detection of gait events using inertial sensor signals and a discrete wavelet transform approach," *Periodicals of Engineering and Natural Sciences*, vol. 12, no. 3, pp. 577–594, 2024, doi: 10.21533/pen.v12.i3.57.

[23] W. J. Jaimes, J. F. Mantilla, S. A. Salinas, and H. J. Navarro, "Modeling and Simulation of a Lower Limb Exoskeleton with Computed Torque Control for Gait Rehabilitation," in *2021 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*, 2021, pp. 1–6. doi: 10.1109/GMEPE/PAHCE50215.2021.9434854.

[24] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," *IEEE Access*, vol. 12, pp. 26839–26874, 2024, doi: 10.1109/ACCESS.2024.3365742.

[25] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023, doi: 10.1038/s41586-023-06291-2.

[26] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 2023, doi: 10.1038/s41591-023-02448-8.

[27] M. Sallam, "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," *Healthcare*, vol. 11, no. 6, 2023, doi: 10.3390/healthcare11060887.

[28] X. Meng, X. Yan, K. Zhang, D. Liu, X. Cui, Y. Yang, M. Zhang, C. Cao, J. Wang, X. Wang, *et al.*, "The application of large language models in medicine: A scoping review," *iScience*, vol. 27, no. 5, 2024, doi: 10.1016/j.isci.2024.109713.

[29] J. Haltaufderheide and R. Ranisch, "The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs)," *npj Digital Medicine*, vol. 7, no. 1, p. 183, 2024, doi: 10.1038/s41746-024-01157-x.

[30] S. Tian, Q. Jin, L. Yeganova, P.-T. Lai, Q. Zhu, X. Chen, Y. Yang, Q. Chen, W. Kim, D. C. Comeau, *et al.*, "Opportunities and challenges for ChatGPT and large language models in biomedicine and health," *Briefings in Bioinformatics*, vol. 25, no. 1, p. bbad493, 2024, doi: 10.1093/bib/bbad493.

[31] H. C. Lucas, J. S. Upperman, and J. R. Robinson, "A systematic review of large language models and their implications in medical education," *Medical Education*, vol. 58, no. 11, pp. 1276–1285, 2024, doi: https://doi.org/10.1111/medu.15402.

[32] OpenAI, "ChatGPT: Optimizing Language Models for Dialogue," 2022. [Online]. Available: https://openai.com/blog/chatgpt

[33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/

[34] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," *arXiv preprint arXiv:1411.5595*, 2014, [Online]. Available: https://arxiv.org/abs/1411.5595

[35] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013, [Online]. Available: https://arxiv.org/abs/1301.3781

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[37] D. Zhou, Y. Shi, B. Kang, W. Yu, Z. Jiang, Y. Li, X. Jin, Q. Hou, and J. Feng, "Refiner: Refining Self-attention for Vision Transformers," *arXiv preprint arXiv:2106.03714*, 2021, [Online]. Available: https://arxiv.org/abs/2106.03714

[38] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, *et al.*, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, 2025, doi: 10.1145/3703155.

[39] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, 2023, doi: 10.1145/3571730.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018, [Online]. Available: https://arxiv.org/abs/1810.04805

[41] H. Touvron, J. Martin, K. Stone, P. Albert, A. Hannun, G. Synnaeve, Y. LeCun, A. Baevski, T. Wolf, J. Spisak, *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv preprint*

arXiv:2307.09288, 2023, [Online]. Available: https://arxiv.org/abs/2307.09288

[42] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[43] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[44] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, *et al.*, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," *arXiv preprint arXiv:2501.12948*, 2025, [Online]. Available: https://arxiv.org/abs/2501.12948

[45] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *arXiv preprint arXiv:1910.10683*, 2019, [Online]. Available: https://arxiv.org/abs/1910.10683

[46] V. Gulati, S. G. Roy, A. Moawad, D. Garcia, A. Babu, J. D. Poot, and O. M. Teytelboym, "Transcending Language Barriers: Can ChatGPT Be the Key to Enhancing Multilingual Accessibility in Health Care?," *Journal of the American College of Radiology*, vol. 21, no. 12, pp. 1888–1895, 2024, doi: 10.1016/j.jacr.2024.05.009.

[47] A. Latif and J. Kim, "Evaluation and Analysis of Large Language Models for Clinical Text Augmentation and Generation," *IEEE Access*, vol. 12, pp. 48987–48996, 2024, doi: 10.1109/ACCESS.2024.3384496.

[48] S. Sandmann, S. Riepenhausen, L. Plagwitz, and J. Varghese, "Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks," *Nature Communications*, vol. 15, no. 1, 2024, doi: 10.1038/s41467-024-46411-8.

[49] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, *et al.*, "Towards Expert-Level Medical Question Answering with Large Language Models," *arXiv preprint arXiv:2305.09617*, 2023.

[50] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models," *arXiv preprint arXiv:2303.04671*, 2023, [Online]. Available: https://arxiv.org/abs/2303.04671

[51] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," *arXiv preprint arXiv:2301.12597*, 2023, [Online]. Available: https://arxiv.org/abs/2301.12597

[52] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining," *arXiv preprint arXiv:2210.10341*, 2023, [Online]. Available: https://arxiv.org/abs/2210.10341

[53] E. Bolton, D. Hall, M. Yasunaga, T. Lee, C. Manning, and P. Liang, "Stanford CRFM Introduces PubMedGPT 2.7B," *Stanford HAI News*, 2022, [Online]. Available: https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b

[54] Y. Li, B. Cao, W. Zhang, Z. Xu, H. Zheng, Y. Han, S. Liu, X. Huang, Y. Zhang, R. Zhang, *et al.*, "Uni-MoE: Scaling Unified Multimodal LLMs with Mixture of Experts," *arXiv preprint arXiv:2405.11273*, 2024, [Online]. Available: https://arxiv.org/abs/2405.11273

[55] Z. Zong, B. Ma, D. Shen, G. Song, H. Shao, D. Jiang, H. Li, and Y. Liu, "MoVA: Adapting Mixture of Vision Experts to Multimodal Context," *arXiv preprint arXiv:2404.13046*, 2024.

