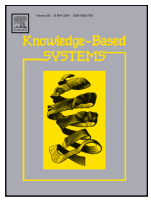
Knowledge-Based Systems 328 (2025) 114215

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Reviewing clinical knowledge in medical large language models: Training and beyond

Qiyuan Li [a,b], Haijiang Liu [a,b], Caicai Guo [a,b], Chao Gao [a], Deyu Chen [c], Meng Wang [d], Feng Gao [a,b], Frank van Harmelen [e], Jinguang Gu [a,b,*]

[a] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, 430065, Hubei, China
[b] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan, 430065, Hubei, China
[c] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, 430074, Hubei, China
[d] School of Cyber Science and Engineering, Wuhan University, Wuhan, 430072, Hubei, China
[e] Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, 1081 HV, The Netherlands

ARTICLE INFO

Keywords:
Large language models
Clinical knowledge
Medical academic
Medical practice

ABSTRACT

The large-scale development of large language models (LLMs) in medical contexts, such as diagnostic assistance and treatment recommendations, necessitates that these models possess accurate medical knowledge and deliver traceable decision-making processes. Clinical knowledge, encompassing the insights gained from research on the causes, prognosis, diagnosis, and treatment of diseases, has been extensively examined within real-world medical practices. Recently, there has been a notable increase in research efforts aimed at integrating this type of knowledge into LLMs, encompassing not only traditional text and multimodal data integration but also technologies such as knowledge graphs (KGs) and retrieval-augmented generation (RAG). In this paper, we review the various initiatives to embed clinical knowledge into training-based, KG-supported, and RAG-assisted LLMs. We begin by gathering reliable knowledge sources from the medical domain, including databases and datasets. Next, we evaluate implementations for integrating clinical knowledge through specialized datasets and collaborations with external knowledge sources such as KGs and relevant documentation. Furthermore, we discuss the applications of the developed medical LLMs in the industrial sector to assess the disparity between models developed in academic settings and those in industry. We conclude the survey by presenting evaluation systems applicable to relevant tasks and identifying potential challenges facing this field. In this review, we do not aim for completeness, since any ostensibly "complete" review would soon be outdated. Our goal is to illustrate diversity by selecting representative and accessible items from current research and industry practices, reflecting real-world situations rather than claiming completeness. Thus, we emphasize showcasing diverse approaches.

1. Introduction

Language models in clinical medicine are becoming increasingly widespread, covering various scenarios such as disease diagnosis [1–3], medical consultation [4], medical literature analysis [5], and patient management [6]. These models assist medical decision-making by processing vast amounts of medical data, including electronic health records (EHRs), medical images, and research literature, significantly enhancing the efficiency and accessibility of healthcare services [7,8]. However, most current models are trained on general corpora or limited medical data, which may not fully capture the complexity and specificity of clinical medicine [9]. This limitation can lead to outputs lacking medical rigor (for example, overlooking individualized treatment differences) and potentially misleading results [10]. Therefore, medical accuracy, ethical compliance, and clinical practicality have become core considerations in developing language models for the healthcare field.

In this work, we systematically review the applicability of integrating clinical knowledge with large language models (LLMs) in clinical medicine, focusing on the techniques and practices adopted.[1] In addition to analyzing unimodal medical text LLMs, we explore the latest ad-

[1] Reviewed papers are on GitHub: https://github.com/vicky-yuan/survey-CKinMLLMs

* Corresponding author at: School of Computer Science and Technology, Wuhan University of Science and Technology.
E-mail addresses: vickyuan@wust.edu.cn (Q. Li), alecliu@ontoweb.wust.edu.cn (H. Liu), guocaicai@ontoweb.wust.edu.cn (C. Guo), gchao@ieee.org (C. Gao), deyuchen@hust.edu.cn (D. Chen), wang_meng@whu.edu.cn (M. Wang), feng.gao86@wust.edu.cn (F. Gao), frank.van.harmelen@vu.nl (F. van Harmelen), simon@wust.edu.cn (J. Gu).

https://doi.org/10.1016/j.knosys.2025.114215
Received 31 May 2025; Received in revised form 30 July 2025; Accepted 30 July 2025
Available online 5 August 2025

vances in multimodal clinical medicine LLMs (e.g., models that combine text and medical imaging), given their increasing prevalence in medical tasks [11]. Based on definitions from clinical medicine and bioethics, we first clarify the standards that language models in the medical field should include evidence-based support, patient privacy protection, and diagnostic interpretability [12,13]. Subsequently, we integrate relevant research covering the following aspects: medical-specific knowledge bases and datasets (pre-training and fine-tuning), academic medical LLM construction mechanisms (text, multimodal, agents, KG + LLM), practical applications in the medical LLMs industry, and evaluation frameworks (e.g., automated and human-like assessments). Additionally, we analyze the critical role of human-AI collaboration in medical settings, such as balancing model automation with clinical decision-making authority [14]. Finally, we discuss the gap between current research and practical implementation, proposing future directions, including optimizing patient care experiences, improving medical resources and services, and enhancing patient psychological support and cross-cultural communication.

While recent surveys tend to concentrate on specific domains, such as assessing clinical efficiency [15], optimizing EMR models [16], addressing limitations in medical education [17], and technical pathways [18–21], our review takes a broader perspective. We survey and analyze the literature on medical LLMs that incorporate text, imaging, and temporal data, synthesizing insights from clinical research, practical applications, and their intersection with natural language processing (NLP). Our review aims to provide a technical roadmap and ethical framework for developing medical language models that are safe, reliable, and aligned with clinical needs.

Fig. 1 shows the review architecture and chapter-wise key findings. The key contributions and research goals of this review are as follows:

1. We review 160 + papers to analyze multiple approaches for embedding clinical knowledge into LLMs, including training-based methods, knowledge graph (KG) support, and retrieval-augmented generation (RAG) techniques (Section 3). Additionally, we examine accessible data sources commonly used for clinical research, including databases and relevant datasets, while discussing their strengths and limitations (Section 2).
2. We provide an overview of medical LLMs' real-world industry applications, conducting comparative analyses with academic LLMs to identify differences and potential improvement opportunities (Section 4).
3. We investigate the evaluation principles and methodologies for assessing medical LLMs, as well as the currently mainstream assessment datasets for evaluating medical capabilities, and conduct an in-depth analysis of the limitations in existing evaluation paradigms (Section 5).
4. We further explore challenges in integrating clinical knowledge into LLMs, including ensuring traceability, mitigating biases, and enhancing real-world applicability, concluding with forward-looking perspectives on future development pathways (Section 6).

Literature collection strategy. We search the literature, designing core keywords such as clinical knowledge, LLMs, KGs, agents, and evaluation. As our study focuses on clinical LLMs in both medical and computational domains, we consider papers published in leading medical journals (npj Digital Medicine, JMIR, Nature Medicine), premier computer science conferences (ACL, EMNLP, NeurIPS), and interdisciplinary journals (Nature Communications, ACM Transactions, PMLR). The emergence of clinically-adapted LLMs has gained significant momentum post-2019, with most benchmark studies published after this period, so we focus on clinical LLM benchmarks from 2019 onward. We also include recent submissions to Arxiv to capture the latest developments in clinical NLP, as journal publication cycles typically span 1–3 years. Specifically, since the release of ChatGPT (OpenAI) in 2022, more works have been published, which has increased the application of LLMs in medical NLP, marking a paradigm shift in the field.

Based on the above, our review further covers three dimensions of diversity to map the clinical LLM landscape: 1) Task Type: a variety of clinical tasks such as diagnosis, prediction, and classification and scenarios such as different modalities in different departments; 2) Model Architecture: includes various anchor LLMs such as Llama [22] and Bloom [23], as well as technical routes such as plain text, multimodality, KG enhancement, and agent framework; 3) Data Features: covering public standard datasets and specific field datasets, considering different scales, modalities, and disease coverage.

2. Clinical databases and datasets

In this chapter, we will present reliable sources for clinical knowledge. First, we introduce open-source clinical databases. Then, we will

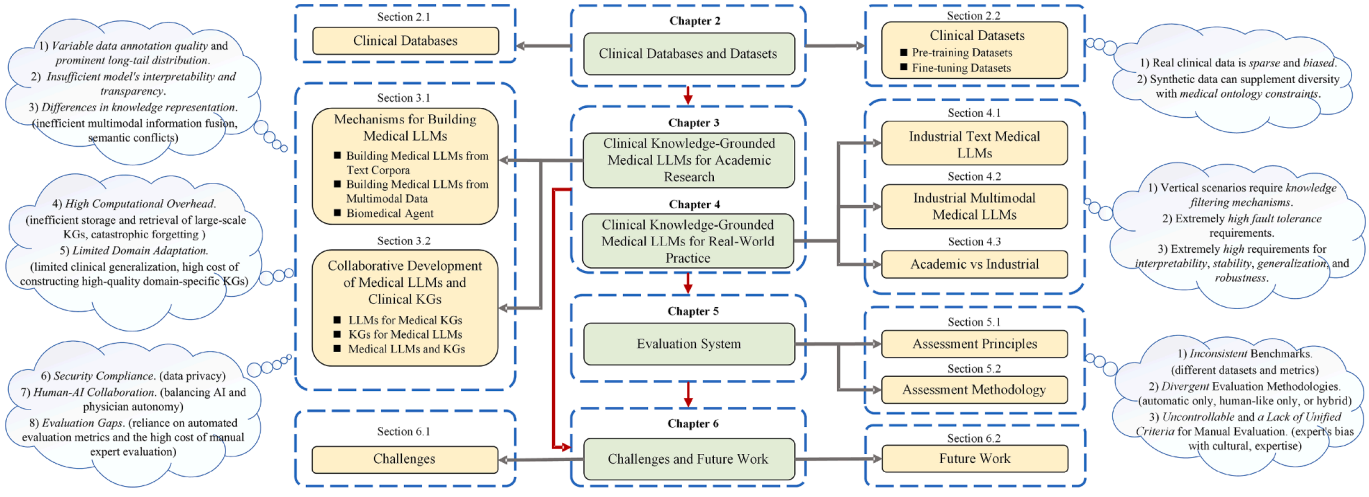


Fig. 1. Review architecture with chapter-wise key findings. **Central** represents the paper's overall structure, while the **bubbles** on both sides display the key findings. **Central:** *Green* represents the main chapter structure, while *yellow* indicates the subsections. *Red arrows* illustrate the paper's structural flow, with chapters beginning with databases and datasets (Chapter 2), followed by an exploration of academic research (Chapter 3) and real-world practice (Chapter 4), and then extend to the evaluation system (Chapter 5). Finally, we analyze challenges in academic research, practical applications, and the evaluation system, and provide an outlook for future developments (Chapter 6). **Bubbles:** The three left bubbles present the key findings of Chapter 3, and the three bubbles on the right (top, middle, and bottom) correspond to the key findings of Chapters 2, 4, and 5, respectively.

Table 1
Examples of accessible clinical databases commonly used in clinical research. "Text Corpora" indicates that the knowledge is stored as text, and "Multimodal" includes not only text but also other modalities such as images and videos. "NHS" stands for "National Health Service", "UMLS" stands for "Unified Medical Language System", "HPO" stands for "Human Phenotype Ontology".

