International Journal of Machine Learning and Cybernetics (2025) 16:1015–1040
https://doi.org/10.1007/s13042-024-02318-w

ORIGINAL ARTICLE

# Large language models for medicine: a survey

**Yanxin Zheng[1] · Wensheng Gan[1] · Zefeng Chen[1] · Zhenlian Qi[2] · Qian Liang[3] · Philip S. Yu[4]**

Received: 28 March 2024 / Accepted: 6 August 2024 / Published online: 19 August 2024


**Abstract**
To address challenges in the digital economy's landscape of digital intelligence, large language models (LLMs) have been developed. Improvements in computational power and available resources have significantly advanced LLMs, allowing their integration into diverse domains for human life. Medical LLMs are essential application tools with potential across various medical scenarios. In this paper, we review LLM developments, focusing on the requirements and applications of medical LLMs. We provide a concise overview of existing models, aiming to explore advanced research directions and benefit researchers for future medical applications. We emphasize the advantages of medical LLMs in applications, as well as the challenges encountered during their development. Finally, we suggest directions for technical integration to mitigate challenges and potential research directions for the future of medical LLMs, aiming to meet the demands of the medical field better.

**Keywords** Artificial intelligence · Medical large language models · Healthcare applications · Ethical considerations · Potential directions

## 1 Introduction

The medical domain experiences a myriad of challenges, driven by the rapid growth of data and the need for improved patient care and medical research [1]. As traditional methods struggle to cope with the vast volumes of information and intricate medical terminology, artificial intelligence (AI) technologies have emerged as pivotal tools in addressing these complexities [2, 3]. AI methodologies have played a significant role in revolutionizing medical information retrieval and processing. Information needs in medicine refer to the relevant information required by medical professionals, patients, and researchers in areas such as clinical practice, medical research, and health management [4, 5]. This includes case data, medical knowledge, treatment guidelines, drug information, and the latest research findings in disease prevention and health promotion. Information plays a crucial role in the medical field, contributing to accurate diagnoses, effective treatment of diseases, improved patient care, and advancements in medical research [6]. Nevertheless, despite significant advancements, existing medical information retrieval and processing systems encounter formidable obstacles. The sheer magnitude and dynamic nature of medical data present challenges for conventional search engines and databases, impairing their ability to deliver

✉ Yanxin Zheng
DuoDuOoozyx@gmail.com

✉ Wensheng Gan
wsgan001@gmail.com

Zefeng Chen
czf1027@gmail.com

Zhenlian Qi
qzlhit@gmail.com

Qian Liang
Liangqian0914@163.com

Philip S. Yu
psyu@uic.edu

[1] College of Cyber Security, Jinan University, Guangzhou 510632, China

[2] School of Information Engineering, Guangdong Eco-Engineering Polytechnic, Guangzhou 510520, China

[3] Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University), Shenzhen 518020, China

[4] Department of Computer Science, University of Illinois Chicago, Chicago, IL 60607, USA

swift and accurate results. Moreover, the intricate domain-specific language and evolving medical knowledge further exacerbate these limitations, often resulting in inaccuracies and inefficiencies in information retrieval [7].

In recent years, the emergence of large language models (LLMs) [8, 9] has represented a significant breakthrough in the fields of artificial intelligence [10] and data sciences. LLMs, based on deep learning, are specifically designed for processing and generating natural language text. They acquire language patterns and knowledge through pre-training on massive text datasets, resulting in strong performance across various natural language processing (NLP) tasks. The development of LLMs can be attributed to two crucial factors [11]. Firstly, the availability of large-scale pre-training datasets enables LLMs to acquire extensive language knowledge and patterns by leveraging a vast amount of text data from the Internet. Secondly, advancements in computational resources and technology have provided the necessary foundation for training and inference with large-scale language models. Extensive training empowers LLMs with robust reasoning capabilities, enabling them to generalize effectively and adapt to text data from diverse domains and tasks. Consequently, LLMs can infer and generate coherent and natural text, effectively handling complex semantics and language structures. By extensively learning language knowledge during the pre-training phase, LLMs retain certain capabilities to tackle unseen tasks or domains. Their zero-shot learning ability equips them with a level of versatility and adaptability. To some extent, LLMs fulfill conversational functionalities. One notable example is the generative pre-trained Transformer (GPT) model series, which adopts the Transformer architecture [12]. Transformer is a deep learning model based on attention mechanisms and demonstrates exceptional proficiency in NLP tasks. Based on the aforementioned factors, LLMs typically exhibit characteristics such as large-scale pre-training, strong generalization, contextual understanding, text generation, zero-shot learning, and certain interactive capabilities. Due to technological advancements and increased accessibility, Model-as-a-Service (MaaS) [13] has served people and vital tools and drivers of innovation across various applications, including intelligent customer service [14], healthcare [15], finance [16], investment [17], education and training [18], artistic creativity [19], and so on. As a result, the widespread use of LLMs in the medical field holds immense potential for providing intelligent support and assistance to improve healthcare quality and efficiency.

Fortunately, LLMs can effectively tackle the challenges of medical information. Firstly, by leveraging deep learning [20] and NLP [21] techniques, LLMs can deeply understand and semantically reason with medical texts. They can comprehend medical terminology, contextual relationships, and semantic structures, thereby enabling more accurate retrieval and processing of medical information. Secondly, LLMs can integrate diverse medical data sources, including medical literature, clinical guidelines, and case reports, offering comprehensive and multifaceted information. They can extract knowledge and insights from vast datasets, providing healthcare professionals and patients with more comprehensive and precise information support. Thirdly, LLMs can actively track medical literature and the latest research advancements, promptly delivering up-to-date information to healthcare professionals and patients. Furthermore, they can provide personalized recommendations and advice tailored to user needs and preferences, augmenting the relevance and practicality of the information. Therefore, researching LLMs in the medical field is necessary and crucial.

The process of training a medical LLM typically involves seven steps [22]: data collection, data preprocessing, model selection and architecture design, model training, hyperparameter tuning, validation and evaluation, and model deployment and application. During the data collection phase, a large-scale corpus of text data in the medical domain is gathered, encompassing medical literature, case reports, clinical guidelines, drug information, and more. The collected data is then preprocessed, and an appropriate model architecture is selected and designed to handle the training and processing of the medical data. The model is trained, and its parameters are iteratively optimized to enhance its proficiency in the specialized domain. Finally, the model's performance is evaluated using a validation set employing metrics such as perplexity, text quality, and accuracy. Once the model training and evaluation are complete, they can be deployed in practical applications to fulfill the information needs of medical professionals and patients. By utilizing LLMs trained with specialized data inputs and multiple training iterations, these models acquire professional judgment capabilities. As a result, LLMs can meet the demands of medical professionals and patients for accurate, timely, and reliable medical information, thereby improving the quality and efficiency of medical decision-making. For instance, considering a scenario where a doctor faces a rare case and is uncertain about the diagnosis and treatment options, the doctor can employ an LLM to quickly search for relevant medical literature, case reports, and expert opinions, gaining more insights and guidance. Leveraging the vast pre-training data and semantic understanding capabilities of LLM, the model assists the doctor in comprehending the case's characteristics, pathophysiology, and potential treatment options. Moreover, the medical LLM can provide the doctor with assessments of the most recent clinical guidelines, drug information [23], and treatment plans [24], facilitating informed decision-making. Similarly, for patients, suppose an individual has been diagnosed with a rare disease and desires to acquire more information about the condition, treatment options, and lifestyle recommendations. The patient can utilize one LLM

to input relevant keywords or questions, enabling them to access medical knowledge and research findings. The model searches for and presents the latest research discoveries, expert opinions, and information published by authoritative organizations, aiding the patient's understanding of the disease's symptoms, diagnostic methods, treatment options, and prognosis. Through interaction with the LLM, the patient can obtain easily understandable and accurate explanations, enhancing their comprehension of their condition and enabling more accurate communication with healthcare providers regarding their symptoms. Additionally, this empowers them to adopt scientifically sound treatment plans. Numerous medical LLMs have been exploited and are expected to apply to life and production [25].

Therefore, exploring the applications, limitations, and potential advancements of LLMs in the medical domain is crucial. This up-to-date survey aims to provide a comprehensive overview of the utilization of LLMs in medicine, including their benefits, challenges, and emerging trends. By analyzing existing literature and research, we seek to offer insights into the current state of LLMs in medicine, compare different approaches, and identify areas for future research and development. To refine the contributions of this paper and provide more insights for further studies, we compare the similarities and differences between this paper and other review articles on the same subject, shown in Table 1, which makes our contribution clearer. This paper's contributions can be summarized as follows:

- Comprehensive coverage. We provide an up-to-date and exhaustive survey of medical LLMs, including the advances in theory, methods, and applications.
- Progressive review. We reviewed the development stages of LLM, manifesting the respective advantages and disadvantages.
- Novel classification. According to various application fields, some representative medical LLM products are introduced, along with their training framework and process.
- Comprehensive discussion. We extensively explored the current trends in the medical grand model, delving into the opportunities and future directions it presents to assist in subsequent efforts in related domains.

*Organization*: The arrangements for this paper are as follows: In Sect. 2, we review the history of LLM and discuss applications of the medical LLM. We make a comparison of the different products of MedGPT in Sect. 3. Furthermore, we demonstrate the duality of the LLM in medicine in Sect. 4. We highlight the opportunities and provide promising directions in Sect. 5. Finally, we conclude this paper in Sect. 6. The outline of this article is shown in Fig. 1. Moreover, in this paper, many acronyms are used to represent concepts, medical terms, types of models, and frequently studied models. Table 2 provides the most common and important terms used in our paper.

## 2 Related concepts

### 2.1 Developments of LLMs

The emergence of LLMs was not a straightforward process but rather a winding and progressive journey. In the early stages, studies focused on developing generative models capable of generating text [32], images [33], audio [34], and other AI-generated content [35]. However, the performance of these models was limited due to the lack of large-scale datasets and powerful computational resources. Later, with the advent of pre-training models [36, 37], several studies began utilizing large-scale datasets for pre-training to extract statistical patterns and semantic representations from the data. The model could acquire initial parameters through pre-training and then serve as a fine-tuning model [38]. The above approach laid the foundation for subsequent model training, while numerous technical challenges in training and optimization awaited resolution. Subsequently, recurrent neural networks (RNNs) [39] were introduced as the foundational models for text generation. Although these models performed well in generating short texts and small-scale data, they encountered issues such as handling long sequences, vanishing gradients, and low computational efficiency. Nevertheless, with advancements in deep learning and computational resources, LLMs started demonstrating their immense potential [40]. Particularly, the introduction of the Transformer architecture enabled models to better capture contextual information and semantic representations of text, leading to breakthroughs in tasks such as text generation, dialogue systems, and natural language understanding. The aforementioned progressive development can be summarized in three stages: generative models, pre-training models, and autoregressive models. Meanwhile, the characteristics of LLM compared to the traditional NLP technology [41] are shown in Fig. 2.

#### 2.1.1 Generative models

Generative models [42] are probabilistic models that aim to generate new samples by generating a probability distribution that is similar to the original distribution. Through such a learning process, the model could generate new samples similar to the statistical patterns and patterns learned from large-scale datasets. In generative models, language models play an important role. The goal of a language model is to estimate the probability of a given word or a given

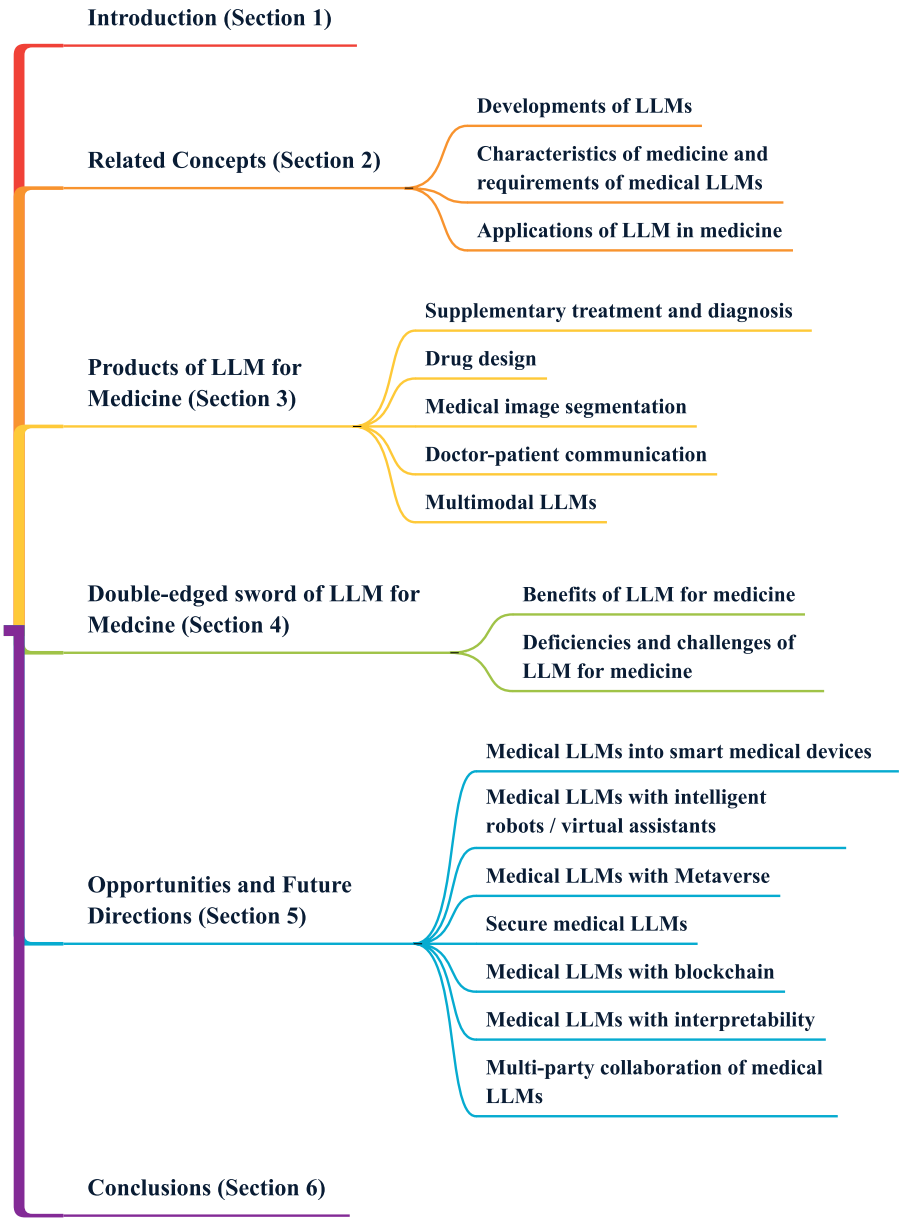


**Fig. 1** The outline of this article

sequence of words by training the model to learn the probability distribution of lexical representation in the sequences. This allows the model to predict the next word based on the previously observed words, thereby generating coherent text. Generative models generate new samples by computing the probability distribution of all the data. There are numerous typical models, such as N-gram models [43], hidden Markov models (HMMs) [44], and long short-term memory networks (LSTMs) [45].

