Review

Large Language Models in Healthcare and Medical Domain: A Review

Zabir Al Nazi [1] and Wei Peng [2,*]

1 Department of Computer Science and Engineering, University of California Riverside, Riverside, CA 92521, USA; znazi002@ucr.edu
2 Department of Psychiatry and Behavioral Sciences, Stanford University, 1070 Arastradero Road, Palo Alto, CA 94303, USA
* Correspondence: wepeng@stanford.edu

Abstract: The deployment of large language models (LLMs) within the healthcare sector has sparked both enthusiasm and apprehension. These models exhibit the remarkable ability to provide proficient responses to free-text queries, demonstrating a nuanced understanding of professional medical knowledge. This comprehensive survey delves into the functionalities of existing LLMs designed for healthcare applications and elucidates the trajectory of their development, starting with traditional Pretrained Language Models (PLMs) and then moving to the present state of LLMs in the healthcare sector. First, we explore the potential of LLMs to amplify the efficiency and effectiveness of diverse healthcare applications, particularly focusing on clinical language understanding tasks. These tasks encompass a wide spectrum, ranging from named entity recognition and relation extraction to natural language inference, multimodal medical applications, document classification, and question-answering. Additionally, we conduct an extensive comparison of the most recent state-of-the-art LLMs in the healthcare domain, while also assessing the utilization of various open-source LLMs and highlighting their significance in healthcare applications. Furthermore, we present the essential performance metrics employed to evaluate LLMs in the biomedical domain, shedding light on their effectiveness and limitations. Finally, we summarize the prominent challenges and constraints faced by large language models in the healthcare sector by offering a holistic perspective on their potential benefits and shortcomings. This review provides a comprehensive exploration of the current landscape of LLMs in healthcare, addressing their role in transforming medical applications and the areas that warrant further research and development.

Keywords: large language model; healthcare; medicine; natural language generation; natural language processing; machine learning applications; ChatGPT; generative AI; medical AI

1. Introduction

Deep learning provides an intelligent way to understand human behaviors, emotions and human healthcare [1–4]. Recent developments in clinical language understanding have ushered in the potential for a paradigm shift in the healthcare sector. These advancements hold the promise of ushering in a new era characterized by the deployment of intelligent systems designed to bolster decision-making, expedite diagnostic processes, and elevate the quality of patient care. In essence, these systems have the capacity to serve as indispensable aids to healthcare professionals as they grapple with the ever-expanding body of medical knowledge, decipher intricate patient records, and formulate highly tailored treatment plans. This transformative potential has ignited considerable enthusiasm within the healthcare community [5–7].

The immense value of large language models (LLMs) lies in their ability to process and synthesize colossal volumes of medical literature, patient records, and the ever-expanding body of clinical research. Healthcare data [8,9] are inherently complex, heterogeneous,

and often overwhelming in scale. LLMs act as a powerful force multiplier, aiding healthcare professionals struggling with information overload. By automating the analysis of medical texts, extracting crucial insights, and applying that knowledge, LLMs are poised to drive groundbreaking research and enhance patient care, significantly improving and contributing to the progression of the healthcare and medical domain.

Notably, this surge of enthusiasm is attributable in part to the exceptional performance of state-of-the-art large language models (LLMs) such as OpenAI's GPT-3.5 and GPT-4 [10,11], and Google's Bard. These models have exhibited remarkable proficiency in a wide spectrum of natural language understanding tasks, highlighting their pivotal role in healthcare. Their ability to comprehend and generate human-like text is poised to play a transformative role in healthcare practices, where effective communication and information processing are of paramount importance [12].

The trajectory of natural language processing (NLP) has been characterized by a series of noteworthy milestones, with each development building upon the strengths and limitations of its predecessors. In its nascent stages, recurrent neural networks (RNNs) laid the foundation for contextual information retention in NLP tasks. However, their inherent limitations in capturing long-range dependencies became evident, thus necessitating a shift in the NLP paradigm.

The pivotal moment in NLP's evolution came with the introduction of transformers, a groundbreaking architecture that addressed the challenge of capturing distant word relationships effectively. This innovation was a turning point, enabling more advanced NLP models. These advancements provided the impetus for the emergence of sophisticated language models such as Llama 2 [13] and GPT-4, which, underpinned by extensive training data, have elevated NLP to a level of understanding and text generation that closely approximates human-like language.

Within the healthcare domain, tailored adaptations of models such as BERT, including BioBERT and ClinicalBERT [14,15], have been introduced to tackle the intricacies of clinical language. The introduction of these models addressed the unique challenges posed by medical text, which frequently features complex medical terminology, lexical ambiguity, and variable usage. However, introducing LLMs into the highly sensitive and regulated domain of healthcare demands careful consideration of ethics, privacy, and security. Patient data must be rigorously protected while ensuring that LLMs do not perpetuate existing biases or lead to unintended harm. Nevertheless, the potential for LLMs to enhance healthcare practices, improve patient outcomes, and spearhead innovative research avenues continues to stimulate ongoing investigation and growth in this rapidly evolving field.

As we navigate this dynamic field, our review aims to function as a comprehensive guide, offering insights to medical researchers and healthcare professionals seeking to optimize their research endeavors and clinical practices. We seek to provide a valuable resource for the judicious selection of LLMs tailored to specific clinical requirements. Our examination encompasses a detailed exploration of LLMs within the healthcare domain, elucidating their underlying technology, diverse healthcare applications, and facilitating discussions on critical topics such as fairness, bias mitigation, privacy, transparency, and ethical considerations. By highlighting these critical aspects, this review aims to illustrate the importance of integrating LLMs into healthcare in a manner that is not only effective but also ethical, fair, and equitable, ultimately fostering benefits for both patients and healthcare providers.

This review paper is organized into distinct sections that systematically address the integration, impact, and limitations of large language models (LLMs) in healthcare:

- Section 2 provides a foundational understanding of LLMs, covering their key architectures such as transformers, foundational models, and multi-modal capabilities.
- In Section 3, the focus shifts to the application of LLMs in healthcare, discussing their use cases and the metrics for assessing their performance within clinical settings.
- Section 4 critically examines the challenges associated with LLMs in healthcare, including issues related to explainability, security, bias, and ethical considerations.

- This paper concludes by summarizing the findings, highlighting the transformative potential of LLMs while acknowledging the need for careful implementation to navigate their limitations and ethical implications.

2. Review of Large Language Models

Large language models have emerged as a notable advancement in the field of natural language processing (NLP) and have attracted considerable interest in recent times [10,16]. These models exhibit notable attributes such as their considerable number of parameters, pretraining on vast collections of textual data, and fine-tuning for specific downstream objectives [13,17,18]. By leveraging these key characteristics, large language models demonstrate exceptional performance across a wide range of NLP tasks. This section presents a comprehensive discussion of the concept, architecture, and pioneering examples of large language models. Furthermore, we explore the pretraining methodology and the significance of transfer learning in facilitating these models to achieve exceptional performance across diverse tasks [19].

Large Language models built upon the transformer architecture have been specifically engineered to enhance the efficiency of natural language data processing in comparison to earlier iterations. The transformer architecture, as proposed by [20], utilizes a self-attention mechanism to capture the contextual relationships between words in a sentence. This mechanism facilitates the model's ability to assign varying degrees of significance to distinct words during the prediction process, rendering it especially suitable for handling long-range dependencies in language.

The key aspects of large language models encompass their substantial magnitude [21,22], of pretraining on vast text corpora [13,23] and subsequent fine-tuning tailored towards specific tasks [24]. These models possess a substantial number of parameters, ranging from hundreds of millions to billions, which allows them to effectively capture intricate patterns and nuances within language. Pretraining is commonly conducted on diverse datasets devoid of task-specific annotations, enabling the model to acquire knowledge from a broad spectrum of linguistic instances and develop a comprehensive grasp of language. Following pretraining, the model undergoes a further fine-tuning process using smaller datasets that are appropriate to the task at hand. This allows the model to successfully adapt to and perform well on specific natural language processing (NLP) tasks.

The progression of natural language processing (NLP) has been characterized by a series of significant advancements. At the outset, recurrent neural networks (RNNs) facilitated the retention of context in natural language processing (NLP) tasks. Nevertheless, recurrent neural networks (RNNs) were found to have several shortcomings around effectively capturing long-range dependencies. The advent of transformers has had a transformative impact by effectively addressing the challenge of capturing distant word relationships. Subsequently, large language models such as Llama 2 [13] and GPT-4 [11] emerged; powered by extensive training data, they significantly advanced NLP capabilities in understanding and generating human-like text. This progression signifies a continuous cycle of innovation, with each stage building upon the strengths and limitations of its predecessor. In the subsequent section, we delineate significant phases of development within the continuum of progress in the landscape of natural language processing (NLP).

In the healthcare domain, specialized adaptations of BERT such as BioBERT [14] and ClinicalBERT [15] have been introduced to address a variety of challenges in comprehending clinical language. GPT-3 (Generative Pretrained Transformer 3), developed by OpenAI, is one of the largest language models to date, with 175 billion parameters [10]. Recently, OpenAI introduced GPT-3.5 and its successor GPT-4 (OpenAI, 2023) [11], which alongside Google AI's Bard have emerged as cutting-edge large language models (LLMs) displaying remarkable capabilities across diverse applications, including healthcare and medicine [6].

2.1. Transformers

The transformer architecture, introduced in "Attention is All You Need" [20], has revolutionized natural language processing. The primary novelty of this model is its utilization of a self-attention mechanism, allowing it to assess the importance of input tokens by considering their relevance to the given task. In this setup, multiple attention heads work in parallel, allowing the model to focus on various aspects of the input, whereas positional encoding conveys relative token positions. Given an input sequence X of length N, the self-attention mechanism computes attention scores $A(i, j)$ between all token pairs (i, j). Three learned matrices, the Query (Q), Key (K), and Value (V), are obtained by linear projections of X.

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

Here, d_k represents the dimension of key vectors. The softmax function normalizes the scores. The output for each token is then computed as a weighted sum of the value vectors for all tokens j. Multi-head attention extends this mechanism by computing multiple attention sets in parallel before concatenating and linearly transforming them to form the final output.

Transformers consist of stacked encoder–decoder blocks, helping models adapt to diverse tasks. Training occurs via unsupervised or semi-supervised learning on vast text corpora using gradient-based optimization. Transformers have become foundational in natural language processing due to their capacity to handle sequential data, capture long-range dependencies, and adapt to various tasks with minimal fine-tuning. This extends beyond text, finding applications in healthcare, recommendation systems, image generation, and other domains.