Name	Content	Modality
Drugs[a]	Drugs.com provides accurate and independent information on over 24,000 prescription drugs, over-the-counter medicines, and products.	Text Corpora
DrugBank[b]	The latest release of DrugBank Online contains 16,568 drug entries, including 2761 approved small molecule drugs, 1611 approved biologics, 135 nutraceuticals, and over 6723 experimental drugs. Additionally, 5303 non-redundant protein sequences are linked to these drug entries.	Text Corpora
NHS Medicine[c]	NHS Medicine provides detailed information on 288 prescription drugs, over-the-counter medicines, drug-drug interaction, and side effects.	Text Corpora
Medline[d]	The flagship bibliographic resource of the National Library of Medicine cites over 31 million journal articles across the life sciences, particularly on biomedical topics.	Text Corpora
Embase[e]	Embase has over 32 million entries and over 2900 proprietary journal indexes. Its unique Emtree thesaurus includes a MeSH thesaurus, 56,000 specialized search terms, and 230,000 synonyms.	Text Corpora
UMLS[f]	UMLS includes about 2 million medical concepts and a medical vocabulary of more than 5 million words to standardize medical terminologies. Its subsystems, like MeSH[g] and RxNorm[h], are widely used in medical literature retrieval, bioinformatics research, and EMR systems.	Text Corpora
HPO[i]	HPO currently contains over 13,000 terms and over 156,000 annotations to hereditary diseases.	Text Corpora
UpToDate Clinical Advisor[j]	UpToDate Clinical Advisor includes 12,400+ clinical topics covering 25 specialties, 9,800+ graded recommendations, 37,000+ image profiles, 220+ medical calculators, 7,600+ English-language drug monographs, and 544,000+ Medline references.	Multimodal
ClinicalKey[k]	Clinicalkey comprises over 676 literature resources, 1005 classic books, 63,699 medical videos, 4.64 million images, 5000 clinical guidelines, 210,000 diagnostic trials, 2562 drug monographs, 99.01 million patient educations, 50 North American clinics, 339 operating videos, 555 clinical spotlights, and 22 million Medline abstracts.	Multimodal
NHS Health [l]	NHS Health provides an overview of 1216 diseases, treatments, and information on healthy lifestyles.	Text Corpora

[a] https://www.drugs.com/
[b] https://go.drugbank.com/
[c] https://www.nhs.uk/medicines/
[d] https://www.nlm.nih.gov/medline/index.HTML
[e] https://www.embase.com/
[f] https://www.nlm.nih.gov/research/umls/index.HTML
[g] https://www.nlm.nih.gov/mesh/
[h] http://www.nlm.nih.gov/research/umls/rxnorm/
[i] https://hpo.jax.org
[j] https://www.uptodate.cn/home
[k] https://www.clinicalkey.com/#!/
[l] https://www.nhs.uk/conditions/

introduce the clinical datasets by categorizing them into pre-training and fine-tuning stages from the perspective of constructing medical LLMs. This will establish a solid data foundation for building academic medical LLMs.

2.1. Clinical databases

Clinical databases systematically summarize medical expertise and real-life clinical experience, enabling LLMs to diagnose conditions, guide clinical decisions, and provide personalized treatments while continuously updating medical knowledge. In this way, they lay a foundation for LLMs to learn domain knowledge. Table 1 lists examples of commonly accessible clinical databases.

We can see various knowledge bases for different types and scopes of knowledge. For instance, structured drug information sources like Drugs, DrugBank, and NHS Medicine can support ontology construction and knowledge extraction, providing a foundation for clinical decision-making. Medical literature databases, such as Medline and Embase, provide essential supplementary sources, offering the latest research findings and treatment protocols while continuously updating the knowledge base. Biomedical information sources like UMLS, HPO, and UpToDate, which are known for their high precision, fine granularity, and large-scale coverage, enhance the depth of medical knowledge. Furthermore, multi-source heterogeneous databases like ClinicalKey and NHS Health form a complete medical knowledge system that addresses the needs of real medical scenarios.

The clinical knowledge base (KB) is the **cornerstone** of medical LLMs, providing *expertise* and *enabling updating* of medical knowledge. This dynamic learning process facilitates the application of medical LLMs in clinically assisted decision-making and personalized treatment recommendations.

2.2. Clinical datasets

Building on the limited data available in each clinical database, medical LLMs often require a more comprehensive range of knowledge and information from real medical data. Following the use of datasets during LLMs development, we categorize them into pre-training and fine-tuning datasets.

2.2.1. Pre-training datasets

Pre-training of general LLMs on extensive text corpora enables broad language understanding, while medical LLMs require large amounts of corpora to learn vast medical knowledge for domain-specific pre-training to master specialized knowledge and clinical comprehension. The corpora can be obtained from medical-related literature, books, journals, websites, and other related resources. Since textual content often does not fully cover fine-grained medical knowledge, such as lesion features and organ images, some pre-training datasets also combine

Table 2
Examples of common pre-training datasets for medical LLMs. These datasets often contain vast textual information ("Text Corpora") or multimodal data, which is designed to provide a wide range of knowledge for medical LLMs during pre-training.

Modality	Dataset	Data Scale	Typical Model
Text Corpora	PubMed[a]	>36M literatures	PubMedBERT [24]
	MedDialog [25]	704.73M tokens	OphGLM [26]
	EHRs [27]	>82B tokens	GatorTron [27]
	ChiMed-CPT [28]	3GB	Qilin-Med [28]
	GAP-REPLAY [29]	48.1B tokens	Meditron [29]
	The Pile [30]	>109.5 GiB	BioMedLM [31]
Multimodal	Medical TextBooks (MTB) [32]	584M tokens	Med-Flamingo [32]
	PMC-OA [33]	1.6M medical image-text	PMC-CLIP [33]
	MIMIC-CXR [34]	>377k medical image-text	MedCLIP [35]
	MIMIC-III [36]	>53.4k EHRs	ClinicalBERT [37]
	MIMIC-IV [38]	>40k EHRs	
	MIMIC-CXR-JPG [39]	>377k medical images	LLaMA-Care [41]
	MIMIC-IV-Note [40]	>2.65M medical texts	

[a] https://pubmed.ncbi.nlm.nih.gov/download/

multimodal information, such as aligned text-image data. Table 2 shows examples of commonly used pre-training datasets.

Text corpora. Text corpora serve as the primary source through which LLMs acquire medical knowledge and specialized linguistic features during pre-training. Most medical pre-training datasets [28,29] contain text from medical literature (PubMed), authentic consultation dialogues [25], and clinical practice records [27]. In addition to the above datasets constructed specifically for medical scenarios, several relevant medical-themed datasets exist in general pre-trained datasets [30] that can also be used to build medical LLMs, such as BioMedLM-2.7B.

Multimodal data. Multimodal data can assist models in acquiring knowledge of medical images and modeling the text-image relationships through pre-training. The data is often obtained from official sources such as textbooks [32] and literature [33] and real-world scenarios such as radiology departments [34,36,38,40]. To create these multimodal medical pre-training datasets, images should be pre-processed by denoising, resampling, enhancing, and normalizing.

It is recommended that during data construction, both text corpora and multimodal data should go through a systematic **quality control process** during preprocessing, including *data anomaly cleaning*, *representation normalization*, and *privacy encryption* to ensure data quality, consistency, and security. These preprocessing measures not only enhance the reliability and usability of the data but also provide standardized inputs for subsequent modeling and analysis, thus enhancing the robustness and interpretability of the results.

2.2.2. Fine-tuning datasets

Unlike the pre-training phase, which learns general representations, the fine-tuning phase adapts pre-trained models to specific tasks by updating full or partial parameters with curated datasets, typically categorized as task-specific (e.g., labeled classification data) or instruction-based (e.g., natural language prompts with expert responses), enhancing both task-specific prediction accuracy and generalization capability.

Task-specific tuning data. Table 3 demonstrates the examples of fine-tuning datasets used in the medical for nine tasks: medical examination (ME), medical question answering (MQA), medical dialogues (MD), medical visual question answering (MVQA), medical visual question generation (MVQG), image-to-text retrieval & text-to-image retrieval (I2T & T2I), report summarization (RS), report generation (RG), and medical image classification (MIC).

Table 3
Examples of common fine-tuning datasets for medical LLMs. These datasets are often used to improve or evaluate the model performance on specific tasks, and they often support multi-task analysis.

Modality	Dataset	Task								
		ME	MQA	MD	MVQA	MVQG	I2T & T2I	RS	RG	MIC
Text Corpora	CMExam [42]	✓								
	MedQA [43]	✓								
	PubMedQA [44]		✓							
	cMedQA2 [45]		✓							
	MedQuAD [46]		✓							
	CMD.		✓	✓						
	MedDialog-CN [25]			✓						
	MultiMedQA [47]	✓	✓							
	MMedBench [48]	✓	✓							
Multimodal	PathVQA [49]				✓					
	VQA-RAD [50]				✓					
	VQA-med-2018 [51]				✓					
	VQA-med-2019 [52]				✓					
	VQA-med-2020 [53]				✓	✓				
	VQA-med-2021 [54]				✓	✓				
	SLAKE [55]				✓					
	PMC-15M [56]						✓			
	ChiMed-VL [57]				✓					
	MultiMedBench [58]		✓		✓			✓	✓	✓

In the context of text-only datasets, we have compiled a series of tasks that primarily aim to enhance model comprehension of textual information and the presentation of knowledge.

1. **Medical Examination**: Medical exams are crucial for verifying the professional competence of medical students. They rely on real-world data, cover a broad range of knowledge, use standardized assessments (including standard answers and scoring criteria), provide constantly updated information resources, and involve multidisciplinary cross-fertilization, which has proven effective for evaluating medical LLMs. Examples include the CMExam [42] and MedQA [43] datasets.
2. **Medical Question Answering**: Medical question answering, as one of the most common healthcare practices, often involves patients obtaining medical encyclopedia knowledge, disease differentiation, and medication recommendations. Researchers have designed fine-tuned datasets such as PubMedQA, cMedQA2, and MedQuAD to enhance the model's performance in these scenarios.
3. **Medical Dialogues**: Unlike medical QA, medical dialogues are the primary means of daily treatment and typically involve multiple QA pairs with larger contexts. Doctors use them to gather information about patients' conditions and provide medical advice. Researchers have constructed medical dialogue datasets like CMD.[2] and MedDialog-CN to enhance the ability of intelligent treatment and assisted decision-making.

In the context of multimodal datasets, researchers have developed a variety of tasks aimed at assessing models' capabilities to extract knowledge from diverse multimodal sources and subsequently represent that knowledge in an equally multimodal manner.

1. **Medical Visual Question Answering**: With the development of multimodal models, more complex tasks like MVQA appear. Medical images capture diverse patient conditions and fine-grained pathological features, playing a vital role in clinical diagnosis. To improve models' ability to extract and align multimodal medical information, researchers have developed several specialized QA datasets, including PathVQA, VQA-RAD, the VQA-Med series, SLAKE, and ChiMed-VL.
2. **Medical Visual Question Generation**: Apart from answering questions, medical visual question generation aims to produce clinically relevant questions based on given medical images automatically. This task requires the model to analyze visual medical content and generate natural questions that reflect the image's clinical features and semantic information. Current benchmarks (VQA-Med-2020/2021) specialize in radiology question generation.
3. **Image-text Retrieval**: Image-text retrieval as a traditional multimodal medical task involves detecting a model's multimodal alignment. This requires the model to retrieve the most relevant text or image data from a large dataset when given either modality. To facilitate this task, researchers created the PMC-15M dataset for model fine-tuning.
4. **Report Summarization**: Implementing LLMs in medical scenarios requires real-life situation evaluations. Medical report summarization, as one of these scenarios, evaluates a model's ability to synthesize and condense clinical information by analyzing medical reports, incorporating imaging data, laboratory results, and physician notes to produce concise yet clinically complete summaries that retain essential diagnostic details. Datasets like MIMIC-III facilitate this by providing standardized benchmarks for developing medical summarization models.
5. **Report Generation** Medical report generation automatically generates corresponding diagnostic reports based on medical images. The core challenge lies in accurately extracting pathological information from visual data and translating it into structured clinical reports to support physician decision-making. The MIMIC-CXR dataset can be used to develop and evaluate these report generation systems.
6. **Medical Image Classification**: Medical image classification focuses on categorizing anatomical regions or pathological findings in clinical images. The task involves automatically identifying and classifying different tissue types (e.g., bones, muscles) and pathological regions (e.g., tumors, inflammation) in CT/MRI scans to assist in clinical diagnosis. Datasets like PAD-UFES-20 and CBIS-DDSM can support the development of diagnostic classification models.

In general, most fine-tuning datasets focus on a specific single task. Recent studies proposed several datasets covering the capabilities of multiple different tasks. MultiMedBench, a multimodal, multitask benchmark for developing general biomedical AI, evaluates medical LLMs across 14 individual tasks using 12 de-identified datasets. It's more than 1 million samples that span various medical topics, including medical issues, radiology reports, pathology, dermatology, chest X-rays, mammography, and genomics.