*The N-gram model* [43] has a relatively simple model structure and computational approach. Based on the previous $n$-1 words, the $N$-gram model predicts the next word, making it effective for short text generation and simple language modeling tasks. However, the $N$-gram model only considers the preceding $n$-1 words, failing to capture longer contextual information. Moreover, in cases of sparse data, the model may encounter problems of missing data and inefficient estimation. As a result, the $N$-gram model cannot capture long-distance dependencies and is sensitive to data sparsity issues.

*HMMs* [44] are statistical models for defining transition and emission probabilities between states and observations. Because of the Markov chain structure [46] and conditional independence hypothesis, HMMs can effectively model the probability distribution of sequence data. Thus, they perform well in the tasks of labeling [47] and sequence generation [48]. However, due to the Markov property for state transitions, they ignore longer contextual information. Also, they are limited and even suffer from problems with overfitting and data drift when the distributional assumptions of the

**Table 1** Contributions and gaps of existing papers

| References | Year | One-sentence summary | Application of LLM | Comparison | Threat concerns | Solution |
|---|---|---|---|---|---|---|
| [26] | 2023 | The future landscape of LLMs in medicine | ✗ | ✗ | Ethics, training data security | ✗ |
| [27] | 2023 | LLMs in medicine | Chatbots | ✓ | Recency, accuracy, coherence, interpretability, ethics | ✓ |
| [28] | 2023 | LLMs encode clinical knowledge | ✗ | ✓ | Evaluation data expansion, clinical accuracy, human evaluation framework, fairness, ethics | ✗ |
| [29] | 2023 | Embracing LLMs for medical applications: Opportunities and challenges | ✗ | ✗ | Central role, accuracy, cost, ethics, data privacy, regulatory framework | ✓ |
| [30] | 2023 | A survey of LLMs in medicine: Progress, application, and challenge | Health Support, Report Generation, Medical robotics, Translation, Education | ✓ | Evaluation benchmarks, domain data limitations, new knowledge adaptation, behaviour alignment, ethics | ✓ |
| [31] | 2023 | Chatgpt and LLM chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine | Academic medicine | ✓ | Accuracy, ethics | ✗ |
| Our work | 2024 | LLMs in medicine: A survey | Supplementary diagnosis, drug design, medical image segmentation, doctor-patient communication, multimodal | ✓ | Computational resources, model efficiency, data-related, practical applications, fairness, privacy, accountability, patient autonomy | ✓ |

**Table 2** Important terms of acronyms and corresponding full form

| Acronym | Full Form |
|---|---|
| LLM | Large language model |
| RNN | Recurrent neural network |
| HMM | Hidden Markov model |
| LSTM | Long short-term memory network |
| ELMo | Embeddings from language model |
| GPT | Generative pre-trained transformer |
| BERT | Bidirectional encoder representations from transformer |
| CBOW | Continuous bag of words model |
| GRU | Gated recurrent unit |
| Med-PaLM | Medical prompt language model |
| MSA | Multiple sequence alignment |
| PLM | Protein language model |
| WCE | Wireless capsule endoscopy |
| CFG | Category-guided feature generation module |
| MedLSAM | Medical localize and segment anything model |
| MLLM | Multimodal large language model |
| EHR | Electronic health record |

training data are not strictly met. This happens when the distribution of the training data is different from the distribution of the data used in real life.

*LSTMs* [45] are a special type of RNN that is used for solving the problems of gradient disappearance and gradient explosion [49] during long sequence training. Gate mechanisms remember important information and forget unimportant information in sequence data. They are suitable for processing sequence data [50], generating text, and capturing long-distance dependencies. However, they also have limitations. For example, when dealing with more complex language modeling tasks, both larger-scale data and computational resources are required due to the shortage of parallel processing.

### 2.1.2 Pre-training models

The pre-training models [37] aim to address the challenges in the generative models, such as generating text with semantic coherence, generating reasonable text, and capturing semantic relationships and information in context. Pre-training models are based on neural networks. They extract semantic representations of text from large-scale corpora by learning language knowledge. This approach enables the model to have a better understanding of the data and extract features. Common pre-training models include Word-2Vec [51], Embeddings from Language Models (ELMo) [52], Generative Pre-trained Transformer (GPT) [53], Bidirectional Encoder Representations from Transformers (BERT) [54], and others. Note that the pre-training model has broader applications and can be used for various tasks

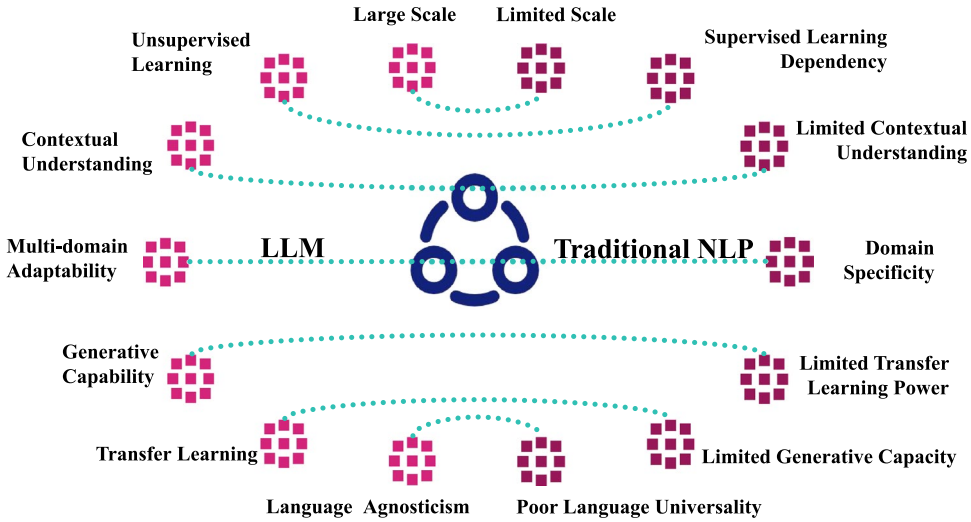


**Fig. 2** Characteristics of LLM versus the traditional NLP

such as word embedding [55], text classification [56], named entity recognition [57], data mining [58], and more.

*Word2Vec* [51] is an early milestone in pre-training models, as proposed in 2013. It is a pre-training model based on the distributed hypothesis. Word2Vec utilizes two training methods: the Continuous Bag of Words (CBOW) model [59] and the Skip-gram model [60]. The CBOW model aims to predict the center word based on the surrounding context words, while the skip-gram model predicts the context words given the center word. A neural network model is trained to map words to fixed-dimensional representations in a continuous vector space. This representation allows for capturing the semantic information of words and calculating the similarity between words. However, Word2Vec can only represent the semantics of individual words but cannot capture contextual information or handle the semantics of polysemous words.

*ELMo* [52] is a pre-training model proposed by the Allen Institute for Artificial Intelligence in 2018. It employs a bi-directional language model to generate vector representations of words. During the pre-training process, ELMo trains two language models simultaneously, one from left to right and the other from right to left, to acquire contextually relevant word representations. Ultimately, ELMo combines these representations to form the final vector representation for each word. This context-sensitive representation enables ELMo to better capture word polysemy and contextual relevance. However, ELMo, based on bi-directional language models, has slower training and inference processes and cannot handle sentence-level semantic relationships.

*GPT* [53] was first introduced by OpenAI in 2018 and has been continuously improved in subsequent versions. It utilizes the transformer model structure [61] and is trained in an autoregressive manner. GPT constructs its model by stacking multiple transformer encoder layers, where each layer can leverage the contextual information from the preceding text. Through large-scale unsupervised pre-training, it learns language models that can be fine-tuned for various downstream tasks, such as text classification and machine translation. GPT aims to address text generation and language understanding challenges. GPT is capable of generating coherent text and comprehending contextual information when trained in an autoregressive manner. However, GPT still has room for improvement as it lacks precise control over input sentences, and the generated results may occasionally lack accuracy.

*BERT* [54] is a pre-training model introduced by Google in 2018. BERT uses a bi-directional transformer model, which simultaneously predicts both left and right words based on context. Unlike GPT, BERT takes into account the contextual information from both the left and right sides. During the pre-training process, BERT predicts the masked words by masking a portion of the input words. Furthermore, BERT incorporates sentence-level pre-training tasks, such as predicting sentence continuity. This bi-directional training approach allows BERT to have a better understanding of contextual information and handle polysemous words and ambiguous sentences, thereby providing more accurate semantic representations. However, BERT cannot directly handle text-generation tasks.

#### 2.1.3 Autoregressive models

The autoregressive model is part of pre-training models, although they differ in their objectives. The autoregressive model aims to generate coherent text by predicting the next word or character based on context information. In contrast, the objective of the pre-training model is to learn language

representations for fine-tuning or transfer learning in subsequent tasks. The autoregressive model is trained through supervised learning and requires a large amount of labeled data. Conversely, the pre-training model is unsupervised and utilizes large-scale unlabeled text data for pre-training, without specific annotations. The autoregressive model excels in text generation tasks and is commonly used in natural language generation [62], and similar tasks.

In the autoregressive model, text generation is accomplished by sequentially predicting the next word or character. The model generates the next element in the sequence using the current context information and adds it to the generated sequence. This recursive generation approach enables the model to maintain contextual coherence. Representative models in this stage include RNN and its variants, such as LSTM and gated recurrent units (GRU) [63]. On the other hand, the pre-training model GPT is based on the transformer architecture and utilizes autoregressive training. This model combines the advantages of both the autoregressive model and the pre-training model.

*Rerrent Neural Network (RNN)* [64] is one of the earliest autoregressive models applied to natural language processing tasks. RNN can handle variable-length sequential data and capture contextual information within the sequence. It has achieved certain accomplishments in language modeling and sequence generation tasks. The basic idea is to capture contextual information within the sequence by recursively passing hidden states. However, traditional RNNs suffer from the problem of vanishing or exploding gradients when dealing with long-term dependencies.

*Gate Recurrent Unit (GRU)* [63] is another improved RNN structure that simplifies LSTM's gating mechanism. Compared to LSTM, GRU has fewer parameters, higher computational efficiency, and comparable performance on certain tasks. However, GRU still cannot fully address the issue of long-term dependencies and may perform poorly on tasks that require handling long-distance dependencies.

*Transformer* [65] is a model that uses a self-attention mechanism and introduces the encoder-decoder structure for sequence-to-sequence tasks. The transformer does not rely on recursive structures and can be computed in parallel, considering all positions in the input sequence simultaneously, regardless of sequence length. Therefore, it is more efficient and effective in handling long texts and modeling long-distance dependencies. However, transformer also suffers from the drawback of higher computational complexity, especially when dealing with longer sequences, which consume significant computational resources. Additionally, the transformer is sensitive to positional information within the sequence and may require additional encoding to retain positional information. GPT is an autoregressive model based on the Transformer architecture.

## 2.2 Characteristics of medicine and requirements of medical LLMs

*Compassionate care.* Medicine emphasizes compassionate care for patients [66], including establishing trust, respecting patient rights and dignity, and addressing patients' psychological and social aspects. Therefore, when interacting with patients, medical LLMs need to embody characteristics of compassionate care, such as respecting patient privacy and addressing emotional and social needs.

*Interpretability.* Medicine encompasses a broad range of knowledge, and medical decisions are made based on this knowledge. Given the critical nature of medical decision-making, results from medical LLMs need to possess good interpretability [67]. This ensures that both healthcare professionals and patients can understand the model's reasoning process and conclusions, thereby enhancing trust.

*Practice-oriented.* Medicine is a practice-oriented discipline [68], where the application of medical knowledge is typically validated through clinical practice. To ensure the models' practicality and applicability, the design and application of medical LLMs must align with medical practices, by taking into account real-world clinical scenarios and healthcare services.

*Team collaboration.* Collaboration within healthcare teams is essential in the field of medicine [69]. Various healthcare professionals, including doctors, nurses, laboratory technicians, and rehabilitation specialists, need to work together to provide comprehensive healthcare services. Recognizing the collaborative nature of the medical field, medical LLMs should be capable of collaborating with other healthcare professionals, supporting integrated decision-making and services within multidisciplinary healthcare teams.

*Ethical challenges.* Medicine involves numerous ethical considerations, encompassing issues such as patient privacy, bioethics, and treatment decisions. Healthcare professionals must confront and address these ethical challenges to ensure fairness and ethical practices in medical care. Similarly, the design and application of medical LLMs should adhere to medical ethical principles [70], emphasizing patient privacy and ensuring fairness and ethical decision-making processes.

*Uncertainty and complexity.* The human body is an incredibly complex system, with numerous factors influencing health and disease, leading to inherent uncertainty in medical practice. When faced with complex cases, medical professionals must make comprehensive judgments [71]. Therefore, medical LLMs should effectively handle this uncertainty, providing probabilistic results and decision-making capabilities.

*Diverse fields.* Medicine encompasses a variety of specialized domains, including internal medicine, surgery, obstetrics, gynecology, and others. Medical LLMs need to

provide tailored support based on the specialized knowledge of different fields, thereby assisting doctors in making accurate diagnoses and treatment decisions within specific domains. Medicine involves addressing a variety of ailments and causes, including genetic diseases, infectious diseases, chronic conditions, and more. Medical LLMs must offer personalized medical recommendations for different ailments and causes, enhancing the specificity of patient treatment plans. The medical process encompasses multiple stages, including prevention, diagnosis, treatment, and rehabilitation. Medical LLMs need to provide support at each stage, such as formulating personalized prevention plans, aiding in clinical decision-making, and offering rehabilitation advice, thereby comprehensively serving patients throughout their entire medical journey. As shown in Table 3, medical LLMs can provide assistance tailored to different medical fields.

