Application of large language models in medicine

Fenglin Liu[1,*,†], Hongjian Zhou[1,*,†], Boyang Gu[2,*], Xinyu Zou[3,*], Jinfa Huang[4,*], Jinge Wu[5,*], Yiru Li[6], Sam S. Chen[7], Yining Hua[8], Peilin Zhou[9], Junling Liu[10], Chengfeng Mao[11], Chenyu You[12], Xian Wu[13], Yefeng Zheng[13], Lei Clifton[14], Zheng Li[15,†], Jiebo Luo[4,†], David A. Clifton[1,16,†]

† These authors jointly supervised this work. * These authors contributed equally.
[1]Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[2]Department of Computing, Imperial College London, London, UK
[3]Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada
[4]Department of Computer Science, University of Rochester, Rochester, NY, USA
[5]Institute of Health Informatics, University College London, London, UK
[6]Western University, London, Ontario, Canada
[7]Department of Kinesiology, University of Georgia, Athens, GA, USA
[8]Harvard T.H. Chan School of Public Health, Boston, MA, USA
[9]Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China [10]Alibaba, Hangzhou, China
[11]Massachusetts Institute of Technology, Cambridge, MA, USA [12]Stony Brook University, Stony Brook, NY, USA
[13]Tencent Jarvis Research Center, Beijing, China
[14]Nuffield Department of Population Health, University of Oxford, Oxford, UK [15]Amazon, Palo Alto, CA, USA
[16]Oxford-Suzhou Centre for Advanced Research, Suzhou, China
{david.clifton,fenglin.liu}@eng.ox.ac.uk,

ABSTRACT

Large language models (LLMs), such as ChatGPT, have received great attention owing to their capabilities for understanding and generating human language. Despite a trend in researching the application of LLMs in supporting different medical tasks (such as enhancing clinical diagnostics and providing medical education), a comprehensive assessment of their development, practical applications and outcomes in the medical space is still missing. Thus, this Review aims to provide an overview of the development and deployment of LLMs in medicine, including the challenges and opportunities they face. In terms of development, we discuss the principles of existing medical LLMs, including their basic model structures, number of parameters, sources and scales of data used for model development. In terms of deployment, we compare different LLMs across various medical tasks and with state-of-the-art lightweight models.

Key points

- Existing medical LLMs, ranging from 110 million to 520 billion parameters, are mainly developed through pre-training, fine-tuning and prompting methods using large-scale medical corpora from different sources.
- Their performance is mostly evaluated based on exam-style QA tasks. Combining different fine-tuning and prompting methods enables LLMs to achieve comparable or even better results than experts.
- LLMs perform poorly in non-QA tasks without pre-set options, thus requiring further improvements before integration into real clinical decision-making processes.
- Medical LLMs are being adapted to various clinical applications but large-scale clinical trials are still missing.
- Mitigating hallucinations, establishing robust data, benchmarks, metrics and addressing ethical, safety, and regulatory concerns through interdisciplinary collaborations will help accelerate the integration of LLMs into clinic practice.

Shot summary

Large language models (LLMs) have received great attention owing to their capabilities to understand and generate human language. This Review aims to provide an overview of the development and deployment of LLMs in medicine, including the challenges and opportunities they face.

1 [H1] Introduction

The recently emerged general large language models (LLMs)[1,2], such as PaLM[3], LLaMA[4,5] , GPT-series[6,7] and ChatGLM[8] , have advanced the state-of-the-art in various natural language processing (NLP) tasks, including text generation, text summarization and question answering (QA), as well as to adapt them to the medical domain (**Box 1**) [9,10]. For example, PaLM[3] and GPT-4[7], MedPaLM-2[10] and MedPrompt[11] have achieved a competitive accuracy of 86.5 and 90.2 compared to human experts (87.0[12]) in the United States Medical Licensing Examination (USMLE) [13], respectively. Based on publicly available general LLMs (such as LLaMA[5]) a wide range of medical LLMs including ChatDoctor[14], MedAlpaca[15], PMC-LLaMA[12], BenTsao[16] and Clinical Camel[17], have been introduced to assist medical professionals to improve patient care[18,19]. Despite these advances, limitations remain; first, many of these models mainly focus on medical dialogue and QA tasks, but their utility in clinical practice (such as electronic health records (EHRs)[20], discharge summary generation[19], health education[21] and care planning[10]) is often overlooked [18]. Moreover, current LLMs often fail to provide practical guidelines and are tested on a small number of users.

In this Review, we first analyse the principles underpinning current medical LLMs, providing detailed descriptions of their architecture, parameter scales and the datasets used during their development (**Box 1**). Next, we evaluate their performance across ten biomedical NLP tasks, including both discriminative and generative tasks. We then explore the practical deployment of medical LLMs in clinical settings by providing guidelines tailored for seven distinct clinical application scenarios. Finally, we discuss challenges such as the risk of generating factually inaccurate yet plausible outputs (hallucination) and the ethical, legal and safety implications, as well as propose promising research directions. We argue for a comprehensive evaluation framework that assesses the trustworthiness of medical LLMs to ensure their responsible and effective use in the healthcare domain. We also maintain a regularly updated list of practical guides on medical LLMs at: https://github.com/AI-in-Health/MedLLMsPracticalGuide.

2 [H1] The principles of medical LLMs

Existing medical LLMs are mainly pre-trained from scratch, fine-tuned from existing general LLMs or directly obtained through prompting to align the general LLMs to the medical domain (**Table 1**). In this section, we introduce the principles of medical LLMs in terms of pre-training, fine-tuning and prompting (**Figure 1**).

2.1 [H2] Pre-training

Pre-training typically involves training an LLM on a large corpus of medical texts, including both structured and unstructured text such as EHRs[22], clinical notes[20] and medical literature[23]. PubMed, MIMIC-III clinical notes[24] and PubMed Central (PMC) literature are three widely used medical corpora for medical LLM pre-training (**Table 1**) [25]. Pre-training medical LLMs typically involves refining the following training objectives: masked language modelling, next sentence prediction and next token prediction (**Box 1**). For example, BERT-series models (such as BioBERT[26], PubMedBERT[27], ClinicalBERT[23] and GatorTron[20], which are originally derived from the general domain BERT or RoBERTa models) mainly adopt the masked language modelling and the next sentence prediction for pre-training, whereas GPT-series models (such as BioGPT[28], and GatorTronGPT[22]) mainly adopt the next token prediction for pre-training (**Box 1**). After pre-training, medical LLMs can learn rich medical knowledge that can be leveraged to achieve strong performance on different medical tasks.

2.2 [H2] Fine-tuning

Training a medical LLM from scratch is expensive and time-consuming; one solution is to fine-tune the general LLMs with medical data using methods such as Supervised Fine-Tuning (SFT), Instruction Fine-Tuning (IFT) and Parameter-Efficient Fine-Tuning (PEFT)[10,15,17] (**Table 1**).

[H3] Supervised fine-tuning (SFT) aims to leverage high-quality medical corpus, which can be physician-patient conversations [14], medical QA[15] and knowledge graphs[16,29]. The developed SFT data serves as a continuation of the pre-training data to further pre-train the general LLMs with the same training objectives, such as next token prediction. DoctorGLM[30] and ChatDoctor[14] are obtained by fine-tuning ChatGLM[8]and LLaMA[4] on the physician-patient dialogue data, respectively. MedAlpaca [15] based on the general LLM Alpaca is fine-tuned using over 160,000 medical QA pairs sourced from diverse medical corpora. Clinicalcamel[17] combines physician-patient conversations, clinical literature and medical QA pairs to refine the LLaMA-2 model[5]. In particular, Qilin-Med [29] and Zhongjing[31] are obtained by incorporating the knowledge graph to perform fine-tuning on the Baichuan[31] and LLaMA[4], respectively.

[H3] Instruction fine-tuning (IFT) generates instruction-based training datasets[32,33] which typically comprise instruction-input-output triples, such as instruction-QA.

To ensure high quality of training data and generalizability to different medical instructions and scenarios, MedPaLM-2 [10] invited qualified medical professionals for input, BenTsao [16] developed knowledge-based instruction data from the knowledge graph (cMeKG [35]), whereas MedAlpaca [15] incorporated both medical dialogues and QA pairs for instruction fine-tuning.

Table 1. Summary of medical-domain LLMs. M: million, B: billion.

Domains	Model Development	Models	# Params	Data Scale	Data Source
Medical-domain LLMs (Sec. 2)	Pre-training (Sec. 2.1)	BioBERT[26]	110M	18B tokens	PubMed+PMC
		PubMedBERT[27]	110M/340M	3.2B tokens	PubMed+PMC
		SciBERT[170]	110M	3.17B tokens	Literature
		NYUTron[73]	110M	7.25M notes, 4.1B tokens	NYU Notes[73]
		ClinicalBERT[23]	110M	112k clinical notes	MIMIC-III[24]
		BioM-ELECTRA[171]	110M/335M	-	PubMed
		BioMed-RoBERTa[172]	125M	7.55B tokens	S2ORC[172]
		BioLinkBERT[173]	110M/340M	21GB	PubMed
		BlueBERT[174]	110M/340M	>4.5B tokens	PubMed+MIMIC-III[24]
		SciFive[175]	220M/770M	-	PubMed+PMC
		ClinicalT5[176]	220M/770M	2M clinical notes	MIMIC-III[24]
		MedCPT[177]	330M	255M articles	PubMed
		DRAGON[178]	360M	6GB	BookCorpus[178]
		BioGPT[28]	1.5B	15M articles	PubMed
		BioMedLM[179]	2.7B	110GB	Pile[179]
		OphGLM[180]	6.2B	20k dialogues	MedDialog[180]
		GatorTron[620]	8.9B	>82B tokens+6B tokens 2.5B tokens+0.5B tokens	EHRs[20] +PubMed Wiki+MIMIC-III[24]
		GatorTronGPT[22]	5B/20B	277B tokens	EHRs[22]
	Fine-tuning (Sec. 2.2)	DoctorGLM[30]	6.2B	323MB dialogues	CMD
		BianQue[181]	6.2B	2.4M dialogues 96k EHRs + 100k dialogues	BianQueCorpus[181] MD-EHR[182] +MedDialog[182]
		ClinicalGPT[182]	7B	192 medical QA	VariousMedQA[182]
		Qilin-Med[29]	7B	3GB	ChiMed[29]
		ChatDoctor[14]	7B	110k dialogues	HealthCareMagic+jClinig
		BenTsao[16]	7B	8k instructions	CMeKG-8K[35]
		HuatuoGPT[183]	7B	226k instructions&dialogues	Hybrid SFT[183]
		Baize-healthcare[46]	7B	101K dialogues	Quora+MedQuAD[46]
		BioMedGPT[184]	10B	>26B tokens	S2ORC[184]
		MedAlpaca[15]	7B/13B	160k medical QA	Medical Meadow[15]
		AlpaCare[41]	7B/13B	52k instructions	MedInstruct-52k[41]
		Zhongjing[31]	13B	70k dialogues	CMtMedQA[31]
		PMC-LLaMA[12]	13B	79.2B tokens	Books+Literature[12] +MedC-I[12]
		CPLLM[47]	13B	109k EHRs	eICU-CRD[47] +MIMIC-IV
		Med42[185]	7B/70B	250M tokens	PubMed+ MedQA[13] + OpenOrca
		MEDITRON[186],	7B/70B	48.1B tokens	PubMed+Guidelines[186]
		OpenBioLLM	8B/70B	-	-
		MedLlama3-v20	8B/70B	-	-
		Clinical Camel[17]	13B/70B	70k dialogues+100k articles 4k medical QA	ShareGPT+PubMed MedQA[13]
		MedPaLM-2[10]	340B	193k medical QA	MultiMedQA[10]
		Med-Flamingo[36]	-	600k pairs	Multimodal Textbook[36] +PMC-OA[36] VQA-RAD[36] +PathVQA[186]
		LLaVA-Med[37]	-	660k pairs	PMC-15M[37] +VQA-RAD[37] SLAKE[37] +PathVQA[186]
		MAIRA-1[39]	-	337k pairs	MIMIC-CXR[187]
		RadFM[40]	-	32M pairs	MedMD[40]
		Med-Gemini[188]	-	-	MedQA-R&RS[38] ++MultiMedQA[10] +MIMIC-III[24] +MultiMedBench[188]
	Prompting (Sec. 2.3)	CodeX[189]	GPT-3.5 / LLaMA-2	Chain-of-Thought (CoT)[49]	-
		DeID-GPT[50]	ChatGPT / GPT-4	Chain-of-Thought (CoT)[49]	-
		ChatCAD[81]	ChatGPT	In-Context Learning (ICL)	-
		Dr. Knows[74]	ChatGPT	ICL	UMLS
		MedPaLM[9]	PaLM (540B)	CoT & ICL	MultiMedQA[10]
		MedPrompt[11]	GPT-4	CoT & ICL[49]	-
		Chat-Orthopedist[63]	ChatGPT	Retrieval-Augmented Generation (RAG)	PubMed+Guidelines +UpToDate +Dyname
		QA-RAG[62]	ChatGPT	RAG	FDA QA[62]
		Almanac[61]	ChatGPT	RAG & CoT	Clinical QA[61]
		Oncology-GPT-4[185]	GPT-4	RAG & ICL	Oncology Guidelines from ASCO and ESMO