2.2. Large Foundational Models

The advent of large foundational models, exemplified by GPT-3 (Brown et al., 2020) [10] and Stable Diffusion (Rombach et al., 2022) [25], has ushered in a transformative era in the field of machine learning and generative artificial intelligence. Researchers have introduced the term "foundation model" to delineate machine learning models that undergo training on extensive, diverse, and unlabeled datasets, endowing them with the ability to adeptly tackle a broad spectrum of general tasks. These encompass tasks related to language comprehension, text and image generation, and natural language dialogue.

Large foundational models are massive AI architectures trained on extensive quantities of unlabeled data, predominantly employing self-supervised learning methods. This approach to training yields models of exceptional versatility, enabling them to excel across a wide spectrum of tasks, ranging from image classification and natural language processing to question-answering while consistently delivering outstanding levels of accuracy.

These models particularly shine in tasks demanding generative capabilities and human interaction, including the creation of marketing content or intricate artwork based on minimal prompts. Nevertheless, adapting and integrating these models into enterprise applications may present specific challenges [26].

2.3. Multimodal Language Models

Multimodal Large Language Models (MLLMs) represent a groundbreaking advancement in the fields of artificial intelligence (AI) and natural language processing (NLP). In contrast to conventional language models focused solely on textual data, MLLMs possess the unique ability to process and generate content across multiple modalities, including text, images, audio, and video. This novel approach significantly expands the capabilities of AI applications, allowing machines to not only comprehend and generate text but also to interpret and integrate information from various sensory inputs. The integration of multiple modalities enables MLLMs to bridge the gap between human communication and

machine understanding, making them versatile tools with the potential to transform diverse fields. This theoretical introduction highlights the transformative potential of MLLMs and their central role in pushing the boundaries of artificial intelligence, affecting areas such as image and speech recognition, content generation, and interactive AI applications [27].

MLLMs are designed to process and integrate information from multiple data sources, such as text, images, and audio, to perform a variety of tasks. These models leverage deep learning techniques to understand and generate content across different modalities, enhancing their applicability in real-world scenarios. For instance, Visual ChatGPT combines text and visual inputs to address complex queries [28], while systems such as BLIP-2 utilize a Qformer to integrate visual features with textual data for enhanced image–text interactions [29]. MLLMs are particularly effective in tasks such as visual question answering (VQA), where they can interpret and respond to queries based on visual content. The integration of modalities allows these models to offer more comprehensive responses and handle a broader range of interactions than single-modality models. The iterative training processes, often involving stages of freezing certain components while fine-tuning others, enable these models to maintain robust language capabilities while adapting to new modalities and tasks.

Figure 1 displays a typical MLLM architecture comprising an encoder E_M, connector C, and large language model (LLM). Additionally, a generator G can be integrated with the LLM to produce outputs beyond text, such as other modalities. The encoder processes inputs such as images, audio, or video into features, which the connector refines to enhance the LLM's comprehension capabilities. Connectors in these systems come in three main varieties: projection-based, query-based, and fusion-based. The first two types utilize token-level fusion, converting features into tokens that are then combined with text tokens, whereas fusion-based connectors perform feature-level fusion directly within the LLM [27].

Recently, the integration of the Mixture of Experts (MoE) architecture into MLLMs has significantly advanced their capabilities. This approach employs multiple specialized submodels, each of which is fine-tuned for specific types of data or tasks such as image recognition or language processing. By selectively activating the most relevant experts based on the input and task, MoE allows MLLMs to dynamically adapt to the demands of multimodal data integration. This enhances the precision of the model in handling complex multimodal interactions and optimizes computational resources. Models such as MoVA [30] and MoE-LLaVA [31] leverage MoE strategies effectively, improving performance while maintaining manageable computational costs during both training and inference phases. The adaptability and efficiency of MoE within MLLMs contribute significantly to their scalability and efficacy in real-world applications across varied tasks and data types [32].
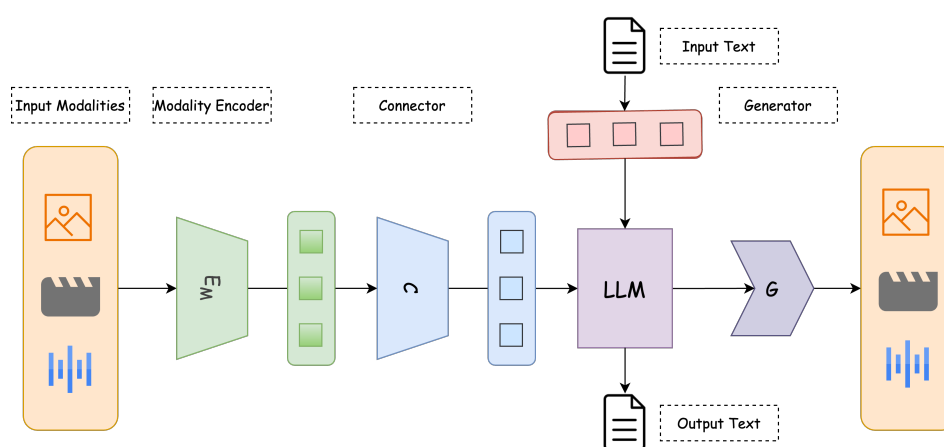

Figure 1. Schematic representation of a standard multimodal large language model (MLLM) architecture.

3. Large Language Models in Healthcare and Medicine

Language models have become a revolutionary force in the constantly changing world of healthcare and medicine, revolutionizing how medical researchers and practitioners engage with data, patients, and the huge corpus of medical knowledge [33]. The use of language models in the medical field has undergone a significant metamorphosis from the early days of simple rule-based systems, feature extraction, and keyword matching to the arrival of cutting-edge technologies such as transformers and LLMs such as GPT-v4 [11]. These language models have overcome the constraints of conventional methods, enabling more complex natural language generation and interpretation.

Several pioneering large language models have significantly influenced the landscape of NLP. The emergence of the transformer architecture [20] marked a significant milestone in the realm of natural language processing, leading to the emergence of expansive pretrained language models such as BERT [34] and RoBERTa [35].

BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. (2018) [34], revolutionized NLP by pretraining a deep bidirectional model on a large corpus, helping it to outperform previous models on various tasks. RoBERTa (Robustly Optimized BERT Pretraining Approach) by Liu et al. (2019) [35] demonstrated that further pretraining improvements and optimization could significantly enhance the performance of BERT.

In this section, we first discuss the current large language models specifically for medical applications in Section 3.1. Then, in Section 3.2 we cover the use cases of various LLMs designed mainly for patients, experts, and medical materials.

3.1. Large Language Models for Medical and Healthcare Applications

Figure 2 provides a comprehensive overview of the progression in biomedical language model (LM) development from 2019 to 2023, emphasizing a logarithmic growth in model complexity and parameter count.
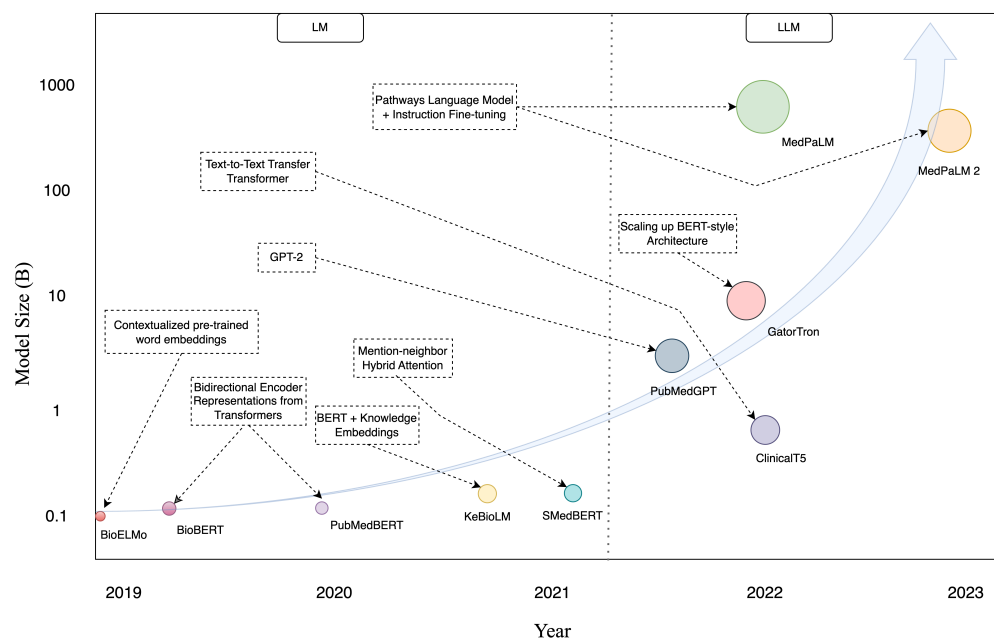

Figure 2. Scale of medical language models: size comparison.

It describes the evolutionary trajectories of various domain-specific adaptations of prominent models such as BioBERT, and GPT-2 along with the inception of more advanced systems such as MedPaLM. The sizes of the illustrated models are proportional to their parameter volumes, showcasing a consistent trend towards larger and more capable models. This culminates in the emergence of LLMs by 2023, which signifies a pivotal shift towards

architectures with substantially heightened computational requirements and potential performance in biomedical text analysis and generation tasks.

On the other hand, Table 1 provides an insightful overview of leading large language models within the healthcare domain. Recently, "BioMistral" was published as a a collection of open-source pretrained large language models for the medical domain. In 2023, "Med-PaLM 2" and "Radiology-Llama2" emerged as key players, addressing the medical question-answering and radiology tasks, respectively. The "DeID-GPT" model extends its capabilities to de-identification, while "Med-HALT" specializes in hallucination testing. Simultaneously, "ChatCAD" offers valuable support in the realm of computer-aided diagnosis. "BioGPT" showcases versatility by handling classification, relation extraction, and question answering. "GatorTron" excels in semantic textual similarity and medical question answering, whereas "BioMedLM" narrows its focus to biomedical question answering. "BioBART" demonstrates prowess in dialogue, summarization, entity linking, and NER. "ClinicalT5" tackles classification and NER, while "KeBioLM" specializes in biomedical pretraining, NER, and relation extraction. Before the advent of language models or transformers, convolutional and recurrent neural networks represented the state of the art in the field. These models collectively represent remarkable strides in healthcare NLP, providing accessible source code or models for further exploration and practical application.

Table 1. Summary of large language models in the healthcare space.

