Review

Generative AI in Medicine and Healthcare: Moving Beyond the ‘Peak of Inflated Expectations’

Peng Zhang , Jiayu Shi , Maged N. Kamel Boulos *

Posted Date: 4 September 2024

doi: 10.20944/preprints202409.0311.v1

Keywords: generative AI; large language models; AI chatbots; ChatGPT; artificial intelligence; retrieval-augmented generation; medicine; healthcare; human health; AI regulation

Review

Generative AI in Medicine and Healthcare: Moving Beyond the 'Peak of Inflated Expectations'

Peng Zhang [1], Jiayu Shi [1] and Maged N. Kamel Boulos [2,*]

[1] Department of Computer Science and Data Science Institute, Vanderbilt University, Nashville, TN 37240, USA
[2] School of Medicine, University of Lisbon, 1649-028 Lisbon, Portugal
* Correspondence: mnkboulos@ieee.org

Abstract: The rapid development of specific-purpose Large Language Models (LLMs), such as Med-PaLM, MEDITRON-70B, and Med-Gemini, has significantly impacted healthcare, offering unprecedented capabilities in clinical decision support, diagnostics, and personalized health monitoring. This paper reviews the advancements in medicine-specific LLMs, the integration of Retrieval-Augmented Generation (RAG) and prompt engineering, and their applications in improving diagnostic accuracy and educational utility. Despite the potential, these technologies present challenges, including bias, hallucinations, and the need for robust safety protocols. The paper also discusses the regulatory and ethical considerations necessary for integrating these models into mainstream healthcare. By examining current studies and developments, this paper aims to provide a comprehensive overview of the state of LLMs in medicine and highlight the future directions for research and application. The study concludes that while LLMs hold immense potential, their safe and effective integration into clinical practice requires rigorous testing, ongoing evaluation, and continuous collaboration among stakeholders.

Keywords: generative AI; large language models; AI chatbots; ChatGPT; artificial intelligence; retrieval-augmented generation; medicine; healthcare; human health; AI regulation

1. Introduction

In July 2024, twenty months after the initial public launch of OpenAI's ChatGPT in November 2022, Gartner, the firm behind the well-known Hype Cycle methodology, declared in a new research report that generative AI has passed the 'peak of inflated expectations' and is moving into the 'trough of disillusionment.' This phase is expected to lead to the 'slope of enlightenment,' ultimately resulting in a 'plateau of productivity' as the technology matures, becomes mainstream, and its real-world benefits begin to materialize [1]. Applications utilizing generative AI and large language models (LLMs) in patient management, such as diagnosis, are considered SaMD/AIaMD (software as a medical device/AI as a medical device) and fall under established MDR (medical device regulation) provisions. However, as of August 2024, no application of this nature has been approved by regulatory bodies like the FDA (Food and Drug Administration, US), MHRA (Medicines and Healthcare products Regulatory Agency, UK), EMA (European Medicines Agency, EU), or corresponding agencies worldwide.

While these tools show promise in certain scenarios, such as assisting in difficult diagnoses [2], the current generation of generative AI and LLMs, including medically trained models such as Google's Med-PaLM 2 [3], are not yet ready for mainstream clinical use. Passing a medical licensing exam with high scores [4]—something that these models can accomplish—does not equate to readiness for safe use in routine patient care [5–7]. This is due to several key limitations inherent in the technology.

One of the most significant issues is the phenomenon of AI "hallucinations," where the models generate plausible sounding but factually incorrect or nonsensical information. A recent study by Aljamaan et al. developed a Reference Hallucination Score specifically for medical AI chatbots,

 doi:10.20944/preprints202409.0311.v1

highlighting the significance of accurately detecting and mitigating hallucinations to ensure the reliability and safety of these tools in clinical environments [8]. This issue of hallucinations, combined with the often inconsistent, unpredictable, and fluctuating (stochastic) performance of these models, and their proneness to bias, underscores their lack of real human-like intelligence. These models operate on vast amounts of data but lack the nuanced understanding and contextual awareness that human practitioners possess. As a result, their unregulated use in critical medical settings could lead to dangerous oversights or errors.

For these tools to be safely and effectively integrated into mainstream healthcare, substantial technological advancements are necessary. These advancements, a number of which will be briefly presented later in this article, would need to address the current limitations and ensure that AI models can reliably support clinical decision-making without introducing undue risk.

Despite these challenges and the current lack of regulatory approvals, some early clinician adopters are already using these tools in practice. This premature adoption is particularly concerning given the potential for critical information oversight—such as missing a patient's drug allergies due to an AI error [7]. The dangers of such oversights have led to growing calls from within the medical community for stringent regulations to govern the use of AI in healthcare. For instance, there have been increasing demands for clear guidelines and rules to prevent medical mistakes caused by AI, highlighting the urgent need for a regulatory framework that ensures patient safety while enabling innovation [9,10].

This article extends upon our previous review [11] and highlights recent advancements from August 2023. By examining the potential advantages, challenges, and ethical considerations of applying generative AI models in medicine and healthcare, this study aims to contribute to the ongoing dialogue on harnessing AI's capabilities responsibly for the betterment of medical practice and patient well-being.

2. Background: A Brief Technology Overview

In this section, we explore the foundational concepts, advancements, and applications of Natural Language Processing (NLP) in healthcare, highlighting the transformative impact of Generative Pre-trained Transformer (GPT) models and related technologies.

Natural Language Processing (NLP) aims to enable machines to comprehend, interpret, and generate human language meaningfully and contextually. Tasks within NLP span language translation, sentiment analysis, speech recognition, text summarization, and question answering, each bridging human communication with computational understanding. In healthcare, NLP has been potentially transformative, extracting valuable insights from vast amounts of unstructured clinical data, such as Electronic Health Records (EHRs), medical literature, and patient-generated content. It excels, though not without some limitations, in converting unstructured clinical notes into structured data, identifying medical conditions, medications, and lab tests through named entity recognition, and detecting adverse drug events by analyzing identified drugs and their interactions. Additionally, NLP models can contribute to early disease detection, facilitating timely interventions and improving patient outcomes.

Among the notable developments in NLP are the Generative Pre-trained Transformer (GPT) models developed by OpenAI. These models leverage the transformer architecture introduced by Vaswani et al. in 2017, which excels in processing sequential data, especially text, using self-attention mechanisms. Self-attention allows the model to understand the importance of and relationships between words in a sentence, improving contextual language understanding and semantic comprehension [12]. The evolution of GPT models has been rapid, beginning with GPT-1 in June 2018, demonstrating the potential of pre-training large-scale Transformer models on extensive text data [13]. GPT-2 followed in February 2019 with 1.5 billion parameters, raising concerns about generating human-like text, including fake news [14]. GPT-3, released in June 2020, significantly advanced the field with 175 billion parameters, showing unparalleled language generation capabilities [15]. GPT-4, introduced in March 2023, further improved performance and included multimodal capabilities, accepting both image and text inputs [16].

 doi:10.20944/preprints202409.0311.v1

There are many general-purpose models on offer today, including Llama 2 [17], GPT-4o (omni with voice mode) [18], LangChain [19], Claude [20], and Mistral [21], among others. Generative AI's capabilities extend beyond text to produce synthetic content such as images, videos, and audio. Tools such as Image FX [22] and DALL-E [23] use language prompts to generate images (Figure 1). GPT-4V is capable of handling multimodal inputs such as both text and images for tasks like visual description and object localization [24]. These models can create visuals to aid in conveying information, which can make them valuable in clinical and public health scenarios requiring rapid and clear communication.

Prompt engineering is crucial for optimizing the performance of generative AI models. It involves designing and refining input prompts to elicit the most accurate and relevant responses from the models. Effective prompt engineering can significantly enhance the usability and reliability of AI systems in healthcare settings [25]. Retrieval-Augmented Generation (RAG) represents another advancement in AI, enhancing the capabilities of LLMs by integrating external knowledge retrieval mechanisms [26]. RAG combines retrieval-based models, which fetch relevant documents or data, with generative models like GPT-3 and GPT-4, which generate coherent responses based on retrieved information. This approach addresses some limitations of pure generative models, such as hallucinations, by grounding responses in real, retrieved data [27]. Techniques like semantic entropy analysis also help detect and reduce hallucinations, increasing the reliability of LLMs in clinical applications [28]. Additionally, the recently introduced "Thermometer" method offers a novel approach to preventing AI models from becoming overly confident in incorrect answers by continuously evaluating their confidence levels across different predictions. This method helps in reducing overconfidence-related errors, making AI models more reliable and safer for clinical use [29].