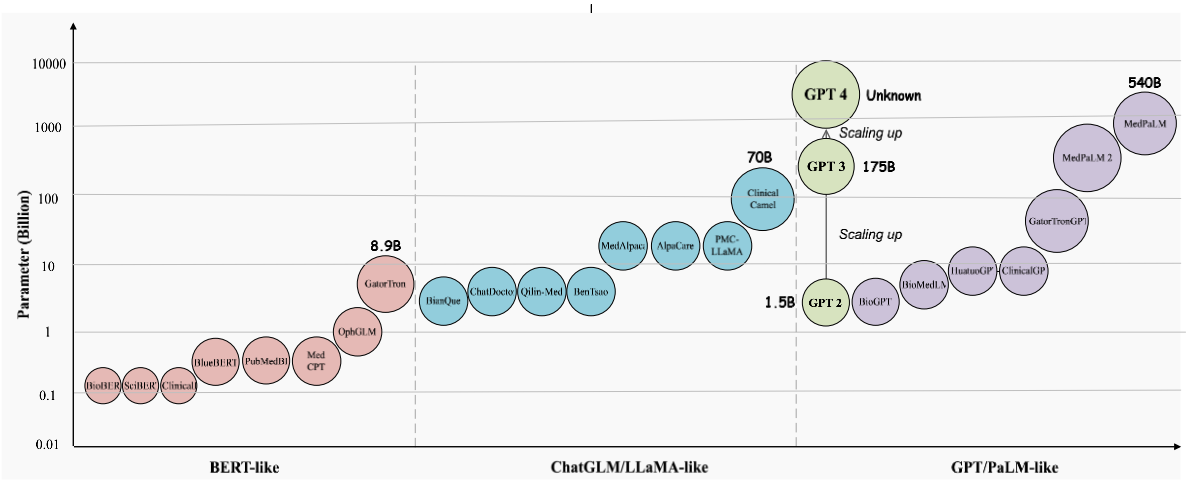

Figure 1. Model size for different medical LLMs. Data from Table 1 was used to illustrate the development of model sizes for medical LLMs in different model architectures, that is, BERT, ChatGLM/LLaMA and GPT/PaLM.

Multimodal LLMs such as Med-Flamingo [36], LLaVA-Med [37] and Med-Gemini [38] have expanded the capabilities of LLMs to process complex and multimodal medical data; for example, Med-Flamingo[36] undergoes IFT on medical image-text data, learning to identify abnormalities and generate diagnostic reports; LLaVA-Med's[37] two-stage IFT process involves aligning medical concepts across visual and textual modalities, followed by fine-tuning on different medical instructions; whereas Med-Gemini's[38] IFT uses a curated dataset of medical instructions and multimodal data, enabling it to comprehend complex medical concepts, procedures and diagnostic reasoning. Meanwhile, MAIRA-1[39] and RadFM[40] are two multimodal LLMs specifically designed for radiology applications. MAIRA-1[39] undergoes IFT on a dataset of radiology instructions and corresponding medical images to analyse radiological images and generate accurate diagnostic reports. By contrast, RadFM[40] leverages a pre-training approach on a large corpus of radiology-specific image-text data, followed by IFT on different sets of radiology instructions. These models' multimodal IFT approaches enable them to bridge the gap between visual and textual medical information, perform a wide range of medical tasks accurately and generate context-aware responses to complex medical queries. The goal of IFT is to improve the model's ability to follow various human and task instructions, align their outputs with the medical domain and produce a specialized medical LLM. Thus, the main difference between SFT and IFT is that the former focuses on injecting medical knowledge into a general LLM through continued pre-training, therefore improving its ability to understand the medical text and accurately predict the next token. By contrast, IFT aims to improve the model's 'instruction following' ability and adjust its outputs to match the given instructions rather than accurately predicting the next token as in SFT[32]. As a result, SFT emphasizes the 'quantity' of training data whereas IFT emphasizes their 'quality and diversity'. Combination of both fine-tuning methods have also been attempted[31,29,41] to simultaneously improve LLM's ability to understand medical knowledge and follow human instructions, resulting in improved overall task performance. For example, Zhongjing[31] first collected a large amount of real medical corpus for SFT and then incorporated multi-turn dialogue as instruction data to perform IFT.

[H3] Parameter-efficient fine-tuning (PEFT) aims to substantially reduce computational and memory requirements for fine-tuning general LLMs. The main concept is to keep most of the parameters in pre-trained LLMs unchanged by fine-tuning only the smallest subset of parameters (or additional parameters). Commonly used PEFT techniques include Low-Rank Adaptation (LoRA)[42], Prefix Tuning[43] and Adapter Tuning[44].

In contrast to fine-tuning full-rank weight matrices, LoRA preserves the parameters of the original LLMs and only adds trainable low-rank matrices into the self-attention module of each Transformer layer [42]. Thus, LoRA can substantially reduce the number of trainable parameters and improve the efficiency of fine-tuning while preserving the ability of the fine-tuned LLM to capture the characteristics of the tasks. Prefix Tuning instead adds a small set of continuous task-specific vectors (that is, 'prefixes') to the input of each Transformer layer[43,1]. These prefixes serve as the additional context to guide the generation of the model without changing the original pre-trained parameter weights. Adapter Tuning introduces small neural network modules, known as adapters, into each Transformer layer of the pre-trained LLMs [45]. These adapters are fine-tuned while keeping the original model parameters frozen [45] for flexible and efficient fine-tuning. Compared to full-rank fine-tuning, the number of trainable parameters introduced by adapters is relatively small, yet they enable the LLMs to adapt to clinical text classification and understanding tasks effectively, achieving similar performance with fewer trainable parameters. In general, PEFT is valuable for developing domain-specific LLMs (such as medical) owing to its ability to reduce computational demands while maintaining the model performance. For example, medical LLMs DoctorGLM [30], MedAlpaca [15], Baize-Healthcare [46], Zhongjing[31], and Clinical Camel[17] adopted the LoRA[42] to align the general LLMs to the medical domain.

2.3 [H2] Prompting

Fine-tuning considerably reduces computational costs compared to pre-training, but still requires model training and collections of high-quality datasets. By contrast, the 'prompting' methods align general LLMs to the medical domain without training any model parameters (such as MedPrompt [11] and MedPaLM [9]). Popular prompting methods include In-Context Learning (ICL), Chain-of-Thought (CoT) prompting, Prompt Tuning and Retrieval-Augmented Generation (RAG).

[H3] In-context learning (ICL) aims to give direct instructions to prompt the LLM to perform a task. ICL consists of four processes: task understanding, context learning, knowledge reasoning and answer generation. First, the model must understand the specific requirements and goals of the task. Second, the model learns to understand the contextual information related to the task with argument context. Then, it uses the model's internal knowledge and reasoning capabilities to understand the patterns and logic in the example. Finally, the model generates the task-related answers. The advantage of ICL is that it does not require a large amount of labelled data for fine-tuning. Based on the type and number of input examples, ICL can be divided into three categories[48]: one-shot Prompting where only one example and task description can be entered; few-shot Prompting which allows the input of multiple instances and task descriptions and Zero-shot Prompting where only task descriptions can be entered. ICL enables the LLMs to make task predictions based on contexts augmented with a few examples and task demonstrations. It allows the LLMs to learn from these examples or demonstrations to perform the task and follow the given examples to give corresponding answers [6]. For example, MedPaLM [9] substantially improves the task performance by providing the general LLM, PaLM [3], with a small number of task examples such as medical QA pairs.

[H3] Chain-of-thought (CoT) prompting further improves the accuracy and logic of model output compared with ICT. Specifically, through prompting words, CoT aims to prompt the model to generate intermediate steps or paths of reasoning when dealing with downstream (complex) problems [49]. Moreover, CoT can be combined with few-shot prompting by giving reasoning examples, thus enabling medical LLMs to give reasoning processes when generating responses. CoT improves model performance for tasks involving complex reasoning (such as medical QA)[9,10]. Medical LLMs such as DeID-GPT[50], MedPaLM [9] and MedPrompt [11] use CoT prompting to assist them in simulating a diagnostic thought process, thus providing more transparent and interpretable predictions or diagnoses. In particular, MedPrompt [11] directly prompts a general LLM, GPT-4[7], to outperform the fine-tuned medical LLMs on medical QA without training any model parameters.

[H3] Prompt tuning aims to improve the model performance by employing both prompting and fine-tuning techniques [51]. The prompt tuning method introduces learnable prompts (that is, trainable continuous vectors) which can be optimized or adjusted during the fine-tuning process to better adapt to different medical scenarios and tasks. Thus, they provide a more flexible way of prompting LLMs than the 'prompting alone' methods that use discrete and fixed prompts. In contrast to traditional fine-tuning methods that train all model parameters, prompt tuning only tunes a very small set of parameters (that is, less than 3% of the total model parameters) associated with the prompts themselves, instead of extensively training the model parameters.

MedPaLM [9] and MedPaLM- 2 [10] have combined all the above prompting methods resulting in an 'Instruction Prompt Tuning' to improve performances on various medical QA datasets. Using the MedQA dataset for the USMLE, MedPaLM-2 [10] achieves a competitive overall accuracy of 86.5% compared to human experts (87.0%), surpassing MedPaLM [9] by a large margin (19%).

[H3] Retrieval-augmented generation (RAG) improves the performance of LLMs by integrating external knowledge into the generation process to minimize hallucinations, obscure reasoning processes and reliance on outdated information [52]. RAG consists of three main components: retrieval, augmentation and generation. The retrieval component uses various indexing strategies and input query processing techniques to search and rank relevant information from an external knowledge base. The retrieved external data is then augmented into the LLM's prompt, providing additional context and grounding for the generated response. By directly updating the external knowledge base, RAG mitigates the risk of amnesia (that is, catastrophic forgetting [53]) associated with model weight modifications, making it particularly suitable for domains with low error tolerance and rapidly evolving information, such as the medical field. In contrast to traditional fine-tuning methods, RAG incorporates new medical information without compromising the model's previously acquired knowledge. MIRAGE [54] is the first benchmark to be proposed based on medical information RAG, including 7,663 questions from five medical QA datasets.

In RAG, retrieval can be achieved by calculating the similarity between the textual embeddings of the question and document chunks. Textual embeddings are vector representations of text, which are encoded by embedding models (such as AngIE [55], Voyage [56] and BGE [57]) into numerical vectors that capture the semantic information of the text. These embedding models serve as the foundation for ensuring that the retrieval step accurately captures the semantic meaning of the question, retrieve accurate documents, and ultimately enhance the quality of the generated responses. In addition to embedding, the retrieval process can be optimized by adaptive retrieval, recursive retrieval and iterative retrieval, among others [58,59,60]. Adaptive retrieval dynamically adjusts retrieval strategies (for example, prioritize specific document chunks) based on the complexity or specificity of the question, improving relevance and efficiency. Recursive retrieval refines retrieval results in multiple rounds, where the retrieved documents are re-evaluated and refined in successive rounds to progressively narrow down the results. Iterative retrieval incorporates prior retrieval results to improve subsequent retrieval attempts. For example, Almanac[61] is a large language framework augmented with retrieval capabilities for medical guidelines and treatment recommendations,

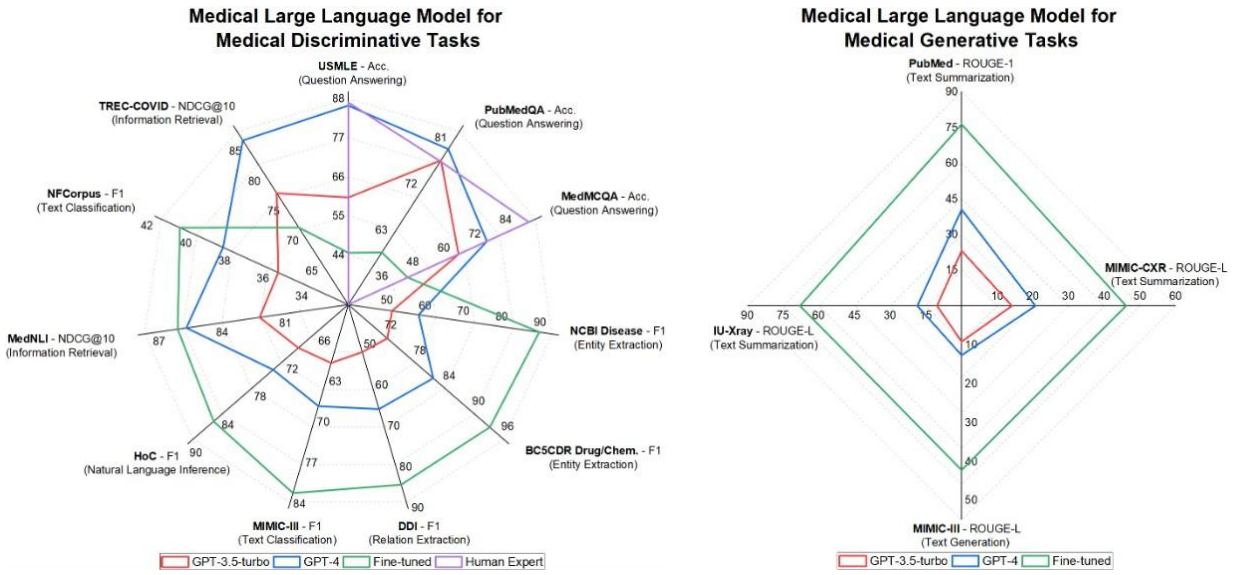

Figure 2. Performance comparison. Performance (Dataset-Metric (Task)) comparison between the GPT-3.5 turbo, GPT-4, state-of-the-art task-specific lightweight models (Fine-tuned), and human experts, on seven medical tasks across eleven datasets. All data presented in our Figures originates from published and peer-reviewed literature (**Supplementary information**).

surpassing ChatGPT on clinical scenario evaluations, particularly in terms of completeness and safety. Another example is QA-RAG[62] which employs RAG with LLM for pharmaceutical regulatory tasks, where the model searches for relevant guideline documents and provides answers based on the retrieved guidelines. Chat-Orthopedist[63] , a retrieval-augmented LLM, assists people with adolescent idiopathic scoliosis to prepare for discussions with clinicians by providing accurate and comprehensible responses to patient inquiries, leveraging the adolescent idiopathic scoliosis domain knowledge from PubMed clinical papers, Scoliosis Research Society's (SRS) practice guidelines, and UpToDate.com.