Method	Year	Task	Institution	Source Code
BioMistral [36]	2024	Medical Question Answering	Avignon Université, Nantes Université	model (https://huggingface.co/BioMistral/BioMistral-7B, accessed on 8 July 2024)
Med-PaLM 2 [37]	2023	Medical Question Answering	Google Research, DeepMind	
Radiology-Llama2 [38]	2023	Radiology	University of Georgia	
DeID-GPT [39]	2023	De-identification	University of Georgia	code (https://github.com/yhydhx/ChatGPT-API, accessed on 8 July 2024)
Med-HALT [40]	2023	Hallucination test	Saama AI Research	code (https://github.com/medhalt/medhalt, accessed on 8 July 2024)
ChatCAD [41]	2023	Computer-aided diagnosis	ShanghaiTech University	code (https://github.com/zhaozh10/ChatCAD, accessed on 8 July 2024)
BioGPT [42]	2023	Classification, relation extraction, question answering, etc.	Microsoft Research	code (https://github.com/microsoft/BioGPT, accessed on 8 July 2024)
GatorTron [43]	2022	Semantic textual similarity, natural language inference, and medical question answering	University of Florida	code (https://github.com/uf-hobi-informatics-lab/GatorTron, accessed on 8 July 2024)
BioMedLM	2022	Biomedical question answering	Stanford CRFM, MosaicML	code (https://github.com/stanford-crfm/BioMedLM, accessed on 8 July 2024)
BioBART [44]	2022	Dialogue, summarization, entity linking, and NER	Tsinghua University, International Digital Economy Academy	code (https://github.com/GanjinZero/BioBART, accessed on 8 July 2024)
ClinicalT5 [45]	2022	Classification, NER	University of Oregon, Baidu Research	model (https://huggingface.co/xyla/Clinical-T5-Large, accessed on 8 July 2024)
KeBioLM [46]	2021	Biomedical pre-training, NER, and relation extraction	Tsinghua University, Alibaba Group	code (https://github.com/GanjinZero/KeBioLM, accessed on 8 July 2024)
CRNN [47]	2017	Relation classification	Indian Institute of Technology	code (https://github.com/desh2608/crnn-relation-classification, accessed on 8 July 2024)
LSTM RNN [48]	2017	Named entity recognition	Wuhan University	code (https://github.com/lvchen1989/BNER, accessed on 8 July 2024)

3.2. Use Cases of Large Language Models in Healthcare

In recent years, the emergence of large language models has catalyzed a transformative shift in the healthcare landscape, offering unprecedented opportunities for innovation and advancement. Their ability to comprehend and generate text that resembles that of humans has demonstrated remarkable potential across a wide range of healthcare applications [49]. The applications of large language models in the healthcare sector are experiencing rapid growth. As illustrated in Figure 3, these models are being utilized for

clinical decision support, medical record analysis, patient engagement, health information dissemination, and more. Their implementation holds the prospect of improving diagnostic accuracy, streamlining administrative procedures, and ultimately enhancing the efficiency, personalization, and comprehensiveness of healthcare delivery. This section delves into a comprehensive exploration of the multifaceted applications of large language models in healthcare; shedding light on their profound implications, these applications bear on the trajectory of medical practices and the eventual outcomes experienced by patients.
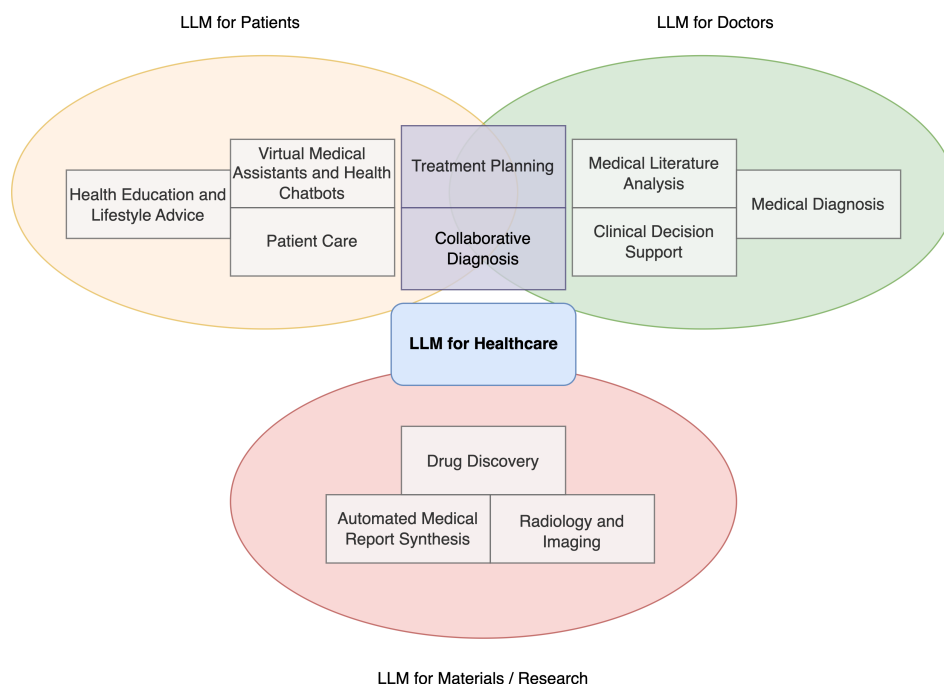

Figure 3. Applications of large language models in healthcare.

- **Medical Diagnosis:** Certain clinical procedures may depend on the use of data analysis, clinical research, and recommendations [50,51]. LLMs can potentially contribute to medical diagnosis by conducting analyses on patient symptoms, medical records, and pertinent data, potentially aiding in the identification of potential illnesses or conditions with a certain degree of accuracy. Large language models have the potential to contribute to several aspects, such as clinical decision assistance, clinical trial recruiting, clinical data administration, research support, patient education, and other related areas [52,53]. Corroborating this perspective, methodologies utilizing transformer models such as BERT, RoBERTa, and DistilBERT were used for the purpose of predicting COVID-19 diagnosis based on textual descriptions of acute alterations in chemosensation [54]. Similarly, a number of alternative investigations have been undertaken within the literature, including strategies using large language models for the diagnosis of Alzheimer's disease [55] and dementia [56]. Furthermore, a corpus of literature has emerged advocating the integration of large language model chatbots to cater to analogous objectives [57–60].

- **Patient Care:** Large language models have emerged as transformative tools with the capacity to significantly enhance the realm of patient care [61]. Through the provision of personalized recommendations [62], customized treatment strategies, and continual monitoring of patients' advancement throughout their medical journey [63], LLMs offer the promise of revolutionizing healthcare delivery. By harnessing the capabilities of LLMs, healthcare providers can ensure a more personalized and patient-centric approach to care. This technology enables the delivery of precise and well-informed medical guidance [64], aligning interventions with patients' distinct requirements and circumstances. The effective use of LLMs within clinical practice not only en-

hances patient outcomes, it enables healthcare professionals to make data-driven decisions, leading to enhanced patient care. As LLMs continue to advance, the potential for augmenting patient care through personalized recommendations and ongoing monitoring remains a promising trajectory in modern medicine [65]. In essence, LLMs represent a pivotal leap forward, holding the capacity to reshape the landscape of patient care by fostering precision, adaptability, and patient-centeredness [66].

- **Clinical Decision Support**: Language models (LMs) have evolved into crucial decision support tools for healthcare professionals. By analyzing extensive medical data, LMs can provide evidence-based recommendations, enhancing diagnostic accuracy, treatment selection, and overall patient care. This fusion of artificial intelligence with healthcare expertise holds immense promise for improved medical decision-making. A body of existing research has illuminated promising prospects for the application of language models within clinical decision support, particularly within the domains of radiology [67], oncology [68], and dermatology [69].

- **Medical Literature Analysis**: Large language models (LLMs) exhibit remarkable efficiency in comprehensively reviewing and succinctly summarizing extensive volumes of medical literature. This capability aids both researchers and clinicians in maintaining topicality with cutting-edge developments and evidence-based methodologies, ultimately fostering informed and optimized healthcare practices. In a fast-evolving field like healthcare, the ability to maintain currency with the latest advancements is paramount, and LLMs can play a pivotal role in ensuring that healthcare remains at the forefront of innovation and evidence-based care delivery [70,71].

- **Drug Discovery**: Large language models have a significant impact in facilitating drug discovery through their capacity to scrutinize intricate molecular structures, discern promising compounds with therapeutic potential, and forecast the efficacy and safety profiles of these candidates [72,73]. Chemical language models have exhibited notable achievements in the domain of de novo drug design [74]. In a corresponding study, the authors explored the utilization of pretrained biochemical language models to initialize targeted molecule generation models, comparing one-stage and two-stage warm start strategies as well as evaluating compound generation using beam search and sampling. They ultimately demonstrated that warm-started models outperformed baseline models and that the one-stage strategy exhibited superior generalization in terms of docking evaluation and benchmark metrics, while beam search proved more effective than sampling for assessing compound quality [75].

- **Virtual Medical Assistants and Health Chatbots**: LLMs can also serve as the underlying intelligence for health chatbots, which are revolutionizing the healthcare landscape by delivering continuous and personalized health-related support. Such chatbots can offer medical advice, monitor health conditions, and even extend their services to encompass mental health support, a particularly pertinent aspect of healthcare given the growing awareness of mental well-being [57,60].

- **Radiology and Imaging:** By integrating visual and textual data, multimodal visual language models hold significant promise for augmenting medical imaging analysis. Radiologists can benefit from these models as they facilitate the early identification of abnormalities in medical images and contribute to the generation of more precise and comprehensive diagnostic interpretations, ultimately advancing the accuracy and efficiency of diagnostic processes in the field of medical imaging [67,76–81].

- **Automated Medical Report Synthesis from Imaging Data:** Automated medical report generation from images is crucial for streamlining the time-consuming and error-prone task faced by pathologists and radiologists. This emerging field at the intersection of healthcare and artificial intelligence (AI) aims to alleviate the burden on experienced medical practitioners and enhance the accuracy of less experienced ones. The integration of AI with medical imaging facilitates the automatic drafting of reports, encompassing abnormal findings, relevant normal observations, and patient history. Early efforts employed data-driven neural networks, combining convolutional and

recurrent models for single-sentence reports; however, limitations arose in capturing the complexity of real medical scenarios [5]. Recent advances have leveraged LLMs such as ChatCAD [67], enabling more sophisticated applications. ChatCAD enhances medical image CAD networks, yielding significant improvements in report generation. ChatCAD+ further addresses writing style mismatches, ensuring universality and reliability across diverse medical domains by incorporating a template retrieval system for consistency with human expertise [41]. In [82], the authors used a pretrained language model (PLM) and in-context learning (ICL) to generate clinical notes from doctor–patient conversations. These integrated systems signify a pivotal advancement in automating medical report generation through the strategic utilization of LLMs.