Instruction-based tuning data. Besides the above training data, instruction fine-tuning has also attracted the field's attention, involving instructional datasets to enhance the model's generalization among tasks.

The general approach is to automatically generate instruction data using pre-trained models such as the Self-Instruct framework [59], BELLE [60], or GPT [61]. These models are used to imitate the responses of real users for various tasks, generating instruction datasets that can understand and execute human instructions. For instance, ChatDoctor [62] has created prompt templates using the fine-tuned LLaMA-7B [22], which incorporates the alpaca instruction dataset and the HealthCareMagic100k doctor-patient dialogue dataset. These templates retrieve external knowledge databases during dialogues, enhancing the model's accuracy. Huatuo [63] utilized the LLaMA-7B to generate a dataset for Chinese medical instructions, improving the model's effectiveness in MQA.

Unlike real data, instruction data can utilize LLMs to generate data with controllable scale, coverage, and privacy protection requirements, although the data generated by the model often prefers learned samples. The quality of these synthetic data can be verified to a certain extent through manual evaluation or automated evaluation on downstream tasks [64–67].

However, there are inherent limitations compared to real data: **1) Distribution Differences:** not fully capturing the complex distribution, long-tail phenomena and noise of real data; **2) Information Loss:** implicit knowledge or context contained in real data that the model has not learned; and **3) Authenticity and Complexity:** the difficulty in simulating the uniqueness and complexity of real clinical cases. Consequently, models in real-world applications may face challenges regarding **1) Reliability:** feasibility in real clinical environments; **2) Generalization:** the ability to generalize to unseen real data; and **3) Potential Bias:** including biases in the model itself, biases in the prompt design process, and selective biases in data screening criteria.

We encourage the use of LLMs to generate synthetic data based on real data when real data is scarce. Still, synthetic data should be regarded as a valuable supplementary resource rather than a perfect substitute for real data.

2.3. Takeaways

As we summarize the above clinical datasets in Table A.1, in terms of data type, authenticity, diversity, coverage, update frequency, us-

[2] https://github.com/Toyhom/Chinese-medical-dialogue-data

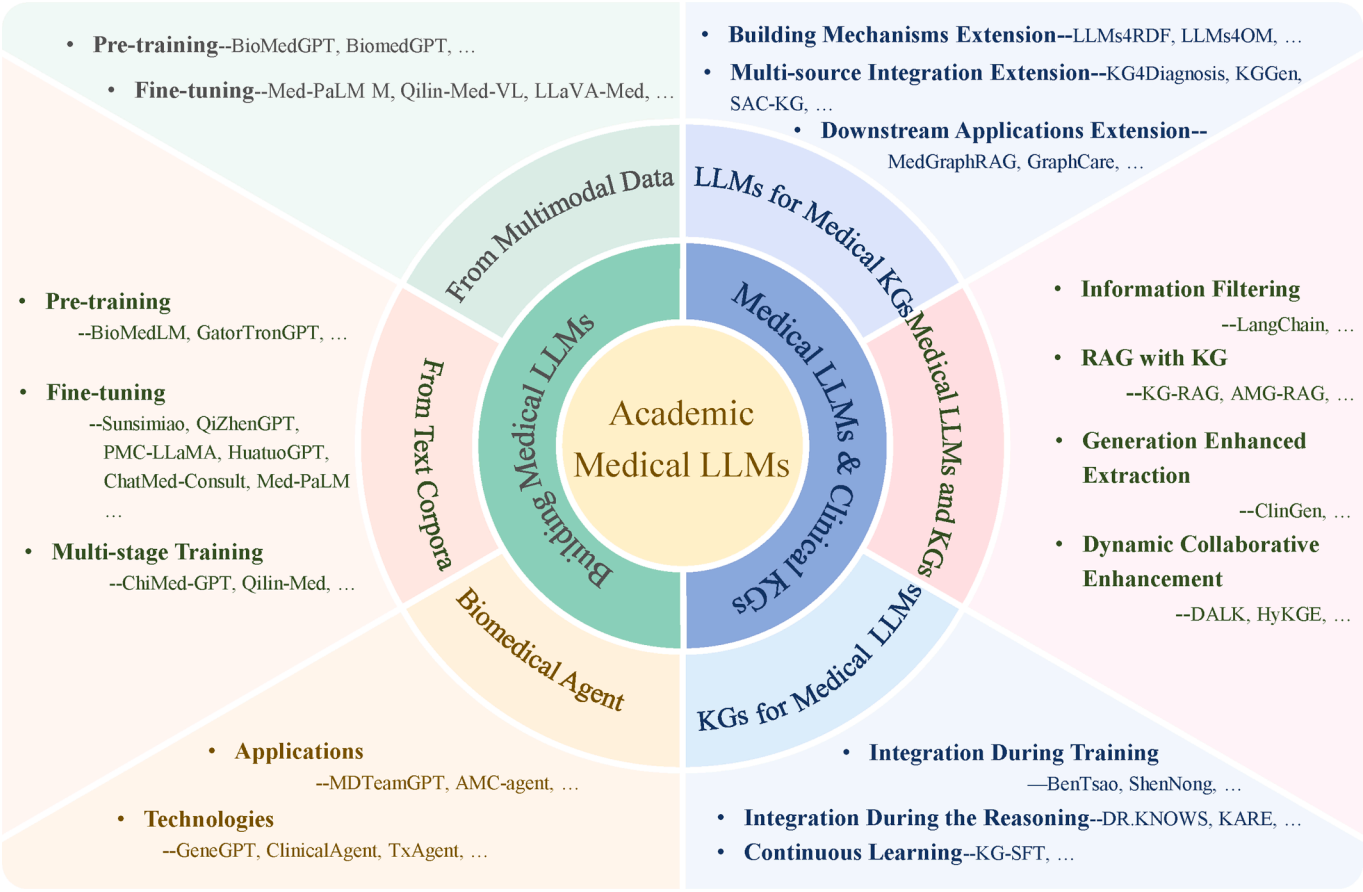


Fig. 2. Overview of the development and integration of Medical LLMs and Clinical KGs, highlighting key processes such as pretraining, fine-tuning, multi-stage training, multimodal data integration, and downstream applications, including dynamic collaboration and RAG.

ability, reliability, and language, the available datasets exhibit the following characteristics: **1) Data Sources and Types:** Most datasets are based on real data and primarily contain textual information, while image data is mainly concentrated in radiology. **2) Coverage:** The coverage of the datasets can be divided into two categories: one encompasses a broad range of medical foundational knowledge, while the other focuses on specific specialty knowledge. **3) Time Span and Update Frequency:** The datasets vary widely in their timeframe and are generally updated less frequently, except for PubMed, which is kept steadily updated, and other datasets with irregular update cycles. **4) Accessibility:** Most datasets are open-source and available, while a few require permission to access. **5) Data Reliability:** Most datasets are considered reliable, as they are based on real data. However, the reliability of a few datasets is partially reliable due to some data being generated by models like GPT-3.5/GPT-4. **6) Language Distribution:** The datasets are predominantly in English, followed by bilingual (English and Chinese), while datasets in other languages are relatively scarce. Additionally, we present the data quantity in Table A.2.

As a result, current medical KBs and training datasets **do not have sufficient multilingual support**. Most existing resources are only available in English or Chinese, with few aligned datasets, which can hinder the applications of medical AI systems for non-native doctors and patients. Furthermore, the **integrity of disease knowledge in these KBs and datasets may not be sufficient**. Information on certain rare or endemic diseases, such as *Gaucher disease*, *Pompe disease*, and *Fabry disease*, is insufficient for model learning, and updating the latest research results and treatments is also costly. In real healthcare scenarios, these shortcomings may impact diagnostic accuracy, treatment effectiveness, and healthcare quality (see Chapter 6). We presented examples of these datasets on GitHub[3] to provide further insight into their format and applications.

3. Clinical knowledge-grounded medical LLMs for academic research

Unlike traditional language models, LLMs such as GPT-3 [68], GLM [69], LLaMA, and Bloom [23] are trained on large-scale textual data (e.g., Wikipedia[4] and BookCorpus [70]) with parameter sizes reaching billions or more. They show strong generalizability across multiple domains through pre-training and fine-tuning [71].

LLMs have attracted extensive research in healthcare [19,27]. They process medical literature and clinical records, supporting medical research, diagnosis, and treatment by integrating clinical knowledge to assist professionals in understanding patient conditions and making decisions. Below, we will systematically introduce the mechanisms of building medical LLMs and their collaborative development with clinical KGs.

Fig. 2 provides an overview of clinical knowledge-grounded medical LLMs for academic research, including textual, multimodal, agent-based, and KG + LLM models, along with their corresponding classifications and methodological examples. We also provide a timeline of medical LLM development and an evolution diagram of the relevant anchor models to enhance understanding. As shown in Fig. 3, the development of medical LLMs and multimodal systems has rapidly advanced,

[3] https://github.com/vicky-yuan/survey-CKinMLLMs
[4] https://huggingface.co/datasets/wikipedia

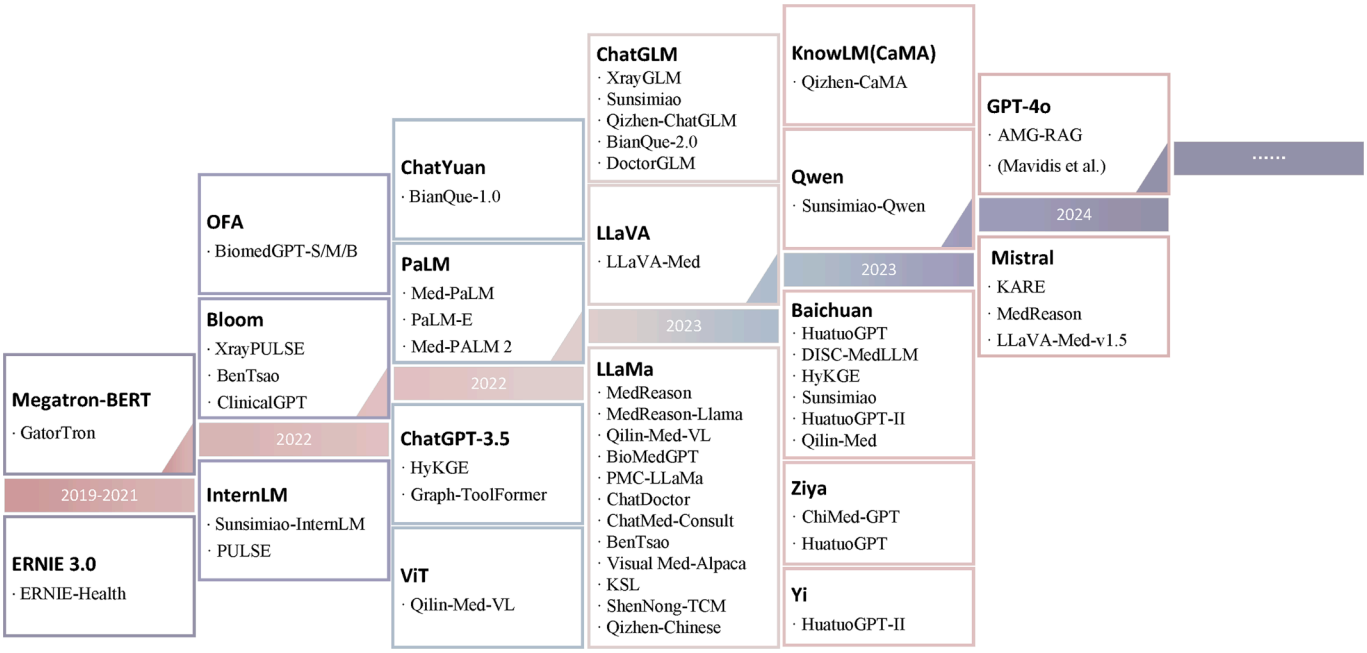


Fig. 3. Evolution timeline of typical medical LLM examples, grouped by their anchor models. Model in **BOLD** represents the foundational model name. Models below are derived from them.

with diverse applications in healthcare. Foundational models such as ERNIE-Health, ChatGPT-3.5, and Med-PaLM are highlighted alongside specialized adaptations like BioMedGPT, XrayGLM, and HuatuoGPT-II. Spanning from 2019 to 2024, the timeline shows significant progress, including multimodal frameworks (e.g., Qilin-Med-VL, LLaVA-Med) and region-specific models (e.g., Qizhen, ShenNong). The figure underscores the integration of textual, visual, and knowledge-grounded reasoning to address medical challenges.