[56] B. Lin, Z. Tang, Y. Ye, J. Huang, J. Zhang, Y. Pang, P. Jin, M. Ning, J. Luo, and L. Yuan, "MoE-LLaVA: Mixture of Experts for Large Vision-Language Models," *arXiv preprint arXiv:2401.15947*, 2024.

[57] D. Dai, C. Deng, C. Zhao, R. X. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, *et al.*, "DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models," *arXiv preprint arXiv:2401.06066*, 2024, [Online]. Available: https://arxiv.org/abs/2401.06066

[58] J. Li, X. Wang, S. Zhu, C.-W. Kuo, L. Xu, F. Chen, J. Jain, H. Shi, and L. Wen, "CuMo: Scaling Multimodal LLM with Co-Upcycled Mixture-of-Experts," *arXiv preprint arXiv:2405.05949*, 2024.

[59] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[60] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *arXiv preprint arXiv:1901.08746*, 2019.

[61] E. Bolton, A. Venigalla, M. Yasunaga, D. Hall, B. Xiong, T. Lee, R. Daneshjou, J. Frankle, P. Liang, M. Carbin, *et al.*, "BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text," *b*, 2024.

[62] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *arXiv preprint arXiv:2007.15779*, 2020.

[63] Q. Lu, D. Dou, and T. H. Nguyen, "ClinicalT5: A Generative Language Model for Clinical Text," *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5436–5443, 2022.

[64] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, A. B. Costa, M. G. Flores, *et al.*, "GatorTron: A Large Language Model for Electronic Health Records," *npj Digital Medicine*, 2022, doi: 10.1038/s41746-022-00704-2.

[65] J. Goswami, K. K. Prajapati, A. Saha, and A. K. Saha, "Parameter-efficient fine-tuning large language model approach for hospital discharge paper summarization," *Applied Soft Computing*, vol. 157, p. 111531, 2024, doi: https://doi.org/10.1016/j.asoc.2024.111531.

[66] T. F. Tan, K. Elangovan, L. Jin, Y. Jie, L. Yong, J. Lim, S. Poh, W. Y. Ng, D. Lim, Y. Ke, *et al.*, "Fine-tuning Large Language Model (LLM) Artificial Intelligence Chatbots in Ophthalmology and LLM-based evaluation using GPT-4," *arXiv preprint arXiv:2402.10083*, 2024, [Online]. Available: https://arxiv.org/abs/2402.10083

[67] M. Song, J. Wang, Z. Yu, J. Wang, L. Yang, Y. Lu, B. Li, X. Wang, X. Wang, Q. Huang, *et al.*, "PneumoLLM: Harnessing the power of large language model for pneumoconiosis diagnosis," *Medical Image Analysis*, vol. 97, p. 103248, 2024, doi: https://doi.org/10.1016/j.media.2024.103248.

[68] O. Unlu, J. Shin, C. J. Mailly, M. F. Oates, M. R. Tucci, M. Varugheese, K. Wagholikar, F. Wang, B. M. Scirica, A. J. Blood, *et al.*, "Retrieval Augmented Generation Enabled Generative Pre-Trained Transformer 4 (GPT-4) Performance for Clinical Trial Screening," *medRxiv*, 2024, doi: https://doi.org/10.1101/2024.02.08.24302376.

[69] Q. Jin, B. Dhingra, W. W. Cohen, and X. Lu, "Probing Biomedical Embeddings from Language Models," *arXiv preprint arXiv:1904.02181*, 2019.

[70] H.-C. Shin, Y. Zhang, E. Bakhturina, R. Puri, M. Patwary, M. Shoeybi, and R. Mani, "BioMegatron: Larger Biomedical Domain Language Model," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4700–4706.

[71] Z. Yuan, Y. Liu, C. Tan, S. Huang, and F. Huang, "Improving Biomedical Pretrained Language Models with Knowledge," *arXiv preprint arXiv:2104.10344*, 2021.

[72] G. Miolo, G. Mantoan, and C. Orsenigo, "ELECTRAMed: a new pre-trained language representation model for biomedical NLP," *arXiv preprint arXiv:2104.09585*, 2021.

[73] I. De la Iglesia, A. Atutxa, K. Gojenola, and A. Barrena, "EriBERTa: A Bilingual Pre-Trained Language Model for Clinical Natural Language Processing," *arXiv preprint arXiv:2306.07373*, 2023, [Online]. Available: https://arxiv.org/abs/2306.07373

[74] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains," *arXiv preprint arXiv:2402.10373*, 2024.

[75] I. García-Ferrero, R. Agerri, A. Atutxa Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, *et al.*, "Medical mT5: An Open-Source Multilingual Text-to-Text LLM for The Medical Domain," *arXiv preprint arXiv:2404.07613*, 2024, [Online]. Available: https://arxiv.org/abs/2404.07613

[76] P. López-Úbeda, T. Martín-Noguerol, C. Díaz-Angulo, and A. Luna, "Evaluation of large language models performance against humans for summarizing MRI knee radiology reports: A feasibility study," *International Journal of Medical Informatics*, vol. 187, p. 105443, 2024, doi: https://doi.org/10.1016/j.ijmedinf.2024.105443.

[77] C. Zakka, R. Shad, A. Chaurasia, A. R. Dalal, J. L. Kim, M. Moor, R. Fong, C. Phillips, K. Alexander, E. Ashley, *et al.*, "Almanac — Retrieval-Augmented Language Models for Clinical Medicine," *NEJM AI*, vol. 1, no. 2, 2024, doi: https://doi.org/10.1056/aioa2300068.

[78] S. Hong, L. Xiao, X. Zhang, and J. Chen, "ArgMed-Agents: Explainable Clinical Decision Reasoning with LLM Discussion via Argumentation Schemes," *arXiv preprint arXiv:2403.06294*, 2024.

[79] R. Lian, V. Hsiao, J. Hwang, Y. Ou, S. E. Robbins, N. P. Connor, C. L. Macdonald, R. S. Sippel, W. A. Sethares, and D. F. Schneider, "Predicting health-related quality of life change using natural language processing in thyroid cancer," *Intelligence-Based Medicine*, vol. 7, p. 100097, 2023, doi: https://doi.org/10.1016/j.ibmed.2023.100097.