## 2.3 Applications of LLM in medicine

Medical LLM exhibits potential capability in various fields, including dentistry [22], radiology [78], nuclear [79], clinical [28], and drug design [80]. It employs common effects and unique functions in different fields. Generally speaking, medical LLMs can manage medical records and medical documents automatically. Compared to the traditional training model, the medical LLM showed advantages in training data, such as BioBERT [81], BioGPT [82], etc. To illustrate in detail, take BioBERT as an example. BioBERT, pre-trained on a vast biomedical corpus, exhibits exceptional adaptability to domain-specific contexts and excels in comprehending and processing intricate biomedical terminologies and entities. Its performance surpasses that of other models in various biological and medical text mining tasks, particularly in three key areas: First, BioBERT demonstrates a 0.62% improvement in the F1 score for recognizing entities within biomedical texts, including genes, proteins, and diseases, compared to alternative models. Secondly, it achieves a noteworthy 2.80% enhancement in the F1 score for extracting relationships between entities from biomedical texts relative to other models. Thirdly, BioBERT exhibits a substantial 12.24% improvement in Mean Reciprocal Rank for answering biomedical-related queries compared to its counterparts. Additionally, BioBERT's structural similarity to the widely used BERT model makes it easier to use and allows for smooth integration into a wide range of biological and medical text-mining tasks. As a result, it can be inferred that the LLM possesses a commendable degree of generalization capability on different data.

For example, in dentistry, there are two main LLM deployment methods: automated dental diagnosis and cross-modal dental diagnosis [22]. It points out how a multi-modal LLM AI system works for dentistry clinical application and exhibits the whole process. It can handle unstructured

**Table 3** Various medical fields and LLMs support

| Medical field | Patient characteristics | Typical disease types | Support by medical LLMs |
|---|---|---|---|
| Internal medicine [72] | Age, family medical history | Diabetes, hypertension, cardiovascular diseases | Formulate personalized prevention plans, assist in clinical decision-making, provide patient rehabilitation advice, and generate drug interaction assessments |
| Surgery [73] | Surgical history, trauma history | Surgical procedures, trauma | Assist in pre- and post-surgery treatment plans, offer rehabilitation advice, and provide real-time surgical guidance |
| Obstetrics and gynecology [74] | Women's age, obstetric and reproductive history | Pregnancy management, gynecological diseases | Provide pregnancy health advice, assist in the diagnosis and treatment of gynecological diseases and generate personalized contraception recommendations |
| Infectious diseases [75] | Immunization status, travel history | Infectious diseases, bacterial infections | Develop infection prevention strategies, assist in diagnosing infectious cases and generate tailored antibiotic prescriptions |
| Genetic medicine [76] | Family genetic history, genetic information | Genetic diseases | Provide genetic diagnosis and treatment recommendations for genetic diseases, and offer personalized genetic counselling |
| Chronic diseases [77] | Lifestyle, long-term medication history | Hypertension, diabetes, chronic obstructive pulmonary disease | Develop long-term treatment plans, provide advice on managing chronic diseases and offer personalized dietary and exercise recommendations |

data, and extract and integrate information for doctors and patients. Furthermore, multiple documents assist medical LLM in efficiently providing treatment methods based on patients' backgrounds. With access to abundant professional training data and the ability to capture context information, medical LLMs aid in diagnosis. Not only does LLM have the ability to read text content, but it can also give interpretations for images. For instance, in radiology [83], medical LLM strengthens communications between doctors and patients by generating simplified reports. These reports are generally considered correct by most participating radiologists. Furthermore, medical LLM can analyze medical image reports, providing deeper interpretation and background knowledge to assist doctors in diagnosing diseases more accurately [84]. Recent research has shown that medical LLMs, trained with the latest data, can provide clinical decision support based on the latest research findings, thereby benefiting clinical medicine. If combined with image interpretations, medical LLM may apply to clinical nuclear medicine to provide comprehensive information on imaging results, similar to clinical radiology [85]. In addition, the application of medical LLMs in drug design can accelerate the process of drug target discovery and help predict different characteristics of drugs [80]. The applications of LLM in medicine are shown in Fig. 3.

## 3 Products of LLM for medicine

According to incomplete statistics since 2023, the number of domestic and international large AI models released in the medical field has exceeded 30, and their application scenarios cover many aspects such as medical scientific research, drug research and development, smart diagnosis and treatment, medical equipment operation and maintenance, and hospital management. Except for the pre-trained data, the medical LLM also exhibits excellent capability of generalizability reflecting its outstanding experiment results in various application fields. In the following, some well-known LLM products for medicine are summarized in detail, the specific applications of various LLMs in each link are illustrated in Fig. 4. In addition, the classification and information of different medical LLMs are shown in Table 4.

### 3.1 Supplementary treatment and diagnosis

*BenTsao (Huatuo)* [90]. BenTsao, originally named Huatuo GPT, was trained by the research team at the Harbin Institute of Technology. BenTsao is a large-scale Chinese language model in the biomedical field, built upon the open-source LLaMa-7B model [110]. The BenTsao model incorporates structured and unstructured medical knowledge from the Chinese Medical Knowledge Graph (CMeKG), which provides medical information about diseases, drugs, symptoms, and more. The CMeKG includes over 10,000 diseases, nearly 20,000 drugs, over 10,000 symptoms, and structured knowledge descriptions for 3,000 diagnostic and therapeutic techniques. This allows for extensive knowledge associations between diseases, symptoms, drugs, and diagnostic and therapeutic techniques, with 1.56 million concept relationships and attribute triplets. The BenTsao team samples instances from the knowledge graph [111, 112] based on specific task knowledge and uses the OpenAI API (GPT 3.5) [113] to construct question-answer data around the medical knowledge base. They employ various prompt forms to fully utilize the knowledge, creating over 8,000 instruction data points for the instruction dataset used in supervised fine-tuning. For medical question-answering tasks, BenTsao introduces a new evaluation metric, the system usability

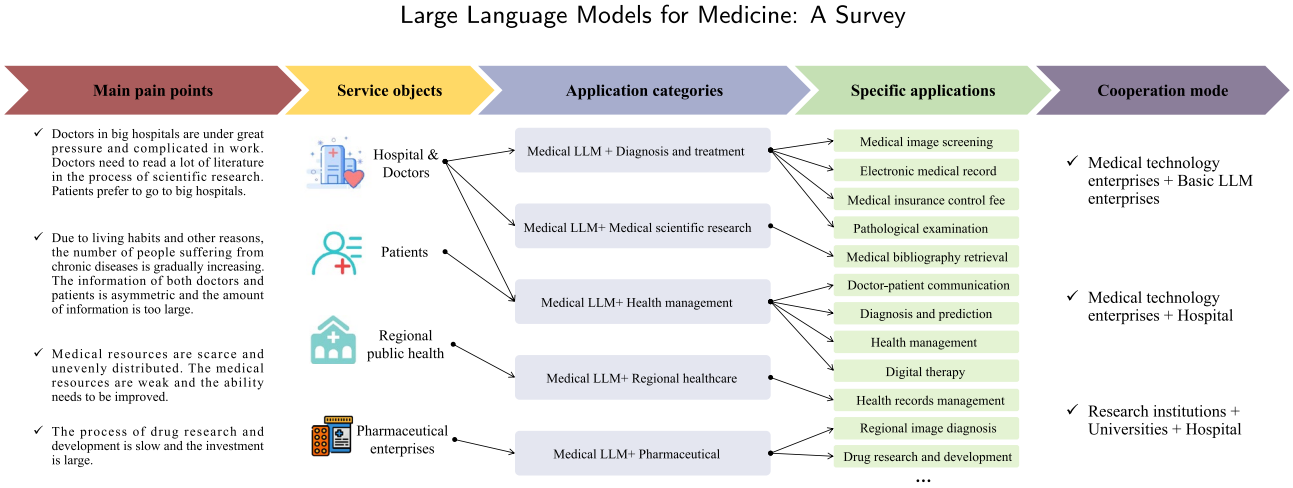


**Fig. 3** The pain points of medical treatment and the applications of medical LLMs (According to the 2023 Medical and Health AI Large Model Industry Research Report by iyiou: https://www.iyiou.com/research)

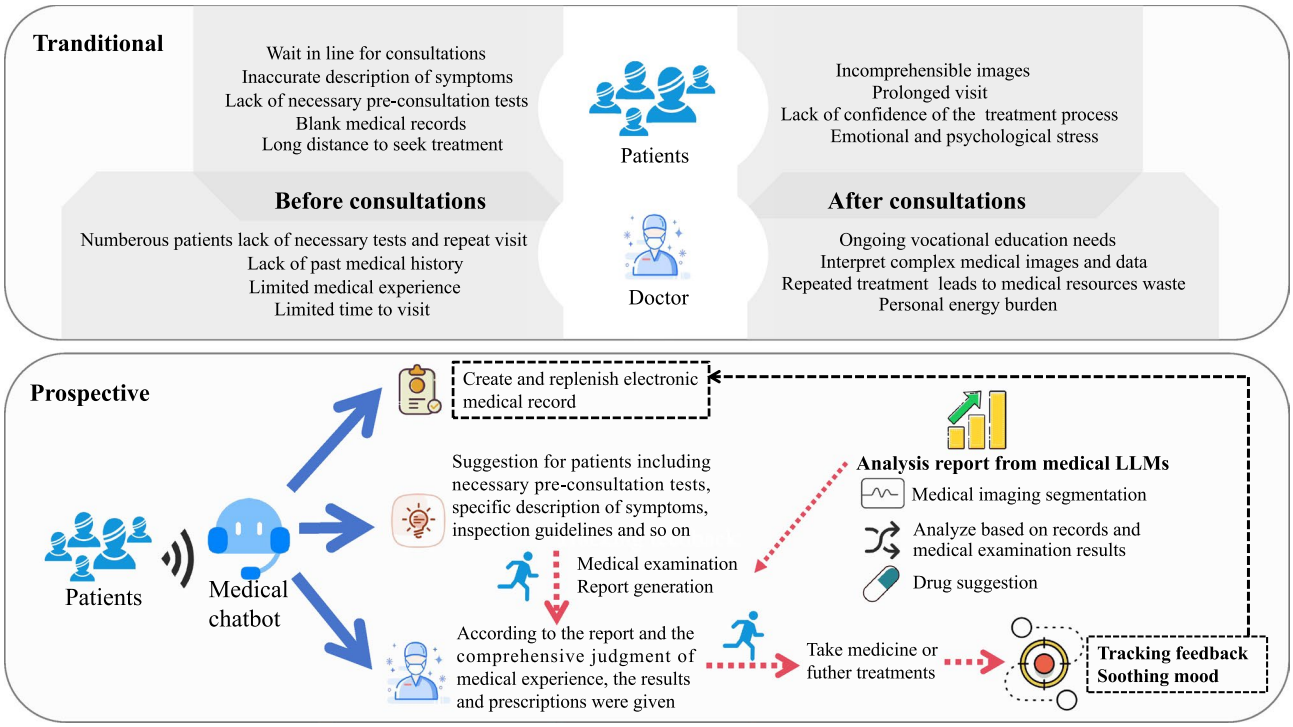


**Fig. 4** Applications of LLMs in each link in medicine

scale (SUS). The SUS metric consists of three dimensions: safety, usability, and smoothness. Safety evaluates whether the generated responses could mislead users and pose a threat to their health. Usability evaluates the degree to which the generated responses reflect medical expertise, while smoothness measures the fluency of the generated responses.

*Med-PaLM* [28]. In July 2023, Google unveiled a new evaluation benchmark for its fine-tuned medical prompt language model (Med-PaLM), known as MultiMedQA, aimed at assessing the clinical capabilities of LLMs in the field of medicine. This benchmark encompasses questions and answers from various domains, such as medical exams and research, intending to test the model's performance in handling clinically relevant issues. Based on prompt-based fine-tuning, Med-PaLM's responses were compared to consumer medical queries with those of clinical professionals within a human-assessed framework. The results demonstrated outstanding performance by Med-PaLM, confirming the effectiveness of prompt-based fine-tuning. Particularly noteworthy is Med-PaLM's achievement on the MedQA dataset, where it surpassed the "passing" score in the style of U.S. Medical Licensing Examination (USMLE) questions for the first time, achieving a score of 67.2%. However, the study highlighted that, despite significant progress, there is still room for improvement, especially when comparing model responses to those of clinical professionals. This suggests the potential for further enhancements in the quality of medical questions answered by the model.

## 3.2 Drug design

*PanGu drug model* [92]. It is a deep learning-based model designed for a variety of drug development applications. The inspiration for the model's design derives from the various representations of molecules that students learn in chemistry courses, such as molecular formulas and structural formulas. Moreover, it learns how to transform between these representations. By utilizing an asymmetric conditional variational autoencoder from graph to sequence, the PanGu drug model can effectively represent molecular features and enhance the performance of downstream drug discovery tasks. During the pre-training phase, the PanGu drug model was trained on 1.7 billion small molecules. Such extensive training allows the model to capture molecules' rich features and learn intricate relationships between them. In various drug discovery tasks, encompassing molecular property prediction, compound-target interactions, drug-drug interactions, chemical reaction yield prediction, molecular generation, and molecular optimization, the PanGu drug model has consistently achieved state-of-the-art results. Note that the PanGu drug model encompasses several distinctive functionalities. The PanGu molecule generator can generate novel compounds with analogous physicochemical

**Table 4** Classification and information of different medical LLMs

| Domain | LLM | Year | Paper | Source |
|---|---|---|---|---|
| Supplementary treatment and diagnosis | MedGPT | 2021 | [86] | https://medgpt.co/home |
| | LLM-Mini-CEX | 2023 | [87] | – |
| | WiNGPT | 2023 | – | https://github.com/winninghealth/WiNGPT2 |
| | SkinGPT-4 | 2023 | [88] | – |
| | DoctorGLM | 2023 | [89] | https://github.com/xionghonglin/DoctorGLM |
| | BenTsao (Huatuo) | 2023 | [90] | https://github.com/SCIR-HI/Huatuo-Llama-Med-Chinese |
| | ClinicalGPT | 2023 | [91] | – |
| Drug design | PanGu Drug Model | 2023 | [92] | http://www.pangu-drug.com/ |
| | HelixFold-Single | 2023 | [93] | https://github.com/PaddlePaddle/PaddleHelix/tree/dev/apps/protein_folding/helixfold-single |
| | TransAntivirus | 2023 | [94] | https://github.com/AspirinCode/TransAntivirus. |
| | OpenBioMed | 2023 | [95] | https://github.com/PharMolix/OpenBioMed |
| Medical image segmentation | DSI-Net | 2021 | [96] | https://github.com/CityU-AIM-Group/DSI-Net |
| | MedLSAM | 2023 | [97] | https://github.com/openmedlab |
| | Lvit | 2023 | [98] | https://github.com/HUANGLIZI/LViT |
| | MedCLIP-SAM | 2024 | [99] | – |
| Doctor-patient communication | BioMedLM – PubMed GPT | 2022 | [100] | https://www.mosaicml.com/blog/introducing-pubmed-gpt |
| | ChatDoctor | 2023 | [101] | https://github.com/Kent0n-Li/ChatDoctor |
| | Disc-medllm | 2023 | [102] | https://github.com/FudanDISC/DISC-MedLLM |
| | BianQue | 2023 | [103] | https://github.com/scutcyr/BianQue |
| | MeChat | 2023 | [104] | https://github.com/qiuhuachuan/smile |
| | PMC-LLaMA | 2023 | [105] | https://github.com/chaoyi-wu/PMC-LLaMA |
| Multimodal | OpenMEDLab | 2023 | – | https://github.com/openmedlab |
| | Med-MLLM | 2023 | [106] | – |
| | PeFoMed | 2024 | [107] | https://github.com/jinlHe/PeFoMed |
| Health management | CIDRS | 2021 | [108] | – |
| | GatorTron | 2022 | [109] | https://catalog.ngc.nvidia.com/orgs/nvidia/teams/clara/models/gatortron_og |
| | CareGPT | 2023 | – | https://github.com/WangRongsheng/CareGPT |
| | Bianshi | 2023 | – | https://www.a-eye.cn/technology.html#Model |

properties, thereby expanding existing compound databases. These features are invaluable for drug research, providing a broader pool of candidate compounds for further research and screening. Additionally, the PanGu molecule optimizer can refine the chemical structure of initial molecules to enhance specific molecular properties, offering a potent tool for drug design and optimization. To enhance user accessibility, the PanGu drug model has developed an automated multi-objective optimization web application.