The rapid advancement of GPT models and the integration of technologies like multimodal content generation, prompt engineering, and RAG carry the potential of reshaping healthcare communication and decision-making. However, the integration of AI in healthcare raises ethical and safety considerations. Chatbots, while potentially beneficial for patient communication, must be carefully managed to ensure they contribute positively to medical practice and patient well-being, and prevent any harmful outcomes. For instance, there is a risk that chatbots could mislead individuals with depression, potentially exacerbating their condition [30]. Ensuring the ethical and safe deployment of AI involves rigorous evaluation, transparent communication of limitations, and continuous monitoring to prevent misuse and harm. These issues are discussed in detail later in this paper.

 doi:10.20944/preprints202409.0311.v1

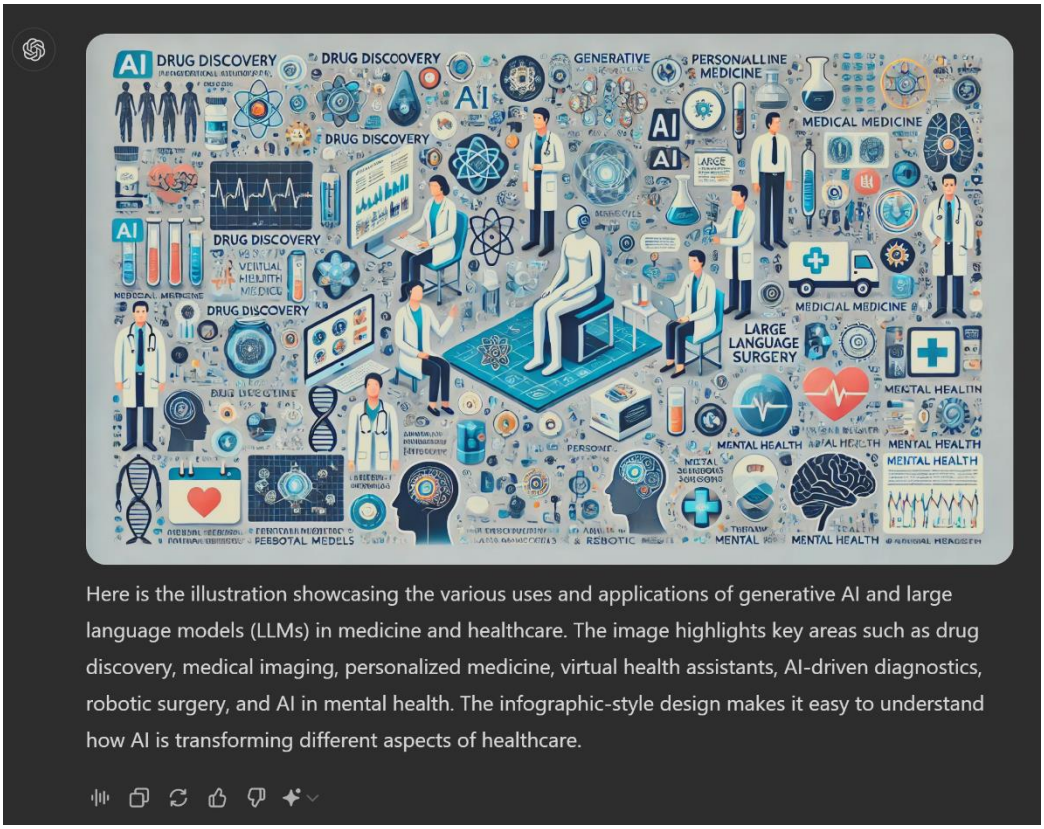

Figure 1. AI-generated image and text description in response to the prompt "generate an image illustrating the different generative AI and LLM uses and applications in medicine and healthcare." Note the malformed and misspelled text towards the top right part of the image (it was probably meant to read "personalized medicine"). This is a common observation with current models. Generator: OpenAI's DALL·E 3, September 1, 2024; Requestor: Maged N. Kamel Boulos; License: Public Domain (CC0).

3. Applications and Implications of Generative AI and LLMs in Medicine

This paper is meant as an incremental update to our original 2023 review [11]. As such, many of the applications, application examples, and issues discussed in the previous review will not be repeated here. For a more comprehensive overview of the subject, interested readers are advised to consult both papers together as one extended text.

3.1. Can We Trust Generative AI? Is It Clinically Safe and Reliable?

The integration of generative AI in healthcare offers transformative potential, from streamlining clinical tasks and improving diagnostic accuracy to providing health insights and aiding in disease prediction. However, as rightly noted by Hager et al. [6], LLMs are not yet ready for autonomous clinical decision-making. Significant concerns have been raised regarding LLM output quality variability, bias, ethical issues, and the need for robust regulatory frameworks, all of which must be satisfactorily addressed to ensure the safe and effective use of these technologies [6].

A number of studies have shown that generative AI cannot reliably extract and summarize information from unstructured clinical notes, e.g., [31,32] , and guidance is being issued by concerned bodies for that purpose; for example, the Royal Australian College of General Practitioners (RACGP) guidance on using AI scribes in practice [34] and the Australian Health Practitioner Regulation Agency (Ahpra) and National Boards guidance on meeting professional obligations when using AI in healthcare [35]. The RACGP guidance reminds general practitioners that they are liable for errors within patient health records even if they were generated by an AI scribe and that they must obtain patient consent to use an AI scribe in the consultation. Ahpra's guidance clearly states that "practitioners must apply human judgment to any output of AI."

 doi:10.20944/preprints202409.0311.v1

In an effort to address the challenge of "faithfulness hallucinations" in AI-generated medical summaries, Mendel AI, a US-based startup specializing in healthcare AI, in collaboration with UMass Amherst, used Hypercube, a tool for the automated detection of LLM hallucinations, to mitigate the high costs and time associated with a completely manual review of LLM-generated medical record summaries. Hypercube integrates medical knowledge bases, symbolic reasoning and NLP to detect hallucinations, allowing for an initial automated detection step before human expert review. If left undetected, these hallucinations can pose significant clinical risks to patients, and lead to misdiagnoses and inappropriate treatments [32].

The remainder of this subsection explores the regulatory, ethical, and practical challenges associated with implementing generative AI in clinical settings and public health, emphasizing the importance of comprehensive standards, human oversight, and collaboration among stakeholders.

3.1.1. Regulation, Regulatory Frameworks, and Ethics

Goodman et al. underscored the potential of LLM-generated summaries to streamline clinical tasks and reduce physician burnout. However, significant concerns were raised regarding their variability, potential for bias, and the lack of US FDA oversight. The authors emphasized the necessity for comprehensive standards, rigorous testing, and regulatory clarifications to ensure the safe and effective use of LLMs in clinical settings. It was pointed out that LLMs can exhibit sycophancy bias, tailoring responses to perceived user expectations, which may lead to confirmation biases in clinical decision-making. Additionally, small errors in summaries, such as adding words not present in the original data, can significantly impact clinical outcomes. The authors advocated for comprehensive standards and careful prompt engineering to mitigate these issues and ensure the safe use of LLMs in clinical environments [35].

Bharel et al. highlighted the critical importance of ethical considerations and robust regulatory frameworks for the safe and effective implementation of AI in public health. It called for comprehensive guidelines to address issues such as data privacy, bias, and accountability in AI-driven healthcare solutions. Moreover, it stressed the need for collaboration between policymakers, healthcare professionals, and AI developers to create regulations that ensure transparency and public trust in AI technologies [36].

The regulatory landscape for AI as a medical device (AIaMD) [37] is evolving, with some voices questioning the suitability of current regulatory frameworks. Google MedLM [38], for example, poses unique regulatory challenges (for any clinical application based on it) as it may be classified as software of unknown provenance (SoUP), complicating its path to approval [39]. Future clinical solutions incorporating Google MedLM and similar developments will need to meet stringent regulatory requirements set by agencies worldwide, potentially necessitating new or updated regulatory paradigms [40]. These concerns are echoed in recent discussions about the need to update regulatory frameworks that were originally optimized for earlier generations of AI. Howell et al. describe three epochs of AI with fundamentally different capabilities and risks: AI 1.0 or rule-based/symbolic AI, AI 2.0 or deep learning models, and AI 3.0 covering foundation models and generative AI [41]. As of 7 August 2024, the US FDA has approved 950 SaMDs that fall under AI 1.0/AI 2.0 but none that belong to AI 3.0 [42].