3.3. Explainable AI Methods for Interpreting Healthcare LLMs

Large language models have significantly advanced the healthcare domain, enhancing tasks such as medical diagnosis and patient monitoring. However, the complexity of these models necessitates interpretability for reliable decision-making [83]. This section discusses "eXplainable and Interpretable Artificial Intelligence" (XIAI) and examines recent XIAI methods by their functionality and scope. Despite challenges such as the difficulty in quantifying interpretability and the lack of standardized evaluation metrics, opportunities exist in integrating XIAI to add interpretability for LLMs in healthcare. Notable XIAI methods include SHAP [84], which quantifies feature contributions, LIME [85,86], which generates interpretable models through input perturbations, t-SNE for visualizing high-dimensional data [87], attention mechanisms that highlight key features [15], and knowledge graphs that structure contextual relationships [88], all of which provide crucial insights into model decision-making processes.

Existing research delves into explainability for LLMs in the healthcare domain. For instance, Yang et al. (2023) [89] investigate different prompting strategies using emotional cues and expert-written examples for mental health analysis with LLMs. This study shows that models such as ChatGPT can generate near-human-level explanations, enhancing interpretability and performance. Additionally, ArgMedAgents (Hong et al., 2024) [90] is a multi-agent framework designed for explainable clinical decision reasoning through interaction, utilizing the Argumentation Scheme for Clinical Discussion and a symbolic solver to provide clear decision explanations. Furthermore, Gao et al. (2023) proposed enhancing LLM explainability for automated diagnosis by integrating a medical knowledge graph (KG) from the Unified Medical Language System (UMLS) by using the DR.KNOWS model to interpret complex medical concepts. Their experiments with real-world hospital data demonstrate a transparent diagnostic pathway. Similarly, TraP-VQA [88], a novel vision–language transformer for Pathology Visual Question Answering (PathVQA) employs Grad-CAM and SHAP methods to offer visual and textual explanations, ensuring transparency and fostering user trust.

We have compiled a list in Table 2 detailing XIAI attributes and summarizing recent research works focused on explainability methods for LLMs in the healthcare domain. This table includes evaluations of various models, highlighting their unique contributions to enhancing interpretability and reliability in medical applications. Each entry outlines the task, method, XAI attributes, and evaluation metrics, offering a clear overview of the advancements and effectiveness of XIAI techniques in improving decision-making processes in healthcare.

Table 2. Summary of recent XIAI methods for LLMs in healthcare.

Method	Year	Task	XIAI Attributes	XIAI Evaluation Metric
MentaLLaMA [code (https://github.com/SteveKGYang/MentalLLaMA, accessed on 8 July 2024)] [91]	2024	Mental health analysis	Prompt-based (ChatGPT w/ task-specific instructions)	BART-score, Human Eval
ArgMed-Agents [90]	2024	Clinical decision reasoning	Prompt-based (Self-argumentation iterations + symbolic solver)	Pred. accuracy with LLM evaluator
Diagnostic reasoning prompts [92]	2024	Medical Question Answering (MedQA)	Prompt-based (Bayesian, differential diagnosis, analytical, and intuitive reasoning)	Expert Evaluation, Inter-rater agreement
SkinGEN [93]	2024	Dermatological diagnosis	Visual explanations (Stable Diffusion), interactive framework.	Perceived explainability ratings
DR. KNOWS [88]	2023	Automated diagnosis generation	Knowledge Graph (explainable diagnostic pathway)	
Human-AI Collaboration [94]	2023	Clinical decision making	Salient features, counterfactual explanations	Agreement Level, Usability Questionnaires
ChatGPT [89]	2023	Mental health analysis	Prompt-based (emotional cues and expert-written few-shot examples)	BART-score, Human Eval
CHiLL [95]	2023	Clinical predictive tasks, Chest X-ray report classification	Interpretable features, linear models	Expert Evaluation, Clinical Judgement Alignment
Trap-VQA [96]	2022	Pathology Visual Question Answering (PathVQA)	Grad-CAM, SHapley Additive exPlanations	Qualitative Evaluation
Vision Transformer [97]	2021	Covid-19 diagnosis	Saliency maps	Visualisation
ClinicalBERT [code (https://github.com/kexinhuang12345/clinicalBERT, accessed on 8 July 2024)] [15]	2019	Predicting hospital readmission	Attention weights	Visualisation

3.4. Future Trajectories of Large Language Models in Healthcare

As large language models (LLMs) continue to integrate into the healthcare sector, future developments promise to revolutionize patient care and medical research. A particularly promising avenue involves enhancing LLMs' capabilities to interpret and generate not only textual but also biomolecular data [98]. This advancement could significantly improve applications in genomics and personalized medicine, enabling these models to predict individual responses to treatments based on genetic profiles, thereby advancing the precision of medical interventions. Furthermore, incorporating adaptive learning capabilities in real time could transform LLMs into dynamic aids during surgical procedures or emergencies, where they might analyze data from medical devices on-the-fly [99] to offer critical decision support.

Another innovative trajectory for LLMs in healthcare is the development of federated learning systems [100]. Such systems could facilitate the secure and privacy-preserving propagation of medical knowledge across institutions, improving model robustness and applicability across varied demographic groups without direct data sharing. This approach can enhance the privacy and security of patient dataas well as enable a collective intelligence that could lead to more generalized healthcare solutions.

The potential of large language models (LLMs) in healthcare extends into the realms of explainable medical AI [101] and the utilization of multimodal models incorporating sensor data. By integrating LLMs with wearable technologies [102], these advanced models can serve as continuous health monitors in non-clinical settings.

To further advance explainable medical AI, LLMs can be instrumental in deciphering the complexities of medical conditions and treatment outcomes. By processing and interpreting multimodal data, including sensor readings, these models can contribute to a deeper understanding of patient health on a granular level. This may aid in the development of precise targeted therapies, improving patient outcomes and enhancing the transparency of medical decisions.

LLMs are poised to revolutionize the healthcare domain by enhancing diagnostic accuracy, personalizing treatment plans, and optimizing operational efficiencies. By integrating LLMs into electronic health record systems, healthcare providers can more accurately diagnose conditions through natural language processing techniques that analyze clinical notes and patient histories. Moreover, LLMs assist in generating personalized treatment recommendations by analyzing vast datasets that include genetic information, clinical outcomes, and patient preferences. Furthermore, these models streamline administrative tasks by automating documentation, coding, and billing processes, thus reducing operational costs and allowing medical staff to focus more on patient care. As generative AI advances, its transformative impact on the healthcare sector is becoming increasingly significant. This technology is poised to revolutionize areas such as clinical trials, personalized medicine, and drug discovery. Additionally, its applications extend to enhancing natural language processing and understanding, improving medical imaging, and supporting virtual assistants in patient care. Generative AI also plays a crucial role in illness detection and screening, facilitating more accurate diagnostics. Moreover, it is being integrated into medical conversation tasks, voice generation, video generation, and image synthesis and manipulation within healthcare settings [103]. These innovations are not only improving the efficiency of medical services but are also paving the way for new methods of patient interaction and treatment planning. As these applications continue to mature, LLMs will become integral in transforming healthcare services into more efficient, accurate, and personalized systems.

3.5. Performance Evaluation and Benchmarks

The medicine and healthcare industries largely acknowledge the potential of artificial intelligence to drive substantial progress in the delivery of healthcare. However, empirical evaluations have demonstrated that numerous artificial intelligence systems do not successfully achieve their desired translation goals, primarily because of intrinsic deficiencies that become evident only after implementation [104,105]. In order to optimize the utilization of LLMs within healthcare settings, it is imperative to develop evaluation frameworks that possess the capacity to thoroughly evaluate their safety and quality. It is important to note that certain highly effective models such as ChatGPT and PaLM 2 [106] are now not publicly available. The absence of accessibility gives rise to notable problems pertaining to transparency, which is a crucial factor in the medical domain and hinders the capacity to thoroughly examine the structure and results of the model. Consequently, this impedes endeavors to recognize and address biases and hallucinations. Thorough research is necessary to understand the specific performance characteristics and ramifications of utilizing publicly accessible, pretrained language models in addressing the challenges in the healthcare and medical domains. Language models that have been pretrained using medical data also encounter comparable difficulties. Therefore, the careful choice and implementation of suitable performance metrics to evaluate the language model assume great significance.

In Table 3, we present a comprehensive catalog of performance metrics, including but not limited to the F1 score, BLEU, GLUE, and ROGUE, which constitute the standard evaluative criteria employed for the rigorous assessment of large language models op-

erating within the healthcare and medical domain. This compendium of metrics serves as a valuable reference, encapsulating the quantitative and qualitative measures utilized to gauge the efficacy, proficiency, and suitability of these models in diverse healthcare applications [105].

Table 3. Evaluation metrics for language models in the healthcare domain.

Eval. Metric	Description	References	Key Highlights
Perplexity	Perplexity, a probabilistic metric, quantifies the uncertainty in the predictions of a language model. Lower values indicate higher prediction accuracy and coherence.	[107]	-
		[108]	The federated learning model achieved a best perplexity value of **3.41** for English.
		[109]	The Transformer model achieved a test perplexity of **15.6** on the PSVG dataset, significantly outperforming the LSTM's perplexity of **20.7**.
		[88]	The lowest perplexity achieved was $\mathbf{3.86 \times 10^{-13}}$
BLEU	The BLEU score assesses the quality of machine translation by comparing it to reference translations.	[110]	The best BLEU-1 score achieved was **13.9** by the ClinicalGPT model.
		[111]	T-5 (fine-tuned) model achieved the best BLEU-1 score of **26.63**.
GLEU	GLEU score computes mean scores of various n-grams to assess text generation quality.	[110]	The best GLEU score achieved was **2.2** by the Bloom-7B model.
		[111]	T-5 (fine-tuned) model achieved the best GLEU score of **11.38**.
ROUGE	ROUGE score evaluates summarization and translation by measuring overlap with reference summaries.	[110]	The best ROGUE-L score achieved was **21.3** by the ClinicalGPT model.
		[111]	T-5 (fine-tuned) model achieved the best ROGUE-L score of **24.85**.
Distinct n-grams	Measures the diversity of generated responses by counting unique n-grams.	[111]	On the Huatuo-26M dataset, the fine-tuned T5 model achieved Distinct-1 and Distinct-2 scores of **0.51** and **0.68**, respectively.
F1 Score	The F1 score balances precision and recall, measuring a model's accuracy in identifying positive instances and minimizing false results.	[112]	The GatorTron-large model achieved the best F1 score of **0.9627** for medical relation extraction.
		[43]	The GatorTron-large model achieved the best F1 score of **0.9000** for clinical concept extraction and **0.9627** for medical relation extraction.
		[113]	The multicenter Transformers-based model achieved an overall F1 score of **84.77%** on the PsyNIT dataset.
		[73]	The BERT-D2 model achieved an F1 score of **81.97%** on the DDI Extraction 2013 corpus.
BERTScore	BERTScore calculates similarity scores between tokens in candidate and reference sentences, using contextual embeddings.	[114]	-
		[82]	The Longformer-Encoder-Decoder (LED$_{\text{large-PubMed}}$) model achieved the best BERTScore F1 of **70.7**.
Human Evaluation	Involves expert human assessors rating the quality of model-generated content, providing qualitative insights into its performance.	[115]	The median performance for all human SCORE users was **65%**, whereas ChatGPT correctly answered **71%** of multiple-choice SCORE questions and **68%** of Data-B questions.