*Large-scale protein language model (HelixFold-Single)* [93]. Compared to traditional methods dependent on multiple sequence alignment (MSA), HelixFold-Single is an innovative protein structure prediction method to overcome limitations and time constraints. The HelixFold-Single method mainly integrates a large-scale protein language model (PLM) with key components derived from AlphaFold2. Initially, self-supervised learning was employed to learn an approach to pre-train the model on the original structures of billions of proteins and create a robust protein language model. This pre-trained model serves as a substitute for conventional MSA. It enables the acquisition of shared evolutionary information from protein sequences. Subsequently, the pre-trained PLM was amalgamated with several pivotal components of AlphaFold2 to form an end-to-end differentiable model. This model can predict the three-dimensional coordinates of a protein's native structure, thereby facilitating protein structure prediction. In contrast to traditional methodologies, the HelixFold-Single approach obviates the need for time-intensive MSA searches, thereby conferring significant advantages in terms of time efficiency. Furthermore, through validation on the CASP14 and CAMEO datasets, as for targets characterized by a substantial number of homologous protein families, the HelixFold-Single method has been demonstrated to achieve accuracy comparable to MSA-based methods. Additionally, the method exhibits shorter runtimes for tasks requiring numerous predictions,

underscoring its potential applicability in high-throughput scenarios.

### 3.3 Medical image segmentation

*Deep Synergistic Interaction Network (DSI-Net)* [96]. DSI-Net is a deep learning approach designed for computer-aided diagnosis systems for gastrointestinal diseases, focusing on analyzing Wireless Capsule Endoscopy (WCE) images. In this system, WCE images are utilized to assist doctors in detecting and diagnosing lesions in patients. Traditional methods typically handle image classification and segmentation tasks separately, neglecting their interrelatedness and complementarity. Based on the backbone network DeepLabV3+, DSI-Net enhances overall performance by maximally exploiting the information exchange between these two tasks through the joint use of the classification branch, coarse segmentation branch, and fine segmentation branch. In the classification task, DSI-Net introduces the Lesion Location Mining module, which accurately highlights the lesion area to improve classification results. It enhances lesion detection and localization accuracy by mining ignored lesion areas and erasing background regions erroneously classified. For segmentation tasks, DSI-Net proposes the Category-Guided Feature Generation module (CFG), which utilizes category prototypes learned in the classification task to improve pixel representations, resulting in more accurate segmentation outcomes. By leveraging category information from the classification branch, the CFG module can generate features relevant to specific categories, thereby enhancing segmentation accuracy. Additionally, DSI-Net incorporates a task interaction loss to strengthen mutual supervision between classification and segmentation tasks, ensuring consistency in their predicted results. Through this approach, DSI-Net effectively leverages the interdependencies between classification and segmentation tasks to enhance overall performance.

*Medical localize and segment anything model (MedLSAM)* [97]. MedLSAM primarily focuses on localization and segmentation tasks in the field of medical image analysis. Traditional image segmentation methods often require layer-by-layer annotation, which is time-consuming and challenging when dealing with large-scale medical image datasets. MedLSAM aims to address this issue. The MedLSAM model's core idea is to combine localization and segmentation tasks, resulting in automated segmentation through self-supervised learning. It introduces a 3D localization base model called MedLAM, which can accurately locate target anatomical structures in the images. To train the MedLAM model, a dataset containing a large number of CT scan images is used, and training is conducted through self-supervised learning tasks. By performing extreme point annotation on a small number of templates, the MedLSAM model can automatically identify and locate the target anatomical regions in unlabeled data. Once the localization task is completed, the MedLSAM model generates 2D bounding boxes, which are then used for precise segmentation with existing image segmentation models. This combined localization and segmentation approach makes the segmentation process more efficient and accurate. Extensive experimental evaluations were conducted on two 3D datasets containing multiple different organs to assess the MedLSAM model. The experimental results demonstrate that the model achieves outstanding performance in both localization and segmentation tasks, with a lower dependency on extreme point annotation.

### 3.4 Doctor-patient communication

*PubMed GPT* [100]. It is a sophisticated LLM, tailored for the biomedical domain. It leverages its advanced natural language processing capabilities to facilitate a broad spectrum of academic research and clinical applications. The GPT−3.5 architecture was used to develop the PubMed GPT model, representing an enhanced iteration of the encoder-decoder structure based on the transformer model. This model encompasses multiple layers, each comprising self-attention mechanisms and feedforward neural networks. Within the encoder component, input text is processed through multiple layers of self-attention. The self-attention mechanism empowers the model to automatically assign weights to different positions in the input sequence, thereby capturing contextual information more effectively. In each self-attention layer, the model computes attention scores to determine the significance of each position to others. Using these scores, the model performs a weighted summation of the inputs, allowing for a more refined focus on contextual information relevant to the current position during input processing. In the decoder, a similar self-attention mechanism is applied to manage the output sequence. To preserve the sequence's order, the model incorporates positional encoding. This technique embeds positional information into the input sequence, which allows the model to discern between words at different positions. By combining self-attention mechanisms and positional encoding, the decoder can conscientiously consider the order and context of the input sequence during output generation. Significantly, the GPT−3.5 architecture is characterized by an extensive parameter count, boasting 175 billion trainable parameters [113]. This expansive parameterization equips the model to adeptly capture intricate semantic and grammatical rules, delivering heightened proficiency in text generation and comprehension. Before training the PubMed GPT model, an extensive corpus of biomedical literature data was utilized as pre-training data. By pre-training on self-supervised tasks using this data, the model assimilated the semantics and grammatical rules inherent

to the biomedical field, establishing a nuanced understanding of biomedical terms and concepts. This comprehensive training regimen endows the model with superior accuracy and domain-specific acumen for text-processing tasks within the biomedical realm.

*ChatDoctor* [101]. ChatDoctor is a fine-tuned medical chat model designed to provide high-quality medical advice and guidance. The model is built upon the LLM for medicine and artificial intelligence (LLaMA) infrastructure and has been fine-tuned using a dataset of 100,000 patient-doctor dialogues from online medical consultation platforms. To ensure patient confidentiality, this dataset undergoes rigorous privacy protection and personal information cleaning. One of the ChatDoctor model's key innovations is the self-retrieval mechanism for information. This mechanism allows the model to retrieve and leverage real-time information from online resources like Wikipedia and offline medical databases. This capability enables the model to better comprehend patient queries and deliver accurate and reliable medical advice. By combining an LLM's powerful language understanding capabilities with real-time knowledge supplementation through self-retrieval, the ChatDoctor model can provide patients with more comprehensive and personalized medical consultations. Through fine-tuning with real-world patient-doctor interaction dialogue data, the ChatDoctor model has shown significant improvements in understanding patient needs and providing recommendations. When answering medical queries, the model's performance has undergone rigorous evaluation and comparison, demonstrating higher accuracy and reliability compared to other existing medical LLMs.

### 3.5 Multimodal LLMs

*Med-MLLM* [106]. It is a medical multimodal large language model (MLLM) designed to address the challenges posed by future pandemics. By learning extensive medical knowledge from unlabeled data, including image understanding, semantic text, and clinical phenotypes, the model can be rapidly deployed and adapted to rare diseases, such as emerging pandemics. The Med-MLLM framework supports the processing of visual modalities (chest X-rays, CT scans, etc.) and text modalities (medical reports, clinical notes, etc.) in medical data, making it applicable to clinical tasks that require simultaneous handling of both visual and textual data. The design objective of Med-MLLM is to overcome challenges faced by traditional neural network models in scenarios involving rare diseases, where there is a lack of sufficient labeled data. Due to the high cost and time-consuming nature of collecting and labeling data for rare diseases, traditional methods encounter difficulties. To address this issue, Med-MLLM employs large-scale pre-training techniques, shortening the model's deployment time and enabling rapid responses to the emergence of future rare diseases. The Med-MLLM framework employs a structured training approach, including pre-training image encoders and text encoders for handling visual and textual data. To further enhance the model's performance, a soft image-text alignment loss is introduced for pre-training visual and text encoders. This multimodal pre-training approach enables Med-MLLM to simultaneously process visual, textual, and multimodal inputs, demonstrating accurate and robust performance in tasks such as COVID-19 reporting, diagnosis, and prognosis. The study also conducted extensive evaluations of Med-MLLM, testing its performance on COVID-19 pandemic data and showcasing its accuracy and robustness in decision-support tasks with limited labeled data. Additionally, they studied 14 other common chest diseases and tuberculosis, finding that even with only 1% labeled data, Med-MLLM exhibited competitive performance in these tasks.

*PeFoMed* [107]. PeFoMed is a model designed for medical visual question answering (Med-VQA), employing a parameter-efficient fine-tuning approach aimed at enhancing the applicability of LLMs in multimodal environments. The model's design objective is to address the limitations of traditional classification tasks in handling open-ended questions by adopting a generative task to answer questions. PeFoMed's core idea is to use a pre-trained MLLM as the base model and adjust it using parameter-efficient fine-tuning techniques to meet the specific requirements of the Med-VQA task. During training, the model freezes the visual encoder and LLM, updating only the visual projection layer and low-rank adaptation layer, significantly reducing the number of parameters that need to be trained and lowering the computational resource demands. To enhance performance, PeFoMed employs a two-stage fine-tuning strategy with specific prompt templates. In the first stage, the model undergoes fine-tuning with large-scale multimodal data from a general domain to acquire basic capabilities for multimodal tasks. In the second stage, the model is fine-tuned with data consisting of medical images and text pairs in order to excel in the Med-VQA task. Manual evaluations were used to assess the model's performance and compare it to other models. The results demonstrate that PeFoMed achieved an overall accuracy of 81.9% on closed-ended questions, a 26% absolute accuracy improvement over the benchmark model GPT-4v. PeFoMed's contributions are notable in several aspects. It not only introduces a parameter-efficient fine-tuning method, but also enables LLMs to adapt to the Med-VQA task under limited resource and dataset conditions. Moreover, it designs specific prompt templates and a two-stage fine-tuning strategy to improve the model's performance and adaptability. Through experiments on public benchmark datasets, PeFoMed has proven to exhibit outstanding performance among generative Med-VQA models.

To indicate the medical LLMs discussed in this section, the key characteristics and performance metrics of these medical LLMs have been summarized in Table 5.

## 4 Double-edged sword of LLM for medcine

As we delve into the double-edged sword of LLMs for medicine, it becomes imperative to explore both the benefits they offer in reshaping healthcare paradigms and the deficiencies and challenges they confront in their integration into the healthcare ecosystem. Let us navigate through the multifaceted landscape of LLMs to discern their impact on the healthcare landscape. The benefits and challenges are shown in Table 6.

### 4.1 Benefits of LLM for medicine

LLMs hold significant promise for strengthening the healthcare landscape, offering advantages that can revolutionize medical practices. LLMs have multifaceted capabilities that span from augmenting diagnostic precision and enabling personalized treatment regimens to facilitating real-time access to cutting-edge medical knowledge. These advantages emphasize their pivotal role in reshaping healthcare paradigms and fostering enhanced patient-centric care. Here are several key benefits of LLM for medicine.

#### 4.1.1 Enhanced capabilities of diagnosis and prediction

*Early detection and prediction.* LLMs play an essential role in the timely detection of disease indicators or factors predisposing individuals to health risks [114]. For example, through analysis of extensive clinical records [109] and various imaging modalities, LLMs possess the capability to prognosticate the likelihood of cardiac events or anticipate the probability of diabetic complications in patients. Their comprehensive assessment amalgamates multifaceted data points, allowing for nuanced risk stratification and early intervention strategies tailored to individual patient profiles.

*Prediction of treatment efficacy.* LLMs can anticipate and assess the efficacy of particular treatment protocols customized to individual patients. For instance, by harnessing the wealth of patient records and genomic information, LLMs can prognosticate survival rates or gauge the anticipated responses to specific cancer therapies [115]. By amalgamating diverse datasets, LLMs enable precise treatment prognoses, facilitating informed decision-making tailored to the unique characteristics of each patient's medical profile. This capability offers invaluable insights into treatment outcomes, contributing to more targeted and effective therapeutic strategies in oncology and beyond.

#### 4.1.2 Knowledge integration and real-time information access

*Latest medical advancements.* LLMs amalgamate a wealth of global medical literature, comprehensive clinical trial data, and expert insights, offering a dynamic platform for real-time updates in medical knowledge. For instance, LLMs can dynamically refine and revise cancer treatment protocols to align with the most recent clinical trial findings and evolving therapeutic guidelines [116]. By continuously synthesizing and analyzing an extensive array of sources, LLMs facilitate an agile and responsive framework for medical professionals, ensuring that treatment strategies stay abreast of the rapidly evolving landscape of evidence-based medicine. This adaptability helps to optimize patient care by incorporating the most current and relevant information into clinical decision-making processes.