Blumenthal and Patel argue that the current approaches to assuring AI safety and efficacy are optimized for older forms of AI and that the regulation of clinical generative AI may require the development of new regulatory paradigms that can specifically address the complexities of LLMs [40]. It is noteworthy that the two articles by Blumenthal and Patel [40] and Howell et al. [41], whilst representing important contributions to the ongoing debate on the subject of medical generative AI regulation, feature Google employees among their authors. Caution should be exercised against potential conflicts of interest, especially when influential companies like Google are involved in shaping new regulation. This is not to say that these companies should not be involved in formulating new regulation. They should be included as a key (but not sole) stakeholder, and their contribution is vital, as without these companies and their very costly generative AI infrastructure and research, we would not have gotten to where we are today (e.g., [43]). However, it should always be

 doi:10.20944/preprints202409.0311.v1

remembered that current regulation focuses on claims, efficacy and safety—principles that do not and should never change across AI generations. Any new clinical generative AI regulation should not attempt to "dilute" existing quality and safety principles and standards in any way, but should rather evolve to maintain or improve upon the high standards that SaMD/AIaMD regulation adheres to today. Reform is inevitable and more agile regulatory changes are coming [37] that can better address the complexities of generative AI and the fast pace of developments in this field, protect patients, and encourage responsible innovation.

Blumenthal and Patel [40] are joined by Derraz et al., who similarly argue that current regulatory frameworks are a "de facto blocker to AI-based personalized medicine" and that new regulatory thinking and approaches are necessary to overcome this situation [44]. In the same vein, Freyer et al. call for a new regulatory framework that recognizes the capabilities and limitations of generative AI applications, but stress that this should be done while also enforcing existing regulations [45].

In August 2024, the UNESCO published its "*Consultation Paper on AI Regulation - Emerging Approaches Across the World*" in which it described nine non-mutually-exclusive regulatory approaches [46]. Medical AI regulatory frameworks (current and future) in different countries will often combine two or more of these approaches:

1. Principles-based approach: Offers core principles guiding the ethical and responsible creation and use of AI systems, emphasizing human-centered processes and respect for human rights.
2. Standards-based approach: Transfers state regulatory authority to organizations that develop technical standards to interpret and enforce mandatory rules.
3. Agile and experimentalist approach: Creates adaptable regulatory frameworks, like sandboxes, that allow businesses to test new AI models and tools under flexible regulations with government oversight.
4. Facilitating and enabling approach: Fosters an environment that promotes the development and use of ethical and human rights-compliant AI by all stakeholders.
5. Adapting existing laws approach: Updates sector-specific and general laws to improve the current regulatory system for AI.
6. Access to information and transparency mandates approach: Mandates transparency measures to ensure public access to basic information about AI systems.
7. Risk-based approach: Implements requirements based on the risk levels associated with using AI in various contexts.
8. Rights-based approach: Sets obligations to protect individuals' rights and freedoms when using AI.
9. Liability approach: Establishes accountability and penalties for the misuse of AI systems.

Schmidt et al. compiled a collection of 141 binding policies applicable to AI in healthcare and population health in the EU and 10 European countries. They concluded that specific AI regulation is still nascent, and that the combination in place today of existing data, technology, innovation, and health and human rights policies is already providing a baseline regulatory framework for AI in health, but needs additional work to address specific regulatory challenges [47].

The European Union's AI Act, which came into force on 1st August 2024 [48], introduces an additional regulatory layer requiring manufacturers to address AI-specific risks and ethical considerations in any medical application or device incorporating AI or machine learning functionalities. This act underscores the importance of aligning with both standard medical device regulations (MDR) and the new AI Act-specific requirements to ensure the safe and effective deployment of AI technologies in healthcare [49,50]. It is worth noting that the EU AI Act was the subject of strong lobbying efforts by big tech companies and EU member states to weaken much of its power through an overreliance on self-regulation and self-certification among other things [51].

The evaluation of AI-based clinical interventions is closely related to their regulation. Evaluation provides much of the evidence required by the governing regulatory frameworks to secure regulatory body approval for the mainstream use of these interventions. However, executing the ideal clinical trial for an AI-based intervention has always proved challenging, and because of this, we have hundreds of medical algorithms (non-generative-AI-based) that received approval on the basis of limited clinical data, which is far from ideal. The testing and evaluation of such interventions

 doi:10.20944/preprints202409.0311.v1

is inherently tricky; for example, a perfectly good algorithm can fail if clinicians (or patients) ignore its suggestions. Other evaluation challenges include AI bias and informed patient consent [52]. Generative AI algorithms are no exception.

This situation has led Coiera and Fraile-Navarro to propose a shift in the focus of generative AI evaluation from reliance on 'premarket assessment' to 'real-world postmarket surveillance'. They argue that traditional scientific methods may not be sufficient to evaluate generative AI, and that viewing it as a cybersocial ecosystem rather than as a specific technology may help with its global performance analysis, such as evaluating resilience and sustainability under changing conditions or tasks [53].

With the fast pace of generative AI and LLM developments, the continual mapping ongoing research on their applications in medicine and healthcare has become necessary to inform and regularly update the corresponding ethical frameworks, ensuring that these technologies are adopted responsibly and effectively. Ong et al. discussed the benefits of LLMs in medical research, education, and clinical tasks, emphasizing the associated challenges and ethical concerns, including data privacy, cognitive and automation biases, and accountability. They went on to propose a bioethical framework based on the principles of beneficence, nonmaleficence, autonomy, and justice to ensure responsible use of LLMs in medicine. They highlighted the importance of human oversight, transparency, and the need for regulations to mitigate risks while harnessing the benefits of these advanced AI technologies [54,55].

Haltaufderheide and Ranisch conducted a systematic review on the ethics of ChatGPT and other LLMs in medicine, identifying four broad LLM application categories (covering health professionals and researchers, patient support, clinical applications, and public health) and a number of recurring ethical issues related to epistemic values (reliability, transparency, hallucinations), therapeutic relationships, and privacy, among others. They also noted the recurrent calls for ethical guidance and human oversight in this area, and proposed shifting the focus of the ethical guidance discussion towards establishing clear definitions of acceptable human oversight in various applications, taking into account the diversity of settings, risks, and performance standards that are involved [56].

3.1.2. Bias and Harm

The use of LLMs in healthcare offers significant benefits but also raises concerns about bias and potential harm. Several studies have explored the performance and biases of LLMs compared to human clinicians.

Levkovich et al. compared ChatGPT-3.5 and ChatGPT-4 with primary care physicians in diagnosing and treating depression. Key findings include that ChatGPT models recommended psychotherapy for mild depression in over 95% of cases, significantly more than primary care physicians, and showed no significant gender or socioeconomic biases, unlike primary care physicians. ChatGPT favored antidepressants alone more often, while physicians preferred combining antidepressants with anxiolytics/hypnotics. This study suggested that ChatGPT aligns well with clinical guidelines and reduces biases, but further research is needed to ensure AI reliability and safety in clinical settings [57].

While some studies like the one above suggest that ChatGPT may exhibit more powerful and less biased capabilities compared with humans, the potential harms of LLMs must be addressed. Omiye et al. examined whether four LLMs (Bard, ChatGPT, Claude, GPT-4) spread harmful race-based medical content. They found that all LLMs tested showed instances of promoting outdated race-based medical practices, and their responses varied with question phrasing. This inconsistency and propagation of debunked racist medical ideas highlight the need for caution in their clinical use and call for further evaluation and adjustments [58].

Advances in reducing bias are also being made. Singhal et al. discussed the evaluation of Med-PaLM 2, highlighting its better performance and lower risk of harm compared to its predecessor. Med-PaLM 2 was tested on adversarial datasets to identify its limitations and potential biases. Although progress has been made in reducing biases and harmful outputs, further validation and safety studies are necessary for real-world applications [59].

 doi:10.20944/preprints202409.0311.v1

Kim et al. evaluated biases in responses from AI chatbots (ChatGPT-4 and Bard) and clinicians. Their study showed that both AI chatbots and clinicians displayed biases based on patient demographics, and that ChatGPT and Bard varied in treatment recommendations, sometimes aligning with or diverging from clinician biases. Specific vignettes showed discrepancies in diagnoses and treatments, indicating biases in both AI and human decision-making. They concluded that while AI chatbots can assist in medical decision-making, their responses are not bias-free, necessitating further research to prevent perpetuating health disparities [60].