3 [H1] Medical tasks

In this section, we will introduce two popular types of medical machine learning tasks: generative and discriminative tasks, including ten representative tasks that further build up clinical applications (**Figure 2**). A detailed definition of the task and performance comparisons across different LLMs can be found in the **Supplementary information**.

3.1 [H2] Discriminative tasks

Discriminative tasks categorize or differentiate data (for example, structured or unstructured text) into specific classes or categories based on given input data. The representative tasks include QA, Entity Extraction, Relation Extraction, Text Classification, Natural Language Inference, Semantic Textual Similarity and Information Retrieval (**Supplementary information**). The typical input for discriminative tasks can be medical questions, clinical notes, medical documents, research papers and patient EHRs. The output can be labels, categories, extracted entities, relationships or answers to specific questions, which are often structured and categorized information derived from the input text. In existing LLMs, the discriminative tasks are widely studied and used to make predictions and extract information from input text. Entity extraction can automatically identify and categorize critical information (that is, entities) such as symptoms, medications, diseases, diagnoses and lab results from patient EHRs, thus assisting in organizing and managing patient data. Entity linking is a subsequent step to entity extraction, where the identified entities are mapped to entries in a structured knowledge base or a standardized terminology system. For example, an extracted entity like 'diabetes' can be linked to a specific concept in SNOMED CT, UMLS (Unified Medical Language System) or ICD codes. This process ensures that the extracted entities are standardized, enabling interoperability across different systems and facilitating more precise medical analysis and data management.

3.2 [H2] Generative tasks

Different from discriminative tasks that focus on understanding and categorizing the input text, generative tasks require a model to generate new, fluent and accurate text based on given inputs. These tasks include medical text summarization [64,65], medical text generation[28] and medical text simplification [66].

For medical text summarization, the input and output are typically long and detailed medical text (for example, 'Findings' in radiology reports) and a concise summarized text (such as the 'Impression' in radiology reports), respectively. In medical text generation (for example, discharge instruction generation [67]), the input can be medical conditions, symptoms, patient

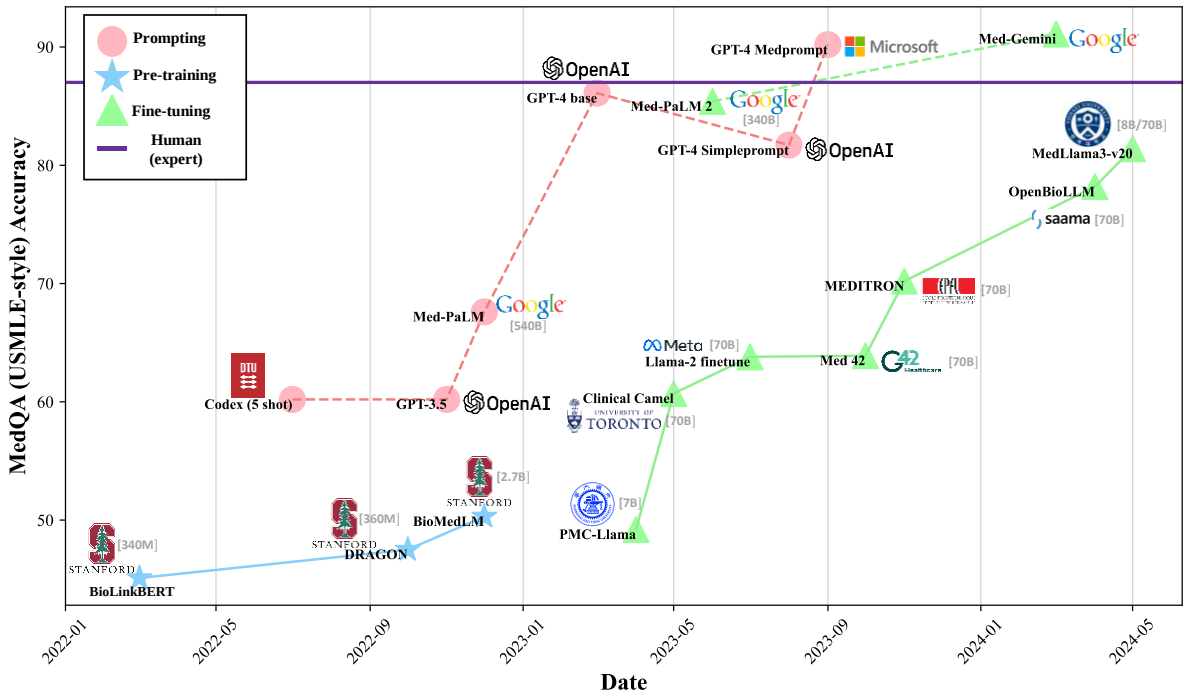

Figure 3. Development of medical LLMs over time. Illustration of the development of different medical LLMs over time by assessing the scores of the United States Medical Licensing Examination (USMLE) from the MedQA dataset. Solid and dashed lines represent open-source and closed-source models, respectively.

demographics or even a set of medical notes or test results. The output can be a diagnosis recommendation of a medical condition, personalized instructional information or health advice for the patient to manage their condition outside the hospital.

Medical text simplification [66] aims to generate a simplified version of the complex medical text by, for example, clarifying and explaining medical terms and therefore improve readability. Specifically, complicated or opaque words are replaced, complex syntactic structures improved and rare concepts explained [68].

3.3 [H2] Performance comparisons

General LLMs such as GPT-3.5-turbo and GPT-4[7] have achieved strong performance on existing medical machine learning tasks and even outperform existing task-specific fine-tuned models and human experts for QA tasks (MedQA (USMLE) [13], PubMedQA [69], and MedMCQA[70]) (**Figure 2, 3**). However, existing general LLMs perform worse than the task-specific fine-tuned models on non-QA discriminative tasks (**Figure 2**). For example, task-specific fine-tuned model BioBERT [26] achieves an F1 score of 89.36 compared with 56.73 by GPT-4 on the entity extraction task. A higher F1 score indicates better performance in balancing precision (correctly identified entities) and recall (completeness of identified entities). Notably, GPT-4 underperforms against task-specific lightweight models on all datasets in generative tasks. We hypothesize that general LLMs are strong in QA because these are close-ended tasks; that is, the correct answer is already provided by multiple candidates. By contrast, most non-QA tasks are open-ended where the model must predict the correct answer from a large pool of possible candidates, or even without any candidate provided.

Real-world clinical practice often involves answering open-ended questions without pre-set options and is therefore different from the structured nature of exam-taking. Thus, medical LLMs should be evaluated not only on medical QA tasks, but also non-QA ones.

4 [H1] Clinical applications

Current medical LLMs are still in the research and development stage with few clinical trials ongoing[71]. In this section, we discuss the application of large language models in medicine in detail (**Figure 4, Table 2**).

4.1 [H2] Medical decision-making

Medical decision-making, including diagnosis, prognosis, treatment suggestion, risk prediction and clinical trial matching, relies heavily on the synthesis and interpretation of vast amounts of information from different sources, such as patient medical

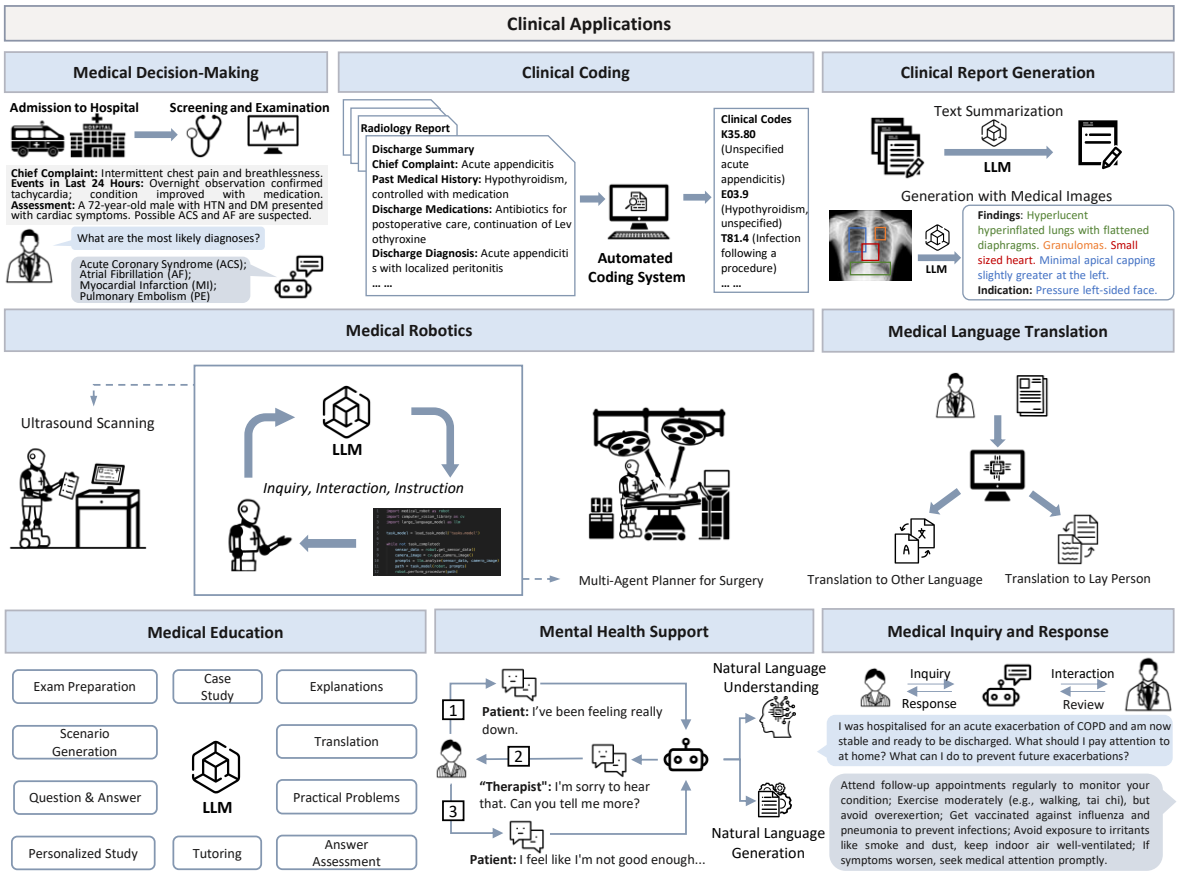

Figure 4. Application of LLMs in medicine. Integrated overview of potential applications[74,111,117,191,192] of large language models in medicine.

histories, clinical data and the latest medical literature. LLMs can rapidly process and understand such data and potentially assisting healthcare professionals in making more informed and legally sound decisions across a wide range of clinical scenarios [72,18]. For example, in medical diagnosis, LLMs can assist practitioners in analysing medical data from tests and self-described symptoms to conclude the most likely health problem[72]. Similarly, LLMs can support treatment planning by providing personalized recommendations based on the latest clinical evidence and patient-specific information[18]. Furthermore, LLMs can contribute to prognosis and risk prediction by identifying patterns and risk factors from large-scale patient data, enabling more accurate and timely interventions[73] .