*Clinical decision support.* LLMs serve as real-time decision support systems, significantly augmenting the depth and precision of diagnostic processes and treatment strategies for clinicians [117]. By swiftly processing vast volumes of patient data alongside the latest medical insights, LLMs offer immediate and comprehensive guidance to healthcare professionals. This real-time assistance bolsters the accuracy of diagnoses and treatment plans, empowering clinicians to make more informed and nuanced decisions at the point of care. The seamless integration of cutting-edge knowledge into clinical practice elevates the quality of healthcare delivery, ensuring that medical interventions align with the most current and validated information available in the field.

#### 4.1.3 Personalized treatment and drug development

*Tailored treatment plans.* Leveraging individual patient data, LLMs can specialize in crafting meticulously tailored treatment blueprints encompassing an array of personalized interventions [118]. These comprehensive plans cover everything from selecting the most appropriate medications to fine-tuning dosage parameters and even assisting in surgical strategies. For instance, drawing insights from detailed genomic profiles and comprehensive medical histories, LLMs excel at charting precise, individualized therapeutic routes for cancer patients. By scrutinizing intricate genetic information and patient-specific medical trajectories, LLMs contribute to the development of highly targeted therapy plans, optimizing treatment efficacy while minimizing potential adverse effects. This personalized approach stands as a testament to the potential of LLMs to revolutionize patient-centric care in oncology and beyond.

*Drug development and precision medicine.* LLMs excel at forecasting medication effectiveness and anticipating potential side effects, thereby expediting the drug development process and nurturing the realm of precision medicine. For

**Table 5** Key characteristics and performance metrics of the discussed medical LLMs

| LLM | Key characteristics | Performance metrics |
|---|---|---|
| BenTsao (Huatuo) [90] | HuaTuo is a LLaMA-based model that has been supervised and fine-tuned with generated QA instances using medical knowledge from the Chinese medical knowledge graph (CMeKG), especially for the Chinese language | The authors proposed a novel metric called SUS (safety, usability, and smoothness) to evaluate LLMs in the biomedical domain |
| Med-PaLM [28] | Med-PaLM is an instruction prompt tuned model, a parameter-efficient approach for aligning large language models to new domains using a few exemplars | Med-PaLM performs encouragingly, but remains inferior to clinicians |
| PanGu Drug Model [92] | PanGu Drug Model uses a novel graph-to-sequence asymmetric conditional variational autoencoder architecture to learn molecule representation from both molecular structures and formulas | PanGu Drug Model achieved state-of-the-art results in 20 drug discovery tasks, including molecule property prediction, molecule generation, and molecule optimization |
| HelixFold-Single [93] | HelixFold-Single combines a large-scale protein language model with the essential geometric learning components from AlphaFold2 to enable end-to-end MSA-free protein structure prediction from primary sequences | HelixFold-Single achieves competitive accuracy with MSA-based methods like AlphaFold2 on proteins with sufficient homologous sequences while consuming much less time for prediction |
| DSI-Net [96] | DSI-Net is a deep synergistic interaction network that jointly performs classification and segmentation of wireless capsule endoscope (WCE) images by leveraging complementary information between the two tasks through modules like lesion location mining and category-guided feature generation | DSI-Net achieves superior classification and segmentation performance on public datasets compared to state-of-the-art methods |
| MedLSAM [97] | MedLSAM introduces a 3D localization foundation model called MedLAM that can directly localize any target anatomical part within a body using just a few template scans | MedLSAM not only aligns closely with the performance of SAM and its specialized medical adaptations that use manual prompts but achieves this with minimal reliance on extreme point annotations across the entire dataset |
| PubMed GPT [100] | PubMed GPT leverages the advanced GPT−3.5 architecture and trained on a vast corpus of biomedical literature data to deliver heightened proficiency in text generation and comprehension tasks within the biomedical realm | The extensive pre-training on biomedical data and expansive parameterization equip PubMed GPT with superior accuracy and domain-specific acumen for biomedical text-processing tasks |
| ChatDoctor [101] | ChatDoctor uses a large dataset of 100,000 patient-doctor dialogues, and it incorporates a self-directed information retrieval mechanism to access and utilize real-time information | The fine-tuning of the model significantly improved its ability to understand patient needs and provide informed advice. By equipping the model with self-directed information retrieval, the accuracy of its responses was substantially improved |
| Med-MLLM [106] | Med-MLLM is a medical multimodal large language model that can learn broad medical knowledge from unlabelled data | The experiments conducted on Med-MLLM show that it can make accurate and robust COVID-19 decision-support with little labeled data, across three different tasks and three different languages |
| PeFoMed [107] | PeFoMed is a parameter-efficient framework for fine-tuning multimodal large language models specifically for medical imaging applications like medical visual question answering and medical report generation | PeFoMed utilizes an evaluation metric using a 5-point Likert scale and its weighted average value to measure the quality and coherence of the generated medical reports, with the scale ratings labeled by both humans and the GPT-4 model |

**Table 6** Benefits and challenges medical LLM

| Domain | Decriptions |
|---|---|
| Benefits | LLM brings about more accurate diagnosis and prediction in medicine, promoting early disease detection and personalized treatment planning |
| | LLM integrates the latest medical advancements, providing real-time decision support and updated medical knowledge to physicians, thus optimizing clinical decision-making processes |
| | LLM can personalize treatment plans and facilitate drug development, offering patients more precise treatment options and medication choices |
| | LLM enhances patient management and healthcare processes, improving medical efficiency and the quality of patient care |
| | LLM supports medical education and dissemination of healthcare knowledge, fostering continuous learning and improvement among medical students and practitioners |
| Challenges | LLM faces fundamental challenges in medical applications due to high computational resource demands and low efficiency |
| | Issues such as data imbalance, privacy protection, and data quality pose challenges to the applications of LLM in healthcare |
| | LLM encounters practical application challenges in clinical validation, multilingual adaptation, and multicultural adaptation |
| | The design of LLM needs to consider ethical concerns such as bias and fairness, privacy protection, transparency, and accountability |

instance, they undertake the pivotal task of predicting how a drug would perform in synergy [119], significantly contributing to the meticulous design of more targeted and refined clinical trials. By delving into extensive datasets and intricate biological markers, LLMs offer invaluable insights into the anticipated efficacy of medications across diverse patient populations. This predictive capability not only accelerates the drug discovery phase but also aids in tailoring clinical trials to specific subsets of patients, fostering a more nuanced and individualized approach toward developing new therapeutic interventions.

#### 4.1.4 Patient management and healthcare process optimization

*Personalized patient management.* LLMs specialize in crafting highly customized health management strategies by intricately analyzing biological markers and lifestyle data. For example, they meticulously design preventive and management protocols tailored specifically for patients dealing with cardiovascular diseases [120]. By integrating detailed genomic profiles with comprehensive lifestyle habits, LLMs formulate precise plans aimed at preventing the onset or progression of cardiovascular conditions [121]. Leveraging this amalgamation of genetic predispositions and individual behaviors, they develop personalized strategies encompassing dietary recommendations, exercise regimens, and medication plans. This tailored approach not only addresses immediate health concerns but also empowers patients with personalized insights to proactively manage their cardiovascular health and mitigate future risks.

*Optimization of healthcare processes.* LLMs are important parts of enhancing efficiency in medical processes for improved patient care. They work by meticulously identifying potential bottlenecks within diagnostic and treatment workflows. For instance, they are able to recognize critical points of congestion or inefficiency and assist in diagnostic assessments [122]. Following this identification, LLMs provide valuable optimization suggestions for streamlining and expediting patient care delivery. Their insights could range from refining scheduling protocols to suggesting workflow modifications that enhance the overall efficacy and promptness of healthcare services. By targeting inefficiencies and optimizing workflows, LLMs significantly contribute to the seamless and expedited delivery of quality healthcare, ultimately benefiting patient outcomes.

#### 4.1.5 Clinical education and dissemination of medical knowledge

*Medical education and training.* The advanced LLMs serve as instrumental tools in the realm of medical education, providing a rich repository of the most recent medical knowledge and comprehensive case studies tailored for both aspiring medical students and seasoned practitioners. Their dynamic functionalities enable the simulation of real-world clinical scenarios, offering an immersive learning experience that significantly contributes to the enhancement of clinical decision-making skills [123]. By leveraging LLMs, medical students gain access to a vast array of intricate case studies and real-time medical data, replicating authentic patient encounters. This hands-on exposure allows them to sharpen their analytical skills, expand their medical knowledge base, and refine their diagnostic and treatment planning abilities within a safe, simulated environment. Moreover, for practicing healthcare professionals, LLMs serve as invaluable resources for continuous learning and staying updated with the latest advancements in medicine. They provide a platform for honing diagnostic acumen and exploring diverse treatment approaches through interactive scenarios.

Ultimately, these educational tools foster a culture of ongoing learning and skill development. It ensures that, in an ever-evolving healthcare landscape, medical practitioners are equipped with the expertise to deliver optimal patient care.

*Dissemination of medical knowledge.* LLMs empower patients by providing direct, comprehensive disease prevention and management guidance through user-friendly healthcare applications. For instance, applications leverage the vast knowledge base of LLMs to give patients tailored advice on disease prevention strategies [124]. They offer insights into lifestyle modifications, preventive measures, and early warning signs associated with various medical conditions. This guidance is personalized, considering individual health profiles and promoting proactive health management among users. Moreover, LLMs assist in delivering meticulous disease management recommendations to patients already diagnosed with specific conditions. They provide detailed information on treatment adherence, and such information is expected to improve [125]. By harnessing the extensive medical knowledge encapsulated within LLMs, these healthcare applications bridge the gap between medical expertise and patient understanding. They empower individuals to make informed decisions regarding their health, fostering a proactive approach toward disease prevention and self-care management.

## 4.2 Deficiencies and challenges of LLM for medicine

For healthcare, the potential of LLMs, powered by artificial intelligence, is undeniably transformative. These models can analyze vast volumes of medical data, aiding in diagnosis, treatment recommendations, and medical research. However, when integrating into the healthcare ecosystem, they also encounter hurdles. The use of LLMs in healthcare poses a multifaceted challenge that necessitates careful consideration and solutions, from computational demands to ethical considerations. These challenges can be categorized into four fundamental domains: fundamental requirements, data-related issues, practical applications, and ethical concerns. Thus, delving into these challenges can help us better understand the complexities of deploying LLMs in the healthcare field.

### 4.2.1 Fundamental requirements

*Computational resources.* LLMs in healthcare often lack the computational efficiency required for real-time processing. These resource-intensive models can result in longer response times, rendering them less suitable for applications that demand immediate decisions, such as emergency diagnosis and treatment. Furthermore, models like GPT-3 [126], GPT-4 [127], or custom medical models require powerful computing hardware, typically in the form of graphics processing units (GPUs) or specialized hardware like tensor processing units (TPUs). With an enormous number of parameters, these models demand significant computational power for both training and inference. Acquiring and maintaining the necessary computational resources can incur prohibitively high costs. This encompasses not only the hardware itself but also the electricity and cooling systems essential for efficient operation. Healthcare institutions and research organizations must allocate budgets to cover these substantial expenses.

*Model efficiency.* In terms of data utilization, LLMs must effectively leverage the available information, especially in healthcare, where data can be scarce and sensitive. Inefficient models may necessitate more data than is practically obtainable, or they may fail to yield meaningful insights from limited datasets. Additionally, LLMs are susceptible to overfitting when training on limited data [128], which can undermine their ability to generalize to new and diverse medical cases. The process of optimizing inference is a critical component of LLMs. While model training demands substantial resources, the efficiency of inference holds equal importance. Healthcare applications require models to deliver timely responses, making the speed of inference a critical consideration. Efficient inference methods are essential to meet the demands of real-time decision-making in healthcare settings.

### 4.2.2 Data-related issues

*Data privacy.* Healthcare data contains sensitive patient information, such as medical records, imaging data, and personal identifiers. Mishandling patient data can result in data breaches, identity theft, and legal repercussions. Ensuring that LLMs handle patient data with strict encryption, access controls, and audit trails is crucial to protecting patient confidentiality.

*Data quality.* Healthcare data comes in diverse formats, including structured electronic health records (EHRs) , unstructured text reports, medical images, and time-series data. Ensuring data quality across these varied data types is challenging due to potential inconsistencies, errors, and data source variations. For example, in medical image analysis, it is critical to ensure the precision of annotations, such as identifying tumors in medical images [129]. Inaccurate annotations can lead to erroneous model training and subsequent misdiagnoses.

*Interpretability and trustworthiness.* In the healthcare domain, LLMs' interpretability and explainability are critical for their acceptance and utility among healthcare professionals. These models must provide clear explanations for decisions, especially in diagnoses and treatment recommendations. By understanding the reasoning behind these decisions, healthcare professionals can evaluate the model's

outputs, identify biases or limitations, and make informed decisions based on the information provided. The transparency offered by interpretability enhances collaboration, allowing healthcare professionals to engage in meaningful discussions with the model, seek clarification, and ultimately improve healthcare decision-making. The interpretability and explainability of LLMs in healthcare not only build trust but also ensure their effective support in delivering high-quality care to patients.

#### 4.2.3 Practical applications

*Clinical validation.* Conducting robust clinical trials to validate healthcare models is a challenging endeavor, primarily due to the inherent variability within medical data. Patient characteristics, medical conditions, and treatment approaches exhibit significant differences, underscoring the need to effectively address this diversity when validating the model's performance. Furthermore, scaling clinical validation to encompass a broad and diverse patient population while involving multiple healthcare institutions is a complex undertaking. It entails not only logistical challenges but also the critical task of ensuring that the model consistently performs effectively across diverse clinical settings and demographics.

*Multilingual and multicultural adaptation.* Adapting healthcare models to multiple languages necessitates a substantial investment in data collection, translation, and cross-lingual model training. This is due to the inherent variations in language structure and medical terminology across different languages, which can pose significant challenges. Furthermore, integrating cultural nuances into the model's understanding and recommendations can be a complex and nuanced process. This is because healthcare practices, beliefs, and patient expectations can vary significantly from one culture to another, requiring precise model adaptations for different cultural contexts. Adding to the complexity, healthcare practices and medical standards may differ across various regions. Consequently, adapting models to address these variations while ensuring consistent performance across diverse clinical settings presents a formidable undertaking.