3.1.3. Model Explainable Capacity and Performance

Understanding the explainability and performance of AI models in healthcare is crucial for their reliable and effective use. This subsection delves into the methodologies for visualizing AI model predictions and compares the triage performance of LLMs and untrained doctors.

A recent study by Lang et al. introduced a novel workflow for understanding the visual signals in medical images that AI classifiers use for predictions [61]. The approach involves several key steps:

- Train a classifier: Assess whether the medical imagery contains relevant signals for the given task.
- StylEx model: Train a StyleGAN-based[†] image generator guided by the classifier to detect and visualize critical visual attributes.
- Attribute visualization: Modify the attributes to create counterfactual visualizations.
- Expert review: Interdisciplinary experts review these visualizations to form hypotheses about the underlying mechanisms.

This method was applied across eight prediction tasks and three medical imaging modalities (retinal fundus photographs, external eye photographs, and chest radiographs). The study identified attributes capturing clinically known features and unexpected confounders, and new physiologically plausible attributes that could indicate novel research directions. These findings enhance the understanding of AI models, improve model assessments, and aid in designing better datasets. The study concludes by presenting a generative AI-based framework that generates visual explanations for AI predictions in medical imagery, aiming to improve trust in AI models, facilitate novel scientific discoveries, and provide a comprehensive understanding of the visual features influencing AI predictions [61].

While AI explainability is a highly desirable feature, it is not always possible to achieve due to technology limitations, and AI can still be made useful without it. We should remember in this context that there are many other things, treatments, and therapies we routinely use in healthcare today but do not fully understand or cannot fully explain, with reliable evidence, how exactly they work (mechanism of action), yet we continue using them without waiting for such explanations of how they work to become available before we can successfully harness them for the benefit of our patients [62].

Unlike explainability, which might not always be available, understanding the performance of a given LLM is essential and doable. For example, a recent study by Masanneck et al. compared the triage performance of LLMs, specifically ChatGPT, and untrained doctors in emergency medicine. The evaluation used 124 anonymized case vignettes, triaged according to the Manchester Triage System. The results showed that GPT-4-based ChatGPT model and untrained doctors had substantial agreement with professional raters, outperforming GPT-3.5-based ChatGPT. The study concluded that while these LLMs do not yet match professional raters, their performance is comparable to that of untrained doctors. Future improvements are anticipated with technological advancements and specific training of these models [63].

3.1.4. Limitations

[†] StyleGAN is a generative adversarial network (GAN) first introduced by NVIDIA researchers in December 2018.

 doi:10.20944/preprints202409.0311.v1

The integration of LLMs in clinical decision-making is not without its limitations. Hager et al. evaluated the performance of a number of Meta Llama 2 LLM derivatives, both generalist (Llama 2 Chat, Open Assistant (OASST), and WizardLM) and medical-domain aligned derivatives (Clinical Camel and MEDITRON—more on these in subsection 3.4). They used a curated dataset of 2,400 patient cases for their evaluation and identified several key issues [6]:

- Diagnostic accuracy: LLMs performed worse than clinicians in diagnosing diseases and frequently failed to follow established diagnostic guidelines. This limitation is critical as accurate diagnosis is the cornerstone of effective medical treatment and patient care.
- Interpretation of data: LLMs struggled with interpreting lab results, following instructions accurately, and handling varying quantities and orders of information. This issue is particularly concerning in clinical settings where precise and context-aware interpretation of data is essential.
- Autonomous decision-making: The current capabilities of LLMs are insufficient for autonomous clinical decision-making without extensive human supervision. This limitation suggests that while LLMs can assist clinicians, they are not yet ready to replace human decision-making in critical healthcare environments.
- Integration into clinical workflows: Improvements in fine-tuning and evaluation frameworks are needed to better integrate LLMs into clinical workflows. This includes developing more robust training data, improving model transparency, and ensuring that AI-generated recommendations can be easily understood and validated by human clinicians.

Ando et al. compared the quality of ChatGPT responses to anesthesia-related medical questions in English and Japanese. They found that English LLM responses were superior in quality to Japanese responses when assessed by bilingual anesthesia experts [64]. Indeed, generative AI models do not process or understand natural language the way humans do [65]. Tokenization, the process of breaking down raw text into smaller units, or tokens, to be processed by an LLM, is a major reason behind some of the strange LLM outputs and limitations observed today including their worse performance in non-English languages. Unless there is a significant breakthrough in tokenization, it appears that new model architectures will be the key to resolving this issue in the future [66].

In summary, while generative AI and LLMs hold immense potential to transform healthcare through enhanced clinical decision-making, improved diagnostic accuracy, and reduced physician burnout, significant challenges remain. Concerns about bias, variability, and ethical implications necessitate robust regulatory frameworks and comprehensive standards. Explainability and performance of AI models must be enhanced through innovative methodologies, and the limitations of current models need to be addressed through continuous improvements in training and evaluation. The safe and effective integration of these advanced technologies in healthcare requires ongoing research, collaboration among stakeholders, and stringent oversight to ensure they contribute positively to patient care and clinical practice.

3.2. Applications of Prompt Engineering in Medicine

The application of prompt engineering in medicine involves optimizing the input prompts given to generative AI models to enhance their performance, reliability, and utility in clinical settings. Prompt engineering is a critical component in ensuring that AI systems like LLMs generate accurate, relevant, and contextually appropriate responses, particularly in sensitive areas such as healthcare.

A Forbes article [67] discussed the potential risks doctors face when relying on generative AI to summarize medical notes. While AI can improve efficiency by creating concise summaries of patient records, it may also lead to critical information being misinterpreted or omitted, jeopardizing patient care. The article emphasizes the importance of prompt engineering, ongoing research to establish standards, and ensuring that AI-generated summaries are always reviewed by qualified healthcare professionals to maintain accuracy and reliability (cf. Rumale et al. [32]).

Effective prompt engineering is crucial for optimizing the performance of generative AI models. Goodman et al. highlighted the importance of designing and refining input prompts to significantly enhance the usability and reliability of AI systems in healthcare settings. For example, specific

 doi:10.20944/preprints202409.0311.v1

prompt strategies can help guide AI models to produce more accurate and relevant responses, which is vital for clinical decision-making [35].

A recent study by Patel et al. assessed how different prompt engineering techniques influence GPT-3.5's ability to answer medical questions. The study compared direct prompts, Chain of Thought (CoT), and modified CoT approaches using 1,000 questions generated by GPT-4 and 95 real USMLE Step 1 questions. The analysis revealed no significant differences in accuracy among the prompt types, with success rates of approximately 61.7% for direct prompts, 62.8% for CoT, and 57.4% for modified CoT on USMLE questions. This finding suggested that while prompt engineering techniques like CoT are designed to enhance reasoning, they do not significantly impact the model's performance on medical calculations or clinical scenarios. Consequently, simpler prompting methods can be as effective as more complex ones, potentially simplifying the integration of AI tools like ChatGPT into medical education and clinical practice [68].

3.3. Applications of RAG in Medicine

The use of Retrieval-Augmented Generation (RAG) in medicine enhances the capabilities of AI models by integrating them with domain-specific knowledge bases and user-provided content. This section explores notable applications of RAG in medical AI, demonstrating how it can improve diagnostic accuracy, educational utility, and overall user experience.

Aeyeconsult [69] is an ophthalmology chatbot developed using GPT-4, LangChain, and Pinecone. Ophthalmology textbooks were processed into embeddings and stored in Pinecone, allowing user queries to be converted, compared to stored embeddings, and answered by GPT-4. This method ensures that responses are accurate and based on verified sources. Aeyeconsult provided citations to the sources used to answer user queries, enhancing transparency and trust. In a comparative study, Aeyeconsult outperformed ChatGPT-4 on the OKAP (Ophthalmic Knowledge Assessment Program) dataset, achieving 83.4% correct answers compared to 69.2%. Aeyeconsult also had fewer instances of no answer and multiple answers. Both systems performed best in General Medicine, with Aeyeconsult achieving 96.2% accuracy. Although Aeyeconsult's weakest performance was in Clinical Optics at 68.1%, it still outperformed ChatGPT-4 in this category (45.5%). These results highlighted the effectiveness of integrating RAG approaches to enhance the performance of medical chatbots.