For example, Dr. Knows[74] can integrate knowledge graphs from UMLS to improve diagnosis prediction and provide treatment suggestions. This approach involves fine-tuning an encoder-decoder LLM T5 [75] with extracted diagnoses as prompts and employing zero-shot prompting for LLMs like ChatGPT. Alternatively, models like DDx PaLM-2 [76] , based on instruction fine-tuning general LLMs (such as Google's PaLM-2) with extensive medical datasets MultiMedQA [10] and MIMIC-III [24], enables interactive diagnosis assistance, where humans and LLMs can iteratively communicate to arrive at a final diagnosis. NYUTron [73] is pretrained and supervised fine-tuned on various NYU hospitals and can perform three clinical tasks (in-patient mortality prediction, comorbidity index prediction and readmission prediction) and two operational tasks (insurance claim denial prediction and inpatient length of stay (LOS) prediction). Similarly, Foresight [77] is trained on UK hospital patient data and can be used for forecasting the risk of disorders, differential diagnoses and suggest medications. For clinical trial matching, TrialGPT [78] predicts criterion-level eligibility with faithful explanations, reducing screening time for human experts. Ongoing clinical trials (NCT06002425)[79] across Germany, Italy, China and the United States (US) are examining the accuracy and efficiency of LLMs in clinical decision-making and treatment recommendations for gastrointestinal cancers, or whether medical laypeople make better decisions when using LLMs (DRKS00033775)[80].

Evaluating LLM-based medical diagnosis systems requires task-specific approaches. For general diagnostic accuracy, metrics like the area under the curve (AUC) , which measures classification performance, as well as precision, recall and F1 score are used with annotated datasets [76,77,78] . Diagnostic information can also be evaluated using text summarization and medical concept extraction performance [74]. One limitation of using LLMs is their heavy reliance on subjective text inputs from patients, moreover, as LLMs are text-based, they cannot analyse medical images, an essential component of diagnostic assessments[79].

Table 2. Summary of medical LLMs tailored to various clinical applications. M: million, B: billion. PT: pre-training. FT: fine-tuning. ICL: in-context learning. CoT: chain-of-thought prompting. RAG: retrieval-augmented generation.

Application	Model	Architecture	Model Development	# Params	Data Scale	Data Source	Evaluation (Task: Score)
Medical Decision-Making (Sec. 4.1)	Dr. Knows[74]	GPT-3.5	ICL	154B	5820 notes	MIMIC-III[24] +IN-HOUSE[74]	Diagnosis Summarization: 30.72 ROUGE-L
	DDx PaLM-2[76]	PaLM-2	FT & ICL	340B	-	MultiMedQA[10] +MIMIC-III[24]	Differential Diagnosis: 0.591 top-10 Accuracy
	NYUTron[73]	BERT	PT&FT	110M	7.25M notes, 4.1B tokens	NYU Notes[73]	Readmission Prediction: 0.799 AUC In-hospital Mortality Prediction: 0.949 AUC Comorbidity Index Prediction: 0.894 AUC Length of Stay Prediction: 0.787 AUC Insurance Denial Prediction: 0.872 AUC
	Foresight[77]	GPT-2	PT&FT	1.5B	35M notes	King's College Hospital, MIMIC-III South London and Maudsley Hospital	Next Biomedical Concept Forecast: 0.913 F1
	TrialGPT[78]	GPT-4	-	-	184 patients	2016 SIGIR[78] , 2021 & 2022 TREC[78]	Ranking Clinical Trials: 0.733 P@10, 0.817 NDCG@10 Excluding clinical trials: 0.775 AUROC
Clinical Coding (Sec. 4.2)	PLM-ICD[82]	RoBERTa	FT	355M	70,539 notes	MIMIC-II +MIMIC-III[24]	ICD Code Prediction: 0.926 AUC, 0.104 F1
	DRG-LLaMA[83]	LLaMA-7B	FT	7B	25k pairs	MIMIC-IV[187]	Diagnosis-related Group Prediction: 0.327 F1
	ChatICD[84]	ChatGPT	ICL	-	10k pairs	MIMIC-III[24]	ICD Code Prediction: 0.920 AUC, 0.681 F1
	LLM-codex[85]	ChatGPT+LSTM	ICL	-	-	MIMIC-III[24]	ICD Code Prediction: 0.834 AUC, 0.468 F1
Clinical Report Generation (Sec. 4.3)	ImpressionGPT[91]	ChatGPT	ICL & RAG	110M	184k reports	MIMIC-CXR[187] +IU X-ray	Report Summarization: 47.93 ROUGE-L
	RadAdapt[92]	T5	FT	223M, 738M	80k reports	MIMIC-III[24]	Report Summarization: 36.8 ROUGE-L
	ChatCAD[81]	GPT-3	ICL	175B	300 reports	MIMIC-CXR[187]	Report Generation: 0.605 F1
	MAIRA-1[39]	ViT+Vicuna-7B	FT	8B	337k pairs	MIMIC-CXR[187]	Report Generation: 28.9 ROUGE-L
	RadFM[40]	ViT+LLaMA-13B	PT& FT	14B	32M pairs	MedMD[40]	Report Generation: 18.22 ROUGE-L
Medical Robotics (Sec. 4.4)	SuFIA[101]	GPT-4	ICL	-	4 tasks	ORBIT-Surgical[101]	Surgical Tasks: 100 Success Rate
	UltrasoundGPT[99]	GPT-4	ICL	-	522 tasks	-	Task Completion: 80 Success Rate
	Robotic X-ray[102]	GPT-4	ICL	-	-	-	X-ray Surgery: 7.6/10 Human Rating
Medical Language Translation (Sec. 4.5)	Medical mT5[104]	T5	PT	738M, 3B	4.5B pairs	PubMed+EMEA[104] ClinicalTrials [104], etc.	(Multi-Task) Sequence Labeling: 0.767 F1 Augment Mining 0.733 F1
	Apollo[105]	Qwen	PT&FT	1.8B-7B	2.5B pairs	ApolloCorpora[105]	QA: 0.588 Accuracy
	BiMediX[106]	Mistral	FT	13B	1.3M pairs	BiMed1.3M[106]	Question Answering: 0.654 Accuracy
	Biomed-sum[107]	BART	FT	406M	27k papers	BioCiteDB[107]	Abstractive Summarization: 32.33 ROUGE-L
	RALL[108]	BART	FT & RAG	406M	63k pairs	CELLS[107]	Lay Language Generation: N/A
Medical Education (Sec. 4.6)	ChatGPT[113]	GPT-3.5/GPT-4	ICL	-	-	-	Curriculum Generation, Learning Planning
	Med-Gemini[38]	Gemini	FT & CoT	-	-	MedQA-R/RS[38] +MultiMedQA[10] MIMIC-III[24] +MultiMedBench[38]	Text-based QA: 0.911 Accuracy Multimodal QA: 0.935 Accuracy
Mental Health Support (Sec. 4.7)	PsyChat[122]	ChatGLM	FT	6B	350k pairs	Xingling[122] +Smilechat[122]	Text Generation: 27.6 ROUGE-L
	ChatCounselor[119]	Vicuna	FT	7B	8k instructions	Psych8K[119]	Question Answering: Evaluated by ChatGPT
	Mental-LLM[124]	Alpaca, FLAN-T5	FT & ICL	7B, 11B	31k pairs	Dreaddit +DepSeverity +SDCNL CSSRS-Suicide +Red-Sam Twt-60Users +SAD[124]	Mental Health Prediction: 0.741 Accuracy
Medical Inquiry and Response (Sec. 4.8)	Healthcare Copilot [6127]	ChatGPT	ICL	-	-	MedDialog[128]	Inquiry Capability: 4.62/5 (ChatGPT) Conversational Fluency: 4.06/5 (ChatGPT) Response Accuracy: 4.56/5 (ChatGPT) Response Safety: 3.88/5 (ChatGPT)
	Conversational Diagnosis [6190]	GPT-4/LLaMA	ICL	-	40k pairs	MIMIC-IV	Disease Screening: 0.770 Top-10 Hit Rate Differential Diagnosis: 0.910 Accuracy

However, they can help with diagnosis as a logical reasoning tool for improving accuracy in other vision-based models; for example, in ChatCAD [81], images are first fed into an existing computer-aided diagnosis (CAD) model to obtain tensor outputs. These outputs are translated into natural language which is then fed into ChatCAD to summarize results and formulate diagnoses. ChatCAD achieves a recall score of 0.781, substantially higher than that (0.382) of the state-of-the-art task-specific model.

4.2 [H2] Clinical coding

Clinical coding, such as the International Classification of Diseases (ICD), medication coding and procedure coding help standardize diagnostic, procedural and treatment information. These codes are essential for tracking health metrics, treatment outcomes, billing and reimbursement processes; however, their manual entry is time-consuming and prone to errors. LLMs can automate this process by extracting relevant medical terms from clinical notes and assigning corresponding codes, including ICD codes [82,83,84,85], medication codes (such as National Drug Code [86]), and procedure codes (for example, Current Procedural Terminology [87]). For example, PLM-ICD [82] that builds upon the RoBERTa model [88], an optimized version of BERT, when fine-tuned for ICD coding can understand medical terms and achieves strong coding performance on 70,539 notes from the MIMIC-II and MIMIC-III datasets[24]. Other examples include DRG-LLaMA [83], which leverages the LLaMA model and applies PEFT techniques such as LoRA to adapt the model to this task. ChatICD [84] and LLM-codex [85] both use ChatGPT with prompts for ICD coding, with LLM-codex [85] taking a step further by training a language model on top of the ChatGPT responses, demonstrating its strong coding performance in MIMIC-III dataset [24].

ICD coding is typically formulated as a multi-label classification task using the MIMIC-III dataset for training and evaluation. Models are assessed based on their F1 score, AUC, and Precision@k, considering either the top k most frequent labels or the full label set. One challenge of deploying LLMs for clinical coding is the potential for biases and hallucinations; in particular, traditional multi-label classification models can easily constrain their outputs to a predefined list of (usually >1000) candidate codes through a classification neural network. By contrast, generative LLMs could suffer from major hallucinations because the input text is lengthy. As a result, the LLM may assign a code that is not in the candidate list or a non-existent clinical code to the input text. It is therefore essential to establish a proactive mechanism to detect and correct errors before patient EHRs are entered. Most LLMs for clinical coding focus on ICD coding but there is a growing need to expand to other types of clinical coding, such as medication and procedure coding, which are equally important to accurately capture patient information, facilitate billing and reimbursement processes and support clinical decision-making.

4.3 [H3] Clinical report generation

Clinical reports, such as radiology reports [89], discharge summaries and patient clinic letters, refer to standardized documentation that healthcare workers complete after each patient visit [90]. Clinical report generation usually involves text generation/summarization and information retrieval, a large portion of which is often medical diagnostic results. It is typically tedious for overworked clinicians to write clinical reports, and therefore they are often incomplete or error-prone. LLMs can act as an assistant tool to improve efficiency and reduce errors in lengthy reports [91, 92]. Another popular approach involves incorporating a vision-based model to provide complementary information [39,40,81]; the vision model analyses the input medical image and generates an annotation, which serves as a supplementary input to the LLM alongside additional text prompts.

General medical vision-language models such as Med-Gemini [38], LlaVA-Med [37], and Med-Flamingo [36] can serve as foundation models for broad medical domains including, radiology and pathology, with other models trained specifically on radiographs, such as ChatCAD [81], MAIRA-1[39] and RadFM [40] with superior performance in specific subdomains.

LLMs can also leverage textual data for report summarization to generate radiology reports. This can be achieved using either unimodal LLMs, which input a long report and generate a summary, or multimodal LLMs, which input both the long report and the corresponding image to generate a summary. Vision-language models can also be developed for report summarization. For example, ImpressionGPT [91] is a unimodal LLM which uses dynamic prompts and iterative optimization to generate report summaries. RadAdapt [92] systematically evaluates various language models and lightweight adaptation methods, achieving optimal report generation performance with a 36.8 ROUGE [93] score , through pre-training on clinical text and parameter-efficient fine-tuning with LoRA, while also investigating the impact of few-shot prompting.

Evaluation of the performance of LLM-based radiology report generation models relies on the MIMIC-III or MIMIC-IV datasets as they are the largest publicly available free-text EHRs. Common automatic evaluation metrics include BLEU [94]and ROUGE [93]. Additionally, radiology-specific metrics such as RadCliQ [95] have been developed to better assess the quality and accuracy of the generated reports in the context of radiology. A clinical trial (NCT06263855)[96] in the US is currently assessing whether using an LLM would improve the clarity and efficiency of discharge summaries.

LLM-generated reports tend to be less concise compared to human-written ones. Evaluation of these models is challenging because of the specialized nature of the content and the generative nature of the task. Current automatic evaluation methods focus on lexical metrics, which can lead to biased and inaccurate assessments of the contextual information present in the reports [97]. For example, consider two sentences with similar meanings but different wordings: “The patient’s blood glucose level is

within normal limits" and "The patient does not exhibit signs of hyperglycemia". Although both convey the absence of hyperglycemia, lexical evaluation metrics might struggle to capture their semantic equivalence, as they rely on direct word-level comparisons. This discrepancy highlights the need for more sophisticated evaluation techniques that go beyond surface-level similarities, consider the underlying medical context, and can account for the nuances and variations in expressing clinical information.