[80] A. Bartal, K. M. Jagodnik, S. J. Chan, M. S. Babu, and S. Dekel, "Identifying women with postdelivery posttraumatic stress disorder using natural language processing of personal childbirth narratives," *American Journal of Obstetrics & Gynecology MFM*, vol. 5, no. 3, p. 100834, 2023, doi: https://doi.org/10.1016/j.ajogmf.2022.100834.

[81] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," *arXiv preprint arXiv:1904.05342*, 2019, [Online]. Available: https://arxiv.org/abs/1904.05342

[82] M. Balas and E. B. Ing, "Conversational AI Models for ophthalmic diagnosis: Comparison of ChatGPT and the Isabel Pro Differential Diagnosis Generator," *JFO Open Ophthalmology*, vol. 1, p. 100005, 2023, doi: https://doi.org/10.1016/j.jfop.2023.100005.

[83] A. Ćirković and T. Katz, "Exploring the Potential of ChatGPT-4 in Predicting Refractive Surgery Categorizations: Comparative Study," *JMIR Form Res*, vol. 7, 2023, doi: https://formative.jmir.org/2023/1/e51798.

[84] D. J. McInerney, G. Young, J.-W. van de Meent, and B. C. Wallace, "CHiLL: Zero-shot Custom Interpretable Feature Extraction from Clinical Notes with Large Language Models," *arXiv preprint arXiv:2302.12343*, 2023.

[85] U. Naseem, M. Khushi, and J. Kim, "Vision-Language Transformer for Interpretable Pathology Visual Question Answering," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 1681–

1690, 2023, doi: 10.1109/JBHI.2022.3163751.

[86] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, and S. Ananiadou, "Towards Interpretable Mental Health Analysis with Large Language Models," *arXiv preprint arXiv:2304.03347*, 2023.

[87] K. Wei, C. Fritz, and K. Rajasekaran, "Answering head and neck cancer questions: An assessment of ChatGPT responses," *American Journal of Otolaryngology*, vol. 45, no. 1, p. 104085, 2024, doi: https://doi.org/10.1016/j.amjoto.2023.104085.

[88] C. Han, D. W. Kim, S. Kim, S. Chan You, J. Y. Park, S. Bae, and D. Yoon, "Evaluation of GPT-4 for 10-year cardiovascular risk prediction: Insights from the UK Biobank and KoGES data," *iScience*, vol. 27, no. 2, p. 109022, 2024, doi: https://doi.org/10.1016/j.isci.2024.109022.

[89] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare," *arXiv preprint arXiv:2110.15621*, 2021, [Online]. Available: https://arxiv.org/abs/2110.15621

[90] P. Boonyarat, D. J. Liew, and Y.-C. Chang, "Leveraging enhanced BERT models for detecting suicidal ideation in Thai social media content amidst COVID-19," *Information Processing & Management*, vol. 61, no. 4, p. 103706, 2024, doi: https://doi.org/10.1016/j.ipm.2024.103706.

[91] K. K. Bressem, J.-M. Papaioannou, P. Grundmann, F. Borchert, L. C. Adams, L. Liu, F. Busch, L. Xu, J. P. Loyen, S. M. Niehues, *et al.*, "medBERT.de: A comprehensive German BERT model for the medical domain," *Expert Systems with Applications*, vol. 237, p. 121598, 2024, doi: https://doi.org/10.1016/j.eswa.2023.121598.

[92] A. K. Chowdhury, S. R. Sujon, M. S. S. Shafi, T. Ahmmad, S. Ahmed, K. M. Hasib, and F. M. Shah, "Harnessing large language models over transformer models for detecting Bengali depressive social media text: A comprehensive study," *Natural Language Processing Journal*, vol. 7, p. 100075, 2024, doi: https://doi.org/10.1016/j.nlp.2024.100075.

[93] A. K. Fiedler, K. Zhang, T. S. Lal, X. Jiang, and S. M. Fraser, "GPT for Pediatric Stroke Research: A Pilot Study," *Pediatric Neurology*, 2024, doi: https://doi.org/10.1016/j.pediatrneurol.2024.07.001.

[94] M. Guevara, S. Chen, S. Thomas, T. L. Chaunzwa, I. Franco, B. H. Kann, S. Moningi, J. M. Qian, M. Goldstein, S. Harper, *et al.*, "Large language models to identify social determinants of health in electronic health records," *npj Digital Medicine*, vol. 7, no. 1, 2024, doi: 10.1038/s41746-023-00970-0.

[95] J. Huang, D. M. Yang, R. Rong, K. Nezafati, C. Treager, Z. Chi, S. Wang, X. Cheng, Y. Guo, L. J. Klesse, *et al.*, "A critical assessment of using ChatGPT for extracting structured data from clinical notes," *npj Digital Medicine*, vol. 7, no. 1, 2024, doi: 10.1038/s41746-024-01079-8.

[96] A. Landschaft, D. Antweiler, S. Mackay, S. Kugler, S. Rüping, S. Wrobel, T. Höres, and H. Allende-Cid, "Implementation and evaluation of an additional GPT-4-based reviewer in PRISMA-based medical systematic literature reviews," *International Journal of Medical Informatics*, vol. 189, p. 105531, 2024, doi: https://doi.org/10.1016/j.ijmedinf.2024.105531.

[97] M. Li, H. Kilicoglu, H. Xu, and R. Zhang, "BiomedRAG: A Retrieval Augmented Large Language Model for Biomedicine," *arXiv preprint arXiv:2405.00465*, 2024, [Online]. Available: https://arxiv.org/abs/2405.00465

[98] C. Maciejewski, K. Ozierański, A. Barwiołek, M. Basza, A. Bożym, M. Ciurla, M. Janusz Krajsman, M. Maciejewska, P. Lodziński, G. Opolski, *et al.*, "AssistMED project: Transforming cardiology cohort characterisation from electronic health records through natural language processing – Algorithm design, preliminary results, and field prospects," *International Journal of Medical Informatics*, vol. 185, p. 105380, 2024, doi: https://doi.org/10.1016/j.ijmedinf.2024.105380.