NVIDIA's 'Chat with RTX' is a free tool for building a custom LLM using a RAG approach. A recent paper described using it to enhance the educational utility of electronic dermatology textbooks. The resultant custom dermatology AI chatbot can improve the self-study potential of electronic textbooks by going beyond standard search and indexing functions. The chatbot is cloud independent (i.e., runs fully locally on the user's machine), making it affordable and easy to set up without requiring high technical expertise whilst protecting copyrighted and other sensitive data (by not having to upload them to remote servers) and offering more flexibility to users, by allowing them to run the software in places and situations where there is no Internet connection. The demonstrated custom AI chatbot can answer dermatology-related queries, generate quizzes, and cite sources from the imported content, enhancing the learning experience. This approach demonstrates a promising way to maximize the educational value of digital reference materials in dermatology (and other medical disciplines), offering a more interactive and intelligent self-study tool. By integrating user-provided content and leveraging RAG technology, the custom dermatology AI chatbot built with 'Chat with RTX' carries the potential of augmenting the accessibility and effectiveness of dermatology education [70].

Other notable RAG-enabled clinical examples include LiVersa, a liver disease-specific LLM [71] and a recent study by Ye that explored a learning-to-rank approach to enhance RAG-based electronic medical record (EMR) search engines by tailoring search results to users' specific search semantics. Ye's reported method involves user labeling of relevant documents, refining word similarities using medical semantic embeddings, and re-ranking new documents based on identified relevant sentences. The system demonstrated superior performance compared to baseline methods, achieving higher Precision-at-10 (P@10) scores with minimal labeled data. The approach improved the accuracy

 doi:10.20944/preprints202409.0311.v1

of RAG models in extracting and explaining complex medical information, such as cancer progression diagnoses, highlighting the potential of user-tailored learning-to-rank methods to support clinical practice and improve the reliability of AI-generated medical insights [72].

While RAG technology can reduce a model's hallucinations by grounding the generated content in retrieved, verified data, it is not a complete solution to the hallucination problem [73]. RAG mitigates but does not entirely prevent AI hallucinations. This limitation arises because the generative component can still synthesize incorrect or misleading content, even when the retrieval mechanism supplies accurate information. Therefore, while RAG enhances the reliability of LLMs, continuous improvements in both retrieval methods and generative accuracy are necessary for these tools to become fully reliable in clinical applications.

3.4. Medicine-Specific LLMs

The development of medicine-specific LLMs has significantly advanced the field of medical AI, offering specialized capabilities for clinical and personal health applications. This subsection reviews several notable models, their development, performance, and potential implications for healthcare.

Med-PaLM [74], introduced by Google Research in December 2022, laid the groundwork for specialized medical LLMs. Following this, Med-PaLM 2 [59] was also developed by Google Research, with the model published in March 2023. Building on its predecessor, Med-PaLM 2 incorporated medical domain-specific fine-tuning and a novel ensemble refinement prompting strategy to enhance its medical reasoning capabilities. It achieved state-of-the-art results on several medical question-answering benchmarks, including a significant improvement over its predecessor. It scored 86.5% on the MedQA dataset (a multiple-choice question answering based on the United States Medical Licensing Examination (USMLE)), outperforming previous models by a notable margin. Human evaluations indicated that Med-PaLM 2's answers were preferred over physician-generated answers on multiple axes relevant to clinical utility, such as factuality, reasoning, and low likelihood of harm. The model also showed substantial improvements in answering complex medical questions and demonstrated better performance and lower risk of harm compared to its predecessor when tested on adversarial datasets. Despite these findings, the study cautioned that further validation and safety studies are necessary before applying the model to real-world settings [59].

Published in November 2023, MEDITRON-70B [75] is an open-source suite of LLMs tailored for the medical domain, available in two versions with 7 billion and 70 billion parameters. Built on Meta's Llama-2, these models are pretrained on a comprehensive medical corpus, including PubMed articles and medical guidelines. MEDITRON-70B outperforms several state-of-the-art baselines, including GPT-3.5 and Med-PaLM, showing significant gains in medical reasoning tasks. However, despite its strong performance, its creators, Chen et al. [75], advised against using MEDITRON in clinical applications without further testing due to safety concerns.

In April 2024, Google DeepMind introduced the Med-Gemini family of models [76], offering highly capable multimodal models specialized in medicine. Med-Gemini models achieve state-of-the-art performance on 10 out of 14 medical benchmarks, including a 91.1% accuracy on the MedQA (USMLE) benchmark, surpassing previous models like Med-PaLM 2. These models excel in handling complex medical tasks involving text, images, and long-context data, outperforming GPT-4V in several benchmarks. Med-Gemini demonstrates practical applications in generating medical summaries, referral letters, and engaging in multimodal medical dialogues. The models have shown potential to assist in clinical decision-making, medical education, and research, although further rigorous evaluation is needed before deployment in real-world clinical settings.

Introduced in June 2024, Google's Personal Health Large Language Model (PH-LLM) [77] is a version of Gemini fine-tuned for text understanding and reasoning over numerical time-series personal health data, particularly focusing on sleep and fitness applications. This model integrates mobile and wearable device data, often neglected in clinical settings, for continuous and longitudinal personal health monitoring. PH-LLM was evaluated on three benchmark datasets for personalized insights and recommendations, expert domain knowledge, and prediction of self-reported sleep quality outcomes. Working with domain experts, its creators developed 857 case studies in sleep and

 doi:10.20944/preprints202409.0311.v1

fitness to demonstrate model's capabilities. The model performed comparably to human experts in fitness and showed significant improvements in sleep insights after fine-tuning, achieving 79% on sleep medicine exams and 88% on fitness exams, surpassing human expert average scores. This underscores the potential of integrating wearable device data into AI models for personalized health recommendations and monitoring, although further development and evaluation are necessary for application in safety-critical personal health domains.

Around the same time in June 2024, the Personal Health Insights Agent (PHIA) was introduced, also by Google researchers [78]. PHIA is an agent system leveraging LLM capabilities for analyzing and interpreting wearable health data to generate personalized health insights. Utilizing iterative reasoning, code generation, and web search, PHIA addresses personal health queries effectively. This model was evaluated using two benchmark datasets comprising over 4,000 health insights questions for objective and open-ended evaluation. Human and expert evaluations demonstrated PHIA's accuracy in addressing factual and open-ended health queries, outperforming standard LLM baselines. It accurately addressed over 84% of factual numerical questions and more than 83% of crowdsourced open-ended questions, providing personalized health insights that can serve as an aid to individuals in interpreting their wearable data and potentially improving their health behaviors. The results highlight the potential for LLM agents to advance behavioral health, making personalized wellness regimens more accessible.

In a randomized study assessing the differential diagnosis (DDx) accuracy for 302 NEJM (New England Journal of Medicine) Clinicopathological Conference (CPC) series case reports, a specialized LLM (Google Med-PaLM 2) was compared with 20 physicians (US board-certified internists with a median of 9 years of experience). Each case report was assessed by two clinicians randomized to one of two assistive arms: search engines and standard medical resources, or LLM assistance in addition to these tools. Clinicians in both groups also provided a baseline, unassisted DDx prior to using their assigned tools. The study revealed that Med-PaLM 2 performed better on its own than unassisted clinicians. Additionally, when comparing the two groups of assisted clinicians, those supported by Med-PaLM 2 achieved a higher DDx quality score, highlighting the growing potential of LLMs in supporting complex medical decision-making tasks [79].

4. Discussion

The development of medicine-specific Large Language Models (LLMs) such as Med-PaLM, MEDITRON-70B, Med-Gemini, PH-LLM, and PHIA showcases the rapid advancements and potential of these models to enhance healthcare delivery and personalized health monitoring. These models demonstrate significant improvements in medical reasoning, diagnostic accuracy, and practical applications, offering new possibilities for clinical decision support, patient communication, and educational tools.

The integration of Retrieval-Augmented Generation (RAG) and prompt engineering in medical applications, as demonstrated by tools like Aeyeconsult [69] and the custom dermatology AI chatbot built using NVIDIA's 'Chat with RTX' [70], further enhances the capabilities of AI models by combining generative AI with domain-specific knowledge bases. These advancements improve diagnostic accuracy, educational utility, and user experience by providing precise, contextually relevant answers and interactive learning tools.