3.6. Quantitative Performance Comparison of LLMs in the Healthcare Domain

Recent advancements in language models have been benchmarked against diverse datasets to evaluate their capabilities across various domains. One such comprehensive benchmark is the MMLU (Massive Multitask Language Understanding) [116], designed to assess the understanding and problem-solving abilities of language models. The MMLU comprises 57 tasks spanning topics such as elementary mathematics, US history, computer science, and law, requiring models to demonstrate a broad knowledge base and problem-solving skills. As listed in Table 4, this benchmark provides a standardized method for testing and comparing various language models, including OpenAI GPT-4o, Mistral 7b, Google Gemini, and Anthropic Claude 3, among others.

The HumanEval benchmark is used to measure the functional correctness of code generated by LLMs from docstrings. This benchmark evaluates models based on their ability to generate code that passes provided unit tests, using the pass@k metric. If any of the 'k' solutions generated by the model pass all unit tests, the model is considered successful in solving the problem [117]. Table 4 provides a concise summary of the performance of various LLMs on the MMLU and HumanEval (Coding) datasets [118].

Table 4. LLM performance benchmarks.

Organization	Model	MMLU Score	Coding (HumanEval)	Release Date
OpenAI	GPT-4 Opus	88.7	-	May 2024
Anthropic	Claude 3.5 Sonnet	88.7	92.0	June 2024
Anthropic	Claude 3 Opus	86.8	-	March 2024
OpenAI	GPT-4 Turbo	86.4	85.4	April 2024
OpenAI	GPT-4	86.4	90.2	April 2023
Meta	Llama 3 400B	86.1	-	-
Google	Gemini 1.5 Pro	85.9	84.1	May 2024
Google	Gemini Ultra	83.7	-	December 2023
OpenAI	GPT-3.5 Turbo	-	73.2	-
Meta	Llama 3 (70B)	-	81.7	-
Meta	Llama 3 (8B)	-	62.2	-
Google	Gemini 1.5 Flash	-	74.3	-

In the healthcare domain, a variety of LLMs have been developed and evaluated on specific datasets such as MedQA, MedNLI [119], Tox21 [120], and PubMedQA [121]. The GPT-4 (2024) model stands out on the MedQA dataset with an impressive accuracy of 93.06%, significantly outperforming other models such as Med-PaLM 2 (CoT + SC) (2023), which achieves 83.7%, and Meerkat-7B (Ensemble) (2024) with 74.3%. On the MedNLI dataset, BioELECTRA-Base (2021) achieves the highest accuracy of 86.34%, closely followed by CharacterBERT (base, medical) (2020) at 84.95%. The Tox21 dataset highlights elEmBERT-V1 (2023), with an outstanding AUC of 0.961 making it the most effective in predicting chemical properties and toxicity. For the PubMedQA dataset, Meditron-70B (CoT + SC) (2023) and BioGPT-Large (1.5B) (2023) exhibit strong performance with accuracy of 81.6% and 81.0%, respectively [122]. These findings underscore the variability in performance across different healthcare tasks, emphasizing the need for careful selection of models based on specific application requirements [123]. Figure 4 presents a comparative performance analysis of various healthcare LLMs, highlighting their accuracy and AUC metrics across different datasets including MedQA, MedNLI, Tox21, and PubMedQA.
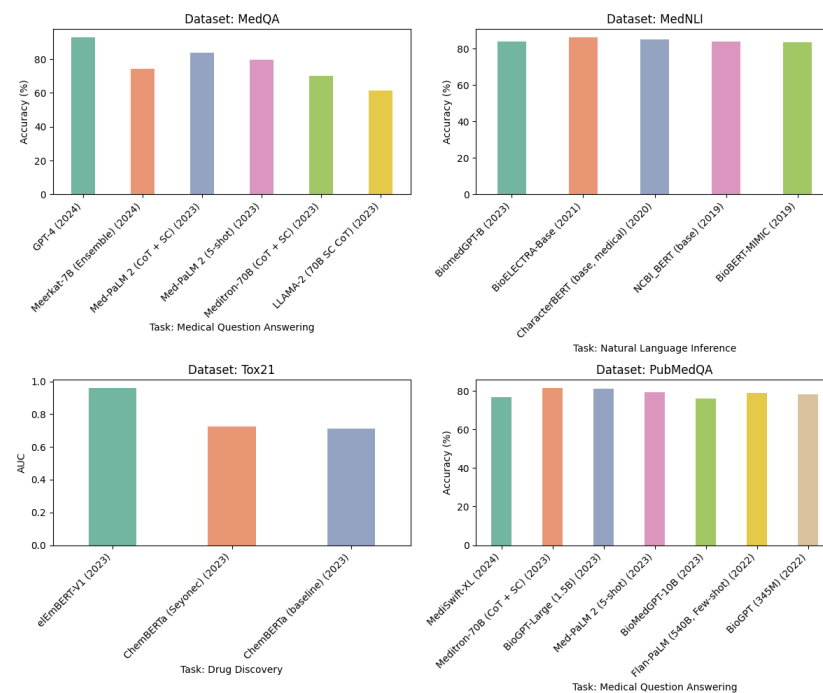
Figure 4. Comparative performance of healthcare LLMs.

4. Limitations and Open Challenges

The integration of large language models (LLMs) in healthcare presents complex challenges, including the need for explainability in model decision-making, robust security and privacy measures to protect sensitive patient data, addressing bias and ensuring fairness in medical AI applications, mitigating the issue of hallucinations where models generate erroneous information, and establishing clear legal frameworks for the responsible use of LLMs in healthcare, As sumarized in Figure 5. All of which demand careful scrutiny and resolution to harness the full potential of these models for improving healthcare outcomes while upholding ethical and legal standards.
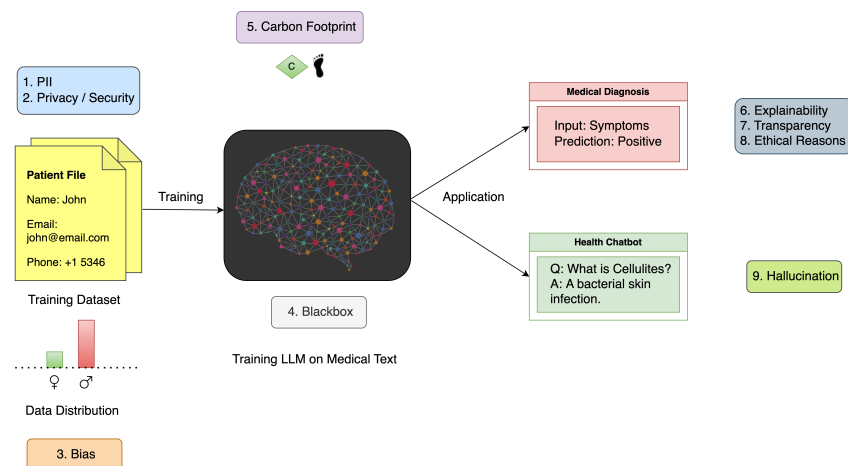

Figure 5. Challenges of large language models in healthcare.

4.1. Model Explainability and Transparency

Large language models face notable challenges when applied to healthcare. Their recommendations often lack transparency due to their opaque nature, which can hinder acceptance among healthcare professionals who prioritize explainability in medical decision-making. Moreover, the presence of biases in the training data may compromise

the accuracy of these models, potentially leading to incorrect diagnoses or treatment recommendations. As such, it is crucial for medical professionals to exercise caution and thoroughly review and validate the recommendations provided by large language models before integrating them into their clinical decision-making processes [124]. In healthcare, the importance of interpretability and explainability for AI models utilized in medical imaging analysis and clinical risk prediction cannot be overstated. Inadequate transparency and explainability have the potential to undermine trustworthiness and hinder the validation of clinical recommendations. Consequently, effective governance underscores the continuous pursuit of transparency and interpretable frameworks, aiming to augment the decision-making process in the realm of healthcare [105]. LLMs often function as "black boxes", meaning that it is challenging to discern the underlying processes leading to their specific conclusions or suggestions. In the healthcare context, where the repercussions of decisions are profound, it becomes imperative for practitioners to grasp the logic behind AI-generated outputs. The persistent endeavor to create models that are more interpretable and transparent remains an enduring challenge within the healthcare domain [125–127].

4.2. Security and Privacy Considerations

Large Language Models are used in medical research, which necessitates careful consideration of data privacy and security issues. Researchers are entrusted with the duty of managing extremely private patient data while enforcing rigorous compliance with current privacy laws. The use of LLMs in this setting raises concerns about a number of aspects of data processing, including data protection, the possibility of re-identification, and the moral application of patient data. One notable issue is the inadvertent inclusion of Personally Identifiable Information (PII) within pretraining datasets, which can compromise patient confidentiality. Additionally, LLMs can make privacy-invading inferences by deducing sensitive personal attributes from seemingly innocuous data, potentially violating individual privacy [128]. Implementing strong measures such as data anonymization, safe data storage procedures, and steadfast adherence to ethical standards are essential to addressing these issues. Together, these steps make up crucial safeguards meant to protect research participants' trust, maintain the integrity of research processes, and protect patient privacy. The importance of these factors is underscored by the necessity of balancing the significant contributions of LLMs in medical research with the critical requirement to protect private patient information [129]. The ability of LLMs to find potentially revealing patterns in large amounts of health data poses a serious privacy risk even when these data are anonymized, necessitating strict regulations and technical protections. More effective anonymization of data is crucial, along with algorithms designed to spot and prevent the re-identification of individuals. Ongoing monitoring of what LLMs produce is vital to ensure that privacy is not accidentally compromised. Implementing these measures can help to guarantee responsible use of sensitive data, allowing LLMs to be used ethically in healthcare while still respecting patient privacy. To ensure the ethical use of LLMs in healthcare, strong governance frameworks must extend beyond basic privacy laws. Proactive policies should anticipate challenges, and experts need to verify LLMs meet ethical guidelines. Engaging patients and healthcare providers in the development process promotes transparency and maintains trust in how health data is used within these systems.

4.3. Bias and Fairness

Researching ways to tackle and reduce biases in language models while also comprehending their ethical ramifications represents a pivotal research domain. It is imperative to create techniques for identifying, alleviating, and forestalling biases in large language models. A primary concern associated with LLMs pertains to the risk of producing misinformation or biased outputs. These models draw from extensive text data encompassing both dependable and unreliable sources, which can inadvertently result in the generation of inaccurate or misleading information. Furthermore, if the training data incorporate

biases, for example gender or racial biases prevalent within the scientific literature, LLMs can perpetuate and magnify these biases in their generated content.