[99] H. Niu, O. A. Omitaomu, M. A. Langston, M. Olama, O. Ozmen, H. B. Klasky, A. Laurio, M. Ward, and J. Nebeker, "EHR-BERT: A BERT-based model for effective anomaly detection in electronic health records," *Journal of Biomedical Informatics*, vol. 150, p. 104605, 2024, doi: https://doi.org/10.1016/j.jbi.2024.104605.

[100] R. Yang, "CaseGPT: a case reasoning framework based on language models and retrieval-augmented generation," *arXiv preprint arXiv:2407.07913*, 2024, [Online]. Available: https://arxiv.org/abs/2407.07913

[101] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, "MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models," in *Proceedings of the ACM Web Conference 2024 (WWW '24)*, 2024, pp. 4489–4500. doi: 10.1145/3589334.3648137.

[102] Y. Gao, R. Li, E. Croxford, J. Caskey, B. W. Patterson, M. Churpek, T. Miller, D. Dligach, and M. Afshar, "Leveraging Medical Knowledge Graphs Into Large Language Models for Diagnosis Prediction: Design and Application Study," *JMIR AI*, vol. 4, p. e58670, 2025, doi: 10.2196/58670.

[103] M. Delsoz, H. Raja, Y. Madadi, A. A. Tang, B. M. Wirostko, M. Y. Kahook, and S. Yousefi, "The Use of ChatGPT to Assist in Diagnosing Glaucoma Based on Clinical Case Reports," *Ophthalmology and Therapy*, 2024, doi: 10.1007/s40123-023-00805-x.

[104] B. Lin, Y. Xu, X. Bao, Z. Zhao, Z. Wang, and J. Yin, "SkinGEN: an Explainable Dermatology Diagnosis-to-Generation Framework with Interactive Vision-Language Models," *30th International Conference on Intelligent User Interfaces (IUI '25)*, 2025, doi: 10.1145/3708359.3712098.

[105] T. Savage, A. Nayak, R. Gallo, E. Rangan, and J. H. Chen, "Diagnostic Reasoning Prompts Reveal the Potential for Large Language Model Interpretability in Medicine," *npj Digital Medicine*, vol. 7, no. 2, p. 20, 2024, doi: 10.1038/s41746-024-01010-1.

[106] M.-Y. Huang, C.-S. Weng, H.-L. Kuo, and Y.-C. Su, "Using a chatbot to reduce emergency department visits and unscheduled hospitalizations among patients with gynecologic malignancies during chemotherapy: A retrospective cohort study," *Heliyon*, vol. 9, no. 5, p. e15798, 2023, doi: https://doi.org/10.1016/j.heliyon.2023.e15798.

[107] M. M. Carlà, G. Gambini, A. Baldascino, F. Boselli, F. Giannuzzi, F. Margollicci, and S. Rizzo, "Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison," *Graefe's Archive for Clinical and Experimental Ophthalmology*, 2024, doi: 10.1007/s00417-024-06470-5.

[108] F. Dennstädt, J. Hastings, P. M. Putora, E. Vu, G. F. Fischer, K. Süveg, M. Glatzer, E. Riggenbach, H.-L. Hà, and N. Cihoric, "Exploring Capabilities of Large Language Models such as ChatGPT in Radiation Oncology," *Advances in Radiation Oncology*, vol. 9, no. 3, p. 101400, 2024, doi: https://doi.org/10.1016/j.adro.2023.101400.

[109] E. A. C. Dronkers, A. Geneid, C. al Yaghchi, and J. R. Lechien, "Evaluating the Potential of AI Chatbots in Treatment Decision-making for Acquired Bilateral Vocal Fold Paralysis in Adults," *Journal of Voice*, 2024, doi: https://doi.org/10.1016/j.jvoice.2024.02.020.

[110] O. K. Gargari, F. Fatehi, I. Mohammadi, S. R. Firouzabadi, A. Shafiee, and G. Habibi, "DIAGNOSTIC ACCURACY OF LARGE LANGUAGE MODELS IN PSYCHIATRY," *Asian Journal of Psychiatry*, p. 104168, 2024, doi: https://doi.org/10.1016/j.ajp.2024.104168.

[111] C. A. Gomez-Cabello, S. Borna, S. M. Pressman, A. Sehgal, B. C. Leibovich, and A. J. Forte, "Large Language Models for Intraoperative Decision Support in Plastic Surgery: A Comparison between ChatGPT-4 and Gemini," *medicina*, vol. 60, no. 6, p. 957, 2024, doi: https://doi.org/10.3390/medicina60060957.

[112] N. Gopalakrishnan, A. Joshi, J. Chhablani, N. K. Yadav, N. G. Reddy, P. K. Rani, R. S. Pulipaka, R. Shetty, S. Sinha, V. Prabhu, *et al.*, "Recommendations for Initial Diabetic Retinopathy Screening of Diabetic Patients Using Large Language Model-Based Artificial Intelligence in Real-Life Case Scenarios," *International Journal of Retina and Vitreous*, vol. 10, p. 11, 2024, doi: 10.1186/s40942-024-00533-9.

[113] J. Y. Hsueh, D. Nethala, S. Singh, W. M. Linehan, and M. W. Ball, "Investigating the clinical reasoning abilities of large language model GPT-4: an analysis of postoperative complications from renal surgeries," *Urologic Oncology: Seminars and Original Investigations*, vol. 42, no. 9, p. 292.e1, 2024, doi: https://doi.org/10.1016/j.urolonc.2024.04.010.

[114] T. L. Wiemken and R. M. Carrico, "Assisting the infection preventionist: Use of artificial intelligence for health care–associated infection surveillance," *American Journal of Infection Control*, vol. 52, no. 6, pp. 625–629, 2024, doi: https://doi.org/10.1016/j.ajic.2024.02.007.

[115] C. Mao, J. Xu, L. Rasmussen, Y. Li, P. Adekkanattu, J. Pacheco, B. Bonakdarpour, R. Vassar, L. Shen, G. Jiang, *et al.*, "AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease," *Journal of Biomedical Informatics*, vol. 144, p. 104442, 2023, doi: https://doi.org/10.1016/j.jbi.2023.104442.