3.1. Mechanisms for building medical LLMs

In the previous section, we examined open-source medical pre-training datasets and fine-tuning datasets to enhance the clinical knowledge of LLM. To achieve optimal outcomes, it is advisable to initially utilize models that demonstrate exceptional performance in either a general or a specific domain. Subsequently, these models can be refined to address particular medical objectives.

To better illustrate the building mechanisms for LLM from clinical knowledge, we classify this process into two categories: *text corpora and multimodal*, based on the representation of clinical knowledge. These two building mechanisms also require customized training objectives for different types of clinical data and application scenarios, enabling the model to perform excellently in specific clinical tasks.

3.1.1. Building medical LLMs from text corpora

Text corpora play a crucial role in medical research and building domain LLMs. This data provides an in-depth analysis and summary of multiple findings and knowledge in medical research. In the following section, we first present the training objectives during pre-training with standard practices. We then present the fine-tuning strategies for various downstream medical tasks. Finally, we introduce several comprehensive works that combine pre-training and fine-tuning into a multi-stage training pipeline to obtain more capable medical LLMs.

Pre-training. **BioMedLM** [31] and **GatorTronGPT** [72] are the main studies using pre-training methods to build medical LLMs with strong multi-task generalization capabilities.

BioMedLM (formerly PubMedGPT) is a biomedical LLM developed by Stanford and MosaicML. Built on GPT-2 [73], it was pre-trained on 50B tokens from PubMed Abstracts/Central. With efficient fine-tuning on downstream datasets, the model achieves an accuracy of 50.3 % on MedQA-USMLE, 74.4 % on PubMedQA, and 95.7 % on BioASQ, demonstrating the capability of LLM, especially in the biomedical domain, and their language generation capabilities in real-world applications.

GatorTronGPT was trained as a generative clinical LLM on 277 billion words of textual data. This included 82 billion words of clinical texts from 126 clinical departments and about 2 million patients at the medical center, and 195 billion words of various general English texts. Their evaluation of the training results based on the GPT-3 in biomedical NLP and medical text generation showed that GatorTronGPT was comparable to humans regarding linguistic readability and clinical relevance. Physicians were unable to distinguish between them.

Fine-tuning. In addition to pre-training, fine-tuning methods are commonly used to construct medical LLMs for specific downstream tasks. As presented in Table 4, these tasks include information extraction (IE), MQA, multi-turn dialogue (MD), ME, and text-to-text (T2T).

Information Extraction automatically derives structured data from unstructured or semi-structured sources. Medical IE task enhances model's quality for applications like case description structurization and knowledge base construction. Traditional IE [74] employs

Table 4
Training tasks for various medical tasks. Each of the tasks uses different training objectives and serves various purposes in real-life applications.

Task	Training Objective	Medical Application Scenarios
IE	Sequence Labelling Generative Extraction	Case Structurization
MQA	Text Generation	Knowledge QA, Assisted Diagnostic Suggestions, Drug Indication Evaluation, Disease Evaluation, Report Interpretation
MD	Long-context Text Generation	Simulated Diagnosis and Treatment, Guidance
ME	Text Classification	Learning Assistance
T2T	Text Generation	Synthesizing Clinical Text Generation

sequential labeling (e.g., BILOU tagging) to identify and extract text patterns. Modern approaches leverage generative modeling, directly serializing structured outputs from input text. SOTA generative extraction approaches [75] typically use a method similar to the UIE [76] architecture.

Medical Question & Answering generates responses to a question based on the model's given context or knowledge stock, enhancing its performance in assisted diagnosis, drug indications, disease assessment, and report interpretation applications. QA tasks with context can also be used to train models using the sequence annotation method described above. However, researchers often prefer to use generative methods (such as Sunsimiao[5] and QiZhenGPT[6]) in conjunction with context-free QA tasks to improve training efficiency and quality.

Multi-turn Dialogue extends medical QA through coherent conversational chains, where dialogue history is aggregated into contextual inputs for models (e.g., HuatuoGPT [77], MedicalGPT[7]). With its longer context length limit for the model, this is more challenging than the medical QA task. Moreover, this task can greatly enhance the model's performance in complex medical scenarios, such as simulated consultation sessions, which is of greater practical value.

Medical Examination requires models to choose from given options based on the provided context from medical exams. Traditional approaches (e.g., BERT-based fine-tuned models [42]) typically process stems and options as separate inputs, determining the most appropriate answer through sequential multi-classification, while generative models (e.g., Transformer-based fine-tuned models [78]) integrate them as a unified input, enabling autonomous answer generation. This task evaluates a model's comprehension of specialized medical knowledge.

Text-to-Text combines text generation tasks outside QA forms, requiring models to process text based on given instructions (e.g., summarizing a medical report [79,80]). Instructions are often declarative sentences, and many tasks allow models to learn requirements through training without explicit instructions. Similar to LLM instruction fine-tuning, these tasks are less computationally expensive.

Table 5 summarizes the construction of medical LLMs for specific downstream tasks using fine-tuning methods. Most medical LLMs obtained through fine-tuning are designed for ME and MQA tasks, which are closely linked to real medical scenarios and, therefore, receive more attention.

We also observe that most medical text LLMs are **fine-tuned from mainstream models** such as LLaMA, ChatGLM-6B, and Bloom. The **diversity in data sources**, including public datasets, domain-specific data, and mixed data, demonstrates these models' adaptability and provides a solid foundation for their application in specific fields. Furthermore, the **variety in task types**, such as QA, MD, and T2T, not only meets diverse healthcare needs but also promotes the practical application of AI technology, potentially reducing medical costs and improving service efficiency.

Multi-stage training. In addition to the methods mentioned above, some researchers construct more comprehensive knowledge-gaining and applicable medical models across tasks by using a combination of pre-training and fine-tuning in multi-stage training, as shown in Table 6.

ChiMed-GPT [89], a Chinese medical LLM based on Ziya-13B-v2 [90], combines *pre-training*, *supervised fine-tuning* (SFT), and *reinforcement learning from human feedback* (RLHF) on data extracted from CMD, to enhance domain adaptation, outperforming existing models that rely solely on SFT in medical *IE, MQA, MD, ME, and T2T* tasks. However, researchers have found that the model exhibits potential biases that urgently need to be addressed.

Qilin-Med [28] is a multi-stage training approach that combines *domain-specific continued pre-training* (DCPT), SFT, and *direct preference optimization* (DPO) with the ChiMed dataset, which includes medical QA, plain text (MedQA-textbooks), KGs, and dialogues. While effective for medical LLMs, limitations include: 1) Chinese medical knowledge reduces the model's global applicability; 2) The multi-stage training risks incorporating human evaluator biases, while BLEU and ROUGE metrics may be insufficient to evaluate medical performance.

3.1.2. Building medical LLMs from multimodal data

Compared to the aforementioned monomodal medical LLMs, multimodal medical LLMs contain more comprehensive knowledge that contributes to more accurate predictions. Moreover, multimodal data contains knowledge with smaller granularity, which can compensate for the deficiencies of a single modality. In the following section, we will conduct an in-depth analysis of the current efforts in the field, presenting the current research status and development trends using the construction methods of strategies such as pre-training and fine-tuning. Table 7 summarises the construction methods of multimodal medical LLMs.

Pre-training. Building multimodal LLMs often requires creating uniform multimodal representations of the same object across different modalities through pre-training methods. Traditional multimodal pre-training methods usually employ contrastive learning (CL) to help models distinguish between different modal representations. As shown in Fig. 4, the core idea is to teach a model's representation by comparing the similarity between positive and negative samples. This is done even if the distance between positive examples becomes closer and the distance between negative examples becomes farther. Models [35] created with CL generally perform better for multimodal image-text retrieval.

Once the LLM is proposed, pre-training the model using full parametric CL methods becomes challenging. This is because pre-training typically necessitates a substantial amount of high-quality data, and full parametric training is hardware-intensive and costly. Therefore, some researchers have proposed new training approaches [91,93,100], as shown in Fig. 5.

Fine-tuning. Similar to constructing medical text LLMs, researchers prefer to fine-tune model parameters to improve the model's performance on specific tasks, which incurs a lower training cost. This approach to fine-tuning can be divided into three main categories: **end-to-end** methods, **BLIP**-derived methods, and **prompt-combination** methods.

The **end-to-end** approach is comparable to the pre-training approach of BioMedGPT mentioned earlier. This involves encoding data from different modalities separately and then aligning the multimodal vector representation using a unified model. For instance, Med-PaLM M, developed by Google Research and DeepMind, is a multimodal generative model that interprets biomedical data. Fine-tuned on the MultiMedBench benchmark, it aligns PaLM-E to the biomedical domain, outperforming or matching SOTA in all tasks. The model demonstrates excellent clinical adaptability and zero-shot generalization, allowing it to reason and make decisions for medical situations for which it has not been explicitly trained.

However, the end-to-end approach lacks the flexibility to solely utilize pre-trained models that support multimodality or combinations of already trained multimodal encoders. To address this issue, the researchers have utilized the **BLIP** training framework, which offers greater flexibility in combining LLMs that support either image processing or text analysis. As shown in Fig. 6, the approach integrates a Q-Former neural network between the image encoder and text model. The multimodal models can be linked by training the Q-Former alone while keeping the image and text models frozen.

The framework introduces three main learning objectives for training tasks: *image-text matching*, *image-based text generation*, and *image-text CL*. Several BLIP-like models have been implemented in healthcare, including Qilin-Med-VL (Clip-ViT and Chinese-LLaMA2, image-text matching and image-based text generation), LLaVA-Med (LLaMA, image-based text generation and image-text CL), XrayPULSE (MedCLIP

[5] https://github.com/X-D-Lab/Sunsimiao
[6] https://github.com/CMKRG/QiZhenGPT
[7] https://github.com/shibing624/MedicalGPT

Table 5
Examples of medical LLMs constructed using fine-tuning methods. These models often utilize an LLM backbone from the general domain as the foundation, i.e., an anchor model, and are fine-tuned for various downstream medical tasks.