#### 4.2.4 Ethical concerns

*Bias and fairness.* Medical language models should be designed to mitigate bias and ensure fairness in medical decision-making. Biased models can lead to disparities in outcomes, disproportionately affecting minority populations. Here are some key ethical considerations. Conduct regular bias audits to identify differences in model performance across various groups. For example, evaluate the model's prediction accuracy across different genders, races, ages, etc. Then, to reduce bias, adjust models based on audit findings. This could involve retraining the model or using more diverse datasets to improve model fairness. Last but not least, transparently report bias evaluation results and corrective actions taken, ensuring users know the model's potential biases and how they are being handled. Addressing bias is crucial for ensuring fairness in medical decision-making. Fair models can help reduce healthcare disparities and improve the overall quality and accessibility of medical services.

*Privacy concerns.* The ethical use of medical models necessitates robust privacy protections. Patient data should be anonymized before use by removing or masking identifiable information to prevent data leaks. Strong encryption techniques should be employed to protect patient data during storage and transmission, ensuring only authorized personnel can access it. Secure storage methods and facilities should be used to safeguard patient data, preventing unauthorized access and potential data breaches. Models should be designed to minimize reliance on sensitive information, access patient data only when necessary, and limit access permissions to ensure only essential personnel handle the data. While protecting privacy, it's also important to maintain the quality of medical insights provided by the models, balancing privacy protection with model performance. Handling privacy concerns is crucial for maintaining patient trust, protecting privacy rights, and ensuring legal and ethical compliance. Ensuring data privacy is fundamental to the safe and effective use of medical models.

*Accountability and transparency.* Ethical considerations in medical LLMs include mechanisms for accountability and transparency. This involves clear documentation of model decision-making processes, ensuring traceability, and providing explanations for specific medical recommendations. Transparent and interpretable models are essential for fostering trust among healthcare professionals and patients. Clear documentation should detail how the model makes decisions, including the data used, algorithms applied, and steps followed in the decision-making process. Ensuring traceability means maintaining a detailed log of inputs, processing steps, and outputs, allowing for a clear audit trail. Explainability requires developing models that can provide clear explanations for their recommendations, highlighting key factors or data points that influenced a decision. This transparency and interpretability build trust, enabling better scrutiny and understanding of AI-driven decisions, and ensuring responsible and ethical use in medical settings.

*Patient autonomy.* Respecting patient autonomy is essential in medical LLMs. Patients should have the right to accept or reject AI-driven recommendations, ensuring these align with their values, preferences, and informed consent. Healthcare providers must communicate AI-generated recommendations clearly and provide enough information

for patients to make decisions. It involves explaining the benefits, risks, and uncertainties associated with AI-driven recommendations, thus supporting patients' ability to make choices that best reflect their personal health goals and values.

*Preventing data misuse.* Ethical guidelines demand safeguarding medical models from misuse or exploitation. Some unintended consequences, such as inappropriate use of patient data, model hacking, or malicious intent, are critical ethical considerations that should be taken into account. Robust security measures should be in place to protect patient data, including encryption, secure storage, and controlled access. Regular security audits and updates are necessary to defend against potential threats. Additionally, clear policies and protocols should be established to address any misuse, ensuring that medical models are used responsibly and ethically. By protecting against data misuse, we uphold the integrity and trustworthiness of medical LLMs.

## 5 Opportunities and future directions

Medical LLMs present extensive opportunities and future directions that will further drive innovation and improvement in medical practices. As shown in Fig. 5, we give some potential opportunities and future directions to develop them, including medical LLMs into smart medical devices [130], medical LLMs with intelligent robots/virtual assistants [131, 132], medical LLMs in Metaverse [133, 134], secure medical LLMs [135], medical LLMs with blockchain [136], and multi-party collaboration of medical LLMs [137].

### 5.1 Medical LLMs into smart medical devices

For real-time health monitoring and data interpretation [138], the integration of LLMs into smart medical devices enables continuous monitoring of patient's physiological parameters. These medical LLMs can promptly interpret transmitted data, providing critical analysis and feedback, thereby facilitating early detection of potential health issues. This integration also should support remote medical consultation and monitoring, allowing healthcare professionals such as physicians and nurses to communicate remotely with patients via voice or text [139]. Effective remote monitoring is enabled by leveraging medical LLMs to offer medical advice and personalized recommendations. Intelligent medication management, facilitated by IoT technology, is another potential application. By integrating IoT devices with medical LLMs, medication packaging or smart medicine boxes can remind patients of medication schedules, provide explanations of medication information, and adjust medication plans based on patient feedback [140], thereby enhancing medication adherence. In clinical settings, it is feasible to use IoT devices in operating rooms and emergency environments. Medical LLMs can interpret real-time surgical data and monitor device feedback, offering immediate intraoperative guidance or emergency advice [141]. This can lead to improvements in surgical efficiency and the speed of emergency decision-making. Furthermore, in

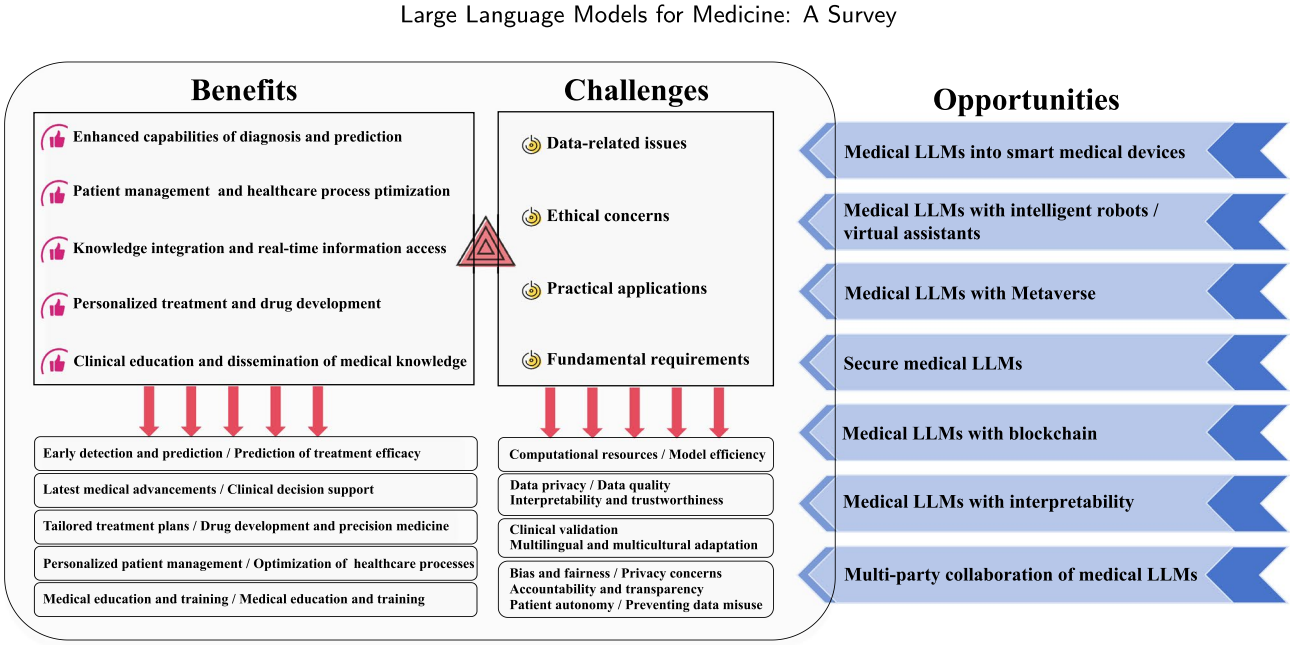


**Fig. 5** Benefits, challenges, and opportunities of medical LLMs

inpatient recovery monitoring and recommendations, leveraging IoT technology to monitor the usage of rehabilitation devices and transmit rehabilitation data to medical LLMs is advantageous. The language model can offer personalized rehabilitation suggestions and monitor recovery progress, enabling healthcare teams, including physicians, therapists, and caregivers, to remotely adjust treatment plans [142]. Integrating these approaches improves the intelligence, real-time capabilities, and personalization of healthcare services and broadens the prospects for the healthcare field.

### 5.2 Medical LLMs with intelligent robots/virtual assistants

In the field of healthcare, intelligent robots and virtual assistants have emerged as essential technological tools for delivering personalized and efficient medical services [131]. The integration of medical LLMs enhances these intelligent entities' semantic understanding and NLP capabilities [143]. Various types of intelligent robots, such as surgical robots [144] and caregiving robots [145], and virtual assistants [146], including chatbots and voice assistants, can better comprehend and respond to patient needs by incorporating LLMs, offering more intelligent medical consultation, appointment scheduling, and health management services [147]. For instance, intelligent robots equipped with deep learning algorithms and medical LLMs can perform emotion recognition and semantic analysis, enabling a more precise understanding of patient language and emotions [146]. This enables them to provide personalized medical advice and services as needed. Similarly, virtual assistants can leverage the capabilities of medical LLMs to automate medical record-keeping and offer clinical decision support for healthcare professionals. This enhances work efficiency and diagnostic accuracy in clinical settings. Thus, the integration of medical LLMs with intelligent robots and virtual assistants promises to bring about more intelligent and efficient medical services, ultimately enhancing the medical experience and treatment outcomes.

### 5.3 Medical LLMs with metaverse

In the Metaverse, virtual medical assistants create an interactive healthcare experience for both doctors and patients. These assistants, powered by medical LLMs, can interact with patients in virtual reality or augmented reality, providing health information, conducting self-diagnoses, and explaining medical concepts. Additionally, leveraging the ability of Metaverse to construct virtual worlds, medical virtual spaces can be established for remote medical consultations and treatments. Doctors and patients can engage in face-to-face medical consultations within this virtual environment, with medical LLMs playing a role in explaining medical conditions and providing treatment recommendations [148]. For medical students and professionals, Metaverse, especially human-centric Metaverse [149], offers opportunities for smart education, virtual practice, enhancing medical training and simulation [150]. Medical LLMs can serve as virtual mentors, offering real-time feedback, explaining surgical procedures, and addressing relevant questions, thus facilitating the training of healthcare professionals. In the long run, medical research and collaboration will also witness a new era within the Metaverse [151]. Future research can focus on collaborating within shared virtual environments, and medical language models can assist in analyzing complex medical literature, providing research recommendations, and advancing medical research. The innovative integration of medical LLMs with Metaverse holds promise for the healthcare sector's future. By embedding medical LLMs into the Metaverse framework, we can establish more comprehensive, intelligent, and personalized healthcare services.

### 5.4 Secure medical LLMs

The use of medical LLMs in the healthcare industry necessitates ensuring data security and privacy. To safeguard the security of medical LLMs, multiple measures must be taken. Firstly, data collection and storage for medical LLMs must adhere to strict privacy protection standards, such as HIPAA [152] and GDPR [153], to ensure the confidentiality of patients' personal information. Secondly, during the training and application process of medical LLMs, secure and controllable technical measures must be adopted to prevent data from being subjected to malicious attacks or misuse. For example, encryption algorithms and access control policies are used to ensure the security of data transmission and access. Additionally, comprehensive permission management and auditing mechanisms must be established for medical LLM applications to ensure that only authorized personnel can access and use relevant data and models. Finally, the applications of medical LLMs must comply with relevant laws, regulations, and ethical guidelines to protect patients' legal rights and medical information security. By comprehensively implementing these measures, medical LLMs can be effectively secured, providing reliable technical support for their applications in the healthcare field.

### 5.5 Medical LLMs with blockchain

Beginning with the utilization of blockchain's smart contract functionality, patients gain control over their medical data, assisted by medical LLMs as interpreters. This empowers selective data sharing and tracking usage. Furthermore, smart contracts manage processes like billing and insurance claims, with LLMs explaining contract content.

Storing medical data on the blockchain enhances security and privacy. Medical LLMs ensure secure access and accurate analysis of patient records, contributing to transparency in clinical trial data. Moreover, constructing a medical knowledge graph on blockchain supports decision-making for doctors and researchers, with medical LLMs providing deeper insights and enhancing research credibility. Additionally, recording pharmaceutical and medical device information on the blockchain ensures authenticity and traceability. Medical LLMs play a crucial role in identifying counterfeit products, thus safeguarding patient safety [154].

### 5.6 Medical LLMs with interpretability

The interpretability of medical LLMs is a crucial topic in the current field of medical artificial intelligence research. In the process of enhancing model transparency and comprehensibility, a focus on both global and local interpretability is significant. In the realm of interpretable deep learning models, innovative methods such as attention mechanisms [155] and interpretable neural networks [156] are continuously emerging to provide more detailed explanations. Furthermore, interpretability directly influences the model's trustworthiness and reliability, prompting the need for research on how transparency and interpretability can enhance the model's reliability and reduce potential medical errors. While providing interpretability, privacy protection for patients must also be considered, involving the adopted privacy protection techniques such as differential privacy [157]. Incorporating the knowledge of domain experts can be utilized to interpret model outputs, thereby increasing the credibility of the model's output. In the context of rule extraction, interpretability is relatively straightforward for decision tree models, allowing for the extraction of rules and information on nodes and branches. Furthermore, addressing uncertainties in model quantification, establishing interpretability standards and evaluation methods, and educating healthcare professionals on understanding and interpreting model outputs are all critical factors for ensuring the success of medical LLMs in terms of interpretability. Considering these aspects comprehensively can make medical LLMs more understandable and acceptable, thereby enhancing their usability and credibility in practical medical applications.

### 5.7 Multi-party collaboration of medical LLMs

The incorporation of medical LLMs within the healthcare domain necessitates concerted efforts from various stakeholders, including governmental bodies, healthcare institutions, patients, and research establishments. Governments wield a pivotal role in shaping policies, regulations, and standards governing the development and implementation of medical LLMs. They possess the capacity to mobilize resources, facilitate data exchange, and provide essential infrastructure support, such as computing capabilities, indispensable for both training and operationalizing these models. Healthcare institutions serve as primary arenas for the deployment of medical LLMs, given their status as providers of medical services. Collaborating with governmental bodies, they can establish platforms conducive to the sharing and integration of medical data, pivotal for refining and optimizing these models. Furthermore, governments and healthcare institutions can not only foster patient involvement in data collection and sharing by instituting mechanisms that promote active engagement, but also harness intelligent health management platforms to deliver personalized healthcare services, thereby improving patient experiences and treatment outcomes. Governments can bolster research establishments through financial support and collaborative frameworks, thereby catalyzing technological innovation and advancement in the realm of medical LLMs. Moreover, partnerships between healthcare and research establishments can drive exploration into the application of medical LLMs across disease diagnosis, prevention, and treatment, fostering innovation in medical technologies. Patients, as recipients of medical care, occupy a pivotal role in the utilization of medical LLMs. Research institutions, meanwhile, play a crucial role in the research and development of these models, often collaborating with healthcare establishments to explore their myriad applications and expedite their translation into practical medical solutions.