[116] M. Jin, Q. Yu, D. Shu, C. Zhang, L. Fan, W. Hua, S. Zhu, Y. Meng, Z. Wang, M. Du, *et al.*, "Health-LLM: Personalized Retrieval-Augmented Disease Prediction System," *arXiv preprint arXiv:2402.00746*, 2024, [Online]. Available: https://arxiv.org/abs/2402.00746

[117] Y. H. Yeo, J. S. Samaan, W. H. Ng, P.-S. Ting, H. Trivedi, A. Vipani, W. Ayoub, J. D. Yang, O. Liran, B. Spiegel, *et al.*, "Assessing the performance of ChatGPT in answer- ing questions regarding cirrhosis and hepatocellu- lar carcinoma," *Clinical and Molecular Hepatology*, vol. 29, no. 3, pp. 721–732, 2023, doi: 10.3350/cmh.2023.0089.

[118] B. B. Whiles, V. G. Bird, B. K. Canales, J. M. DiBianco, and R. S. Terry, "Caution! AI Bot Has Entered the Patient Chat: ChatGPT Has Limitations in Providing Accurate Urologic Healthcare Advice," *Urology*, vol. 180, pp. 278–284, 2023, doi: https://doi.org/10.1016/j.urology.2023.07.010.

[119] Z. W. Lim, K. Pushpanathan, S. M. E. Yew, Y. Lai, C.-H. Sun, J. S. H. Lam, D. Z. Chen, J. H. L. Goh, M. C. J. Tan, B. Sheng, *et al.*, "Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard," *eBioMedicine*, vol. 95, p. 104770, 2023, doi: https://doi.org/10.1016/j.ebiom.2023.104770.

[120] E. Waisberg, J. Ong, M. Masalkhi, N. Zaman, P. Sarker, A. G. Lee, and A. Tavakkoli, "Google's AI chatbot 'Bard': a side-by-side comparison with ChatGPT and its utilization in ophthalmology," *Eye*, vol. 38, pp. 642–645, 2024, doi: 10.1038/s41433-023-02760-0.

[121] J. W. Ayers, A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, A. M. Goodman, C. A. Longhurst, M. Hogarth, *et al.*, "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum," *JAMA Internal Medicine*, vol. 183, no. 6, pp. 589–596, 2023, doi: 10.1001/jamainternmed.2023.1838.

[122] X. Chen, Z. Zhao, W. Zhang, P. Xu, L. Gao, M. Xu, Y. Wu, Y. Li, D. Shi, and M. He, "EyeGPT: Ophthalmic Assistant with Large Language Models," *arXiv preprint arXiv:2403.00840*, 2024, [Online]. Available: https://arxiv.org/abs/2403.00840

[123] C. A. Gomez-Cabello, S. Borna, S. M. Pressman, A. Sehgal, B. C. Leibovich, and A. J. Forte, "Artificial Intelligence in Postoperative Care: Assessing Large Language Models for Patient Recommendations in Plastic Surgery," *Healthcare*, vol. 12, no. 11, p. 1083, 2024, doi: https://doi.org/10.3390/healthcare12111083.

[124] E. Kozaily, M. Geagea, E. R. Akdogan, J. Atkins, M. B. Elshazly, M. Guglin, R. J. Tedford, and R. M. Wehbe, "Accuracy and consistency of online large language model-based artificial intelligence chat platforms in answering patients' questions about heart failure," *International Journal of Cardiology*, vol. 408, p. 132115, 2024, doi: https://doi.org/10.1016/j.ijcard.2024.132115.

[125] T. Lai, Y. Shi, Z. Du, J. Wu, K. Fu, Y. Dou, and Z. Wang, "Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs)," *BioMedInformatics*, vol. 4, no. 1, pp. 8–33, 2024, doi: 10.3390/biomedinformatics4010002.

[126] S. Lang, J. Vitale, T. F. Fekete, D. Haschtmann, R. Reitmeir, M. Ropelato, J. Puhakka, F. Galbusera, and M. Loibl, "Are large language models valid tools for patient information on lumbar disc herniation? The spine surgeons' perspective," *Brain and Spine*, vol. 4, p. 102804, 2024, doi: https://doi.org/10.1016/j.bas.2024.102804.

[127] C. L. Monroe, Y. G. Abdelhafez, K. Atsina, E. Aman, L. Nardo, and M. H. Madani, "Evaluation of responses to cardiac imaging questions by the artificial intelligence large language model ChatGPT," *Clinical Imaging*, vol. 112, p. 110193, 2024, doi: https://doi.org/10.1016/j.clinimag.2024.110193.

[128] R. Olszewski, K. Watros, M. Mańczak, J. Owoc, K. Jeziorski, and J. Brzeziński, "Assessing the response quality and readability of chatbots in cardiovascular health, oncology, and psoriasis: A comparative study," *International Journal of Medical Informatics*, vol. 190, p. 105562, 2024, doi: https://doi.org/10.1016/j.ijmedinf.2024.105562.

[129] C. E. Onder, G. Koc, P. Gokbulut, I. Taskaldiran, and S. M. Kuskonmaz, "Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-023-50884-w.

[130] Z. Rasool, S. Kurniawan, S. Balugo, S. Barnett, R. Vasa, C. Chesser, B. M. Hampstead, S. Belleville, K. Mouzakis, and A. Bahar-Fuchs, "Evaluating LLMs on document-based QA: Exact answer selection and numerical extraction using CogTale dataset," *Natural Language Processing Journal*, vol. 8, p. 100083, 2024, doi: https://doi.org/10.1016/j.nlp.2024.100083.

[131] M. B. Singer, J. J. Fu, J. Chow, and C. C. Teng, "Development and Evaluation of Aeyeconsult: A Novel Ophthalmology Chatbot Leveraging Verified Textbook Knowledge and GPT-4," *Journal of Surgical Education*, vol. 81, no. 3, pp. 438–443, 2024, doi: https://doi.org/10.1016/j.jsurg.2023.11.019.

[132] Z. Zhao, S. Wang, J. Gu, Y. Zhu, L. Mei, Z. Zhuang, Z. Cui, Q. Wang, and D. Shen, "ChatCAD+: Towards a Universal and Reliable Interactive CAD using LLMs," *arXiv preprint arXiv:2305.15964*, 2024.