Model	Anchor Model	Data Source	Task
Sunsimiao	Baichuan-7B[a], ChatGLM-6B[b], InternLM-7B-Chat[c], Qwen-7B[d]	-	MQA, ME
QiZhenGPT	Chinese-LLaMA-Plus-7B[e], KnowLM-13B-Base[f], ChatGLM-6B	Qizhen Medical KB[g]	MQA, ME, T2T
PMC-LLaMA [81]	LLaMA-7B, MedLLaMA-13B[h]	Books + Literature, MedC-I [81]	MQA, ME
ChatMed-Consult [82]	LLaMA-7B	ChatMed_Consult_Dataset [82]	MQA, ME
ShenNong-TCM [83]	Chinese-LLaMA-7B	ShenNong_TCM_Dataset [83]	MQA, T2T
BenTsao[i]	Huozi1.0, Bloom-7B, Chinese-Alpaca-7B, LLaMA-7B	CMeKG-8K[j]	MQA, ME, T2T
Med-PaLM	PaLM [84]	MultiMedQA	MQA, ME, T2T
Med-PaLM 2 [78]	PaLM2 [85]	MultiMedQA	MQA, ME, T2T
ClinicalGPT [86]	Bloom-7B	cMedQA2, cMedQA-KG, MD-HER, MedDialog, MEDQA-MCMLE	MQA, MD, ME
MedicalGPT	Bloomz, LLaMA1/2/3, ChatGLM1/2/3-6B, Baichuan1/2-7B/13B, Cohere, Orion, Mistral, DeepSeek1/3, InternLM2, Qwen1/1.5/2/2.5, XVERSE, Yi	Chinese medical datasets, HuatuoGPT-Hybrid SFT	IE, MQA, ME
HuatuoGPT	Baichuan-7B, Ziya-LLaMA-13B-Pretrain-v1	Hybrid SFT [77]	MD, ME
HuatuoGPT-II [87]	Baichuan2-7B/13B-Base, Yi-34B	Chinese & English Corpus, Huatuo-26M	MQA, MD, ME
BianQue-1.0 [88]	ChatYuan-large-v2	Chinese medical QA instructions, multi-turn dialogues	MD
BianQue-2.0 [79]	ChatGLM-6B	BianQueCorpus [79]	MD, ME, T2T
DoctorGLM [80]	ChatGLM-6B	CMD.	MD, T2T
ChatDoctor	LLaMA-7b	HealthCareMagic, iCliniq	MD, ME
PULSE[k]	Bloomz-7b1-mt, InternLM-20B	-	MQA, ME, T2T

[a] https://github.com/baichuan-inc/baichuan-7B
[b] https://github.com/THUDM/ChatGLM-6B
[c] https://huggingface.co/internlm/internlm-chat-7b
[d] https://huggingface.co/Qwen/Qwen-7B
[e] https://github.com/ymcui/Chinese-LLaMA-Alpaca
[f] https://github.com/zjunlp/KnowLM
[g] http://www.mk-base.com/#/official/home
[h] https://huggingface.co/chaoyi-wu/MedLLaMA_13B
[i] https://github.com/scir-hi/huatuo-llama-med-chinese
[j] https://github.com/king-yyf/CMeKG_tools
[k] https://github.com/openmedlab/PULSE

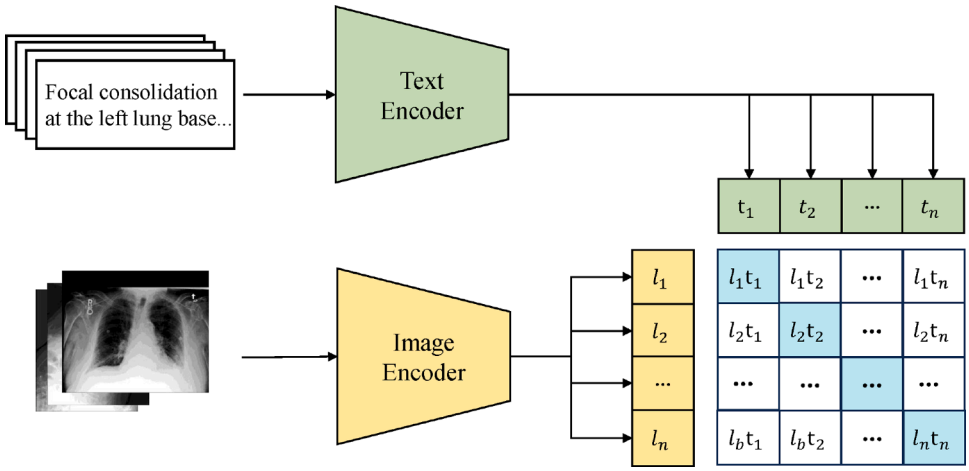


Fig. 4. Contrastive learning multimodal pre-training technique. This type of pre-training has strong transferability and zero-shot capabilities.

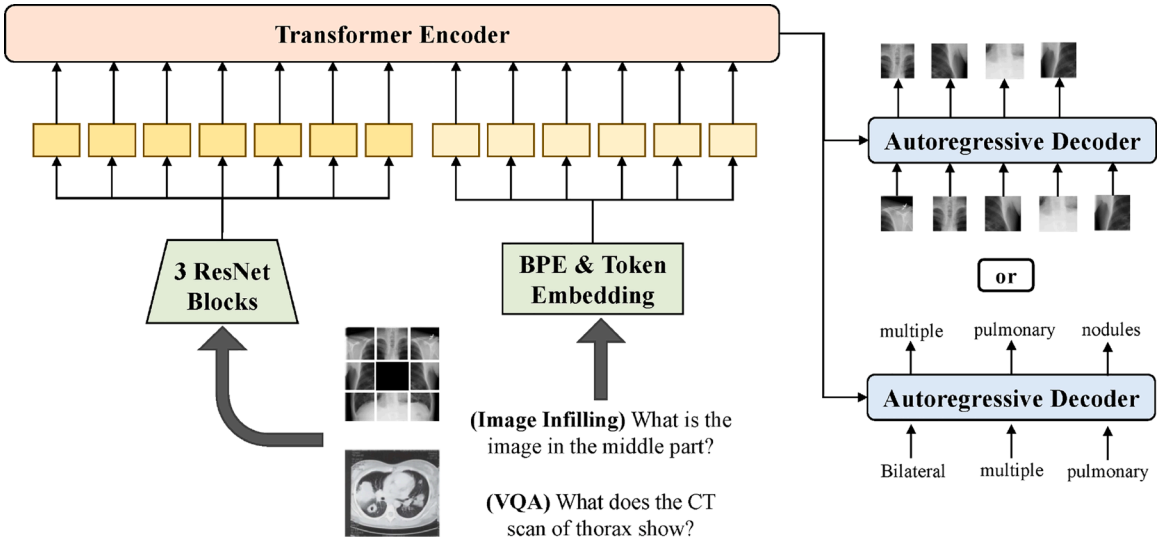


Fig. 5. An end-to-end multimodal pre-training technique. This type of pre-training views the model as a pipeline solution for specific tasks. Therefore, their abilities can be transferred among similar tasks easily.

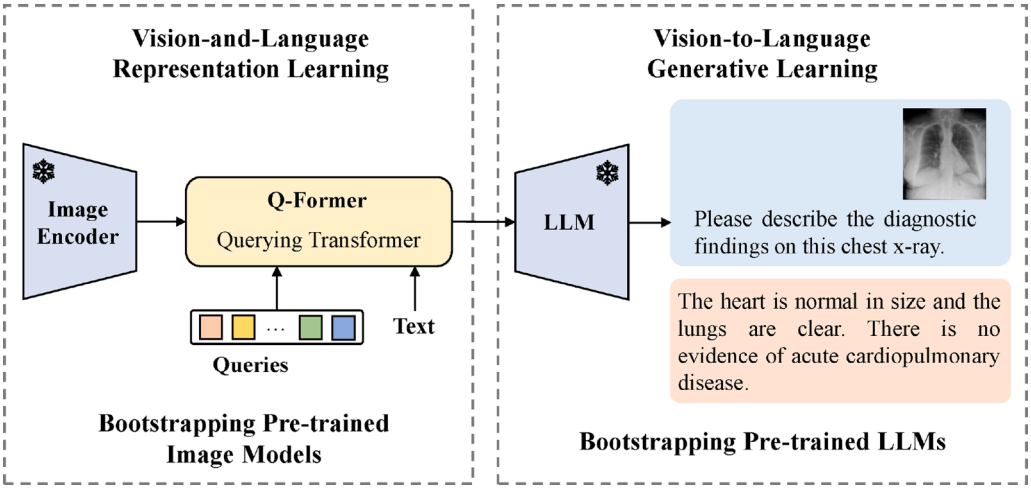


Fig. 6. BLIP fine-tuning method. This type of method treats different parts of the model separately so that we can improve the model's performance without impacting the model's generalization.

Table 6
Pre-training and fine-tuning integrated for Medical LLM examples. These models often integrate clinical data in multiple training stages to ensure the accuracy of the knowledge representations and multi-task utilization.

Model	Anchor Model	Data Source	Task
ChiMed-GPT [89]	Ziya-13B-v2 [90]	CMD.	IE, MQA, MD, ME, T2T
Qilin-Med	Baichuan-7B	ChiMed	MQA, MD, ME

and Q-former adapter, image-based text generation), and XrayGLM (VisualGLM-6B trained using BLIP with ViT and ChatGLM2, image-text matching and image-based text generation).

In addition to the previously mentioned multimodal construction mechanism, another way to construct multimodal models is to link already trained multimodal encoders through prompts known as **prompt-combination**. Fig. 7 below. This method is cheaper and more convenient for training than the abovementioned methods.

Visual Med-Alpaca is a prime example of a multimodal healthcare LLM constructed with prompts. The model is trained using an instruction set developed collaboratively by GPT-3.5-Turbo and human experts. Finally, with hours of fine-tuning and a plug-and-play vision module, it can perform various medical tasks.

In summary, these two building mechanisms of LLMs above have shown improvements in medical tasks. However, they still face several challenges, including quality issues in *data annotation*, *uneven data distribution*, *insufficient model interpretability and transparency*, and limitations in *multimodal information fusion*. During the evaluation of model performance, most methods rely primarily on automated medical task evaluation metrics, while relatively few incorporate manual evaluation. Furthermore, the practical **application value** of these models still requires further validation and evaluation.

Although these LLMs are not yet fully ready for practical application, their great potential cannot be ignored. They offer useful references and insights for practical applications in the medical field and indicate directions for future research and development.

Table 7
Summary of methods for constructing multimodal medical LLM examples. We present these methods according to the modality of their training strategies.

Strategy	Model	Anchor Model	Data Source	Task
Pre-training	BioMedGPT [91]	LLaMA2-Chat-7B	S2ORC [92]	BioMedical QA Molecule QA Protein QA
	BiomedGPT [93]	OFA [94]	14 multimodal biomedical datasets	MVQA, MVQG, RS, RG, MIC
Fine-tuning	Med-PaLM M	PaLM-E [95]	MultiMedBench	MQA, RS, MVQA, RG, MIC
	Qilin-Med-VL	ViT, Chinese-LLaMA2-13B	ChiMed-VL	MVQA
	LLaVA-Med [96]	GPT-4 [97], Mistral-7B	PMC-15M	MQA, MVQA
	XrayPULSE[a]	PULSE	MIMIC-CXR, OpenI	MD
	XrayGLM [98]	VisualGLM-6B	MIMIC-CXR, OpenI	RG, MD
	Visual Med-Alpaca [99]	LLaMA-7B	VariousMedQA [99]	MQA, MVQA, MD

[a] https://github.com/openmedlab/XrayPULSE

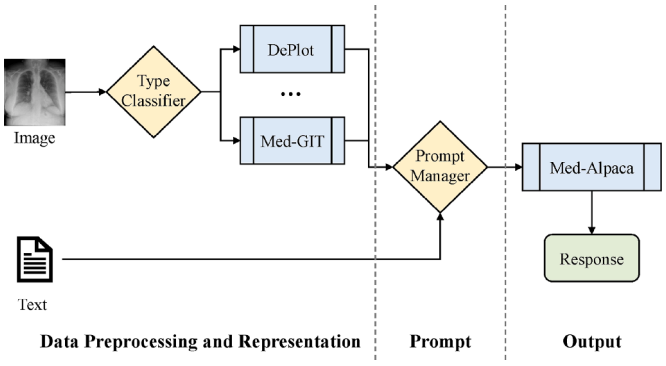


Fig. 7. Prompt-combination fine-tuning method. This type of method combines the excellent performance of models over multiple modalities to create a stronger model through prompt design.

3.1.3. Biomedical agent

In addition to the traditional approach of pre-training and fine-tuning to develop medical LLMs, the ability of complex reasoning and tool invocation possessed by LLMs has led to an increasing interest in using agents, which are LLM derivatives. Agents aim to design and build tools and extensions so that LLMs can accomplish more complex tasks. LLM-based agents often can perceive their surroundings using sensors, make decisions, and take responsive actions using actuators [101], as shown in Fig. 8.

These agents can interact with the environment, humans, or other agents through perception, reasoning, and planning [101–103]. Recent work [104] has shown the potential of combining agents with LLMs, advancing healthcare automation.

Current research on clinical medical agents has developed into a dual focus on expanding application scenarios and fostering technological integration and innovation.

- **Applications**: Studies demonstrate a trend from single-function systems toward diversified clinical scenarios. These include assistive interventions for specific populations (e.g., virtual dialogue agents for autism social skills training [105]), personalized medical consultation for general needs [106,107], and intelligent diagnostic assistance (e.g., simulated depression diagnosis platform [108]).
- **Technologies**: Research primarily concentrates on three dimensions: 1) Enhancing specialized capabilities through domain tool integration (e.g., GeneGPT [109] and OpenMedCalc [110]); 2) Exploring multi-agent collaboration [111,112], multimodal agents [113,114] and evaluation benchmarks [114,115] to improve complex medical task performance; 3) Developing adaptive dialogue systems [116] to optimize human-agent interaction.