By fostering collaboration among governmental bodies, healthcare institutions, patients, and research establishments, the comprehensive integration of medical LLMs into the healthcare landscape can be achieved. This collaborative approach promises to deliver intelligent, personalized, and efficient healthcare solutions, thereby elevating healthcare standards and enhancing patient quality of life.

## 6 Conclusion

In this paper, we comprehensively explore the pivotal role of LLMs in the field of medicine. These models demonstrate significant potential not only in medicine-assisted diagnosis, biopharmaceutical design, and medical image segmentation, but also in achievements related to health management, doctor-patient communication, and multimodal applications of LLMs in medicine. However, challenges persist, encompassing issues such as data privacy, model interpretability, ethical concerns, and technical difficulties in practical implementations. Future research should focus on addressing these challenges to ensure the reliability and safety of models in real clinical environments. Regarding these problems, we proposed several technologies that are possible to combine with medical LLMs to solve them, including

smart medical devices, intelligent robots/virtual assistants, Metaverse, secure issues, blockchain, interpretability, multi-party collaboration of LLMs, and so on. In summary, LLMs bring unprecedented opportunities to the field of medicine. LLMs in medicine are poised to play a greater role in personalized medicine, new drug development, and health management. Nevertheless, it is imperative to prioritize ethical and privacy considerations in this process. We anticipate achieving more significant accomplishments in improving patient quality of life, advancing medical research, and optimizing medical processes. Encouraging collaborative efforts among future researchers and practitioners is essential to drive the development of LLMs in medicine for the benefit of humanity.

**Acknowledgements** This research was supported in part by the National Natural Science Foundation of China (No. 62272196), the Natural Science Foundation of Guangdong Province (No. 2022A1515011861), Guangzhou Basic and Applied Basic Research Foundation (No. 2024A04J9971).

**Author contributions** Yanxin Zheng: paper reading and review, writing original draft. Wensheng Gan: conceptualization, review and editing, supervisor. Zefeng Chen and Zhenlian Qi: conceptualization, review and editing. Qian Liang and Philip S. Yu: review and editing.

**Data availability** This is a review paper, and no data was generated during the study.

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

1. McCue ME, McCoy AM (2017) The scope of big data in one medicine: unprecedented opportunities and challenges. Front Vet Sci 4:194
2. Nilsson NJ (1982) Principles of artificial intelligence. Springer Science & Business Media, Berlin
3. Cao Y, Peng H, Yu PS (2020) Multi-information source Hin for medical concept embedding. Advances in knowledge discovery and data mining. Springer, Berlin, pp 396–408
4. Silberg WM, Lundberg GD, Musacchio RA (1997) Assessing, controlling, and assuring the quality of medical information on the internet: caveant lector et viewor-let the reader and viewer beware. JAMA 277:1244–1245
5. Duggan C, Bates I (2008) Medicine information needs of patients: the relationships between information needs, diagnosis and disease. Quality Saf Health Care 17:85
6. Waitzkin H (1985) Information giving in medical care. J Health Soc Behav 26:81–101
7. Carpineto C, Romano G (2012) A survey of automatic query expansion in information retrieval. ACM Comput Surv 44:1–50
8. Gan W, Qi Z, Wu J, Lin JC-W (2023) Large language models in education: vision and opportunities. In: IEEE international conference on big data, IEEE, pp 4776–4785
9. Shanahan M (2024) Talking about large language models. Commun ACM 67:68–79
10. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, Zhang M, Wang J, Jin S, Zhou E, et al (2023) The rise and potential of large language model based agents: a survey, arXiv preprint arXiv:2309.07864
11. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z, et al (2023) A survey of large language models, arXiv preprint arXiv:2303.18223
12. Lin T, Wang Y, Liu X, Qiu X (2022) A survey of Transformers. AI Open 3:111–132
13. Gan W, Wan S, Yu PS (2023) Model-as-a-service (MaaS): a survey. In: IEEE international conference on big data, IEEE, pp 4636–4645
14. Liu S, Peng C, Wang C, Chen X, Song S (2023) icsBERTs: optimizing pre-trained language models in intelligent customer service. Proc Comput Sci 222:127–136
15. Tarcar AK, Tiwari A, Dhaimodker VN, Rebelo P, Desai R, Rao D (2019) Healthcare NER models using language model pretraining, arXiv preprint arXiv:1910.11241
16. Wu S, Irsoy O, Lu S, Dabravolski V, Dredze M, Gehrmann S, Kambadur P, Rosenberg D, Mann G (2023) BloombergGPT: a large language model for finance, arXiv preprint arXiv:2303.17564
17. Gupta U (2023) GPT-InvestAR: Enhancing stock investment strategies through annual report analysis with large language models, arXiv preprint arXiv:2309.03079
18. Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G, Günnemann S, Hüllermeier E et al (2023) ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 103:102274
19. Roemmele M, Gordon AS (2018) Automated assistance for creative writing with an RNN language model. In: The 23rd international conference on intelligent user interfaces companion, ACM, pp 1–2
20. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444
21. Chowdhary K, Chowdhary K (2020) Natural language processing. Fundamentals of artificial intelligence. Springer, Berlin, pp 603–649
22. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, Yin H, Xu C, Yang R, Zheng Q et al (2023) ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci 15:29
23. Goel A, Gueta A, Gilon O, Liu C, Erell S, Nguyen LH, Hao X, Jaber B, Reddy S, Kartha R, et al (2023) LLMs accelerate annotation for medical information extraction. In: Machine learning for health, PMLR, pp 82–100
24. Wilhelm TI, Roos J, Kaczmarczyk R (2023) Large language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res 25:e49324
25. Minssen T, Vayena E, Cohen IG (2023) The challenges for regulating medical use of ChatGPT and other large language models. J Am Med Assoc 330:315–316
26. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J-N, Laleh NG, Löffler CML, Schwarzkopf S-C, Unger M, Veldhuizen GP et al (2023) The future landscape of large language models in medicine. Commun Med 3:141
27. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. Nat Med 29:1930–1940
28. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S et al (2023) Large language models encode clinical knowledge. Nature 620:172–180

29. Karabacak M, Margetis K (2023) Embracing large language models for medical applications: Opportunities and challenges, Cureus 15
30. Zhou H, Gu B, Zou X, Li Y, Chen SS, Zhou P, Liu J, Hua Y, Mao C, Wu X, et al (2023) A survey of large language models in medicine: progress, application, and challenge, arXiv preprint arXiv:2311.05112
31. Kim JK, Chua M, Rickard M, Lorenzo A (2023) ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine. J Pediatr Urol 19(5):598–604
32. Hu Z, Yang Z, Liang X, Salakhutdinov R, Xing EP (2017) Toward controlled generation of text. In: International conference on machine learning, PMLR, pp 1587–1596
33. SarkarK, Liu L, Golyanik V, Theobalt C (2021) HumanGAN: a generative model of human images. In: International conference on 3D vision, IEEE, pp 258–267
34. Kim S, Lee S-G, Song J, Kim J, Yoon S (2019) FloWaveNet: a generative flow for raw audio. In: International conference on machine learning, PMLR, pp 3370–3378
35. Wu J, Gan W, Chen Z, Wan S, Lin H (2023) AI-generated content (AIGC): a survey, arXiv preprint arXiv:2304.06632
36. Han X, Zhang Z, Ding N, Gu Y, Liu X, Huo Y, Qiu J, Yao Y, Zhang A, Zhang L et al (2021) Pre-trained models: past, present and future. AI Open 2:225–250
37. Zeng F, Gan W, Wang Y, Yu PS (2023) Distributed training of large language models. In: IEEE 29th international conference on parallel and distributed systems, IEEE, pp 840–847
38. Howard J, Ruder S (2018) Universal language model fine-tuning for text classification, In: The 56th annual meeting of the ACL, pp 328–339
39. Grossberg S (2013) Recurrent neural networks. Scholarpedia 8:1888
40. Wu J, Gan W, Chen Z, Wan S, Yu PS (2023) Multimodal large language models: a survey. In: IEEE international conference on big data, IEEE, pp 2247–2256
41. Cambria E, White B (2014) Jumping NLP curves: a review of natural language processing research. IEEE Comput Intell Mag 9:48–57
42. Harshvardhan G, Gourisaria MK, Pandey M, Rautaray SS (2020) A comprehensive survey and analysis of generative models in machine learning. Comput Sci Rev 38:100285
43. Brown PF, Della Pietra VJ, de Souza PV, Lai JC, Mercer RL (1992) Class-based N-gram models of natural language. Comput Sci Rev 18:467–480
44. Blunsom P (2004) Hidden Markov models. Lecture Notes 15:48
45. Yu Y, Si X, Hu C, Zhang J (2019) A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput 31:1235–1270
46. Tweedie RL (2001) Markov chains: structure and applications. Handb Stat 19:817–851
47. Qiao M, Bian W, Da Xu RY, Tao D (2015) Diversified hidden Markov models for sequential labeling. IEEE Trans Knowl Data Eng 27:2947–2960
48. Käll L, Krogh A, Sonnhammer EL (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics 21:i251–i257
49. Philipp G, Song D, Carbonell JG (2017) The exploding gradient problem demystified-definition, prevalence, impact, origin, tradeoffs, and solutions, arXiv preprint arXiv:1712.05577
50. Lippi M, Montemurro MA, Degli-Esposti M, Cristadoro G (2019) Natural language statistical features of LSTM-generated texts. IEEE Trans Neural Netw Learn Syst 30:3326–3337
51. Church KW (2017) Word2Vec. Nat Lang Eng 23:155–162
52. Ethayarajh K (2019) How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings, arXiv preprint arXiv:1909.00512
53. Roumeliotis KI, Tselikas ND (2023) ChatGPT and open-AI models: a preliminary review. Future Internet 15:192
54. Kenton JDM-WC, Toutanova LK (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: The NAACL-HLT, vol 1, ACL, p 2
55. Luo Y, Tang J, Yan J, Xu C, Chen Z (2014) Pre-trained multi-view word embedding using two-side neural network. In: The AAAI conference on artificial intelligence, p 28
56. Zheng J, Cai F, Chen H, de Rijke M (2020) Pre-train, interact, fine-tune: a novel interaction representation for text classification. Inf Process Manage 57:102215
57. Yohannes HM, Amagasa T (2022) Named-entity recognition for a low-resource language using pre-trained language model. In: The 37th ACM/SIGAPP symposium on applied computing, ACM, pp 837–844
58. Gan W, Lin JC-W, Chao H-C, Zhan J (2017) Data mining in distributed environment: a survey. Wiley Interdiscip Rev Data Min Knowl Discov 7:e1216
59. Wang Q, Xu J, Chen H, He B (2017) Two improved continuous bag-of-word models. In: International joint conference on neural networks, IEEE, pp 2851–2856
60. McCormick C (2016) Word2Vec tutorial-the skip-gram model. Available online at: http://www.mccormickml.com
61. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, et al (2020) Transformers: state-of-the-art natural language processing. In: The conference on empirical methods in natural language processing: system demonstrations, ACL, pp 38–45
62. Zhang H, Dang M, Peng N, Van den Broeck G (2023) Tractable control for autoregressive language generation. In: International conference on machine learning, PMLR, pp 40932–40945
63. Dey R, Salem FM (2017) Gate-variants of gated recurrent unit (GRU) neural networks. In: IEEE 60th international midwest symposium on circuits and systems, IEEE, pp 1597–1600
64. Mikolov T, Karafiát M, Burget L, Cernockỳ J, Khudanpur S (2010) Recurrent neural network based language model. Interspeech 2:1045–1048
65. Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y (2021) Transformer in transformer. Adv Neural Inf Process Syst 34:15908–15919
66. Tehranineshat B, Rakhshan M, Torabizadeh C, Fararouei M (2019) Compassionate care in healthcare systems: a systematic review. J Natl Med Assoc 111:546–554
67. Vellido A (2020) The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl 32:18069–18083
68. Zink W, Bernhard M, Keul W, Martin E, Völkl A, Gries A (2004) Invasive techniques in emergency medicine: I. Practice-oriented training concept to ensure adequately qualified emergency physicians. Der Anaesthesist 53:1086–1092
69. Pollack CV Jr, Amin A, Talan DA (2012) Emergency medicine and hospital medicine: a call for collaboration. Am J Med 125:826-e1
70. Muller H, Mayrhofer MT, Van Veen E-B, Holzinger A (2021) The ten commandments of ethical medical AI. Computer 54:119–123
71. Bhise V, Rajan SS, Sittig DF, Morgan RO, Chaudhary P, Singh H (2018) Defining and measuring diagnostic uncertainty in medicine: a systematic review. J Gen Intern Med 33:103–115
72. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R (2024) Large language models in medicine: the potentials and pitfalls: a narrative review. Ann Intern Med 177:210–220
73. Puladi B, Gsaxner C, Kleesiek J, Hölzle F, Röhrig R, Egger J (2023) The impact and opportunities of large language models