However, despite these promising developments, significant challenges remain. The risks associated with bias, hallucinations, and ethical concerns necessitate ongoing research and the establishment of robust safety protocols to ensure the responsible use of AI in medicine. Additionally, the regulatory landscape must evolve to keep pace with these advancements, ensuring that AI tools meet the high standards required for clinical use while safeguarding patient safety and privacy.

To sum up, while generative AI and LLMs hold immense potential to transform healthcare, their successful integration into clinical practice and personal health management will require continuous collaboration among stakeholders, including regulatory bodies, patients/patient advocates, healthcare professionals, industry representatives, the academia, government partners, and international organizations (e.g., IMDRF, the International Medical Device Regulators Forum, and

 doi:10.20944/preprints202409.0311.v1

ICH, the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use) [80]. Success will also be dependent on rigorous testing and a commitment to addressing the ethical and regulatory challenges that lie ahead. The future of AI in medicine is bright, but will require careful stewardship to realize its full potential in improving patient care and healthcare outcomes.

It should be noted that, while this paper attempted to cover as much topic breadth (bird's-eye view) as possible, as an incremental update to our 2023 review on the same subject [11], there are additional innovative applications of generative AI and LLM technology in healthcare that we were not able to discuss without making this paper unduly long. Among these, a few notable applications stand out that deserve mentioning, albeit very briefly, before concluding this review. They include Huma's cloud generative AI builder, which automates or semi-automates the generation and coding of new healthcare applications from text inputs, demonstrating the potential of AI to streamline app development processes [81]. Another application worth noting is the integration of Brain-Computer Interface (BCI) technology with ChatGPT by Synchron, which offers a glimpse into the future of assistive technologies, where generative AI could help users with neurological impairments communicate and control devices more effectively **[82–84]**. Finally, generative AI is also being used in the development of health digital twins [85], as well as in drug discovery [86] and drug repurposing [87].

5. Conclusion

The applications of generative AI in medicine and healthcare have come a long way since the initial public launch in November 2022 of OpenAI's ChatGPT, a general-purpose LLM. ChatGPT and other general-purpose LLMs have greatly and rapidly improved afterwards, and continue to do so, with OpenAI said to be working on a new approach and LLM (codenamed 'Strawberry' and 'Orion' respectively [88]) that will have the ability to do more complex reasoning and context-sensitive solving instead of just mere pattern recognition and word prediction. Furthermore, the medical and healthcare arena witnessed the release of specific-purpose and narrow-focus (medically-trained) LLMs (such as Google Med-PaLM), and the adoption of promising methods such as retrieval-augmented generation and more robust applications involving clusters of multiple specialized LLMs working together (as seen in Hippocratic AI [89]), all of which are intended to improve the efficacy and reliability of generative AI applications in medicine.

As we move beyond the initial hype surrounding generative AI in medicine and healthcare (the 'peak of inflated expectations'), we realize there is still a long way to go to properly regulate these applications and better mitigate or overcome their current limitations in order to make them ready for safe mainstream clinical use and ultimately reach their 'plateau of productivity'.

To supplement this article, given the fast pace of developments in this area, we have made publicly available a Web page at [90] featuring a regularly updated archive of Web pointers to handpicked news, posts and articles about generative AI in medicine and healthcare.

Authors' Contributions: M.N.K.B. conceived the manuscript idea, set its scope and direction, conducted its core literature review, wrote and edited its draft and final versions, and invited P.Z. and J.S. to contribute as co-authors. P.Z and J.S. contributed equally to additional parts of the literature review and to the writing of the initial draft versions of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gartner Research. Hype Cycle for Generative AI, 2024. 31 July 2024. Available online: https://www.gartner.com/en/documents/5636791 (accessed on 30 August 2024).

2. Holohan, M. Mom ChatGPT diagnosis pain. Today, 11 September 2023. Available online: https://www.today.com/health/mom-chatgpt-diagnosis-pain-rcna101843 (accessed on 30 August 2024).
3. Google Cloud. Sharing Google Med-PaLM 2: Medical Large Language Model. Available online: https://cloud.google.com/blog/topics/healthcare-life-sciences/sharing-google-med-palm-2-medical-large-language-model (accessed on 30 August 2024).
4. Gottlieb, S.; Benezra, S. Op-ed: How well can AI chatbots mimic doctors in a treatment setting? CNBC, Published: 18 July 2024. Available online: https://www.cnbc.com/2024/07/18/op-ed-how-well-can-ai-chatbots-mimic-doctors.html (accessed on 30 August 2024).
5. Kim, W. No, you cannot gauge large language models (LLMs) "for their medical proficiency" using multiple-choice questions alone. LinkedIn Commentary, 2024. Available online: https://www.linkedin.com/posts/woojinkim_genai-chatgpt-gpt4-activity-7225200801898487809-QRxW (accessed on 30 August 2024).
6. Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; Rueckert, D. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **2024**. https://doi.org/10.1038/s41591-024-03097-1.
7. Alba, D.; Swetlitz, I. Google Taps AI to Revamp Costly Health-Care Push Marred by Flops. BNN Bloomberg, 2024. Available online: https://www.bnnbloomberg.ca/business/technology/2024/07/30/google-taps-ai-to-revamp-costly-health-care-push-marred-by-flops/ (accessed on 30 August 2024).
8. Aljamaan, F.; Temsah, M.H.; Altamimi, I.; Al-Eyadhy, A.; Jamal, A.; Alhasan, K.; Mesallam, T.A.; Farahat, M.; Malki, K.H. Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study. *JMIR Med. Inform.* **2024**, *12*, e54345. https://doi.org/10.2196/54345.
9. Dudley-Nicholson, J. Doctors Call for AI Rules to Prevent Medical Mistakes. The Standard (Australia), 17 July 2024. Available online: https://www.standard.net.au/story/8698797/doctors-call-for-ai-rules-to-prevent-medical-mistakes/ (accessed on 30 August 2024).
10. Meskó, B.; Topol, E.J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit. Med.* **2023**, *6*, 120. https://doi.org/10.1038/s41746-023-00873-0.
11. Zhang, P.; Kamel Boulos, M.N. Generative AI in medicine and healthcare: promises, opportunities and challenges. *Future Internet* **2023**, *15*(9), 286. https://doi.org/10.3390/fi15090286.
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Advances in Neural Information Processing Systems; Curran Associates, Inc.: 2017; pp. 5998-6008.
13. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
14. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. 2019. Available online: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
15. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. 2020. Available online: https://arxiv.org/abs/2005.14165
16. OpenAI. GPT-4 Technical Report. 2023. Available online: https://cdn.openai.com/papers/gpt-4.pdf
17. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
18. Hello GPT-4o. Available online: https://openai.com/index/hello-gpt-4o (accessed on 30 August 2024).
19. Topsakal, O.; Akinci, T.C. Creating large language model applications utilizing langchain: A primer on developing LLM apps fast. In *International Conference on Applied Engineering and Natural Sciences*; 2023; Volume 1, Issue 1, pp. 1050–1056.
20. Meet Claude [Internet]. Available online: https://www.anthropic.com/claude (accessed on 30 August 2024).
21. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; Casas, D.D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L.R. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825.
22. Google AI Test Kitchen. Image FX. Available online: https://aitestkitchen.withgoogle.com/tools/image-fx(accessed on 30 August 2024).
23. DALL-E 3. Available online: https://openai.com/index/dall-e-3/ (accessed on 30 August 2024).
24. Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.C.; Liu, Z.; Wang, L. The dawn of LMMs: Preliminary explorations with GPT-4V (ision). *arXiv* **2023**, *9*(1), 1. arXiv:2309.17421.
25. Meskó, B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J. Med. Internet Res.***2023**, *25*, e50638.
26. Lewis, P.; Perez, E.; Kiela, D.; Cho, K.; Stenetorp, P. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2020. Available online: https://arxiv.org/abs/2005.11401

27. Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; Weston, J. Retrieval augmentation reduces hallucination in conversation. *arXiv* **2021**, arXiv:2104.07567.
28. Farquhar, S.; Kossen, J.; Kuhn, L.; Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **2024**, *630*, 625–630. https://doi.org/10.1038/s41586-024-07421-0.
29. Zewe, A. Method prevents an AI model from being overconfident about wrong answers. *MIT News*, Massachusetts Institute of Technology, 2024. Available online: https://news.mit.edu/2024/thermometer-prevents-ai-model-overconfidence-about-wrong-answers-0731 (accessed on 30 August 2024).
30. Williamson, S.M.; Prybutok, V. The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation. *Information* **2024**, *15*(6), 299.
31. Burford, K.G.; Itzkowitz, N.G.; Ortega, A.G.; Teitler, J.O.; Rundle, A.G. Use of Generative AI to Identify Helmet Status Among Patients With Micromobility-Related Injuries From Unstructured Clinical Notes. *JAMA Netw Open* **2024** Aug 1;7(8):e2425981. doi: 10.1001/jamanetworkopen.2024.25981.
32. Rumale Vishwanath, P.; Tiwari, S.; Naik, T.G.; Gupta, S.; Thai, D.N.; Zhao, W.; Kwon, S.; Ardulov, V.; Tarabishy. K.; McCallum, A.; Salloum, W. Faithfulness Hallucination Detection in Healthcare AI. In Proceedings of KDD-AIDSH 2024, August 26, 2024, Barcelona, Spain. Available online: https://openreview.net/pdf?id=6eMIzKFOpJ
33. McDonald, K. RACGP issues guidance on AI scribes in practice. Pulse IT, 2024, August 20. Available online: https://www.pulseit.news/australian-digital-health/racgp-issues-guidance-on-ai-scribes-in-practice/?goal=0_b39f06f53f-9a4da8fc00-413088949 (accessed on 30 August 2024).
34. Australian Health Practitioner Regulation Agency (Ahpra) and National Boards. Meeting your professional obligations when using Artificial Intelligence in healthcare. Available online: https://www.ahpra.gov.au/Resources/Artificial-Intelligence-in-healthcare.aspx (accessed on 30 August 2024).
35. Goodman, K.E.; Paul, H.Y.; Morgan, D.J. AI-Generated Clinical Summaries Require More Than Accuracy. *JAMA* **2024**, February 27.
36. Bharel, M.; Auerbach, J.; Nguyen, V.; DeSalvo, K.B. Transforming Public Health Practice With Generative Artificial Intelligence: Article examines how generative artificial intelligence could be used to transform public health practice in the US. *Health Aff.* **2024**, *43*(6), 776–782.
37. UK MHRA. Software and Artificial Intelligence (AI) as a Medical Device. Guidance, Updated 13 June 2024. Available online: https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device (accessed on 30 August 2024).
38. Matias, Y.; Gupta, A. MedLM: generative AI fine-tuned for the healthcare industry. Google Cloud Blog, 2023, December 13. Available online: https://cloud.google.com/blog/topics/healthcare-life-sciences/introducing-medlm-for-the-healthcare-industry (accessed on 30 August 2024).
39. Harvey, H.; Pogose, M. How to get ChatGPT regulatory approved as a medical device. Hardian Health, 2024. Available online: https://www.hardianhealth.com/insights/how-to-get-regulatory-approval-for-medical-large-language-models (accessed on 30 August 2024).
40. Blumenthal, D.; Patel, B. The Regulation of Clinical Artificial Intelligence. *NEJM AI* **2024**, *1*(8), AIpc2400545. https://doi.org/10.1056/AIpc2400545.
41. Howell, M.D.; Corrado, G.S.; DeSalvo, K.B. Three Epochs of Artificial Intelligence in Health Care. *JAMA* **2024**, *331*(3), 242–244. https://doi.org/10.1001/jama.2023.25057.
42. US FDA. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. Available online: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (accessed on 30 August 2024).
43. Amazon Web Services (AWS). Generative AI for Healthcare (White Paper). 2024 June. Available online: https://pages.awscloud.com/rs/112-TZM-766/images/AWS-GenAI-for-HCLS-Whitepaper_062024.pdf (accessed on 30 August 2024).
44. Derraz, B.; Breda, G.; Kaempf, C.; Baenke, F.; Cotte, F.; Reiche, K.; Köhl, U.; Kather, J.N.; Eskenazy, D.; Gilbert, S. New regulatory thinking is needed for AI-based personalised drug and cell therapies in precision oncology. *NPJ Precis Oncol* **2024** Jan 30;8(1):23. doi: 10.1038/s41698-024-00517-w.
45. Freyer, O.; Wiest, I.C.; Kather, J.N.; Gilbert, S. A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digit Health* **2024** Sep;6(9):e662-e672. doi: 10.1016/S2589-7500(24)00124-9.
46. Gutiérrez, J.D. Consultation paper on AI regulation: emerging approaches across the world. UNESCO: Paris, France, 2024 August. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000390979 (accessed on 30 August 2024).
47. Schmidt, J.; Schutte, N.M.; Buttigieg, S.; Novillo-Ortiz, D.; Sutherland, E.; Anderson, M.; de Witte, B.; Peolsson, M.; Unim, B.; Pavlova, M.; Stern, A.D.; Mossialos, E.; van Kessel, R. Mapping the regulatory landscape for artificial intelligence in health within the European Union. *NPJ Digit Med* **2024** Aug 27;7(1):229. doi: 10.1038/s41746-024-01221-6.

48. European Commission. European Artificial Intelligence Act comes into force. Press Release, 1 August 2024. Available online: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_4123 (accessed on 12 August 2024).
49. van Rooijen SB. The EU AI Act's Impact on Medical Devices and MDR Certification. LinkedIn post, July 2024. https://www.linkedin.com/posts/sigridberge_eu-ai-act-impact-on-medical-devices-activity-7222488338342006784-T45R/ (accessed on 30 August 2024).
50. van Rooijen SB. EU AI Act - Healthcare. LinkedIn post, August 2024. https://www.linkedin.com/posts/sigridberge_eu-ai-act-has-come-into-effect-what-does-activity-7224662682841325570-6Zx3/ (accessed on 30 August 2024).
51. Wachter, S. Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond. *Yale Journal of Law & Technology* **2024**;26(3):671-718. https://yjolt.org/limitations-and-loopholes-eu-ai-act-and-ai-liability-directives-what-means-european-union-united.
52. Lenharo, M. The testing of AI in medicine is a mess. Here's how it should be done. *Nature* **2024** Aug;632(8026):722-724. doi: 10.1038/d41586-024-02675-0.
53. Coiera, E.; Fraile-Navarro, D. AI as an Ecosystem — Ensuring Generative AI Is Safe and Effective. *NEJM AI* **2024**;1(9). doi: 10.1056/AIp2400611.
54. Ong, J.C.; Chang, S.Y.; William, W.; Butte, A.J.; Shah, N.H.; Chew, L.S.; Liu, N.; Doshi-Velez, F.; Lu, W.; Savulescu, J.; Ting, D.S. Medical Ethics of Large Language Models in Medicine. *NEJM AI* **2024**, AIra2400038. https://doi.org/10.1056/AIra2400038.
55. Ong, J.C.; Chang, S.Y.; William, W.; Butte, A.J.; Shah, N.H.; Chew, L.S.; Liu, N.; Doshi-Velez, F.; Lu, W.; Savulescu, J.; Ting, D.S. Ethical and Regulatory Challenges of Large Language Models in Medicine. *Lancet Digit. Health* **2024**, *6*(6), e428–e432. https://doi.org/10.1016/S2589-7500(24)00061-X.
56. Haltaufderheide, J.; Ranisch, R. The Ethics of ChatGPT in Medicine and Healthcare: A Systematic Review on Large Language Models (LLMs). *NPJ Digit. Med.* **2024**, *7*(1), 183. https://doi.org/10.1038/s41746-024-01157-x.
57. Levkovich, I.; Elyoseph, Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. *Fam. Med. Community Health* **2023**, *11*(4), e002391. https://doi.org/10.1136/fmch-2023-002391.
58. Omiye, J.A.; Lester, J.C.; Spichak, S.; et al. Large language models propagate race-based medicine. *NPJ Digit. Med.***2023**, *6*, 195. https://doi.org/10.1038/s41746-023-00939-z.
59. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; Schaekermann, M. Towards expert-level medical question answering with large language models. *arXiv* **2023**, arXiv:2305.09617.
60. Kim, J.; Cai, Z.R.; Chen, M.L.; Simard, J.F.; Linos, E. Assessing Biases in Medical Decisions via Clinician and AI Chatbot Responses to Patient Vignettes. *JAMA Netw. Open* **2023**, *6*(10), e2338050. https://doi.org/10.1001/jamanetworkopen.2023.38050.
61. Lang, O.; Yaya-Stupp, D.; Traynis, I.; Cole-Lewis, H.; Bennett, C.R.; Lyles, C.R.; Lau, C.; Irani, M.; Semturs, C.; Webster, D.R.; Corrado, G.S. Using generative AI to investigate medical imagery models and datasets. *EBioMedicine* **2024**, *102*. https://doi.org/10.1016/j.ebiom.2024.105075.
62. Painter A, et al. Explaining Explainable AI (for healthcare). YouTube, 2024, August 15. Available online: https://www.youtube.com/watch?v=d5ZMVIgQ0jM (accessed on 30 August 2024).
63. Masanneck, L.; Schmidt, L.; Seifert, A.; Kölsche, T.; Huntemann, N.; Jansen, R.; Mehsin, M.; Bernhard, M.; Meuth, S.; Böhm, L.; Pawlitzki, M. Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study. *J. Med. Internet Res.* **2024**, *26*, e53297. https://doi.org/10.2196/53297.
64. Ando, K.; Sato, M.; Wakatsuki, S.; Nagai, R.; Chino, K.; Kai, H.; Sasaki, T.; Kato, R.; Nguyen, T.P.; Guo, N.; Sultan, P. A comparative study of English and Japanese ChatGPT responses to anaesthesia-related medical questions. *BJA Open* **2024** Jun 14;10:100296. doi: 10.1016/j.bjao.2024.100296.
65. Greefhorst, A. The 'Artificial Stubbornness' of ChatGPT when Solving a Simple Puzzle: The farmer with his wolf, goat, and cabbage. International Policy Digest, 2024, May 28. Available online: https://intpolicydigest.org/the-artificial-stubbornness-of-chatgpt-when-solving-a-simple-puzzle/ (accessed on 12 August 2024).
66. Wiggers, K. Tokens are a big reason today's generative AI falls short. TechCrunch, 2024, July 6. Available online: https://techcrunch.com/2024/07/06/tokens-are-a-big-reason-todays-generative-ai-falls-short/ (accessed on 9 July 2024).
67. Eliot, L. Doctors relying on generative AI to summarize medical notes might unknowingly be taking big risks. *Forbes*, 2024, February 5. Available online: https://www.forbes.com/sites/lanceeliot/2024/02/05/doctors-relying-on-generative-ai-to-summarize-medical-notes-might-unknowingly-be-taking-big-risks/ (accessed on 30 August 2024).