Takeaways. Despite significant progress in clinical medical agent research, several critical challenges remain to be addressed, such as **limited domain adaptability** (constrained generalization capabilities in complex clinical scenarios like multi-morbidity management), **safety and compliance challenges** (e.g., data privacy concerns), **human-AI collaboration bottlenecks** (e.g., maintaining physician autonomy), and **model interpretability and reliability** issues. Meanwhile, enterprises (e.g., Sinohealth, United Imaging) release medical agent solutions, signaling the sector's transition from "single-point efficiency tools" to "collective intelligence-driven productivity transformation" as the new developmental paradigm.

3.2. Collaborative development of medical LLMs and clinical KGs

In recent years, with the rapid development of KGs and LLMs, combining KGs and LLMs has become a popular research direction [117]. While LLMs excel at contextual adaptation and diverse NLP tasks (e.g., summarization, QA, translation), they face challenges like *hallucination and interpretability* issues due to their implicit parametric knowledge. Conversely, KGs offer structured, interpretable, explicit knowledge but suffer from high construction costs, incompleteness, and limited NLP capabilities. Their combination mitigates these limitations, enhancing both NLP performance and application potential.

In medical AI research, we classify LLM-KG integration approaches into three distinct paradigms: (1) **LLMs for Medical KGs**: LLM empowers medical KG construction, which means using LLM's advantage to enrich and optimize medical KGs. (2) **KGs for Medical LLMs**: Medical KG empowers LLM, integrating structured knowledge from KGs into LLM and enhancing its knowledge understanding and application capabilities in specific fields. (3) **Medical LLMs and KGs**: The mutual synergy between medical KG and LLM is to achieve more efficient knowledge retrieval and language generation through their interaction. Next, we will introduce the relevant research progress in detail.

3.2.1. LLMs help construct medical KGs

Constructing and maintaining traditional KGs is time-consuming and requires much manual input. Each step, from data collection and cleaning to knowledge collation and annotation, as well as regular updating and iteration, requires many annotators and professional knowledge support. LLM provides an effective solution for automating knowledge acquisition, for it has capabilities for processing textual information.

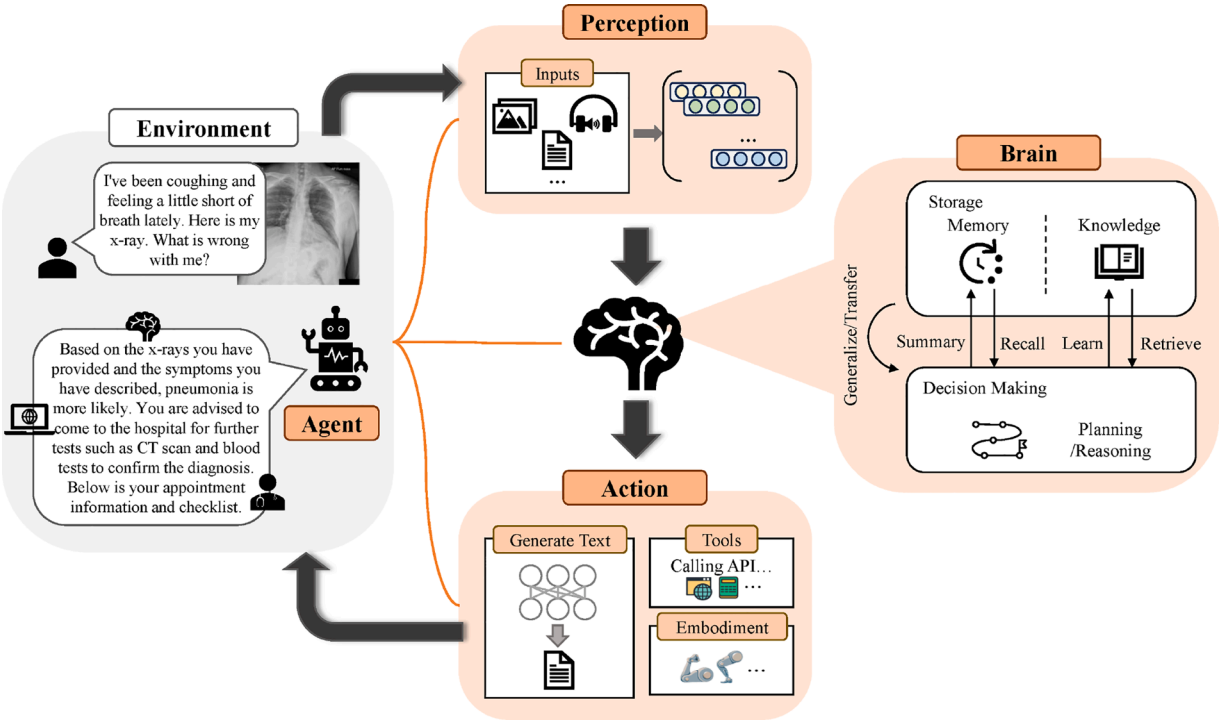


Fig. 8. LLM-based agent framework. It contains three components: perception, brain, and action. The interaction begins in the *Environment*, where users provide inputs such as X-ray images and symptom descriptions. These inputs are processed by the *Perception* module and sent to the *Brain* module. The *Brain* uses stored memory and knowledge to inform decision-making (including planning and reasoning) through generalization and transfer. Finally, the *Agent* executes actions, such as generating text, calling APIs, or performing embodiment, based on these decisions, resulting in feedback in the *Environment* as outputs.

There are three mainstream approaches for constructing medical KGs: building mechanisms, multi-source integration, and downstream applications.

Building mechanisms extension. There have been numerous studies on using LLM to extract, represent, and reason on medical knowledge. Specifically, it includes **1)** Ontology knowledge mapping optimization, where LLMs are used for intelligent RDF KG construction [118] and ontology matching in heterogeneous alignment [119]; **2)** Automated extraction and representation of medical knowledge, where entity recognition is done by ERNIE-Health,[8] and clinical text structuring via GatorTron [27,120]; **3)** Enhancement of complex medical inference capability, such as Graph-ToolFormer's graph-guided reasoning [121], and KSL's multi-hop knowledge retrieval capability [122]. These approaches significantly enhance medical KGs in terms of knowledge organizations, extractions, and inferences.

Multi-source integration extension. Expanding and integrating the knowledge source of KG to additional medical data, such as medical records, medical images, and drug information, can also benefit from LLMs' assistance. For instance, Arsenyan et al. [123] propose an end-to-end approach that harnesses LLMs for the automatic generation of KGs from EMR notes. Additionally, KG4Diagnosis [124] represents an innovative end-to-end framework designed for the construction, diagnosis, treatment, and reasoning related to automated medical knowledge graphs. Meanwhile, frameworks for automated knowledge graph construction, such as KG-Gen [125] and SAC-KG [126], offer new perspectives for developing clinical knowledge graphs.

Downstream applications extension. With the integration of LLMs, KG applications can be extended to more scenarios, such as disease diagnosis and treatment recommendations, because of enhanced relation interpretations. For instance, KG4Diagnosis [124] employs a hierarchical multi-agent framework (GPLLM + Consultant-LLMs) to automate medical KG construction for 362 diseases, thereby improving diagnostic reasoning. Wu et al. developed MedGraphRAG [127], which adopts a three-layer KG architecture (documents → literature → terminology) combined with the U-retrieve strategy, significantly enhancing the accuracy and interpretability of medical QA. Jiang et al. proposed GraphCare [128], which utilizes LLM-augmented personalized KGs to optimize AUROC metrics for mortality and readmission prediction tasks.

3.2.2. KGs enhance medical LLMs' interpretations

While ChatGPT, GPT-4, and other LLMs have shown impressive language comprehension and generation abilities, they are still limited by their pre-trained corpus and model capabilities. They also face challenges such as limited knowledge memory, weak reasoning, and difficulty in updating knowledge. KGs can be a viable solution to address these issues due to their structured and explicit representation of knowledge, natural interpretability, and high-quality information. Additionally, KGs can also represent and generate Chain-of-Thought (CoT), which can further enhance the reasoning ability of LLMs by structuring a better reasoning chain. The approaches with KGs assisting medical LLMs can be broadly classified into three categories: knowledge learning during training, knowledge integration during reasoning, and continuous learning from historical experience.

Knowledge learning during training. LLM training with KGs enables it to understand targeted professional knowledge and concepts. KGs can help researchers ensure the quality and diversity of training data, the rationality of task design, and sometimes even the loss function to improve the LLM's ability to understand and express medical knowledge. For instance, BenTsao [63] uses the medical KG to generate instruction data to fine-tune the model and significantly improve the MQA capability. ShenNong-TCM [83] is based on open-source traditional Chinese

[8] https://huggingface.co/nghuyong/ernie-health-zh

medicine (TCM) KG generation data and trained using the entity-centric self-instruct method, which performs well on the cMedKnowQA dataset. DISC-MedLLM [129] combines medical KG, real-world dialogues, and human feedback methods to optimize the model's performance in medical consultation scenarios.

Knowledge integration during the reasoning. Knowledge integration from medical KGs during reasoning can help LLMs answer medical questions with provided medical evidence. Existing studies have designed effective reasoning algorithms and query strategies to fully utilize the information in the medical KG and improve the LLM's reasoning and answering capabilities. For instance, Gao et al. [130] propose a dynamic knowledge enhancement framework that innovatively incorporates KG information during inference to strengthen the model's comprehension of complex medical concepts. KARE [131] improves medical prediction accuracy through its KG-enhanced RAG mechanism. Additionally, MedReason [132] uses a logical reasoning chain approach to transform clinical Q&A pairs into structured chains with medical KGs, enhancing model performance in complex clinical scenarios through SFT techniques.

Continuous learning from historical experience. Continuously optimizing and improving the performance of LLM can be effectively conducted by recording and analyzing historical data and cases. The key to this phase is to design effective methods for collecting and analyzing experiences and cases to help LLMs continuously learn and improve. For instance, KARE employs a dynamic knowledge retrieval mechanism that constantly updates the KG with patient historical data while enhancing the model's learning of novel clinical patterns through reasoning chains. Similarly, MedGraphRAG utilizes reinforcement learning to optimize its U-Retrieve strategy for medical QA tasks, dynamically adjusting KG subgraph selection weights based on historical retrieval performance. Additionally, KG-SFT [133] adopts a KG-guided supervised fine-tuning approach that preserves historical reasoning paths and incorporates a re-prompt mechanism during new data training, thereby improving the model's continual learning capability for low-resource medical tasks.

3.2.3. Other collaborations

In healthcare AI applications, interpretability, credibility, and traceability of outcomes are essential. This necessitates combining medical knowledge graphs with various LLM technologies for a more significant role. LLMs, often due to their black-box nature, may produce biased results. However, KGs contribute to interpretability, credibility, and traceability, aiding in understanding the workings of LLMs. Therefore, *integrating KG throughout the entire lifecycle of LLMs*, from pre-training to various application stages, is an efficient strategy. This integration enhances the training efficacy of LLMs and improves the practicality and reliability of their inferential results.

Currently, medical KGs and LLMs are constructed collaboratively using various methods such as information filtering, RAG with KG, generation-enhanced extraction, and dynamic collaborative enhancement.

Information filtering. Information filtering refers to extracting relevant knowledge from KGs and then sending it to the LLM for answering. Zhang et al. [134] proposed using LangChain to create a new model that combines KGs and LLMs deeply. Firstly, the input text of questions related to TCM formulas undergoes information filtering, specifically text classification, to determine its relevance. Secondly, LangChain retrieves knowledge related to the text from the KB and inputs it into LLMs such as ChatGPT and ChatGLM, along with the question in the form of prompts. The LLM then generates a professional answer through reasoning. Finally, the answer undergoes knowledge extraction to extract the triples. The extracted triple is then matched with the existing prescription KG to verify its expertise. Furthermore, the nodes in the KG are used as input for the LLM to obtain natural language explanations, achieving bidirectional conversion between the LLM and the KG.