To ensure the reliability and accuracy of information derived from LLMs, researchers must exercise caution and implement rigorous validation and verification processes. LLMs have the potential to amplify pre-existing biases inherent in their training data, particularly those linked to demographics, disease prevalence, and treatment outcomes. Consequently, the generated outputs may inadvertently reflect and perpetuate these biases, posing considerable challenges in achieving equitable and unbiased healthcare outcomes.

In order to address these challenges, researchers must remain vigilant in recognizing and mitigating biases within both the training data and the outputs generated by LLMs. This diligence is crucial for promoting fairness and inclusivity within the realm of biomedical research and healthcare applications, which will ultimately enhance the ethical and equitable utility of LLMs in these domains [129]. Prioritizing bias mitigation in LLMs is essential. Researchers should curate and preprocess training data diligently to reduce inherent biases and address sources of inequality. Routine audits and evaluations are necessary to identify and correct biases in model training and deployment. Collaborative efforts between domain experts, ethicists, and data scientists can establish guidelines and best practices for unbiased LLM development, fostering fairness and inclusivity in biomedical research and healthcare.

4.4. Hallucinations and Fabricated Information

Language models exhibit a proclivity for generating erroneous content, commonly referred to as hallucinations. This phenomenon is characterized by the production of text that appears plausible but lacks factual accuracy. This inherent trait poses a substantial risk when such generated content is employed for critical purposes, such as furnishing medical guidance or contributing to clinical decision-making processes. The consequences of relying on hallucinatory information in healthcare contexts can be profoundly detrimental, potentially leading to harmful or even catastrophic outcomes [130].

The gravity of this issue is exacerbated by the continuous advancement of LLMs, which continually enhance their capacity to generate increasingly persuasive and believable hallucinations. Moreover, LLMs are often critiqued for their opacity, as they provide no discernible link to the original source of information, thereby creating a formidable barrier to the verification of the content they produce. To mitigate these risks, healthcare professionals must exercise extreme caution when utilizing LLMs to inform their decision-making processes and should rigorously validate the accuracy and reliability of the generated information.

Current research endeavors are dedicated to addressing hallucination issues within LLMs in the healthcare and medical domain. The introduction of Med-HALT, a novel benchmark dataset, serves the purpose of evaluating hallucination phenomena in LLMs in medical contexts. Med-HALT encompasses two distinct test categories, namely, reasoning-based and memory-based hallucination assessments. These tests have been meticulously designed to gauge the problem-solving and information retrieval capabilities of LLMs when operating within the medical domain [40].

4.5. Legal and Ethical Reasons

Ethical concerns extend to the generation of potentially harmful content by LLMs, especially when delivering distressing medical diagnoses without providing adequate emotional support. Moreover, the blurring line between LLM-generated and human-written text poses a risk of misinformation dissemination, plagiarism, and impersonation.

To address these challenges, rigorous auditing and evaluation of LLMs is essential, along with the development of regulations for their medical use. Thoughtful selection of training datasets, particularly within the medical domain, is crucial to ensure the responsible handling of sensitive data. These measures collectively strive to strike a balance between harnessing LLMs' potential and safeguarding patient privacy and ethical standards [128].

The European Union's AI Act and the United States' Health Insurance Portability and Accountability Act (HIPAA) are two significant regulatory frameworks impacting the deployment of AI in healthcare. The AI Act introduces comprehensive regulations, including the Artificial Intelligence Liability Directive (AILD), which addresses liability for AI-related damages. This directive ensures that victims are compensated and that preventive measures are cost-effective. The AI Act classifies General-Purpose AI (GPAI) models and imposes specific obligations on providers, including technical documentation, risk assessments, and transparency around training data [131].

In the United States, HIPAA sets stringent standards for the protection of patient data, impacting how LLMs handle sensitive information. Compliance with HIPAA requires robust data encryption, regular security assessments, and strict access controls to protect patient information. These regulations ensure that LLMs used in healthcare settings adhere to high standards of privacy and security, mitigating risks associated with data breaches and unauthorized access.

Other relevant laws and compliance frameworks include the General Data Protection Regulation (GDPR) in the EU, which emphasizes data protection and privacy, and the Medical Device Regulation (MDR), which ensures the safety and efficacy of AI-driven medical devices. These regulations collectively impact the deployment of generative AI in healthcare by ensuring legal accountability, protecting patient data, and promoting ethical standards in AI development and application.

The implementation of regulatory frameworks such as the EU's AI Act, HIPAA, the GDPR, and the MDR significantly impacts the deployment of LLMs and generative AI in healthcare by ensuring transparency, data protection, and patient safety. These regulations necessitate detailed documentation of AI models, advanced data encryption, strict access controls, and rigorous clinical testing, thereby increasing development costs and timelines. However, they also promote reliability, legal accountability, and ethical standards in AI development, fostering trust among users and stakeholders and encouraging the responsible and wider adoption of AI technologies in healthcare [132].

5. Conclusions

In conclusion, the integration of large language models in healthcare showcases immense potential for enhancing clinical language understanding and medical applications. These models offer versatility and sophistication and can be applied to tasks from named entity recognition to question-answering, bolstering decision support and information retrieval. Comparative analyses of state-of-the-art LLMs and open-source options emphasize their significance in healthcare, promoting innovation and collaboration. Performance metrics drive continuous improvement but call for rigorous evaluation standards, considering potential biases and ethical concerns. However, several challenges persist, including the need for robust training data, bias mitigation, and data privacy. The use of LLMs in healthcare necessitates further research and interdisciplinary cooperation. While LLMs promise transformative benefits, their full potential hinges on addressing these challenges while upholding ethical standards. The ongoing journey of LLMs in healthcare demands collective efforts to harness their power for improved patient care while ensuring ethical and responsible application.

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

References

1. Shi, H.; Peng, W.; Chen, H.; Liu, X.; Zhao, G. Multiscale 3D-shift graph convolution network for emotion recognition from human actions. *IEEE Intell. Syst.* **2022**, *37*, 103–110. [CrossRef]
2. Yu, H.; Cheng, X.; Peng, W.; Liu, W.; Zhao, G. Modality unifying network for visible-infrared person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 11185–11195.
3. Li, Y.; Peng, W.; Zhao, G. Micro-expression action unit detection with dual-view attentive similarity-preserving knowledge distillation. In Proceedings of the 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 15–18 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
4. Hong, X.; Peng, W.; Harandi, M.; Zhou, Z.; Pietikäinen, M.; Zhao, G. Characterizing subtle facial movements via Riemannian manifold. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2019**, *15*, 94. [CrossRef]
5. He, K.; Mao, R.; Lin, Q.; Ruan, Y.; Lan, X.; Feng, M.; Cambria, E. A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics. *arXiv* **2023**, arXiv:2310.05694.
6. Wang, Y.; Zhao, Y.; Petzold, L. Are Large Language Models Ready for Healthcare? A Comparative Study on Clinical Language Understanding. *arXiv* **2023**, arXiv:2304.05368.
7. Yu, P.; Xu, H.; Hu, X.; Deng, C. Leveraging generative AI and large Language models: A Comprehensive Roadmap for Healthcare Integration. *Healthcare* **2023**, *11*, 2776. [CrossRef] [PubMed]
8. Peng, W.; Feng, L.; Zhao, G.; Liu, F. Learning optimal k-space acquisition and reconstruction using physics-informed neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20794–20803.
9. Peng, W.; Adeli, E.; Bosschieter, T.; Park, S.H.; Zhao, Q.; Pohl, K.M. Generating realistic brain mris via a conditional diffusion probabilistic model. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, BC, Canada, 8–12 October 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 14–24.
10. Brown, T.; 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. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
11. OpenAI. GPT-4 Technical Report. 2023. Available online: https://arxiv.org/abs/2303.08774 (accessed on 8 July 2024).
12. Zhang, C.; Zhang, C.; Li, C.; Qiao, Y.; Zheng, S.; Dam, S.K.; Zhang, M.; Kim, J.U.; Kim, S.T.; Choi, J.; et al. One small step for generative AI, one giant leap for agi: A complete survey on chatgpt in aigc era. *arXiv* **2023**, arXiv:2304.06488.
13. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
14. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
15. Huang, K.; Altosaar, J.; Ranganath, R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv* **2019**, arXiv:1904.05342.
16. Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A.H.; Riedel, S. Language models as knowledge bases? *arXiv* **2019**, arXiv:1909.01066.
17. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://api.semanticscholar.org/CorpusID:49313245 (accessed on 8 July 2024)
18. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. Palm: Scaling language modeling with pathways. *arXiv* **2022**, arXiv:2204.02311.
19. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
21. Fedus, W.; Zoph, B.; Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* **2022**, *23*, 5232–5270.
22. Du, N.; Huang, Y.; Dai, A.M.; Tong, S.; Lepikhin, D.; Xu, Y.; Krikun, M.; Zhou, Y.; Yu, A.W.; Firat, O.; et al. Glam: Efficient scaling of language models with mixture-of-experts. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 5547–5569.
23. Wang, H.; Li, J.; Wu, H.; Hovy, E.; Sun, Y. Pre-trained language models and their applications. *Engineering* **2023** , *25*, 51–65. [CrossRef]
24. Wei, J.; Bosma, M.; Zhao, V.Y.; Guu, K.; Yu, A.W.; Lester, B.; Du, N.; Dai, A.M.; Le, Q.V. Finetuned language models are zero-shot learners. *arXiv* **2021**, arXiv:2109.01652.
25. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
26. Rawte, V.; Sheth, A.; Das, A. A survey of hallucination in large foundation models. *arXiv* **2023**, arXiv:2309.05922.
27. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A Survey on Multimodal Large Language Models. *arXiv* **2023**, arXiv:2306.13549.

28. Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; Duan, N. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv* **2023**, arXiv:2303.04671.

29. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023 ; pp. 19730–19742.

30. Zong, Z.; Ma, B.; Shen, D.; Song, G.; Shao, H.; Jiang, D.; Li, H.; Liu, Y. Mova: Adapting mixture of vision experts to multimodal context. *arXiv* **2024**, arXiv:2404.13046.

31. Lin, B.; Tang, Z.; Ye, Y.; Cui, J.; Zhu, B.; Jin, P.; Zhang, J.; Ning, M.; Yuan, L. Moe-llava: Mixture of experts for large vision-language models. *arXiv* **2024**, arXiv:2401.15947.

32. Li, J.; Wang, X.; Zhu, S.; Kuo, C.W.; Xu, L.; Chen, F.; Jain, J.; Shi, H.; Wen, L. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *arXiv* **2024**, arXiv:2405.05949.

33. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940 . [CrossRef]

34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.

35. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.

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

37. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. Towards expert-level medical question answering with large language models. *arXiv* **2023**, arXiv:2305.09617.

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

39. Liu, Z.; Yu, X.; Zhang, L.; Wu, Z.; Cao, C.; Dai, H.; Zhao, L.; Liu, W.; Shen, D.; Li, Q.; et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv* **2023**, arXiv:2303.11032.

40. Umapathi, L.K.; Pal, A.; Sankarasubbu, M. Med-halt: Medical domain hallucination test for large language models. *arXiv* **2023**, arXiv:2307.15343.

41. Zhao, Z.; Wang, S.; Gu, J.; Zhu, Y.; Mei, L.; Zhuang, Z.; Cui, Z.; Wang, Q.; Shen, D. ChatCAD+: Towards a Universal and Reliable Interactive CAD using LLMs. *arXiv* **2023**, arXiv:2305.15964.

42. Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.Y. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinform.* **2022**, *23*, bbac409. [CrossRef] [PubMed]

43. Yang, X.; Chen, A.; PourNejatian, N.; Shin, H.C.; Smith, K.E.; Parisien, C.; Compas, C.; Martin, C.; Flores, M.G.; Zhang, Y.; et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv* **2022**, arXiv:2203.03540.

44. Yuan, H.; Yuan, Z.; Gan, R.; Zhang, J.; Xie, Y.; Yu, S. BioBART: Pretraining and evaluation of a biomedical generative language model. *arXiv* **2022**, arXiv:2204.03905.

45. Lu, Q.; Dou, D.; Nguyen, T. ClinicalT5: A generative language model for clinical text. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP, Abu Dhabi, United Arab Emirates, 7–11 December 2022 ; pp. 5436–5443.

46. Yuan, Z.; Liu, Y.; Tan, C.; Huang, S.; Huang, F. Improving biomedical pretrained language models with knowledge. *arXiv* **2021**, arXiv:2104.10344.

47. Raj, D.; Sahu, S.; Anand, A. Learning local and global contexts using a convolutional recurrent network model for relation classification in biomedical text. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 311–321.

48. Lyu, C.; Chen, B.; Ren, Y.; Ji, D. Long short-term memory RNN for biomedical named entity recognition. *BMC Bioinform.* **2017**, *18*, 462 . [CrossRef] [PubMed]

49. Dasgupta, I.; Lampinen, A.K.; Chan, S.C.; Creswell, A.; Kumaran, D.; McClelland, J.L.; Hill, F. Language models show human-like content effects on reasoning. *arXiv* **2022**, arXiv:2207.07051.

50. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. [CrossRef] [PubMed]

51. Chen, Z.; Micsinai Balan, M.; Brown, K. Language models are few-shot learners for prognostic prediction. *arXiv* **2023**, arXiv:2302.12692 .

52. Xue, V.W.; Lei, P.; Cho, W.C. The potential impact of ChatGPT in clinical and translational medicine. *Clin. Transl. Med.* **2023**, *13*, e1206. [CrossRef]

53. Chen, Z.; Balan, M.M.; Brown, K. Boosting Transformers and Language Models for Clinical Prediction in Immunotherapy. *arXiv* **2023**, arXiv:2302.12692.

54. Li, H.; Gerkin, R.C.; Bakke, A.; Norel, R.; Cecchi, G.; Laudamiel, C.; Niv, M.Y.; Ohla, K.; Hayes, J.E.; Parma, V.; et al. Text-based predictions of COVID-19 diagnosis from self-reported chemosensory descriptions. *Commun. Med.* **2023**, *3*, 104. [CrossRef] [PubMed]

55. Mao, C.; Xu, J.; Rasmussen, L.; Li, Y.; Adekkanattu, P.; Pacheco, J.; Bonakdarpour, B.; Vassar, R.; Shen, L.; Jiang, G.; et al. AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease. *J. Biomed. Inform.* **2023**, *144*, 104442. [CrossRef] [PubMed]

56. Agbavor, F.; Liang, H. Predicting dementia from spontaneous speech using large language models. *PLoS Digit. Health* **2022**, *1*, e0000168. [CrossRef] [PubMed]

57. Bill, D.; Eriksson, T. *Fine-Tuning a LLM Using Reinforcement Learning from Human Feedback for a Therapy Chatbot Application*; KTH: Stockholm, Sweden, 2023.

58. Balas, M.; Ing, E.B. Conversational ai models for ophthalmic diagnosis: Comparison of chatgpt and the isabel pro differential diagnosis generator. *JFO Open Ophthalmol.* **2023**, *1*, 100005. [CrossRef]

59. Lai, T.; Shi, Y.; Du, Z.; Wu, J.; Fu, K.; Dou, Y.; Wang, Z. Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models. *arXiv* **2023**, arXiv:2307.11991.

60. Bilal, M.; Jamil, Y.; Rana, D.; Shah, H.H. Enhancing Awareness and Self-diagnosis of Obstructive Sleep Apnea Using AI-Powered Chatbots: The Role of ChatGPT in Revolutionizing Healthcare. *Ann. Biomed. Eng.* **2023**, *52*, 136–138. [CrossRef] [PubMed]

61. Javaid, M.; Haleem, A.; Singh, R.P. ChatGPT for healthcare services: An emerging stage for an innovative perspective. *Benchcouncil Trans. Benchmarks Stand. Eval.* **2023**, *3*, 100105. [CrossRef]

62. Ali, S.R.; Dobbs, T.D.; Hutchings, H.A.; Whitaker, I.S. Using ChatGPT to write patient clinic letters. *Lancet Digit. Health* **2023**, *5*, e179–e181. [CrossRef] [PubMed]

63. Nguyen, J.; Pepping, C.A. The application of ChatGPT in healthcare progress notes: A commentary from a clinical and research perspective. *Clin. Transl. Med.* **2023**, *13*, e1324. [CrossRef] [PubMed]

64. Walker, H.L.; Ghani, S.; Kuemmerli, C.; Nebiker, C.A.; Müller, B.P.; Raptis, D.A.; Staubli, S.M. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *J. Med. Internet Res.* **2023**, *25*, e47479. [CrossRef] [PubMed]

65. Iftikhar, L.; Iftikhar, M.F.; Hanif, M.I. Docgpt: Impact of chatgpt-3 on health services as a virtual doctor. *Paediatrics* **2023**, *12*, 45–55.

66. Yang, H.; Li, J.; Liu, S.; Du, L.; Liu, X.; Huang, Y.; Shi, Q.; Liu, J. Exploring the Potential of Large Language Models in Personalized Diabetes Treatment Strategies. *medRxiv* **2023**. [CrossRef]

67. Wang, S.; Zhao, Z.; Ouyang, X.; Wang, Q.; Shen, D. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv* **2023**, arXiv:2302.07257.

68. Sorin, V.; Barash, Y.; Konen, E.; Klang, E. Large language models for oncological applications. *J. Cancer Res. Clin. Oncol.* **2023**, *149*, 9505–9508. [CrossRef] [PubMed]

69. Matin, R.N.; Linos, E.; Rajan, N. Leveraging large language models in dermatology. *Br. J. Dermatol.* **2023**, *189*, 253–254. [CrossRef] [PubMed]

70. Sallam, M. The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *medRxiv* **2023**. [CrossRef]

71. Tang, L.; Sun, Z.; Idnay, B.; Nestor, J.G.; Soroush, A.; Elias, P.A.; Xu, Z.; Ding, Y.; Durrett, G.; Rousseau, J.F.; et al. Evaluating large language models on medical evidence summarization. *NPJ Digit. Med.* **2023**, *6*, 158. [CrossRef]

72. Liu, Z.; Roberts, R.A.; Lal-Nag, M.; Chen, X.; Huang, R.; Tong, W. AI-based language models powering drug discovery and development. *Drug Discov. Today* **2021**, *26*, 2593–2607. [CrossRef] [PubMed]

73. Datta, T.T.; Shill, P.C.; Al Nazi, Z. Bert-d2: Drug-drug interaction extraction using bert. In Proceedings of the 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, 21–22 January 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.

74. Grisoni, F. Chemical language models for de novo drug design: Challenges and opportunities. *Curr. Opin. Struct. Biol.* **2023**, *79*, 102527. [CrossRef] [PubMed]

75. Uludoğan, G.; Ozkirimli, E.; Ulgen, K.O.; Karalı, N.; Özgür, A. Exploiting pretrained biochemical language models for targeted drug design. *Bioinformatics* **2022**, *38*, ii155–ii161. [CrossRef] [PubMed]

76. Ma, L.; Han, J.; Wang, Z.; Zhang, D. CephGPT-4: An Interactive Multimodal Cephalometric Measurement and Diagnostic System with Visual Large Language Model. *arXiv* **2023**, arXiv:2307.07518.

77. Khader, F.; Mueller-Franzes, G.; Wang, T.; Han, T.; Arasteh, S.T.; Haarburger, C.; Stegmaier, J.; Bressem, K.; Kuhl, C.; Nebelung, S.; et al. Medical Diagnosis with Large Scale Multimodal Transformers–Leveraging Diverse Data for More Accurate Diagnosis. *arXiv* **2022**, arXiv:2212.09162.

78. Thawkar, O.; Shaker, A.; Mullappilly, S.S.; Cholakkal, H.; Anwer, R.M.; Khan, S.; Laaksonen, J.; Khan, F.S. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv* **2023**, arXiv:2306.07971.

79. Liu, J.; Hu, T.; Zhang, Y.; Gai, X.; Feng, Y.; Liu, Z. A ChatGPT Aided Explainable Framework for Zero-Shot Medical Image Diagnosis. *arXiv* **2023**, arXiv:2307.01981.

80. Monajatipoor, M.; Rouhsedaghat, M.; Li, L.H.; Jay Kuo, C.C.; Chien, A.; Chang, K.W. Berthop: An effective vision-and-language model for chest X-ray disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*; Springer: Cham, Switzerland, 2022; pp. 725–734.

81. Roshanzamir, A.; Aghajan, H.; Soleymani Baghshah, M. Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 92. [CrossRef] [PubMed]

82. Giorgi, J.; Toma, A.; Xie, R.; Chen, S.; An, K.; Zheng, G.; Wang, B. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In Proceedings of the 5th Clinical Natural Language Processing Workshop, Toronto, ON, Canada, 9 July 2023; pp. 323–334.