68. Patel, D.; Raut, G.; Zimlichman, E.; et al. Evaluating prompt engineering on GPT-3.5's performance in USMLE-style medical calculations and clinical scenarios generated by GPT-4. *Sci. Rep.* **2024**, *14*, 17341. https://doi.org/10.1038/s41598-024-66933-x.
69. Singer, M.B.; Fu, J.J.; Chow, J.; Teng, C.C. Development and evaluation of Aeyeconsult: a novel ophthalmology chatbot leveraging verified textbook knowledge and GPT-4. *J. Surg. Educ.* **2024**, *81*(3), 438–443. https://doi.org/10.1016/j.jsurg.2023.11.019.
70. Kamel Boulos, M.N.; Dellavalle, R. NVIDIA's 'Chat with RTX' custom Large Language Model and Personalized AI Chatbot Augments the Value of Electronic Dermatology Reference Material. *JMIR Dermatol.* **2024**, *7*, e58396. https://doi.org/10.2196/58396.
71. Ge, J.; Sun, S.; Owens, J.; Galvez, V.; Gologorskaya, O.; Lai, J.; Pletcher, M.; Lai, K. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. Hepatology. 2024 Mar 07; doi: 10.1097/HEP.0000000000000834.01515467-990000000-00791.
72. Ye, C. Exploring a learning-to-rank approach to enhance the Retrieval Augmented Generation (RAG)-based electronic medical records search engines. *Informatics and Health* **2024**, *1*(2), 93–99. https://doi.org/10.1016/j.infoh.2024.07.001.
73. Wiggers, K. Why RAG won't solve Generative AI's hallucination problem [Internet]. TechCrunch, 2024, May 4. Available online: https://techcrunch.com/2024/05/04/why-rag-wont-solve-generative-ais-hallucination-problem/ (accessed on 30 August 2024).
74. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; Payne, P. Large language models encode clinical knowledge. *Nature* **2023**, *620*(7972), 172–180.
75. Chen, Z.; Cano, A.H.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mohtashami, A.; Sallinen, A. Meditron-70b: Scaling medical pretraining for large language models. *arXiv* **2023**, arXiv:2311.16079.
76. Saab, K.; Tu, T.; Weng, W.H.; Tanno, R.; Stutz, D.; Wulczyn, E.; Zhang, F.; Strother, T.; Park, C.; Vedadi, E.; Chaves, J.Z. Capabilities of Gemini models in medicine. *arXiv* **2024**, arXiv:2404.18416.
77. Cosentino, J.; Belyaeva, A.; Liu, X.; Furlotte, N.A.; Yang, Z.; Lee, C.; Schenck, E.; Patel, Y.; Cui, J.; Schneider, L.D.; Bryant, R. Towards a Personal Health Large Language Model. *arXiv* **2024**, arXiv:2406.06474.
78. Merrill, M.A.; Paruchuri, A.; Rezaei, N.; Kovacs, G.; Perez, J.; Liu, Y.; Schenck, E.; Hammerquist, N.; Sunshine, J.; Tailor, S.; Ayush, K. Transforming wearable data into health insights using large language model agents. *arXiv* **2024**, arXiv:2406.06464.
79. McDuff, D.; Schaekermann, M.; Tu, T.; Palepu, A.; Wang, A.; Garrison, J.; Singhal, K.; Sharma, Y.; Azizi, S.; Kulkarni, K.; Hou, L. Towards accurate differential diagnosis with large language models. *arXiv* **2023**, arXiv:2312.00164.
80. World Health Organization. Regulatory considerations on artificial intelligence for health. World Health Organization, 2023 October. ISBN 978-92-4-007887-1 (electronic version). https://iris.who.int/handle/10665/373421.
81. Sharma, S. Huma raises $80M to turn text into healthcare apps with gen AI. 2024. Available online: https://venturebeat.com/ai/huma-raises-80m-to-turn-text-into-healthcare-apps-with-gen-ai/ (accessed on 30 August 2024).
82. Synchron announces brain computer interface chat feature powered by OpenAI. 2024. Available online: https://www.businesswire.com/news/home/20240711493318/en/Synchron-Announces-Brain-Computer-Interface-Chat-Feature-Powered-by-OpenAI (accessed on 31 August 2024).
83. Orrall, J. How this brain implant is using ChatGPT. 2024. Available online: https://www.cnet.com/tech/computing/how-this-brain-implant-is-using-chatgpt/ (accessed on 31 August 2024).
84. What it's like using a brain implant with ChatGPT - Video. CNET, 2024. Available online: https://www.cnet.com/videos/what-its-like-using-a-brain-implant-with-chatgpt/ (accessed on 31 August 2024).
85. Makarov, N.; Bordukova, M.; Rodriguez-Esteban, R.; Schmich, F.; Menden, M.P. Large Language Models forecast Patient Health Trajectories enabling Digital Twins. medRxiv 2024.07.05.24309957; doi: 10.1101/2024.07.05.24309957.
86. Gangwal, A., Lavecchia, A. Unleashing the power of generative AI in drug discovery. *Drug Discov Today* **2024** Jun;29(6):103992. doi: 10.1016/j.drudis.2024.103992.
87. Ghandikota, S.K.; Jegga, A.G. Application of artificial intelligence and machine learning in drug repurposing. Prog Mol *Biol Transl Sci* **2024**;205:171-211. doi: 10.1016/bs.pmbts.2024.03.030.
88. Caswell A. OpenAI to launch new advanced "Strawberry" AI product this fall — what we know so far. Tom's Guide, 2024, August 28. Available online: https://www.tomsguide.com/ai/openai-to-launch-new-advanced-strawberry-ai-product-this-fall-what-we-know-so-far (accessed on 30 August 2024).
89. Hippocratic AI. Foundation Model. Available online: https://www.hippocraticai.com/foundationmodel (accessed on 30 August 2024).

 doi:10.20944/preprints202409.0311.v1

90. Kamel Boulos MN. Generative AI in medicine and health/care: handpicked news, posts and articles from around the Web. Available online: https://healthcybermap.org/HDTs/genai-med.html (accessed on 30 August 2024).