4.4 [H2] Medical robotics

Medical robotics is revolutionizing healthcare by improving surgical procedures and medical imaging [98]. LLMs can complement robotic technology by augmenting their decision-making, communication, interaction and control abilities. For example, surgical robots assisted with LLMs enable minimally invasive procedures with higher accuracy and reduced patient recovery times when compared to traditional surgical robots without LLMs [99,100]. Multi-agent planning systems designed with LLMs involve the coordination of multiple robotic units to perform collaborative tasks, improving surgical accuracy and operational efficiency [100]. Similarly, SuFIA [101] combines the advanced reasoning capabilities of LLMs, specifically GPT-4 Turbo, with perception modules to implement high-level planning and low-level control of surgical robots for tasks such as instrument navigation and tissue manipulation.

In the field of medical imaging, UltrasoundGPT [99] equips ultrasound robots with LLMs and domain-specific knowledge by using an ultrasound operation knowledge database to enable precise motion planning. UltrasoundGPT employs a dynamic scanning strategy based on prompt engineering which enables LLMs to adjust motion planning during procedures. This system demonstrates faster scan completion times and improved image quality compared to conventional ultrasound systems. Similarly, a simplified set of standardized commands and instructions enabled GPT-4 to control a robotic X-ray system named the Brainlab Loop-X[102] device.

The complexity of medical procedures, ethical considerations and patient safety concerns make it difficult to evaluate these systems in real healthcare environments. Thus, most current evaluations rely on simulated data and controlled laboratory settings; for example, SuFIA and Robotic X-ray's performance are assessed using a combination of simulated surgical scenarios and expert human evaluation [101,102]. Similarly, UltrasoundGPT is tested through the assessment of task completion [99]. Moreover, the complex and dynamic nature of shared human-robot workspaces might lead to LLM-powered medical robots to misjudge human intentions or make inappropriate decisions, posing safety risks. Future research could explore safety features such as using sensing technologies and physical design constraints to minimize errors [103].

4.5 [H2] Medical language translation

There are two main areas of medical language translation; the translation of medical terminology from one language to another [104,105,106] and the translation of medical dialogue for ease of interpretation by non-professional personnel [107,108]. Effective medical language translation is essential for providing high-quality healthcare to diverse patient populations.

Multilingual LLMs such as Medical mT5 [104], Apollo [105] and BiMediX [106], which are trained on extensive medical datasets in multiple languages, can be further fine-tuned to translate medical terminology between languages such as English, French, Spanish, Chinese and Arabic. When translating medical dialogue for non-professional understanding, it is crucial to fine-tune LLMs on datasets that encompass both technical medical conversations and their corresponding lay-language explanations. This training approach allows the models to learn the mapping between complex medical jargon and more accessible language. Techniques such as retrieval augmentation, which involves retrieving relevant lay-language explanations from external knowledge sources, can further enhance the quality and clarity of the translated dialogue [107,108]. By integrating domain-specific knowledge from various sources, LLMs can generate more accurate and informative translations that cater to the needs of non-professional audiences.

Evaluating the performance of multilingual LLMs in medical language translation requires a multi-faceted approach. Some models, such as Apollo [105] and BiMediX [106], use multiple choice QA test data with the calculation of accuracy score [105,106]. For generative benchmarks such as summarization [107,108], quantitative metrics like BLEU [94] and ROUGE [93], are commonly used to assess translation quality, but they should be supplemented with domain-specific evaluation criteria. For medical translations, accuracy of terminology, preservation of clinical meaning and consistency across languages are crucial factors. Human evaluation by bilingual medical experts is essential to validate the nuanced understanding of medical concepts across languages. For patient-oriented translations, comprehension tests with lay individuals can assess the effectiveness of jargon simplification.

In both translation and simplification tasks, misinterpretation is a common occurrence that can have damaging consequences. In developing and deploying medical translation and simplification platforms, developers should prioritize professional datasets, such as textbooks and peer-reviewed journals for medical knowledge recall. This way, it will be less likely for misinformation from unreliable sources to skew the output [109]. Another ethical consideration of using LLMs to perform medical translation is the potential for discriminatory wording to be inserted inadvertently into the output. Such wording is difficult to prevent due to potential biases in training data and the multi-step processing of input through various model components [110].

4.6 [H2] Medical education

LLMs can be incorporated into the medical education system by facilitating study through explanations, language translation, answering questions, assisting with medical exam preparation and providing Socratic-style tutoring [111]. Thus, medical education could involve text generation, text simplification, semantic textual similarity and information retrieval, among others. Medical education can be augmented by generating scenarios, problems and corresponding answers by an LLM. Moreover, students can gain a richer educational experience through personalized study modules and case-based assessments, including a wider array of challenges and scenarios beyond those found in standard textbooks [110]. LLMs can also generate feedback on student responses to practical problems, informing them about their areas of weakness in real time [112]. LLMs can also be used to educate the public; medical dialogues are often complex and difficult to understand for the average patient. LLMs can tune the textual output of prompts to use varying degrees of medical terminology for different audiences [110].

Integrating LLMs into medical education can start with existing pre-trained models such as ChatGPT and Med-Gemini [38]. For example, ChatGPT [113] can provide explanations and clarifications on complex medical concepts, whereas Med-Gemini [38], a multimodal model, can analyse medical images and generate detailed reports. Institutions, such as Second Xiangya Hospital of Central South University in China [114] and Carleton University in Canada [115], are exploring the integration of these language models into curricula, leveraging their strengths while ensuring proper oversight and ethical considerations.

To evaluate the effectiveness of integrating LLMs into medical education, a combination of quantitative and qualitative methods should be used. Current research focuses on the QA based evaluation [38]; quantitative metrics can include student performance on assessments, such as exam scores and clinical skills evaluations, comparing outcomes before and after the introduction of LLM-based tools. Qualitative methods, such as surveys and focus groups, can gather feedback from students and educators on the perceived benefits, challenges and areas for improvement in using LLMs for learning and teaching. Additionally, longitudinal studies can track the long-term impact of LLM integration on student learning outcomes, clinical competence and career readiness.

Potential downsides of using LLMs in medical education include the current lack of ethical training and biases in training datasets [21]. These biases, if not addressed, can propagate through the generated outputs, reinforcing stereotypes and potentially leading to discrimination in medical education. The lack of explicit ethical training during LLM development may also result in the generation of content that does not align with the ethical principles and guidelines of the medical profession, such as promoting unethical practices or violating patient privacy. Moreover, LLMs can generate plausible-sounding but factually incorrect information, which can mislead students and healthcare professionals if relied upon without proper verification. This can lead to the propagation of misconceptions, inappropriate treatment strategies, or misdiagnosis [116]. To mitigate these risks, it is essential to establish rigorous fact-checking and validation processes and emphasize the importance of critical thinking, evidence-based practice and the verification of information from multiple reliable sources in medical education.

4.7 [H2] Mental health support

Mental health support involves both diagnosis and treatment; for example, depression is treated through a variety of psychotherapies, including cognitive behaviour therapy, interpersonal psychotherapy and psychodynamic therapy, among others [117]. Many of these techniques are based on patient-doctor conversations, with lengthy and expensive treatment plans. The ability of LLMs to serve as conversation partners could lower the barrier to entry for patients with financial or physical constraints [118], increasing the accessibility to mental health treatments [119].

The willingness and level of self-disclosure has a strong influence on the effectiveness of mental health diagnosis and treatment, including with robots [120]. Instead of pre-training or fine-tuning on general medical data, it is often better to use medical QA data because the LLM's main task will be talking to the patient, which involves back-and-forth conversation in the format of QA [121]. PsyChat [122] is a client-centric LLM dialogue system that provides psychological support comprising five modules: client behaviour recognition, counselor strategy selection, input packer, response generator and response selection. Specifically, the response generator is fine-tuned with ChatGLM-6B on a vast dialogue dataset, Xingling and SmileChat [122]. The system demonstrated improved performance in metrics including empathy, relevance, and therapeutic alignment compared to base LLM ChatGLM. Similarly, ChatCounselor initializes from Vicuna and fine-tunes from an 8k size instruct-tuning dataset collected from real-world counselling dialogue examples [119]. Psy-LLM is meant to be an assistive mental health tool to support the workflow of professional counsellors, tailored for depression or anxiety cases [121]. A clinical trial (NCT06346496 [123]) in China is assessing the effectiveness of using LLMs for depression and anxiety symptoms in young adults over a 28-day period.

Fine-tuning on a variety of datasets can improve LLM's capability on multiple mental-health-specific tasks across different datasets simultaneously [124]. For example, Mental-Alpaca and Mental-FLAN-T5 are instruction fine-tuned on mental health datasets for tasks such as depression detection, stress prediction and suicide risk prediction [124]. Automated evaluations of mental health measure the relevance, coherence and empathy of the generated responses using metrics BLEU and accuracy. Mental health professionals conduct human evaluations through simulated counselling sessions, assessing the clinical appropriateness and therapeutic potential of the models' responses. Various evaluation frameworks have also been introduced

that integrate text generation (conversational response) [121] , QA [119] and mental health prediction [124]. For example, GPT-4 has been used as an evaluator to assess their ChatCounselor against traditional mental health chatbots[119]. The evaluation focuses on criteria including empathy, safety and therapeutic alignment, with ChatCounselor demonstrating superior performance in empathetic understanding and adherence to therapeutic principles.

The biggest risks in using LLMs for mental health support are the lack of emotional understanding and the inappropriate or harmful responses [125]. LLMs might struggle to fully grasp and respond to the complex emotional states and needs of individuals seeking mental health support and might not be able to provide the same level of empathy and human connection required in therapeutic interactions. Moreover, if not properly trained or controlled, LLMs might generate responses that are inappropriate, insensitive or even harmful to individuals in vulnerable emotional states [126]. They might provide advice that is not grounded in evidence-based psychological practices or that goes against established mental health guidelines. Addressing these challenges requires rigorous training of LLMs in evidence-based practices, ethical considerations, and risk assessment protocols, as well as collaboration between mental health professionals and AI researchers.

4.8 [H2] Medical inquiry and response

LLMs are also suitable for tasks such as answering real-time patient inquiries and assisting physicians in documentation [127]. Instead of relying solely on rule-based algorithms or limited datasets, these systems leverage the vast knowledge and reasoning capabilities of LLMs to engage in diagnostic conversations and provide personalized recommendations. For example, Healthcare Copilot [127] combines dialogue management modules, patient history tracking mechanisms, and information processing units, to enable safe patient-LLM interactions, enhance conversations with historical data and summarize consultations. Similarly, Google's Articulate Medical Intelligence Explore (AMIE) [128] uses a self-play-based simulated environment with automated feedback mechanisms to enable the model to learn and adapt across different medical scenarios. Current evaluation often involves the calculation of metrics such as accuracy, precision, recall and F1-score [128]. Multi-dimensional assessments including inquiry capability, conversational fluency, response accuracy and safety using benchmarks and comparisons with human experts or well-established models like ChatGPT have also been conducted [127]. However, these metrics alone are not sufficient and evaluation should also focus on diagnostic accuracy, patient satisfaction and adherence to medical guidelines [129]. ChiCTR2400081938[130] is currently assessing ChatGPT as an online consultant to assist physicians in remote diagnosis and treatment of hypertension in young adults; they focus on patient satisfaction, patient management efficiency, doctor work efficiency and quality of response information for evaluation.

Still, integrating medical LLMs into existing healthcare workflows and infrastructure will require substantial technical and organizational efforts. Privacy and security concerns surrounding patient data must also be carefully considered and addressed. Ensuring transparency, explainability and accountability in the decision-making processes is crucial to maintaining trust and facilitating informed consent from patients[131].

5 [H1] Challenges

In this section, we address the challenges and discuss solutions to the adoption of LLMs in an array of medical applications.

5.1 [H2] Hallucination

Hallucination refers to when the generated output contains inaccurate or nonfactual information. It can be categorized into intrinsic and extrinsic hallucinations [116] ; the former generates outputs logically contradicting factual information, such as wrong mathematical calculations [116]. Extrinsic hallucinations occurs when the generated output cannot be verified, such as 'faking' citations that do not exist or 'dodging' the question. When integrating LLMs into the medical domain, fluent but nonfactual LLM hallucinations can lead to the dissemination of incorrect medical information, causing misdiagnoses and inappropriate treatments.