83. Huang, G.; Li, Y.; Jameel, S.; Long, Y.; Papanastasiou, G. From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? *Comput. Struct. Biotechnol. J.* **2024**, *24*, 362–373. [CrossRef] [PubMed]

84. Thorsen-Meyer, H.C.; Placido, D.; Kaas-Hansen, B.S.; Nielsen, A.P.; Lange, T.; Nielsen, A.B.; Toft, P.; Schierbeck, J.; Strøm, T.; Chmura, P.J.; et al. Discrete-time survival analysis in the critically ill: A deep learning approach using heterogeneous data. *NPJ Digit. Med.* **2022**, *5*, 142. [CrossRef] [PubMed]

85. Zhang, A.Y.; Lam, S.S.W.; Ong, M.E.H.; Tang, P.H.; Chan, L.L. Explainable AI: Classification of MRI brain scans orders for quality improvement. In Proceedings of the 6th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, New York, NY, USA, 2 December 2019; pp. 95–102.

86. Ozyegen, O.; Kabe, D.; Cevik, M. Word-level text highlighting of medical texts for telehealth services. *Artif. Intell. Med.* **2022**, *127*, 102284. [CrossRef] [PubMed]

87. Dobrakowski, A.G.; Mykowiecka, A.; Marciniak, M.; Jaworski, W.; Biecek, P. Interpretable segmentation of medical free-text records based on word embeddings. *J. Intell. Inf. Syst.* **2021**, *57*, 447–465. [CrossRef]

88. Gao, Y.; Li, R.; Caskey, J.; Dligach, D.; Miller, T.; Churpek, M.M.; Afshar, M. Leveraging a medical knowledge graph into large language models for diagnosis prediction. *arXiv* **2023**, arXiv:2308.14321.

89. Yang, K.; Ji, S.; Zhang, T.; Xie, Q.; Kuang, Z.; Ananiadou, S. Towards interpretable mental health analysis with large language models. *arXiv* **2023**, arXiv:2304.03347.

90. Hong, S.; Xiao, L.; Zhang, X.; Chen, J. ArgMed-Agents: Explainable Clinical Decision Reasoning with Large Language Models via Argumentation Schemes. *arXiv* **2024**, arXiv:2403.06294.

91. Yang, K.; Zhang, T.; Kuang, Z.; Xie, Q.; Huang, J.; Ananiadou, S. MentaLLaMA: Interpretable mental health analysis on social media with large language models. In Proceedings of the ACM on Web Conference 2024, Singapore, 13–17 May 2024; pp. 4489–4500.

92. Savage, T.; Nayak, A.; Gallo, R.; Rangan, E.; Chen, J.H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit. Med.* **2024**, *7*, 20. [CrossRef] [PubMed]

93. Lin, B.; Xu, Y.; Bao, X.; Zhao, Z.; Zhang, Z.; Wang, Z.; Zhang, J.; Deng, S.; Yin, J. SkinGEN: An explainable dermatology diagnosis-to-generation framework with interactive vision-language models. *arXiv* **2024**, arXiv:2404.14755.

94. Lee, M.H.; Chew, C.J. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-AI collaborative clinical decision making. *Proc. ACM Hum.-Comput. Interact.* **2023**, *7*, 369. [CrossRef]

95. McInerney, D.J.; Young, G.; van de Meent, J.W.; Wallace, B.C. Chill: Zero-shot custom interpretable feature extraction from clinical notes with large language models. *arXiv* **2023**, arXiv:2302.12343.

96. Naseem, U.; Khushi, M.; Kim, J. Vision-language transformer for interpretable pathology visual question answering. *IEEE J. Biomed. Health Inform.* **2022**, *27*, 1681–1690. [CrossRef]

97. Park, S.; Kim, G.; Oh, Y.; Seo, J.; Lee, S.; Kim, J.; Moon, S.; Lim, J.; Ye, J. Vision Transformer for COVID-19 CXR Diagnosis using Chest X-ray Feature Corpus. *arXiv* **2021**, arXiv:2103.07055.

98. Pan, J. Large language model for molecular chemistry. *Nat. Comput. Sci.* **2023**, *3*, 5. [CrossRef] [PubMed]

99. Liang, J.; Wang, Z.; Ma, Z.; Li, J.; Zhang, Z.; Wu, X.; Wang, B. Online Training of Large Language Models: Learn while chatting. *arXiv* **2024**, arXiv:2403.04790.

100. Che, T.; Liu, J.; Zhou, Y.; Ren, J.; Zhou, J.; Sheng, V.S.; Dai, H.; Dou, D. Federated learning of large language models with parameter-efficient prompt tuning and adaptive optimization. *arXiv* **2023**, arXiv:2310.15080.

101. Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; Du, M. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 20. [CrossRef]

102. Kim, Y.; Xu, X.; McDuff, D.; Breazeal, C.; Park, H.W. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv* **2024**, arXiv:2401.06866.

103. Pahune, S.; Rewatkar, N. Large Language Models and Generative AI's Expanding Role in Healthcare. 2024. Available online: https://www.researchgate.net/profile/Saurabh-Pahune-2/publication/377217911_Large_Language_Models_and_Generative_AI's_Expanding_Role_in_Healthcare/links/659aad286f6e450f19d3f129/Large-Language-Models-and-Generative-AIs-Expanding-Role-in-Healthcare.pdf (accessed on 8 July 2024).

104. Reddy, S.; Rogers, W.; Makinen, V.P.; Coiera, E.; Brown, P.; Wenzel, M.; Weicken, E.; Ansari, S.; Mathur, P.; Casey, A.; et al. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform.* **2021**, *28*, e100444. [CrossRef] [PubMed]

105. Reddy, S. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Inform. Med. Unlocked* **2023**, *41*, 101304. [CrossRef]

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

107. Liao, W.; Liu, Z.; Dai, H.; Xu, S.; Wu, Z.; Zhang, Y.; Huang, X.; Zhu, D.; Cai, H.; Liu, T.; et al. Differentiate chatgpt-generated and human-written medical texts. *arXiv* **2023**, arXiv:2304.11567.

108. Manoel, A.; Garcia, M.d.C.H.; Baumel, T.; Su, S.; Chen, J.; Sim, R.; Miller, D.; Karmon, D.; Dimitriadis, D. Federated Multilingual Models for Medical Transcript Analysis. In Proceedings of the Conference on Health, Inference, and Learning, Honolulu, HI, USA, 23–29 July 2023 ; pp. 147–162.

109. Zhang, Y.; Nie, A.; Zehnder, A.; Page, R.L.; Zou, J. VetTag: Improving automated veterinary diagnosis coding via large-scale language modeling. *NPJ Digit. Med.* **2019**, *2*, 35. [CrossRef] [PubMed]

110. Wang, G.; Yang, G.; Du, Z.; Fan, L.; Li, X. ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. *arXiv* **2023**, arXiv:2306.09968.

111. Li, J.; Wang, X.; Wu, X.; Zhang, Z.; Xu, X.; Fu, J.; Tiwari, P.; Wan, X.; Wang, B. Huatuo-26M, a Large-scale Chinese Medical QA Dataset. *arXiv* **2023**, arXiv:2305.01526.

112. Yang, X.; Chen, A.; PourNejatian, N.; Shin, H.C.; Smith, K.E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A.B.; Flores, M.G.; et al. A large language model for electronic health records. *NPJ Digit. Med.* **2022**, *5*, 194. [CrossRef] [PubMed]

113. Crema, C.; Buonocore, T.M.; Fostinelli, S.; Parimbelli, E.; Verde, F.; Fundarò, C.; Manera, M.; Ramusino, M.C.; Capelli, M.; Costa, A.; et al. Advancing Italian Biomedical Information Extraction with Large Language Models: Methodological Insights and Multicenter Practical Application. *arXiv* **2023**, arXiv:2306.05323.

114. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.

115. Beaulieu-Jones, B.R.; Shah, S.; Berrigan, M.T.; Marwaha, J.S.; Lai, S.L.; Brat, G.A. Evaluating Capabilities of Large Language Models: Performance of GPT4 on Surgical Knowledge Assessments. *medRxiv* **2023**. [CrossRef]

116. Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; Steinhardt, J. Measuring massive multitask language understanding. *arXiv* **2020**, arXiv:2009.03300.

117. Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H.P.D.O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. Evaluating large language models trained on code. *arXiv* **2021**, arXiv:2107.03374.

118. Klu AI. MMLU Benchmark (Massive Multi-Task Language Understanding). 2024. Available online : https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu (accessed on 8 July 2024).

119. Jin, Q.; Dhingra, B.; Cohen, W.W.; Lu, X. Probing biomedical embeddings from language models. *arXiv* **2019**, arXiv:1904.02181.

120. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity prediction using deep learning. *Front. Environ. Sci.* **2016**, *3*, 80. [CrossRef]

121. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.W.; Lu, X. Pubmedqa: A dataset for biomedical research question answering. *arXiv* **2019**, arXiv:1909.06146.

122. Papers with Code. Medical Papers with Code. 2024. Available online : https://paperswithcode.com/area/medical (accessed on 8 July 2024).

123. Lee, J.; Myeong, I.S.; Kim, Y. The Drug-Like Molecule Pre-Training Strategy for Drug Discovery. *IEEE Access* **2023**, *11*, 61680–61687. [CrossRef]

124. Ali, H.; Qadir, J.; Alam, T.; Househ, M.; Shah, Z. In Proceedings of the ChatGPT and Large Language Models (LLMs) in Healthcare: Opportunities and Risks, Mount Pleasant, MI, USA, 16–17 September 2023.

125. Briganti, G. A clinician's guide to large language models. *Future Med. AI* **2023**, *1*, FMAI1. [CrossRef]

126. Bisercic, A.; Nikolic, M.; van der Schaar, M.; Delibasic, B.; Lio, P.; Petrovic, A. Interpretable Medical Diagnostics with Structured Data Extraction by Large Language Models. *arXiv* **2023**, arXiv:2306.05052.

127. Jiang, Y.; Qiu, R.; Zhang, Y.; Zhang, P.F. Balanced and Explainable Social Media Analysis for Public Health with Large Language Models. *arXiv* **2023**, arXiv:2309.05951.

128. Omiye, J.A.; Gui, H.; Rezaei, S.J.; Zou, J.; Daneshjou, R. Large language models in medicine: The potentials and pitfalls. *arXiv* **2023**, arXiv:2309.00087.

129. Thapa, S.; Adhikari, S. ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls. *Ann. Biomed. Eng.* **2023**, *51*, 2647–2651. [CrossRef] [PubMed]

130. Tian, S.; Jin, Q.; Yeganova, L.; Lai, P.T.; Zhu, Q.; Chen, X.; Yang, Y.; Chen, Q.; Kim, W.; Comeau, D.C.; et al. Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health. *arXiv* **2023**, arXiv:2306.10070.

131. Novelli, C.; Casolari, F.; Hacker, P.; Spedicato, G.; Floridi, L. Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. *arXiv* **2024**, arXiv:2401.07348.

132. Hacker, P.; Engel, A.; Mauer, M. Regulating ChatGPT and other large generative AI models. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA, 12–15 June 2023; pp. 1112–1123.