like ChatGPT in oral and maxillofacial surgery: a narrative review, Int J Oral Maxillofac Surg
74. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA (2023) The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol 228:696–705
75. Schwartz IS, Link KE, Daneshjou R, Cortés-Penfield N (2023) Black box warning: large language models and the future of infectious diseases consultation. Clin Infect Dis 78(4):860–866
76. Feldman J, Thomas-Bachli A, Forsyth J, Patel ZH, Khan K (2019) Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise. J Am Med Inform Assoc 26:1355–1359
77. Biswas SS (2023) Role of chat GPT in public health. Ann Biomed Eng 51:868–869
78. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, Cuocolo R, Cannella R, Koçak B (2023) Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions, Diagn Intervent Radiol, Epub–ahead
79. Alberts IL, Mercolli L, Pyka T, Prenosil G, Shi K, Rominger A, Afshar-Oromieh A (2023) Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? Eur J Nucl Med Mol Imaging 50:1549–1552
80. Chakraborty C, Bhattacharya M, Lee S-S (2023) Artificial intelligence enabled ChatGPT and large language models in drug target discovery, drug discovery, and development. Mol Therapy-Nucl Acids 33:866–868
81. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36:1234–1240
82. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu T-Y (2022) BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinf 23:bbac409
83. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, Weber T, Wesp P, Sabel BO, Ricke J et al (2023) ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 33:1–9
84. Lecler A, Duron L, Soyer P (2023) Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. Diagn Intervent Imaging 104:269–274
85. Shaikh F, Dehmeshki J, Bisdas S, Roettger-Dupont D, Kubassova O, Aziz M, Awan O (2021) Artificial intelligence-based clinical decision support systems using advanced medical imaging and radiomics. Curr Probl Diagn Radiol 50:262–267
86. Kraljevic Z, Shek A, Bean D, Bendayan R, Teo J, Dobson R (2021) MedGPT: medical concept prediction from clinical narratives, arXiv preprint arXiv:2107.03134
87. Shi X, Xu J, Ding J, Pang J, Liu S, Luo S, Peng X, Lu L, Yang H, Hu M, et al (2023) LLM-Mini-Cex: Automatic evaluation of large language model for diagnostic conversation, arXiv preprint arXiv:2308.07635
88. Zhou J, He X, Sun L, Xu J, Chen X, Chu Y, Zhou L, Liao X, Zhang B, Gao X (2023) SkinGPT-4: an interactive dermatology diagnostic system with visual large language model. MedRXiv: 2023–2006
89. Xiong H, Wang S, Zhu Y, Zhao Z, Liu Y, Huang L, Wang Q, Shen D (2023) DoctorGLM: Fine-tuning your chinese doctor is not a herculean task, arXiv preprint arXiv:2304.01097
90. Wang H, Liu C, Xi N, Qiang Z, Zhao S, Qin B, Liu T (2023a) Huatuo: tuning llama model with chinese medical knowledge, arXiv preprint arXiv:2304.06975
91. Wang G, Yang G, Du Z, Fan L, Li X (2023b) ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation, arXiv preprint arXiv:2306.09968
92. Lin X, Xu C, Xiong Z, Zhang X, Ni N, Ni B, Chang J, Pan R, Wang Z, Yu F et al (2023) PanGu drug model: learn a molecule like a human. Sci China Life Sci 66:879–882
93. Fang X, Wang F, Liu L, He J, Lin D, Xiang Y, Zhu K, Zhang X, Wu H, Li H et al (2023) A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. Nature Mach Intell 5:1087–1096
94. Mao J, Wang J, Zeb A, Cho K-H, Jin H, Kim J, Lee O, Wang Y, No KT (2023) Transformer-based molecular generative model for antiviral drug design. J Chem Inf Model 64(7):2733–2745
95. Luo Y, Liu XY, Yang K, Huang K, Hong M, Zhang J, Wu Y, Nie Z (2023) Towards unified AI drug discovery with multiple knowledge modalities, arXiv preprint arXiv:2305.01523
96. Zhu M, Chen Z, Yuan Y (2021) DSI-Net: deep synergistic interaction network for joint classification and segmentation with endoscope images. IEEE Trans Med Imaging 40:3315–3325
97. Lei W, Wei X, Zhang X, Li K, Zhang S (2023), MedLSAM: localize and segment anything model for 3d medical images, arXiv preprint arXiv:2306.14752
98. Li Z, Li Y, Li Q, Wang P, Guo D, Lu L, Jin D, Zhang Y, Hong Q (2023) Lvit: language meets vision transformer in medical image segmentation. IEEE Trans Med Imaging 43:96–107
99. Koleilat T, Asgariandehkordi H, Rivaz H, Xiao Y (2024) MedCLIP-SAM: bridging text and image towards universal medical image segmentation, arXiv preprint arXiv:2403.20253
100. Venigalla A, Frankle J, Carbin M (2022) PubMed GPT: a domain-specific large language model for biomedical text, Available online at: https://www.mosaicml.com/blog/introducing-pubmed-gpt
101. Yunxiang L, Zihan L, Kai Z, Ruilong D, You Z (2023) ChatDoctor: a medical chat model fine-tuned on llama model using medical domain knowledge, arXiv preprint arXiv:2303.14070
102. Bao Z, Chen W, Xiao S, Ren K, Wu J, Zhong C, Peng J, Huang X, Wei Z (2023) DISC-MedLLM: Bridging general large language models and real-world medical consultation, arXiv preprint arXiv:2308.14346
103. Chen Y, Wang Z, Xing X, Xu Z, Fang K, Wang J, Li S, Wu J, Liu Q, Xu X, et al (2023) BianQue: balancing the questioning and suggestion ability of health LLMs with multi-turn health conversations polished by ChatGPT, arXiv preprint arXiv:2310.15896
104. Qiu H, He H, hang S, Li A, Lan Z (2023) SMILE: single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support, arXiv preprint arXiv:2305.00450
105. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y (2024) PMC-LLaMA: toward building open-source language models for medicine. J Am Med Inf Assoc, ocae045
106. Liu F, Zhu T, Wu X, Yang B, You C, Wang C, Lu L, Liu Z, Zheng Y, Sun X et al (2023) A medical multimodal large language model for future pandemics. NPJ Digit Med 6:226
107. He J, Li P, Liu G, Zhao Z, Zhong S (2024) PeFoMed: parameter efficient fine-tuning on multimodal large language models for medical visual question answering, arXiv preprint arXiv:2401.02797
108. Wang J, Zhang G, Wang W, Zhang K, Sheng Y (2021) Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. J Cloud Comput 10:4
109. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Costa AB, Flores MG et al (2022) A large language model for electronic health records. NPJ Digit Med 5:194
110. Zeng A, Liu X, Du Z, Wang Z, Lai H, Ding M, Yang Z, Xu Y, Zheng W, Xia X, et al (2022) GLM-130B: an open bilingual pre-trained model, arXiv preprint arXiv:2210.02414

111. Yang Y, Yin X, Yang H, Fei X, Peng H, Zhou K, Lai K, Shen J (2021) KGSynNet: a novel entity synonyms discovery framework with knowledge graph. Database systems for advanced applications. Springer, Berlin, pp 174–190
112. Zhao X, Wu J, Peng H, Beheshti A, Monaghan JJ, McAlpine D, Hernandez-Perez H, Dras M, Dai Q, Li Y et al (2022) Deep reinforcement learning guided graph neural networks for brain network analysis. Neural Netw 154:56–67
113. Koubaa A (2023) GPT-4 versus GPT-3.5: A concise showdown. Available online at: https://www.techrxiv.org/articles/preprint/GPT-4_vs_GPT-3_5_A_Concise_Showdown/22312330
114. Zhao X, Liu H, Dai Q, Peng H, Bai X, Peng H (2023) Multi-omics sampling-based graph transformer for synthetic lethality prediction, In: IEEE international conference on bioinformatics and biomedicine, IEEE, pp 785–792
115. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F (2023) Applications of large language models in cancer care: current evidence and future perspectives. Front Oncol 13:1268915
116. Yuan J, Bao P, Chen Z, Yuan M, Zhao J, Pan J, Xie Y, Cao Y, Wang Y, Wang Z, et al (2023) Advanced prompting as a catalyst: empowering large language models in the management of gastrointestinal cancers, Innov 521
117. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, Sigler C, Knödler M, Keller U, Beule D et al (2023) Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 6:e2343689–e2343689
118. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, Sedoc J, DeRubeis RJ, Willer R, Eichstaedt JC (2024) Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. npj Mental Health Res 3:12
119. Li T, Shetty S, Kamath A, Jaiswal A, Jiang X, Ding Y, Kim Y (2024) CancerGPT for few shot drug pair synergy prediction using large pretrained language models. npj Digit Med 7:40
120. Gala D, Makaryus AN (2023) The utility of language models in cardiology: a narrative review of the benefits and concerns of ChatGPT-4. Int J Environ Res Public Health 20:6438
121. Arslan S (2023) Exploring the potential of ChatGPT in personalized obesity treatment. Ann Biomed Eng 51(9):1887–1888
122. Yang Z, Yao Z, Tasmin M, Vashisht P, Jang WS, Ouyang F, Wang B, Berlowitz D, Yu H (2023) Performance of multimodal gpt-4v on usmle with image: potential for imaging diagnostic support with explanations, medRxiv 2023–10
123. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D (2023) The role of large language models in medical education. Appl Implic 9:e50945
124. Zeng G, Yang W, Ju Z, Yang Y, Wang S, Zhang R, Zhou M, Zeng J, Dong X, Zhang R et al (2020) Meddialog: large-scale medical dialogue datasets. In: The conference on empirical methods in natural language processing. pp 9241–9250
125. Jin H, Chen R, Zhou A, Chen J, Zhang Y, Wang H (2024) GUARD: role-playing to generate natural-language jailbreakings to test guideline adherence of large language models, arXiv preprint arXiv:2402.03299
126. Ye J, Chen X, Xu N, Zu C, Shao Z, Liu S, Cui Y, Zhou Z, Gong C, Shen Y, et al (2023) A comprehensive capability analysis of GPT-3 and GPT-3.5 series models, arXiv preprint arXiv:2303.10420
127. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, et al (2023) GPT-4 technical report, arXiv preprint arXiv:2303.08774
128. Dodge J, Ilharco G, Schwartz R, Farhadi A, Hajishirzi H, Smith N (2020) Fine-tuning pretrained language models: weight initializations, data orders, and early stopping, arXiv preprint arXiv:2002.06305
129. Osborne JD, Flatow J, Holko M, Lin SM, Kibbe WA, Zhu L, Danila MI, Feng G, Chisholm RL (2009) Annotating the human genome with disease ontology. BMC Genom 10:S6
130. Wang L, Lou Z, Jiang K, Shen G (2019) Bio-multifunctional smart wearable sensors for medical devices. Adv Intell Syst 1:1900040
131. Huang R, Li H, Suomi R, Li C, Peltoniemi T (2023) Intelligent physical robots in health care: systematic literature review. J Med Internet Res 25:e39786
132. Preum SM, Munir S, Ma M, Yasar MS, Stone DJ, Williams R, Alemzadeh H, Stankovic JA (2021) A review of cognitive assistants for healthcare: trends, prospects, and future directions. ACM Comput Surv 53:1–37
133. Chen Z, Wu J, Gan W, Qi Z (2022) Metaverse security and privacy: an overview. In: IEEE international conference on big data, IEEE, pp 2950–2959
134. Chen Z, Gan W, Wu J, Lin H, Chen C-M (2024) Metaverse for smart cities: a surveys. Internet Things Cyber-Phys Syst 4:203–216
135. He J, Vechev M (2023) Controlling large language models to generate secure and vulnerable code, arXiv preprint arXiv:2302.05319
136. Roman-Belmonte JM, De la Corte-Rodriguez H, Rodriguez-Merchan EC (2018) How blockchain technology can change medicine. Postgrad Med 130:420–427
137. Chen C, Feng X, Zhou J, Yin J, Zheng X (2023) Federated large language model: a position paper, arXiv preprint arXiv:2307.08925
138. Li Y, Liu C, Zou H, Che L, Sun P, Yan J, Liu W, Xu Z, Yang W, Dong L, et al (2023) Integrated wearable smart sensor system for real-time multi-parameter respiration health monitoring. Cell Rep Phys Sci 4
139. Wu X, Liu C, Wang L, Bilal M (2023) Internet of things-enabled real-time health monitoring system using deep learning. Neural Comput Appl 14565–14576
140. Dou Y, Huang Y, Zhao X, Zou H, Shang J, Lu Y, Yang X, Xiao J, Peng S (2024) ShennongMGS: An LLM-based chinese medication guidance system. ACM Trans Manage Inf Syst
141. Raheja N, Manocha AK (2023) An IoT enabled secured clinical health care framework for diagnosis of heart diseases. Biomed Signal Process Control 80:104368
142. Neo JRE, Ser JS, Tay SS (2024) Use of large language model-based chatbots in managing the rehabilitation concerns and education needs of outpatient stroke survivors and caregivers. Front Digital Health 6:1395501
143. Revell G (2024) Generative AI applications in the health and well-being domain: Virtual and robotic assistance and the need for niche language models (NLMs). Applications of generative AI. Springer, Berlin, pp 189–207
144. Chen K, Du Y, You T, Islam M, Guo Z, Jin Y, Chen G, Heng P-A (2024) LLM-assisted multi-teacher continual learning for visual question answering in robotic surgery, arXiv preprint arXiv:2402.16664
145. Padmanabha A, Yuan J, Gupta J, Karachiwalla Z, Majidi C, Admoni H, Erickson Z (2024) Voicepilot: Harnessing LLMs as speech interfaces for physically assistive robots, arXiv preprint arXiv:2404.04066
146. Dong XL, Moon S, Xu YE, Malik K, Yu Z (2023) Towards next-generation intelligent assistants leveraging LLM techniques. In: The 29th ACM SIGKDD conference on knowledge discovery and data mining. pp 5792–5793
147. Vu MD, Wang H, Li Z, Chen J, Zhao S, Xing Z, Chen C (2024) GPTVoiceTasker: LLM-powered virtual assistant for smartphone, arXiv preprint arXiv:2401.14268

148. Chen Z, Gan W, Sun J, Wu J, Yu PS (2024) Open metaverse: issues, evolution, and future. In: Companion proceedings of the ACM web conference, pp 1351–1360
149. Yang R, Li L, Gan W, Chen Z, Qi Z (2023) The human-centric metaverse: a survey. In: Companion proceedings of the ACM web conference, pp 1296–1306
150. El Saddik A, Ghaboura S (2023) The integration of ChatGPT with the Metaverse for medical consultations. IEEE Consum Electron Mag 13:6–15
151. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider C, Forte AJ (2024) AI and ethics: a systematic review of the ethical considerations of large language model use in surgery research. Healthcare 12:825
152. Marks M, Haupt CE (2023) AI chatbots, health privacy, and challenges to HIPAA compliance. JAMA 330:309–310
153. Lawlor RT (2023) The impact of GDPR on data sharing for European cancer research. Lancet Oncol 24:6–8
154. Heston TF (2024) Prespective chapter: integrating large language models and blockchain in telemedicine, IntechOpen
155. Chen M-Y, Chiang H-S, Sangaiah AK, Hsieh T-C (2020) Recurrent neural network with attention mechanism for language model. Neural Comput Appl 32:7915–7923
156. Singh C, Askari A, Caruana R, Gao J (2023) Augmenting interpretable models with large language models during training. Nature Commun 14:7913
157. Song Y, Zhang J, Tian Z, Yang Y, Huang M, Li D (2024) LLM-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification, arXiv preprint arXiv:2402.16515