RAG with KG. To generate more knowledgeable responses based on LLM and a self-built KB, collaboration between LLM and KG can also be achieved. The KG-RAG framework [135] developed by Soman et al. innovatively combines biomedical KGs (e.g., SPOKE) with mainstream LLMs (Llama2, GPT series), which significantly improves the performance of the model in tasks such as drug-use queries and biomedical Q&A by fusing the explicit and implicit knowledge from KG and LLM, respectively. Furthermore, the AMG-RAG framework [136] achieves automated construction and dynamic updating of medical KGs, thought chain reasoning integration, and external evidence retrieval (e.g., PubMed), which improves the answer accuracy and enhances the interpretability of medical reasoning.

Generation enhanced extraction. By reinforcing the information extraction capability of LLM, the generated natural language responses are extracted to structured knowledge and matched with the professional KG for verification, realizing a deep integration method between LLMs and KGs. Xu et al. [137] conducted in-depth research on using LLMs for clinical text generation tasks and proposed an innovative and resource-efficient method called ClinGen, which integrates knowledge into the entire clinical process. This model utilizes a medical domain-specific KG and an LLM to guide data generation. Research on seven clinical NLP tasks and 16 datasets shows that ClinGen can continuously improve performance and significantly enrich the diversity of generated instances in various tasks.

Dynamic collaborative enhancement. Unlike previous approaches that synergize the LLM with static KGs, Li et al. [138] propose the DALK framework that integrates LLMs with dynamic KGs for Alzheimer's disease (AD) research. By constructing an AD-specific KG from scientific literature and employing hierarchical sampling with novel knowledge retrieval methods, DALK achieves bidirectional enhancement between LLMs and KGs, significantly improving AD-related QA performance as demonstrated in benchmark evaluations. Moreover, the HyKGE framework [139] leverages LLMs' reasoning capabilities to compensate for incomplete user queries, optimizing LLM-KG interactions through: (1) hypothesis-driven exploration of LLMs' zero-shot capabilities to expand KG search directions, and (2) a hierarchical Rerank module that effectively balances knowledge diversity and relevance while filtering noise. Extensive experiments on multiple MQA datasets confirmed HyKGE's superior performance in both accuracy and interpretability.

3.2.4. Takeaways

Integrating KGs and LLMs combines structured knowledge with probabilistic reasoning, enhancing AI's interpretability and factual accuracy. However, current technologies still face several challenges: **1) Knowledge representation discrepancies** (semantic alignment issues during fusion); **2) High computational overhead** (storage and retrieval inefficiencies with large-scale KGs; catastrophic forgetting during dynamic updates); **3) Limited domain adaptability** (high costs associated with building high-quality domain-specific KGs); **4) Lack of unified evaluation standards** (e.g., lack framework to assess KG + LLM knowledge coverage-current tools like MedKGEval [140] are limited to medical KGs). Future research directions may explore dynamic KGs + continual learning LLMs, neuro-symbolic hybrid reasoning, and lightweight architectures to develop more intelligent and explainable AI systems.

4. Clinical knowledge-grounded medical LLMs for real-world practice

Beyond their use in scientific research, medical LLMs that integrate clinical knowledge have been applied in various scenarios, including medical research, drug research and development (R&D), intelligent diagnosis and treatment (D&T), medical equipment maintenance, and hospital management. These applications support diagnosis, imaging, drug innovation, dialogue services, and personalized treatment plans, which

require knowledge of various types of TCM, pharmaceutical molecules, biomedicine, and genomics in multilingual contexts. We classify them into two types based on different modalities: plain text and multimodal.

4.1. Industrial text medical LLMs

Table B.1 presents the successful transition of text-based medical LLMs from academia to industry, where companies adopt medical LLMs trained on extensive medical data based on general LLMs(e.g., GPT), with improved performance in diagnosis, health management, and research, exhibiting that: 1) **Convergent Techniques, Divergent Applications.** While most models (EyeGPT[9], Tongyi Renxin[10]) adopt the progressive "pretraining + domain fine-tuning + RLHF" approach, their applications vary significantly. 2) **Highlight of Intelligent TCM.** Models like DaJing[11] and Gushengtang[12] bridge TCM theory with modern AI, covering diagnosis-to-treatment workflows while maintaining clinical validity through expert-in-the-loop evaluation. 3) **Diversified Implementation Models.** From research tools (WiNEX[13]) to consumer health apps (Xiaoyi[14]) and from hospital partnerships (Zhiyun Health[15]) to O2O pharmacy integrations (DingDangKuaiYao[16]), the diverse product and business models reflect deep industry-technology convergence.

While medical LLMs have demonstrated remarkable efficacy in enhancing operational efficiency (e.g., outpatient documentation) and reducing diagnostic errors, critical challenges remain in clinical compliance, data privacy protection, and cross-institutional data interoperability. Looking ahead, as more models are completing clinical validation (exemplified by Zhiyun Health's chronic disease management system), medical AI is poised to become a core infrastructure of healthcare.

4.2. Industrial multimodal medical LLMs

In addition to text-based medical industrial LLM, the development of multimodal data has led to the emergence of multimodal industrial LLMs. Table C.1 presents successful cases of multimodal medical LLM transitioning from academia to industry. Medical LLMs are diversifying, with companies pursuing distinct technical approaches to develop specialized solutions. Multimodal integration (StoneNeedle [141], MedLinker[17]) and knowledge enhancement, (Tencent MedLLM[18] PanGu [142]) have emerged as dominant technological trends, significantly improving diagnostic accuracy and scenario coverage. Meanwhile, applications have expanded from diagnostic assistance (Medical Sense[19]) and drug discovery (PanGu) to comprehensive health management, (WeiMai[20]) establishing end-to-end service capabilities connecting patients, physicians, and enterprises (01Bot[21]).

While these technological breakthroughs drive healthcare intelligence transformation, challenges remain in data compliance, multimodal generalization, and clinical validation. Future development requires balancing specialization with accessibility through approaches like KG + LLM collaboration and expert-reinforced learning, to continuously enhance the precision and accessibility of medical services, thereby laying the foundation for a digital healthcare ecosystem.

[9] http://eyegpt.com.cn/#/
[10] https://tongyi.aliyun.com/renxin
[11] http://www.dajingtcm.com/dajinggpt
[12] https://www.gstzy.cn/
[13] https://www.winning.com.cn/WiNEX/
[14] https://www.xunfeihealthcare.com/
[15] https://www.zyhealth.com
[16] https://www.ddky.com/
[17] https://www.medlinker.com/
[18] https://healthcare.tencent.com/
[19] https://chat.sensetime.com/
[20] https://www.myweimai.net/
[21] https://01.baidu.com/bot.HTML

4.3. Academic vs. industrial

The anchor models used in practical applications involve analyzing large datasets during the training process, which cover both public and non-public medical data. Moreover, many medical experts must participate in manual feedback supervision and fine-tuning training during the training process. Finally, to ensure the practicality and effectiveness of the model, a series of strict evaluation and validation measures need to be carried out before the model is applied.

These practices highlight significant differences from academic medical LLMs, particularly in terms of the training data, training process, and evaluation methods. Industrial models require more extensive datasets and expert involvement to refine their performance, while academic models rely more on publicly available data and standardized evaluation metrics.

In the following section, we will compare industrial and academic medical LLM in more detail regarding research objectives, data resources, training methods, and application scenarios, as shown in Table 8.

4.4. Takeaways

Based on the differences between academic medical LLMs and industrial medical LLMs, we summarize the potential issues of implementing academic models into real-life applications in the following aspects: **1) Data Resource Limitations:** Academic research often utilizes publicly available datasets, while industry relies on private data with real-world diversity but potential sensitivity issues. Academic models may need to adapt to changes in data quality and scale when transitioning to industry. **2) Balance of Model Performance and Practicality:** Academic models focus on theoretical innovation, while industry requires stability and high interpretability. Therefore, industry application scenarios often require models with stronger generalization ability and robustness to handle various uncertainties in practical operations. **3) Clinical and Regulatory Compliance:** The healthcare industry requires models to meet strict clinical and ethical standards. Therefore, academic research results must undergo a rigorous clinical trial and regulatory approval when translated into industrial products. **4) Interdisciplinary Cooperation Challenges:** Cultural and workflow differences between academia and industry can hinder collaboration, requiring mutual understanding of needs and methodologies. **5) Technology Transfer and Landing Challenges:** Applying academic research to industry involves addressing hardware compatibility and computing resource limitations, as well as modifying and engineering academic research findings to ensure they operate reliably in real clinical environments.

5. Evaluation system

Evaluating models is crucial for verifying their validity and reliability. A robust evaluation framework not only measures a model's performance but also guides its ongoing improvement and progression. This subsection will examine the construction of the evaluation framework, illustrated in Fig. 9.

5.1. Assessment principles

The cornerstone of the evaluation system is a series of principles that together ensure the assessment process's fairness, accuracy, and effectiveness. Let us take a closer look at these key principles:

- **Accuracy**: The overarching principle of model assessment is accuracy, which refers to how closely the model predicts or generates results compared to the true or expected values. For language models, accuracy reflects the correctness and reasonableness of the generated content.

Table 8
Academic medical LLMs versus industrial medical LLMs. We compare and analyze them with research objectives, data resources, training & optimization, performance metrics, application scenarios, data privacy & security, collaboration & open source, and updates & iterations.

	Academic Medical LLMs	Industrial Medical LLMs
Research Objectives	Research and innovation are often pursued to achieve theoretical breakthroughs and technological advancements.	While commercial interests may not be the primary driver, the ultimate goal is to solve practical problems and create commercial value in response to market demand.
Data Resources	Publicly available datasets are frequently used, or data is obtained through research collaborations with limited data volumes.	The organization utilizes a significant amount of data, including patient records and medical images, which are both diverse and voluminous.
Training & Optimization	Focuses on innovations in model structure and algorithms, as well as theoretical performance improvements.	Focuses more on the practical application performance of the model, such as deployment efficiency, stability, and security.
Performance Metrics	Accuracy, Recall, F1 score, etc.	More diversified, including actual application effects, user feedback, business metrics, etc.
Application Scenarios	Demonstrate research results through academic research, published papers, and algorithmic competitions.	Focuses on the effectiveness of models in real clinical applications as well as commercial solutions.
Data Privacy & Security	Often adhere to data privacy and security regulations, but may not be as strict compared to the industry.	Must adhere to strict healthcare data privacy and security regulations.
Collaboration &Open Source	Prefers academic collaboration, paper publication, and open source code.	Prefers commercial collaboration and productization, technology licensing, may not be fully open source.
Updates & Iterations	Relatively slow update frequency, mainly paper publication.	Iterate quickly to adapt to market needs and business changes.