Current solutions to mitigate LLM hallucination can be categorized into training-time correction, generation-time correction and retrieval-augmented correction. The first adjusts model parameter weights by including factually consistent reinforcement learning [132] and contrastive learning [133]. Generation-time correction adds a 'reasoning' process to the LLM inference to ensure reliability, using techniques such as sampling multiple outputs [134] or a confidence score to identify hallucination before the final generation. Retrieval-augmented correction instead uses external resources to mitigate hallucination such as using factual documents as prompts [135] or chain-of-retrieval prompting technique [136]. For example, training-time correction is particularly suitable for specialized medical tasks like radiology reporting where consistent patterns exist in the training data. Generation-time correction works well for general medical consultations where multiple perspectives need to be considered. Retrieval-augmented correction is essential for tasks requiring up-to-date medical knowledge such as treatment recommendations, where external verification against current medical guidelines is crucial.

5.2 [H2] Lack of evaluation benchmarks and metrics

Current benchmarks and metrics often fail to evaluate LLM's overall capabilities, especially in the medical domain. For example, MedQA (USMLE) [13] and MedMCQA [70] offer extensive coverage on QA tasks but fail to evaluate trustworthiness, helpfulness, explainability and faithfulness [97]. Although HealthSearchQA provides some improvement by evaluating LLMs on common health queries that reflect real-world information needs, it still lacks comprehensive assessment of the aforementioned crucial aspects [9]. Benchmarks such as TruthfulQA [137] and HaluEval [138] evaluate metrics such as truthfulness but do not cover the medical domain. Future research is necessary in this space.

5.3 [H2] Domain data limitations

Current training datasets in the medical domain (**Table 1**) remain relatively small compared to those for general-purpose LLMs (**Box 1**). This results in medical-specified LLMs exhibiting extraordinary performance on open benchmarks with extensive data coverage yet falling short on real-life tasks such as differential diagnosis and personalized treatment planning [10]. Although the volume of medical and health data is large, most require extensive ethical, legal and privacy procedures to be accessed. Moreover, these data are often unlabelled and solutions such as human labelling and unsupervised learning [139] are hindered by a lack of human expertise and small margins of error.

Current state-of-the-art approaches [10,14] typically fine-tune the general LLMs on smaller open-sourced datasets to improve their domain-specific performance. Another solution is to generate high-quality synthetic datasets using LLMs to broaden the knowledge coverage; however, training on generated datasets causes models to experience catastrophic forgetting, where they lose their original pretrained knowledge and capabilities due to the limited diversity and context in synthetic data [140].

5.4 [H2] New knowledge adaptation

Once trained, it is expensive and inefficient to inject new knowledge into an LLM through re-training. However, it is sometimes necessary to update on a new adverse effect of a medication or a new disease. Two problems arise during such knowledge updates; the first is how to make LLMs 'forget' the 'old knowledge', as it is almost impossible to remove it all from the training data and the discrepancy between new and old knowledge can cause unintended association and bias [141]. The second problem is the timeliness of the additional knowledge to ensure the model is updated in real-time [142]. Current solutions to knowledge adaptation can be categorized into model editing and RAG. Model editing [143] alters the knowledge of the model by modifying its parameters; however, this method does not generalize well, meaning that its effectiveness is often limited to specific scenarios or model architectures, and it may not perform consistently across different tasks or domains. By contrast, RAG provides external knowledge sources as prompts during model inference; for example, by updating the model's external knowledge memory [144]. Although RAG does not directly solve the 'forget' issue, it addresses the 'timeliness' problem by enabling quick updates of external knowledge without altering the model's core parameters.

5.5 [H2] Behaviour alignment

Behaviour alignment refers to the process of ensuring that the LLM's behaviours align with the objectives of its task, which is often to mimic general human behaviour. For example, ChatGPT demonstrates general conversational capabilities in answering human inquiries [145], but their answers to medical consultations are not as concise and professional as those of human experts [9,10].

Current solutions include instruction fine-tuning, reinforcement learning from human feedback (RLHF) [145] and prompt tuning [44,51]. Instruction fine-tuning [32] refers to improving the performance of LLMs on specific tasks based on explicit instructions [145] to generate better outputs. RLHF uses human feedback to evaluate and align the LLMs' outputs which could then be used as chatbots [146] and decision-making agents [147]. Prompt tuning can also align LLMs to the expected output format; for example, chain of hindsight prompting enables the LLMs to review its initial response, identify potential errors and generate corrected outputs [148].

5.6 [H2] Ethical and safety concerns

Concerns have been raised regarding using LLMs in the medical domain [149], with a focus on ethics, accountability and safety. For example, the scientific community has disapproved of using ChatGPT in writing biomedical research papers [131]. Tracking the accountability of using LLMs as clinical assistants is also challenging [33,150]; for example, prompt injection can cause the LLM to leak personally identifiable information (such as email addresses) from its training data [151]. Such leakage has been attributed to the mismatched generalization between safety and capability objectives, that is, the pre-training of LLMs uses a larger and more varied dataset compared to the dataset used for safety training, resulting in many of the model's capabilities not being covered by safety training [152]. A potential solution is to increase the safety training dataset and develop comprehensive safety training to cover the model's behaviours and capabilities.

5.7 [H2] Regulatory challenges

The regulatory landscape of LLMs presents distinct challenges owing to their large scale, broad applicability and varying reliability

across applications. As LLMs progressively permeate the fields of medicine and healthcare, their versatility allows a single LLM family to facilitate a multitude of tasks across a broad spectrum of interest groups. This is different from previous AI-based medical technologies which were typically tailored to meet specific medical needs and cater to particular interest groups [72,153]. This divergence requires regulators to develop fast and adaptable frameworks to ensure the safety, ethical standards and privacy of LLM-powered medical technologies without compromising innovation. For example, creating a dedicated regulatory category and incorporating patient-centred design principles in LLM development can help ensure decisions align with patient welfare and clinical best practices [153]. Other suggestions include assessing LLMs-enabled applications in real-world settings, obligations of transparency of data and algorithms, adaptive risk assessment and mitigation processes, continuous testing and refinement of audited technologies [153,154,155]. Such proactive regulatory adaptations are crucial to maintaining high standards of safety, ethics and trustworthiness of medical technology.

6 [H1] Outlook

Although LLMs have already made an impact on people's lives through chatbots and search engines, their integration into medical practices are still at an infant stage. In particular, existing benchmarks fall short in evaluating LLMs for clinical applications [156]; traditional benchmarks mainly gauge accuracy in medical QA and do not capture the full spectrum of clinical skills required [9]. Criticisms have been levelled against the use of human-centric standardized medical exams for LLM evaluation, arguing that passing these tests does not necessarily reflect an LLM's proficiency in the nuanced expertise required in real-world clinical settings [9]. Thus, more comprehensive benchmarks should be developed to assess capabilities such as sourcing from authoritative medical references, adapting to the evolving landscape of medical knowledge and communicating uncertainties clearly [9,18]. New benchmarks should also incorporate scenarios that test an LLM's ability through simulation of real-world applications and adjust to feedback from clinicians while maintaining robustness. Moreover, these benchmarks should also assess parameters such as fairness, ethics and equity, which are currently evaluated through basic metrics like demographic parity but require more sophisticated measures incorporating contextual considerations [9]. For example, AMIE uses real physician evaluations and comprehensive criteria including clinical reasoning, patient communication, and professional behavior as reflected in the Objective Structured Clinical Examination (OSCE). However, these benchmarks are still not adaptive, scalable and robust enough for different and personalized applications. Future research could focus on using synthetic alongside real-world data, incorporate clinical guidelines such as patient safety protocols and cost-effectiveness, and develop interactive evaluation systems where clinicians provide real-time feedback and assess model-physician collaboration.

Although LLMs mainly address NLP tasks, multimodal LLMs (MLLMs) or Large Multimodal Models (LMMs) (text and visual data) [157] support a broader range of tasks, such as comprehending the underlying meaning of a meme and generating website codes from images. Several MLLM-based frameworks integrating vision and language, such as Med-Flamingo [36], LLaVA-Med [37] and Med-Gemini [38] adopt medical image-text pairs for fine-tuning, thus enabling the medical LLMs to understand medical images (for example, radiology). For example, integrating vision, audio and language inputs for automated dental diagnosis have shown promising results for clinical assessment[158]. However, only very few medical LLMs [159,160] can process time series data such as electrocardiograms (ECGs) [159] and sphygmomanometers (PPGs) [160], despite their importance in diagnosis and monitoring. MLLMs trained at scale could potentially generalize across various domains and modalities outside of NLP tasks. However, the training of MLLMs at scale faces challenges in aligning and processing multiple modalities simultaneously, leading to computational constraints that result in smaller model sizes compared to single-modality LLMs. Future research could focus on improving processing, representation and learning of multi-modal data and knowledge, cost-effective training of MLLMs, especially more resource-demanding modalities such as videos and images and collection and access to multi-modal clinical data.

Another promising line of research are LLM-based agents [147], which can be seen as autonomous systems that combine LLMs' reasoning capabilities with the ability to interact with external tools and environments to achieve specific goals. These agents use LLMs as controllers to leverage their reasoning capabilities. By integrating LLMs with external tools and multimodal perceptions, these agents can interact with environments, learn from feedback and acquire new skills to solve complex tasks (for example, software design or molecular dynamics simulation) through human-like behaviours, such as role-playing and communication [125,147]. For example, Chat-Orthopedist [63] interacts with external knowledge bases, such as UpToDate.com, acquiring up-to-date adolescent idiopathic scoliosis domain knowledge to provide accurate and comprehensible responses to patient inquiries. However, integrating these agents in the medical domain is challenging, as the medical field involves numerous roles [125,147] and decision-making processes. For example, disease diagnosis often requires series of tests such as CT scans, ultrasounds, electrocardiograms and blood tests. LLMs could be used to model each of these roles/expertise and create collaborative medical agents to provide a more holistic and accurate diagnosis. These agents not only interpret individual medical reports but can also integrate these interpretations to form a cohesive medical opinion. Future research in this space could explore seamless data pipelines that collects data from various devices and transforms them into data format compatible with LLMs; improving communication and collaboration between agents, especially in areas such as ensuring truthfulness during communication, dispute resolution between agents and role-based data security measures; real-time decision-making using

data collected from remote monitoring devices; and adaptive learning to prepare for unforeseen medical and healthcare situations. Finally, current medical LLM research has largely focused on general medicine, likely due to the greater availability of data in this area [10,150]. This has resulted in the under-representation of specialized fields such as 'rehabilitation therapy' or 'sports medicine'.

So far, the medical community has primarily adopted LLMs provided by companies without questioning their data training, ethical protocols or privacy protection. Medical professionals are therefore encouraged to actively participate in creating and deploying medical LLMs by providing relevant training data, defining the desired benefits of LLMs and conducting tests in real-world scenarios to evaluate these benefits [18,81]. Such assessments would help to determine the legal and medical risks associated with LLM use in medicine and inform strategies to mitigate LLM hallucination [161]. Moreover, training 'bilingual' professionals—those versed in both medicine and LLM technology—will be increasingly more important. Future research may explore interdisciplinary frameworks to facilitate the sharing of localized data from rural clinics; 'bilingual education programs' that offer training from both worlds, as demonstrated in emerging medical AI curricula [162] ; developing institutional data management protocols and privacy protection mechanisms to help hospitals and physicians 'guard' patient data from corporations without stifling innovation.

References

1. Zhao, W. X. *et al.* A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
2. Yang, J. *et al.* Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712* (2023).
3. Chowdhery, A. *et al.* Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).
4. Touvron, H. *et al.* Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
5. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
6. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).
7. OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
8. Du, Z. *et al.* Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 320–335 (2022).
9. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
10. Singhal, K. *et al.* Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617* (2023).
11. Nori, H. *et al.* Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452* (2023).
12. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454* (2023).
13. Jin, D. *et al.* What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, 6421 (2021).
14. Li, Y. *et al.* Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *arXiv preprint arXiv:2303.14070* (2023).
15. Han, T. *et al.* Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247* (2023).
16. Wang, H. *et al.* Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975* (2023).
17. Toma, A. *et al.* Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031* (2023).
18. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. medicine* **29**, 1930–1940 (2023).
19. Patel, S. B. & Lam, K. Chatgpt: the future of discharge summaries? *The Lancet Digit. Heal.* **5**, e107–e108 (2023).
20. Yang, X. *et al.* A large language model for electronic health records. *NPJ Digit. Medicine* **5**, 194 (2022).
21. Abd-Alrazaq, A. *et al.* Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Med. Educ.* **9**, e48291 (2023).
22. Peng, C. *et al.* A study of generative large language model for medical research and healthcare. *arXiv preprint arXiv:2305.13523* (2023).
23. Alsentzer, E. *et al.* Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323* (2019).
24. Johnson, A. E. *et al.* Mimic-iii, a freely accessible critical care database. *Sci. data* **3**, 1–9 (2016).
25. Wu, J. *et al.* Clinical text datasets for medical artificial intelligence and large language models—a systematic review. *NEJM AI* **1**, AIra2400012 (2024).
26. Lee, J. *et al.* Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
27. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Comput. for Healthc. (HEALTH)* **3**, 1–23 (2021).
28. Luo, R. *et al.* Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinforma.* **23**, bbac409 (2022).
29. Ye, Q. *et al.* Qilin-med: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089* (2023).
30. Xiong, H. *et al.* Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097* (2023).
31. Yang, S. *et al.* Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *arXiv preprint arXiv:2308.03549* (2023).
32. Zhang, S. *et al.* Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792* (2023).
33. He, K. *et al.* A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694* (2023).
34. Lewis, M. *et al.* Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