[133] T. I. Wilhelm, J. Roos, and R. Kaczmarczyk, "Large Language Models for Therapy Recommendations Across 3 Clinical Specialties: Comparative Study," *Journal of Medical Internet Research*, vol. 25, p. e49324, Oct. 2023, doi: 10.2196/49324.

[134] Z. Liu, Y. Li, P. Shu, A. Zhong, L. Yang, C. Ju, Z. Wu, C. Ma, J. Luo, C. Chen, *et al.*, "Radiology-Llama2: Best-in-Class Large Language Model for Radiology," *arXiv preprint arXiv:2309.06419*, 2023.

[135] M. Alkhalaf, P. Yu, M. Yin, and C. Deng, "Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records," *Journal of Biomedical Informatics*, vol. 156, p. 104662, 2024, doi: https://doi.org/10.1016/j.jbi.2024.104662.

[136] T. Alqahtani, H. A. Badreldin, M. Alrashed, A. I. Alshaya, S. S. Alghamdi, K. bin Saleh, S. A. Alowais, O. A. Alshaya, I. Rahman, M. S. Al Yami, *et al.*, "The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research," *Research in Social and Administrative Pharmacy*, vol. 19, no. 8, pp. 1236–1242, 2023, doi: https://doi.org/10.1016/j.sapharm.2023.05.016.

[137] Z. Chen, Q. Wang, Y. Sun, H. Cai, and X. Lu, "Chat-ePRO: Development and pilot study of an electronic patient-reported outcomes system based on ChatGPT," *Journal of Biomedical Informatics*, vol. 154, p. 104651, 2024, doi: https://doi.org/10.1016/j.jbi.2024.104651.

[138] E. T. Pan and M. Florian-Rodriguez, "Human vs machine: identifying ChatGPT-generated abstracts in Gynecology and Urogynecology," *American Journal of Obstetrics and Gynecology*, vol. 231, no. 2, p. 276.e1, 2024, doi: https://doi.org/10.1016/j.ajog.2024.04.045.

[139] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu, "BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model," *arXiv preprint arXiv:2204.03905*, 2022.

[140] G. Eysenbach, "The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers," *JMIR Medical Education*, vol. 9, 2023, doi: 10.2196/46885.

[141] S. S. Mannam, R. Subtirelu, D. Chauhan, H. S. Ahmad, I. M. Matache, K. Bryan, S. V. K. Chitta, S. C. Bathula, R. Turlip, C. Wathen, *et al.*, "Large Language Model-Based Neurosurgical Evaluation Matrix: A Novel Scoring Criteria to Assess the Efficacy of ChatGPT as an Educational Tool for Neurosurgery Board Preparation," *World Neurosurgery*, vol. 180, p. e765, 2023, doi: https://doi.org/10.1016/j.wneu.2023.10.043.

[142] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Med-HALT: Medical Domain Hallucination Test for Large Language Models," *arXiv preprint arXiv:2307.15343*, 2023.

[143] B. R. Beaulieu-Jones, M. T. Berrigan, S. Shah, J. S. Marwaha, S.-L. Lai, and G. A. Brat, "Evaluating capabilities of large language models: Performance of GPT-4 on surgical knowledge assessments," *Surgery*, vol. 175, no. 4, pp. 936–942, 2024, doi: https://doi.org/10.1016/j.surg.2023.12.014.

[144] B. Quah, C. W. Yong, C. W. M. Lai, and I. Islam, "Performance of large language models in oral and maxillofacial surgery examinations," *International Journal of Oral and Maxillofacial Surgery*, 2024, doi: https://doi.org/10.1016/j.ijom.2024.06.003.

[145] S. Yamaguchi, M. Morishita, H. Fukuda, K. Muraoka, T. Nakamura, I. Yoshioka, I. Soh, K. Ono, and S. Awano, "Evaluating the efficacy of leading large language models in the Japanese national dental hygienist examination: A comparative analysis of ChatGPT, Bard, and Bing Chat," *Journal of Dental Sciences*, 2024, doi: https://doi.org/10.1016/j.jds.2024.02.019.

[146] Y. Lee, L. Tessier, K. Brar, S. Malone, D. Jin, T. McKechnie, J. J. Jung, M. Kroh, and J. T. Dang, "Performance of artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and Bard in the American Society for Metabolic and Bariatric Surgery textbook of bariatric surgery questions," *Surgery for Obesity and Related Diseases*, vol. 20, no. 7, pp. 609–613, 2024, doi: 10.1016/j.soard.2024.04.014.

[147] B. Coskun, G. Ocakoglu, M. Yetemen, and O. Kaygisiz, "Can ChatGPT, an Artificial Intelligence Language Model, Provide Accurate and High-quality Patient Information on Prostate Cancer?," *Urology*, vol. 180, pp. 35–58, 2023, doi: https://doi.org/10.1016/j.urology.2023.05.040.

[148] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, and T. Liu, "HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge," *arXiv preprint arXiv:2304.06975*, 2023, [Online]. Available: https://arxiv.org/abs/2304.06975

[149] R. K. Khanna, J.-B. Ducloyer, A. Hage, A. Rezkallah, E. Durbant, M. Bigoteau, R. Mouchel, R. Guillon-Rolf, L. Le, R. Tahiri, *et al.*, "Evaluating the potential of ChatGPT-4 in ophthalmology: The good, the bad and the ugly," *Journal Français d'Ophtalmologie*, vol. 46, no. 7, pp. 697–705, 2023, doi: https://doi.org/10.1016/j.jfo.2023.07.001.

[150] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, *et al.*, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no. 2, pp. 1–12, 2023, doi: 10.1371/journal.pdig.0000198.

[151] T. Sean, C. Lauren, W. Emma, Y. Antonio, and F. Misha, "Improved Performance of ChatGPT-4 on the OKAP Examination: A Comparative Study with ChatGPT-3.5," *Journal of Academic Ophthalmology*, vol. 15, no. 2, p. 184187, 2023, doi: https://doi.org/10.1055/s-0043-1774399.

[152] L. Finch, V. Broach, J. Feinberg, A. Al-Niaimi, N. R. Abu-Rustum, Q. Zhou, A. Iasonos, and D. S. Chi, "ChatGPT compared to national guidelines for management of ovarian cancer: Did ChatGPT get it right? – A Memorial Sloan Kettering Cancer Center Team Ovary study," *Gynecologic Oncology*, vol. 189, pp. 75–79, 2024, doi: https://doi.org/10.1016/j.ygyno.2024.07.007.

[153] J. Lee, I.-S. Yoo, J.-H. Kim, W. T. Kim, H. J. Jeon, H.-S. Yoo, J. G. Shin, G.-H. Kim, S. Hwang, S. Park, *et al.*, "Development of AI-generated medical responses using the ChatGPT for cancer patients," *Computer Methods and Programs in Biomedicine*, vol. 254, p. 108302, 2024, doi: https://doi.org/10.1016/j.cmpb.2024.108302.

[154] R. C. King, J. S. Samaan, Y. H. Yeo, B. Mody, D. M. Lombardo, and R. Ghashghaei, "Appropriateness of ChatGPT in Answering Heart Failure Related Questions," *Heart, Lung and Circulation*, 2024, doi: https://doi.org/10.1016/j.hlc.2024.03.005.

[155] H. Mohammad-Rahimi, Z. H. Khoury, M. I. Alamdari, R. Rokhshad, P. Motie, A. Parsa, T. Tavares, J. J. Sciubba, J. B. Price, and A. S. Sultan, "Performance of AI chatbots on controversial topics in oral medicine, pathology, and radiology," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 137, no. 5, pp. 508–514, 2024, doi: https://doi.org/10.1016/j.oooo.2024.01.015.

[156] T. P. Nguyen, B. Carvalho, H. Sukhdeo, K. Joudi, N. Guo, M. Chen, J. T. Wolpaw, J. J. Kiefer, M. Byrne, T. Jamroz, *et al.*, "Comparison of artificial intelligence large language model chatbots in answering frequently asked questions in anaesthesia," *BJA Open*, vol. 10, p. 100280, 2024, doi: https://doi.org/10.1016/j.bjao.2024.100280.

[157] Y. Wu, Z. Zhang, X. Dong, S. Hong, Y. Hu, P. Liang, L. Li, B. Zou, X. Wu, D. Wang, *et al.*, "Evaluating the performance of the language model ChatGPT in responding to common questions of people with epilepsy," *Epilepsy & Behavior*, vol. 151, p. 109645, 2024, doi: https://doi.org/10.1016/j.yebeh.2024.109645.

[158] X. Xue, D. Zhang, C. Sun, Y. Shi, R. Wang, T. Tan, P. Gao, S. Fan, G. Zhai, M. Hu, *et al.*, "Xiaoqing: A Q&A model for glaucoma based on LLMs," *Computers in Biology and Medicine*, vol. 174, p. 108399, 2024, doi: https://doi.org/10.1016/j.compbiomed.2024.108399.

[159] Z. Yang, E. Khatibi, N. Nagesh, M. Abbasian, I. Azimi, R. Jain, and A. M. Rahmani, "ChatDiet: Empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework," *Smart Health*, vol. 32, p. 100465, 2024, doi: https://doi.org/10.1016/j.smhl.2024.100465.

[160] B. Chintagunta, N. Katariya, X. Amatriain, and A. Kannan, "Medically Aware {GPT}-3 as a Data Generator for Medical Dialogue Summarization," *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pp. 66–76, 2021, doi: 10.18653/v1/2021.nlpmc-1.9.

[161] K. Roy, M. Gaur, M. Soltani, V. Rawte, A. Kalyan, and A. Sheth, "ProKnow: Process knowledge for safety constrained and explainable question generation for mental health diagnostic assistance," *Frontiers in Big Data*, vol. 5, p. 1056728, 2023, doi: 10.3389/fdata.2022.1056728.

[162] S. Balumuri, S. Bachina, and S. Kamath S, "SB_NITK at MEDIQA 2021: Leveraging Transfer Learning for Question Summarization in Medical Domain," in *Proceedings of the 20th Workshop on Biomedical*

Language Processing, D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, Eds., Online: Association for Computational Linguistics, 2021, pp. 273–279. doi: 10.18653/v1/2021.bionlp-1.31.

[163] B. Sharma, Y. Gao, T. Miller, M. M. Churpek, M. Afshar, and D. Dligach, "Multi-Task Training with In-Domain Language Models for Diagnostic Reasoning," *arXiv preprint arXiv:2306.04551*, 2023, [Online]. Available: https://arxiv.org/abs/2306.04551

[164] Y.-N. Chuang, R. Tang, X. Jiang, and X. Hu, "{SPeC}: A Soft Prompt-Based Calibration on Performance Variability of Large Language Model in Clinical Notes Summarization," *Journal of Biomedical Informatics*, vol. 151, p. 104606, 2024, doi: 10.1016/j.jbi.2024.104606.

[165] S. Yadav, D. Gupta, and D. Demner-Fushman, "{CHQ-Summ}: A Dataset for Consumer Healthcare Question Summarization," *arXiv preprint arXiv:2206.06581*, 2022, doi: 10.48550/arXiv.2206.06581.

[166] S. Yadav, M. Sarrouti, and D. Gupta, "NLM at MEDIQA 2021: Transfer Learning-based Approaches for Consumer Question and Multi-Answer Summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, Eds., Online: Association for Computational Linguistics, 2021, pp. 291–301. doi: 10.18653/v1/2021.bionlp-1.34.

[167] C. Feng, A. Mehmani, and J. Zhang, "Deep Learning-Based Real-Time Building Occupancy Detection Using AMI Data," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4490–4501, 2020, doi: 10.1109/TSG.2020.2982351.

[168] A. Alqahtani, R. Salama, M. Diab, and A. Youssef, "Care4Lang at MEDIQA-Chat 2023: Fine-tuning Language Models for Classifying and Summarizing Clinical Dialogues," in *Proceedings of the 5th Clinical Natural Language Processing Workshop*, T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, and A. Rumshisky, Eds., Toronto, Canada: Association for Computational Linguistics, 2023, pp. 524–528. doi: 10.18653/v1/2023.clinicalnlp-1.55.