35. Byambasuren, O. *et al.* Preliminary study on the construction of chinese medical knowledge graph. *J. Chin. Inf. Process.* **33**, 1–9 (2019).
36. Moor, M. *et al.* Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 353–367 (2023).
37. Li, C. *et al.* Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* **36** (2024).
38. Saab, K. *et al.* Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416* (2024).
39. Hyland, S. L. *et al.* Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668* (2023).
40. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463* (2023).
41. Zhang, X. *et al.* Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558* (2023).
42. Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
43. Li, X. L. & Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).
44. Liu, X. *et al.* P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 61–68 (2022).
45. Houlsby, N. *et al.* Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, 2790–2799 (2019).
46. Xu, C., Guo, D., Duan, N. & McAuley, J. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196* (2023).
47. Shoham, O. B. & Rappoport, N. Cpllm: Clinical prediction with large language models. *arXiv preprint arXiv:2309.11295* (2023).
48. Dong, Q. *et al.* A survey on in-context learning. *arXiv preprint arXiv:2301.00234* (2022).
49. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).
50. Liu, Z. *et al.* Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032* (2023).
51. Lester, B., Al-Rfou, R. & Constant, N. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059 (2021).
52. Gao, Y. *et al.* Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).
53. Luo, Y. *et al.* An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747* (2023).
54. Xiong, G., Jin, Q., Lu, Z. & Zhang, A. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178* (2024).
55. Li, X. & Li, J. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871* (2023).
56. Wang, G. *et al.* Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291* (2023).
57. Chen, J. *et al.* Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2309.07597* (2023).
58. Shao, Z. *et al.* Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294* (2023).
59. Trivedi, H., Balasubramanian, N., Khot, T. & Sabharwal, A. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509* (2022).
60. Asai, A., Wu, Z., Wang, Y., Sil, A. & Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511* (2023).
61. Zakka, C. *et al.* Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI* **1**, AIoa2300068 (2024).
62. Kim, J. & Min, M. From rag to qa-rag: Integrating generative ai for pharmaceutical regulatory compliance process. *arXiv preprint arXiv:2402.01717* (2024).
63. Shi, W. *et al.* Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. In *Proceedings of the 14th ACM International Conference on BCB*, 1–10 (2023).
64. Tang, L. *et al.* Evaluating large language models on medical evidence summarization. *npj Digit. Medicine* **6**, 158 (2023).
65. Van Veen, D. *et al. Adapted large language models can outperform medical experts in clinical text summarization*. *Nat Med* **30**, 1134–1142 (2024).
66. Ondov, B., Attal, K. & Demner-Fushman, D. A survey of automated methods for biomedical text simplification. *J. Am. Med. Informatics Assoc.* **29**, 1976–1988 (2022).
67. Liu, F. *et al.* Retrieve, reason, and refine: Generating accurate and faithful patient instructions. *Adv. Neural Inf. Process. Syst.* **35**, 18864–18877 (2022).

68. Joseph, S. *et al.* Multilingual simplification of medical texts. *arXiv preprint arXiv:2305.12532* (2023).
69. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W. & Lu, X. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146* (2019).
70. Pal, A., Umapathi, L. K. & Sankarasubbu, M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, 248–260 (2022).
71. Omar, M. *et al.* Large language models in medicine: A review of current clinical trials across healthcare applications. *PLOS Digital Health* **3**, e0000662 (2024).
72. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. Ai in health and medicine. *Nat. medicine* **28**, 31–38 (2022).
73. Jiang, L. Y. *et al.* Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
74. Gao, Y. *et al.* Leveraging a medical knowledge graph into large language models for diagnosis prediction. *arXiv e-prints* arXiv–2308 (2023).
75. Chung, H. W. *et al.* Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).
76. McDuff, D. *et al.* Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164* (2023).
77. Kraljevic, Z. *et al.* Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digit. Heal.* **6**, e281–e290 (2024).
78. Jin, Q. *et al.* Matching patients to clinical trials with large language models. *ArXiv* (2023).
79. US National Library of Medicine. ClinicalTrials.gov, https://clinicaltrials.gov/study/NCT06002425 (2024).
80. The Federal Institute for Drugs and Medical Devices. The German Clinical Trials Register (DRKS), https://drks.de/search/en/trial/DRKS00033775 (2024).
81. Wang, S., Zhao, Z., Ouyang, X., Wang, Q. & Shen, D. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257* (2023).
82. Huang, C.-W., Tsai, S.-C. & Chen, Y.-N. Plm-icd: Automatic icd coding with pretrained language models. *arXiv e-prints* arXiv–2207 (2022).
83. Wang, H., Gao, C., Dantona, C., Hull, B. & Sun, J. Drg-llama: tuning llama model to predict diagnosis-related group for hospitalized patients. *npj Digit. Medicine* **7**, 16 (2024).
84. Liu, J., Yang, S., Peng, T., Hu, X. & Zhu, Q. Chaticd: Prompt learning for few-shot icd coding through chatgpt. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 4360–4367 (2023).
85. Yang, Z., Batra, S. S., Stremmel, J. & Halperin, E. Surpassing gpt-4 medical coding with a two-stage approach. *arXiv preprint arXiv:2311.13735* (2023).
86. Food, Administration, D. *et al. National drug code directory* (Consumer Protection and Environmental Health Service, Public Health Service, US Dept. of Health, Education, and Welfare: USGPO, 1976).
87. Elkin, P. L. & Brown, S. H. Current procedural terminology. In *Terminology, Ontology and their Implementations*, 367–370 (Springer, 2023).
88. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
89. Liu, F., Wu, X., Ge, S., Fan, W. & Zou, Y. Exploring and distilling posterior and prior knowledge for radiology report generation. In *IEEE Conference on Computer Vision and Pattern Recognition* (2021).
90. Liu, X. *et al.* Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat. Medicine* **25**, 1467–1469 (2019).
91. Ma, C. *et al.* An iterative optimizing framework for radiology report summarization with chatgpt. *IEEE Transactions on Artif. Intell.* (2024).
92. Van Veen, D. *et al.* Radadapt: Radiology report summarization via lightweight domain adaptation of large language models. *arXiv preprint arXiv:2305.01146* (2023).
93. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of Association for Computational Linguistics (ACL)* (2004).
94. Papineni, K., Roukos, S., Ward, T. & Zhu, W. BLEU: a Method for automatic evaluation of machine translation. In *Proceedings of Association for Computational Linguistics (ACL)* (2002).
95. Yu, F. *et al.* Evaluating progress in automatic chest x-ray radiology report generation. *Patterns* **4** (2023).
96. US National Library of Medicine. ClinicalTrials.gov, https://clinicaltrials.gov/study/NCT06263855 (2024).
97. Xie, Q. *et al.* Faithful ai in medicine: A systematic review with large language models and beyond. *medRxiv* (2023).
98. Dupont, P. E. *et al.* A decade retrospective of medical robotics research from 2010 to 2020. *Sci. robotics* **6**, eabi8017 (2021).
99. Xu, H. *et al.* Enhancing surgical robots with embodied intelligence for autonomous ultrasound scanning. *arXiv preprint arXiv:2405.00461* (2024).
100. Wang, J. *et al.* Large language models for robotics: Opportunities, challenges, and perspectives. *arXiv preprint arXiv:2401.04334* (2024).
101. Moghani, M. *et al.* Sufia: Language-guided augmented dexterity for robotic surgical assistants. *arXiv preprint arXiv:2405.05226* (2024).
102. Killeen, B. D., Chaudhary, S., Osgood, G. & Unberath, M. Take a shot! natural language control of intelligent robotic

x-ray systems in surgery. *Int. J. Comput. Assist. Radiol. Surg.* 1–9 (2024).
103. Weerarathna, I. N., Raymond, D. & Luharia, A. Human-robot collaboration for healthcare: A narrative review. *Cureus* **15** (2023).
104. García-Ferrero, I. *et al.* Medical mt5: an open-source multilingual text-to-text llm for the medical domain. *arXiv preprint arXiv:2404.07613* (2024).
105. Wang, X. *et al.* Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640* (2024).
106. Pieri, S. *et al.* Bimedix: Bilingual medical mixture of experts llm. *arXiv preprint arXiv:2402.13253* (2024).
107. Tang, C., Wang, S., Goldsack, T. & Lin, C. Improving biomedical abstractive summarisation with knowledge aggregation from citation papers. *arXiv preprint arXiv:2310.15684* (2023).
108. Guo, Y., Qiu, W., Leroy, G., Wang, S. & Cohen, T. Retrieval augmentation of large language models for lay language generation. *J. Biomed. Informatics* **149**, 104580 (2024).
109. Chen, Y., Arunasalam, A. & Celik, Z. B. Can large language models provide security & privacy advice? measuring the ability of llms to refute misconceptions. In *Proceedings of the 39th Annual Computer Security Applications Conference*, 366–378 (2023).
110. Karabacak, M. *et al.* The advent of generative language models in medical education. *JMIR Med. Educ.* **9**, e48163 (2023).
111. Biri, S. K. *et al.* Assessing the utilization of large language models in medical education: Insights from undergraduate medical students. *Cureus* **15** (2023).
112. Ahn, S. The impending impacts of large language models on medical education. *Korean J. Med. Educ.* **35**, 103 (2023).
113. Peacock, J., Austin, A., Shapiro, M., Battista, A. & Samuel, A. Accelerating medical education with chatgpt: an implementation guide. *MedEdPublish* **13** (2023).
114. Tian, Q. *et al.* Iteratively refined ChatGPT outperforms clinical mentors in generating high-quality interprofessional education clinical scenarios: a comparative study. *Research Square* **3**, rs-4637356 (2024).
115. Veras, M. *et al.* Usability and efficacy of artificial intelligence chatbots (ChatGPT) for health sciences students: protocol for a crossover randomized controlled trial. *JMIR Research Protocols* **12**, e51873 (2023).
116. Rawte, V., Sheth, A. & Das, A. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922* (2023).
117. Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S. & Torous, J. B. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Can. J. Psychiatry* **64**, 456–464 (2019).
118. Stock, A., Schlögl, S. & Groth, A. Tell me, what are you most afraid of? exploring the effects of agent representation on information disclosure in human-chatbot interaction. *arXiv e-prints* arXiv–2307 (2023).
119. Liu, J. M. *et al.* Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461* (2023).
120. Robinson, N., Connolly, J., Suddrey, G. & Kavanagh, D. J. A brief wellbeing training session delivered by a humanoid social robot: A pilot randomized controlled trial. *arXiv e-prints* arXiv–2308 (2023).
121. Lai, T. *et al.* Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991* (2023).
122. Qiu, H., Li, A., Ma, L. & Lan, Z. Psychat: A client-centric dialogue system for mental health support. *arXiv preprint arXiv:2312.04262* (2023).
123. US National Library of Medicine. ClinicalTrials.gov, https://clinicaltrials.gov/study/NCT06346496 (2024).
124. Xu, X. *et al.* Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM on Interactive, Mobile, Wearable Ubiquitous Technol.* **8**, 1–32 (2024).
125. Ma, Z., Mei, Y. & Su, Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, vol. 2023, 1105 (2023).
126. Chung, N. C., Dyer, G. & Brocki, L. Challenges of large language models for mental health counseling. *arXiv preprint arXiv:2311.13857* (2023).
127. Ren, Z., Zhan, Y., Yu, B., Ding, L. & Tao, D. Healthcare copilot: Eliciting the power of general llms for medical consultation. *arXiv preprint arXiv:2402.13408* (2024).
128. Tu, T. *et al.* Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654* (2024).
129. Hager, P. *et al.* Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Medicine* 1–10 (2024).
130. Chinese Clinical Trial Register (ChiCTR). ChiCTR.org.cn, https://www.chictr.org.cn/showproj.html?proj=220887 (2024).
131. Stokel-Walker, C. Chatgpt listed as author on research papers: many scientists disapprove. *Nature* **613**, 620–621 (2023).
132. Roit, P. *et al.* Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186* (2023).
133. Chern, I.-C. *et al.* Improving factuality of abstractive summarization via contrastive reward learning. *arXiv preprint arXiv:2307.04507* (2023).
134. Manakul, P., Liusie, A. & Gales, M. J. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large