[169] B. Fatani, "ChatGPT for Future Medical and Dental Research," *Cureus*, vol. 15, no. 4, p. e37285, 2023, doi: 10.7759/cureus.37285.

[170] C. Wang, C. Yao, P. Chen, J. Shi, Z. Gu, and Z. Zhou, "Artificial Intelligence Algorithm with ICD Coding Technology Guided by Embedded Electronic Medical Record System in Medical Record Information Management," *Microprocessors and Microsystems*, p. 104962, 2023, doi: https://doi.org/10.1016/j.micpro.2023.104962.

[171] R. F. Teixeira Gomes, L. F. Schuch, M. Domingues Martins, E. F. Honório, R. M. de Figueiredo, J. Schmith, G. N. Machado, and V. C. Carrard, "Use of Deep Neural Networks in the Detection and Automated Classification of Lesions Using Clinical Images in Ophthalmology, Dermatology, and Oral Medicine—A Systematic Review," *Journal of Digital Imaging*, vol. 36, pp. 1060–1070, 2023, doi: 10.1007/s10278-023-00775-3.

[172] S. Hornstein, J. Scharfenberger, U. Lueken, R. Wundrack, and K. Hilbert, "Predicting recurrent chat contact in a psychological intervention for the youth using natural language processing," *npj Digital Medicine*, vol. 7, no. 1, 2024, doi: 10.1038/s41746-024-01121-9.

[173] D. Zarate, M. Ball, M. Prokofieva, V. Kostakos, and V. Stavropoulos, "Identifying self-disclosed anxiety on Twitter: A natural language processing approach," *Psychiatry Research*, vol. 330, p. 115579, 2023, doi: https://doi.org/10.1016/j.psychres.2023.115579.

[174] M. M. de A. Cardoso, J. Machado-Rugolo, L. Thabane, N. C. da Rocha, A. M. P. Barbosa, D. S. Komoda, J. T. C. de Almeida, D. da S. P. Curado, S. A. T. Weber, and L. G. M. de Andrade, "Application of natural language processing to predict final recommendation of Brazilian health technology assessment reports," *International Journal of Technology Assessment in Health Care*, vol. 40, no. 1, p. e19, 2024, doi:

10.1017/S0266462324000163.

[175] V. Nair, N. Katariya, X. Amatriain, I. Valmianski, and A. Kannan, "Adding more data does not always help: A study in medical conversation summarization with PEGASUS," *arXiv preprint arXiv:2111.07564*, 2021, doi: 10.48550/arXiv.2111.07564.

[176] F. P. Hansen, L. B. Soares, E. Shareghi, and A. Søgaard, "Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens," *arXiv preprint arXiv:2206.13349*, 2022, doi: 10.48550/arXiv.2206.13349.

[177] X. Li, Y. Li, L. Qiu, S. Joty, and L. Bing, "Evaluating Psychological Safety of Large Language Models," *arXiv preprint arXiv:2212.10529*, 2022.

[178] S. Y. Feng, V. Khetan, B. Sacaleanu, A. Gershman, and E. Hovy, "CHARD: Clinical Health-Aware Reasoning Across Dimensions for Text Generation Models," *arXiv preprint arXiv:2210.04191*, 2022.

[179] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou, "Large language models in medicine: the potentials and pitfalls," *Annals of Internal Medicine*, vol. 177, no. 2, pp. 210–220, 2024, doi: 10.7326/M23-2772.

[180] A. Stanciu, "Data Management Plan for Healthcare: Following FAIR Principles and Addressing Cybersecurity Aspects. A Systematic Review using InstructGPT," *medRxiv*, 2023, doi: 10.1101/2023.04.21.23288932.

[181] D. Rezaeikhonakdar, "AI Chatbots and Challenges of HIPAA Compliance for AI Developers and Vendors," *Journal of Law, Medicine & Ethics*, no. 4, pp. 988–995, 2023, doi: 10.1017/jme.2024.15.

[182] F. Mireshghallah, T. Hartvigsen, A. L. Beam, and M. Ghassemi, "Privacy preserving strategies for electronic health records in the era of large language models," *npj Digital Medicine*, vol. 8, no. 1, p. 57, 2025, doi: 10.1038/s41746-025-01429-0.

[183] S. B. Atallah, N. R. Banda, A. Banda, and N. A. Roeck, "How large language models including generative pre-trained transformer (GPT) 3 and 4 will impact medicine and surgery," *Techniques in Coloproctology*, vol. 27, pp. 609–614, 2023, doi: 10.1007/s10151-023-02837-8.

[184] Q. Xie, E. J. Schenck, H. S. Yang, Y. Chen, Y. Peng, and F. Wang, "Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond," *medRxiv*, 2023, doi: 10.1101/2023.04.18.23288752.

[185] D. M. Korngiebel and S. D. Mooney, "Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery," *NPJ Digital Medicine*, vol. 4, no. 1, pp. 93–95, 2021, doi: 10.1038/s41746-021-00427-0.

[186] J. Lin, X. Dai, Y. Xi, W. Liu, B. Chen, H. Zhang, Y. Liu, C. Wu, X. Li, C. Zhu, *et al.*, "How Can Recommender Systems Benefit from Large Language Models: A Survey," *arXiv preprint*, 2023, doi: 10.48550/arXiv.2306.05817.

[187] S. Thapa and S. Adhikari, "ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls," *Annals of Biomedical Engineering*, vol. 51, no. 12, pp. 2647–2651, 2023, doi: 10.1007/s10439-023-03284-0.

[188] A. Bisercic, M. Nikolic, M. van der Schaar, B. Delibasic, P. Lio, and A. Petrovic, "Interpretable Medical Diagnostics with Structured Data Extraction by Large Language Models," *arXiv preprint*, 2023, doi: 10.48550/arXiv.2306.05052.

[189] Y. Jiang, R. Qiu, Y. Zhang, and P.-F. Zhang, "Balanced and Explainable Social Media Analysis for Public Health with Large Language Models," *arXiv preprint*, 2023, doi: 10.48550/arXiv.2309.05951.

[190] Z. Al Nazi and W. Peng, “Large Language Models in Healthcare and Medical Domain: A Review,” *Informatics*, vol. 11, no. 3, 2024, doi: 10.3390/informatics11030057.