language models. *arXiv preprint arXiv:2303.08896* (2023).
135. Shuster, K., Poff, S., Chen, M., Kiela, D. & Weston, J. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567* (2021).
136. Dhuliawala, S. *et al.* Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495* (2023).
137. Lin, S., Hilton, J. & Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).
138. Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y. & Wen, J.-R. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv e-prints* arXiv–2305 (2023).
139. Liu, F. *et al.* Auto-encoding knowledge graph for unsupervised medical report generation. In *Advances in Neural Information Processing Systems* (2021).
140. Shumailov, I. *et al.* Model dementia: Generated data makes models forget. *arXiv preprint arXiv:2305.17493* (2023).
141. Hoelscher-Obermaier, J., Persson, J., Kran, E., Konstas, I. & Barez, F. Detecting edit failures in large language models: An improved specificity benchmark. *arXiv preprint arXiv:2305.17553* (2023).
142. Liu, F. *et al.* A medical multimodal large language model for future pandemics. *npj Digit. Medicine* **6**, 226 (2023).
143. Yao, Y. *et al.* Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172* (2023).
144. Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
145. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
146. Glaese, A. *et al.* Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375* (2022).
147. Xi, Z. *et al.* The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864* (2023).
148. Liu, H., Sferrazza, C. & Abbeel, P. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676* **3** (2023).
149. Sallam, M. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, 887 (MDPI, 2023).
150. Tian, S. *et al.* Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings Bioinforma.* **25**, bbad493 (2024).
151. Li, H., Guo, D., Fan, W., Xu, M. & Song, Y. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197* (2023).
152. Wei, A., Haghtalab, N. & Steinhardt, J. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483* (2023).
153. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine* **6**, 120 (2023).
154. Derraz, B. *et al.* New regulatory thinking is needed for ai-based personalised drug and cell therapies in precision oncology. *NPJ Precis. Oncol.* **8**, 23 (2024).
155. Mökander, J., Schuett, J., Kirk, H. R. & Floridi, L. Auditing large language models: a three-layered approach. *AI Ethics* 1–31 (2023).
156. Liu, F. *et al.* Large Language Models Are Poor Clinical Decision-Makers: A Comprehensive Benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13696–13710 (2024).
157. Yin, S. *et al.* A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* (2023).
158. Huang, H. *et al.* Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *Int. J. Oral Sci.* **15**, 29 (2023).
159. Li, J., Liu, C., Cheng, S., Arcucci, R. & Hong, S. Frozen language model helps ecg zero-shot learning. *arXiv preprint arXiv:2303.12311* (2023).
160. Englhardt, Z. *et al.* Exploring and characterizing large language models for embedded system development and debugging. *arXiv preprint arXiv:2307.03817* (2023).
161. Mello, M. M. & Guha, N. Chatgpt and physicians' malpractice risk. In *JAMA Health Forum*, e231938–e231938 (2023).
162. Mekki, Y. M., & Zughaier, S. M. Teaching artificial intelligence in medicine. *Nature Reviews Bioengineering*, 1-2. (2024)
163. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
164. He, P., Liu, X., Gao, J. & Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654* (2021).
165. Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
166. Chiang, W.-L. *et al.* Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality (2023).

167. Jiang, A. Q. *et al.* Mistral 7b. *arXiv preprint arXiv:2310.06825* (2023).
168. Bai, J. *et al.* Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).
169. Tay, Y. *et al.* Ul2: Unifying language learning paradigms. In *International Conference on Learning Representations* (2022).
170. Beltagy, I., Lo, K. & Cohan, A. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).
171. Alrowili, S. & Shanker, V. Large biomedical question answering models with albert and electra. In *CLEF (Working Notes)*, 213–220 (2021).
172. Gururangan, S. *et al.* Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of Association for Computational Linguistics (ACL)* (2020).
173. Yasunaga, M., Leskovec, J. & Liang, P. Linkbert: Pretraining language models with document links. In *Proceedings of Association for Computational Linguistics (ACL)* (2022).
174. Peng, Y., Yan, S. & Lu, Z. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, 58–65 (2019).
175. Phan, L. N. *et al.* Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598* (2021).
176. Lu, Q., Dou, D. & Nguyen, T. Clinicalt5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5436–5443 (2022).
177. Jin, Q. *et al.* Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *arXiv preprint arXiv:2307.00589* (2023).
178. Yasunaga, M. *et al.* Deep bidirectional language-knowledge graph pretraining. *Adv. Neural Inf. Process. Syst.* **35**, 37309–37323 (2022).
179. Venigalla, A., Frankle, J. & Carbin, M. Biomedlm: a domain-specific large language model for biomedical text. *MosaicML. Accessed: Dec* **23**, 2 (2022).
180. Gao, W. *et al.* Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue. *arXiv preprint arXiv:2306.12174* (2023).
181. Chen, Y. *et al.* Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:2310.15896* (2023).
182. Wang, G., Yang, G., Du, Z., Fan, L. & Li, X. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968* (2023).
183. Zhang, H. *et al.* Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075* (2023).
184. Luo, Y. *et al.* Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442* (2023).
185. Ferber, D. *et al.* Gpt-4 for information retrieval and comparison of medical oncology guidelines. *NEJM AI* **1**, AIcs2300235 (2024).
186. Chen, Z. *et al.* Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079* (2023).
187. He, X., Zhang, Y., Mou, L., Xing, E. & Xie, P. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286* (2020).
188. Johnson, A. E. *et al.* Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. data* **6**, 317 (2019).
189. Yang, L. *et al.* Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162* (2024).
190. Liévin, V., Hother, C. E. & Winther, O. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143* (2022).
191. Sun, Z., Luo, C. & Huang, Z. Conversational disease diagnosis via external planner-controlled large language models. *arXiv preprint arXiv:2404.04292* (2024).
192. Dong, H. *et al.* Automated clinical coding: what, why, and where we are? *NPJ digital medicine* **5**, 159 (2022).
193. D'Onofrio, G. *et al.* Emotion recognizing by a robotic solution initiative. *Sensors* **22**, 2861 (2022).
194. Bengio, Y., Ducharme, R. & Vincent, P. A neural probabilistic language model. *Adv. neural information processing systems* **13** (2000).
195. Mikolov, T., Karafiát, M., Burget, L., Cernocky`, J. & Khudanpur, S. Recurrent neural network based language model. In *Interspeech*, vol. 2, 1045–1048 (2010).
196. Sundermeyer, M., Ney, H. & Schlüter, R. From feedforward to recurrent lstm neural networks for language modeling. *IEEE/ACM Transactions on Audio, Speech, Lang. Process.* **23**, 517–529 (2015).
197. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
198. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* **30** (2017).
199. Kaplan, J. *et al.* Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).
200. Hoffmann, J. *et al.* Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* (2022).

Acknowledgements

This work was supported in part by the Pandemic Sciences Institute at the University of Oxford; the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC); an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; the Well-come Trust funded VITAL project; the UK Research and Innovation (UKRI); the Engineering and Physical Sciences Research Council (EPSRC); and the InnoHK Hong Kong Centre for Cerebro-cardiovascular Engineering (COCHE). Fenglin Liu gratefully acknowledges funding from the Clarendon Fund and the Magdalen Graduate Scholarship.

Author Contributions

HZ, FL, and DAC conceived and designed the study. HZ, FL, BG, XZ, JH, and JW conducted the literature review, performed data analysis, and drafted the manuscript. All authors contributed to the interpretation and final manuscript preparation. All authors read and approved the final manuscript.

Competing Interests

The authors declare no competing interests.

Box table Summary of general (large) language models.

Domains	Model Structures	Models	# Params	Pre-train Data Scale
General-domain (Large) Language Models	Encoder-only	BERT[163]	110M/340M	3.3B tokens
		RoBERTa[88]	355M	161GB
		DeBERTa[164]	1.5B	160GB
	Decoder-only	GPT-2[165]	1.5B	40GB
		Vicuna[166]	7B/13B	LLaMA + 70K dialogues
		Alpaca	7B/13B	LLaMA+ 52K IFT
		Mistral[167]	7B	-
		LLaMA[4]	7B/13B/33B/65B	1.4T tokens
		LLaMA-2[5]	7B/13B/34B/70B	2T tokens
		LLaMA-3	8B/70B	15T tokens
		GPT-3[6]	6.7B/13B/175B	300B tokens
		Qwen[168]	1.8B/7B/14B/72B	3T tokens
		PaLM[3]	8B/62B/540B	780B tokens
		FLAN-PaLM[75]	540B	-
		Gemini (Bard)	-	-
		GPT-3.5[145]	-	-
		GPT-4[7]	-	-
		Claude-3	-	-
	Encoder-Decoder	BART[34]	140M/400M	160GB
		ChatGLM[8]	6.2B	1T tokens
		T5[75]	11B	1T tokens
		FLAN-T5[75]	3B/11B	780B tokens
		UL2[169]	19.5B	1T tokens
		GLM[8]	130B	400B tokens

Column "# params" shows the number of parameters, M:million, B: billion.

BOX 1: Background of large language models (LLMs)

The impressive performance of LLMs can be attributed to Transformer-based language models, large-scale pre-training and scaling laws.

Language Models A language model [193, 194, 195] is a probabilistic model that models the joint probability distribution of tokens (meaningful units of text, such as words or sub-words or morphemes) in a sequence, that is, the probabilities of how words and phrases are used in sequences. Therefore, it can predict the likelihood of a sequence of tokens given the previous tokens, which can be used to predict the next token in a sequence or to generate new sequences.

The Transformer architecture The recurrent neural network (RNN)[194,196] has been widely used for language modelling by processing tokens sequentially and maintaining a vector named 'hidden state' that encodes the context of previous tokens. Nonetheless, sequential processing makes it unsuitable for parallel training and limits its ability to capture long-range dependencies, making it computationally expensive and hindering its learning ability for long sequences. The strength of the Transformer[197] lies in its fully attentive mechanism, which relies exclusively on the attention mechanism and eliminates the need for recurrence. When processing each token, the attention mechanism computes a weighted sum of the other input tokens, where the weights are determined by the relevance between each input token and the current token. It enables the model to adaptively focus on different parts of the sequence to learn the joint probability distribution of tokens. Therefore, Transformer not only enables modelling of long-text but also allows highly paralleled training[163], thus reducing training costs.

Large-scale Pre-training The LLMs are trained on massive corpora of unlabelled texts (for example, CommonCrawl, Wiki, and Books) to learn rich linguistic knowledge and language patterns. The common training objectives are *masked language modelling (MLM)* and *next token prediction (NTP)*. In MLM, a portion of the input text is masked, and the model is tasked with predicting the masked text based on the remaining unmasked context, encouraging the model to capture the semantic and syntactic relationships between tokens[163]. In NTP, the model is required to predict the next token in a sequence given the previous tokens[6].

Scaling Laws LLMs are scaled-up versions of Transformer architecture[197] with increased numbers of Transformer layers, model parameters and volume of pre-training data. The "scaling laws"[198,199] predict how much improvement can be expected in a model's performance as its size increases (in terms of parameters, layers, data, or the amount of training computed). The scaling laws proposed by OpenAI[198] shows that to achieve optimal model performance, the budget allocation for model size should be larger than the data.

The scaling laws proposed by Google DeepMind[199] shows that both model and data sizes should be increased in equal scales. The scaling laws guide researchers to allocate resources and anticipate the benefits of scaling models.

General Large Language Models Existing general LLMs can be divided into three categories based on their architecture (Table 1).

Encoder-only LLMs consist of a stack of Transformer encoder layers, employ a bidirectional training strategy that enables them to integrate context from both the left and the right of a given token in the input sequence. This bidirectionality enables the models to achieve a deep understanding of the input sentences[163]. Thus, encoder-only LLMs are particularly suitable for language understanding tasks (such as sentiment analysis or document classification) where the full context of the input is essential for accurate predictions. BERT[163] and DeBERTa[164] are the representative encoder-only LLMs.

Decoder-only LLMs use a stack of Transformer decoder layers and are characterized by their uni-directional (left-to-right) processing of text, enabling them to generate language sequentially. This architecture is trained unidirectionally using the next token prediction training objective to predict the next token in a sequence, given all the previous tokens. After training, the decoder-only LLMs generate sequences autoregressively (that is, token-by-token). The examples are the GPT-series developed by OpenAI[6,7], the LLaMA-series developed by Meta[4,5] and the PaLM[3] developed by Google.

Encoder-decoder LLMs are designed to simultaneously process input sequences and generate output sequences. They consist of a stack of bidirectional Transformer encoder layers followed by a stack of unidirectional Transformer decoder layers. The encoder processes and understands the input sequences while the decoder generates the output sequences[8,75]. Representative examples of encoder-decoder LLMs include Flan-T5 [68], and ChatGLM [8]. Specifically, ChatGLM has 6.2B parameters and is a conversational open-source LLM optimized for Chinese to support Chinese-English bilingual question-answering.

Additional Information

Related Links

PubMed: https://pubmed.ncbi.nlm.nih.gov/
PubMed Central (PMC): https://www.ncbi.nlm.nih.gov/pmc/
Alpaca: https://github.com/tatsu-lab/stanford_alpaca
ChatGPT: https://chat.openai.com/
LLaMA-3: https://github.com/meta-llama/llama3
Bard: https://gemini.google.com/
Claude-3: https://www.anthropic.com/news/claude-3-family
HealthcareMagic: https://www.healthcaremagic.com/
iCliniq: https://www.icliniq.com/
OpenBioLLM: https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B
MedLlama3-v20: https://huggingface.co/ProbeMedicalYonseiMAILab/medllama3-v20
ShareGPT: https://sharegpt.com/