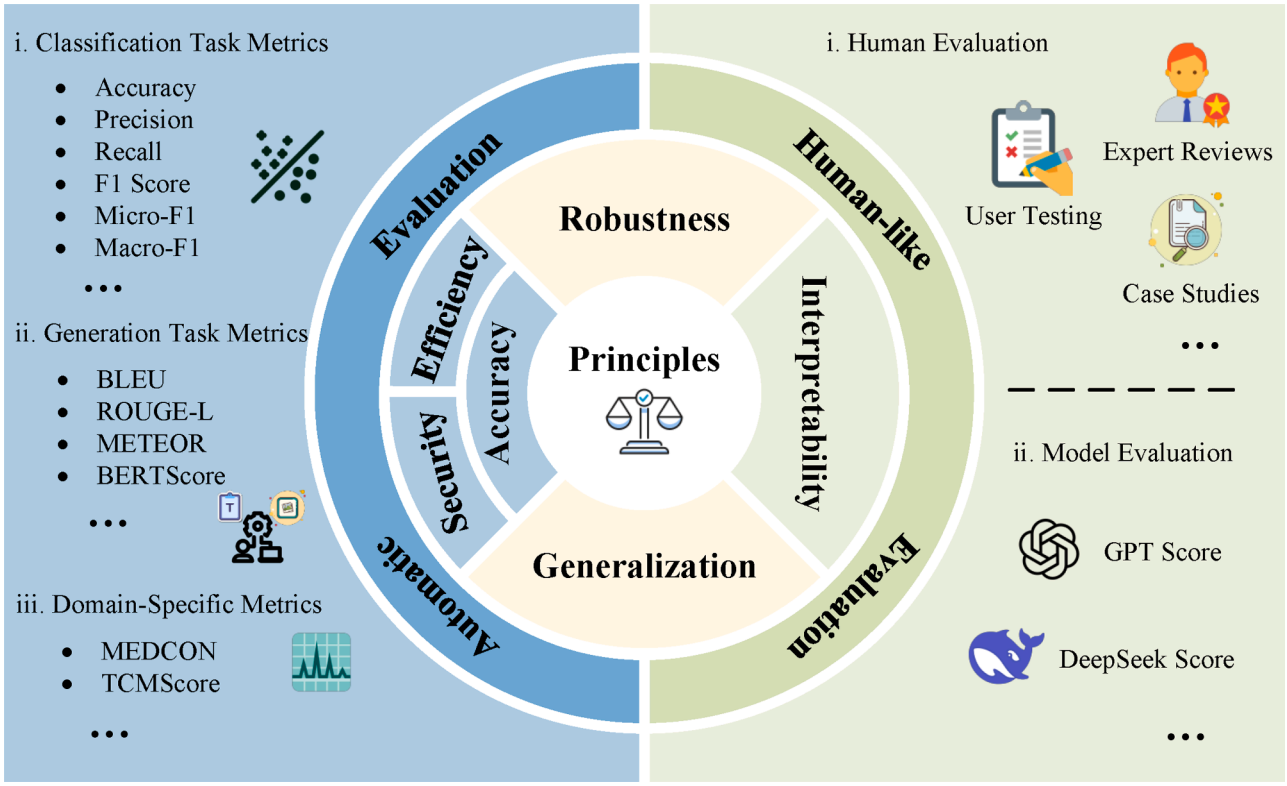


Fig. 9. An evaluation framework for medical LLMs. The framework is built on six principles (Accuracy, Robustness, Generalization, Interpretability, Efficiency, and Security). Evaluation combines automatic and human-like methods: automatic assessment (classification, generation, domain-specific) for accuracy/security/efficiency; human-like assessment (including expert review and model scoring) for interpretability; and hybrid assessment methods for robustness/generalization.

- **Robustness**: The model can maintain efficient performance when faced with different data inputs, ensuring accuracy even in the presence of noise, missing data, or anomalies.
- **Generalization**: The model can maintain good performance when faced with unseen data, indicating a true understanding of the problem rather than just memorizing the training data.
- **Interpretability**: The workings and predictions of the model can be explained and understood, which is especially critical in scenarios involving important decisions.
- **Efficiency**: The time and resources consumed by the model in processing the task. In practice, efficient models can respond to requests faster and save computational resources.

- **Security**: Guarantee that the model is not maliciously exploited, ensuring data confidentiality during processing, and preventing the model's predictions from leading to undesirable consequences.

In addition to the above, other aspects, such as simplicity and novelty, can also be evaluated for the preferences of different models.

5.2. Assessment methodology

Selecting appropriate assessment methods is essential to ensuring the quality and reliability of model evaluation results. When evaluating model performance, considerations should be made to select assessment methods and metrics accurately reflecting the model's actual performance. This section will examine two primary assessment methods: automatic evaluation and human-like evaluation.

5.2.1. Automatic evaluation

Automated quantitative metrics assess model performance and applicability quickly and objectively using algorithms and computational tools. They can process large datasets, reduce human bias, and enable real-time updates. We classify these metrics from three aspects: classification task, generation task, and domain-specific, which provide insights into the model's predictive ability and support quantitative analysis.

Classification task metrics. The following metrics are designed to evaluate the performance of classification models for both binary and multi-classification tasks, applicable to general and medical domains. Accuracy measures the proportion of correct predictions, while Precision indicates the ratio of true positives among predicted positives. Recall represents the fraction of actual positives correctly identified, and the F1-score provides a balanced measure as the harmonic mean of precision and recall. For multi-class scenarios, Micro-F1 calculates a global F1 score by aggregating all classes, whereas Macro-F1 computes class-specific F1 scores before averaging, making it suitable for multi-class classification tasks. Together, these metrics assess a model's predictive accuracy and robustness.

Generation task metrics. The following metrics are used to evaluate the performance of text generation models, especially in tasks such as machine translation, text summarization, and dialogue systems. For instance, BLEU [143] usually evaluates the quality of machine translation. It scores by comparing the overlap between the machine-generated translations and human translations. ROUGE-L [144] assesses the performance of automatic summarization and machine translation. It calculates the longest common subsequence between the predicted and reference text. METEOR [145] also evaluates the quality of machine translation. It combines precision and recall, considering word matches and word order between the candidate and reference translation. BERTScore [146], based on the BERT model, evaluates the quality of text generation tasks such as machine translation and text summarization.

Domain-specific metrics. The above evaluation metrics, while commonly applied to a wide range of fields, their application in medicine need to be appropriately adjusted to adapt to the uniqueness and complexity of medical tasks. For example, the medical concepts-based assessment metric, MEDCON [147], is used to measure the accuracy and consistency of clinical concepts. This metric calculates the F1 score to determine the similarity between the UMLS concept set in the candidate and reference clinical notes. There is also the TCMScore metric [148] for assessing TCM semantic and knowledge coherence, which combines the matching of TCM terms and semantic consistency between the generated and standard analysis. In this case, term matching calculates by adding term diversity to the original F1 score calculated from precision and recall, thus evolving into a Term F1 Score.

5.2.2. Human-like evaluation

We categorize human-like evaluation methods into two types: human evaluation and LLM-based evaluation. Human evaluation involves professionals conducting user testing, expert reviews, and case studies for qualitative insights. The LLM-based evaluation uses models like GPT-4 to simulate expert scoring and analysis. This approach focuses on leveraging expert knowledge for subjective assessments, especially in areas where automatic evaluation is insufficient, such as model interpretability and robustness. Furthermore, human-like evaluation examines model practicality and user acceptance.

Currently, medical academic LLMs emphasize the model's performance in response to specific medical scenarios, such as medical exams and QA. These tasks are often derived from real medical data and evaluated for accuracy using an automatic matching method. Common evaluation datasets include MultiMedQA, MultiMedBench, CMB [149], PromptCBLUE [150], RJUA-QA Datasets [151], EHRNoteQA [152], Aci-bench [147] and TCM-ED [148]. While these evaluation datasets provide diverse medical information, including images, text, and clinical Q&A, and some also introduce human evaluation frameworks to improve data accuracy, they have some limitations. For example, MultiMedQA has limited labeled data based on diagnostic reports and can only accurately assess a limited number of tasks. MultiMedBench has a limited amount of data in the transcriptomics and proteomics domains. CMB, PromptCBLUE, RJUA-QA Datasets, and TCM-ED still have limitations in terms of data diversity, and the EHRNoteQA and Aci-bench are insufficient in task diversity. They all have room for improvement in data scale or disease coverage. Interestingly, MedKGEval [140] pioneers medical knowledge coverage evaluation in LLMs using medical KGs through multi-level tasks (entity/relation/subgraph), quantifying GPT-4o's limited medical knowledge coverage (e.g., only 55.6 % at the relation level) and revealing the inherent limitations of vertical-domain models.

5.3. Takeaways

We systematically summarize the data distribution of academic LLMs across different medical capabilities (see Figs. D.1 for details) and conduct a comparative analysis of these models' performance in medical capabilities (see our GitHub for details). According to existing research, for medical professional knowledge mastery ability, the Med-PaLM 2 model (in English datasets) and HuatuoGPTII (in Chinese datasets) perform better. For medical Q&A ability, Med-PaLM 2 performs better. For medical information extraction ability, GatorTronGPT (in English datasets) and ChiMed-GPT (in Chinese datasets) perform better. For medical dialogue ability, HuatuoGPT performs better. For medical content generation ability, PULSE performs better. For medical visual Q&A ability, LLaVA-Med performs better overall. For medical report summarization, report generation, and image classification ability, Med-PaLM M performs better.

However, our research finds that there are significant differences between models when assessing the same medical capabilities, mainly in the following aspects: **1) Dataset inconsistency** (evaluate using different datasets, making direct comparisons difficult); **2) Variability in evaluation methods and metrics** (metrics are not entirely uniform in automatic evaluation, even within the same dataset); **3) Limitations of human evaluation** (human evaluation systems vary, hindering unified assessments across models). These issues limit fair and systematic comparisons between models with the same capabilities and affect the generalizability and reproducibility of the research findings.

Given the specificities of the medical field, we recommend combining automatic and human evaluations to create a comprehensive assessment framework. This approach leverages the efficiency of automatic evaluation with the depth of human analysis, ensuring thorough, reliable, and effective model performance in practical applications. Future research should focus on the following points:

1) Establish unified evaluation benchmarks and datasets (provide a foundation for horizontal comparisons between different models); **2) Clarify evaluation metrics and methods** (improve the comparability and consistency of evaluation results); **3) Develop a standardized human evaluation system** (reduce subjective bias in the evaluation process).

6. Challenges and future work

6.1. Challenges

Despite promising advances, medical LLMs face persistent challenges, including 1) data scarcity (multilingual/rare disease coverage), 2) model limitations (opaque decisions, multimodal fusion), 3) evaluation gaps (narrow metrics), and 4) imperfect KG integration (catastrophic forgetting). Current research addresses these through synthetic data augmentation, human-AI hybrid annotation, multi-task evaluation frameworks, and incremental KG fine-tuning, though often at the cost of computational overhead or generalizability.

While medical LLMs show promise, real-world implementation faces critical gaps in 1) clinically meaningful evaluation (beyond accuracy to dosage/treatment precision), 2) population diversity (comorbidities, rare etiologies), and 3) interpretability (traceable evidence for clinician trust). Current approaches address these through clinician-in-the-loop validation, multi-criteria evaluation frameworks (e.g., safety-reliability scores), and hybrid diagnostic systems that combine LLM outputs with structured medical knowledge, though often at the cost of increased system complexity or reduced automation efficiency.

In addition to the challenges mentioned above, data privacy, data security, and the corresponding regulatory and ethical issues are important for medical LLM practices [153]. These concerns must be strictly adhered to, whether in academic research or practical application.

6.2. Future work

To address these challenges, further investigations can consider the following paths:

1. In academic research, improving the medical knowledge base (multi-source entity alignment [154], multimodal fusion + causal reasoning [155]) and minimizing factual errors (domain-aware retrieval and adaptive calibration + RAG fine-tuning and cross-modal alignment [156]) can enhance the accuracy of models while strengthening their robustness and reliability.
2. Larger and more diverse evaluation datasets (multilingual/multimodal/multiple diseases) can help in the multi-dimensional assessment of models (three-layer assessment of core-translational-governance [157]) and validate their clinical applicability.
3. Increasing the transparency [158] of the models can improve their interpretability [159]. Additionally, closely integrating the models with clinical practice can ensure their traceability and evidence-based transferability [111,160].

With the accumulation of medical data and optimization of algorithms, medical LLMs are expected to perform even better in real-world applications, providing doctors with more accurate decision-making assistance and patients with higher-quality medical services. Beyond technical advancements, integrating the doctor-patient relationship and providing humanistic care into academic medical LLMs research can enhance overall healthcare quality, such as optimizing patient experiences [161–163], improving medical resources and services [164,165], and providing psychological support [166] and cross-cultural communication [167].

7. Conclusion

Clinical knowledge about the causes, prognosis, diagnosis, and treatment of diseases can improve treatment performances, and promote physical health. Recent LLMs' development opens new possibilities in medical AI. In this review, we gather the building paradigms of medical AI systems including the use of clinical databases, datasets, training pipelines, evaluation systems, as well as methods integrating medical KGs. Finally, we present the differences between academic medical LLMs and industrial ones and summarize the challenges to implementing academic LLMs in real-life medical situations. We also present some of the future directions for academic research and applications. We hope that our review presents an overview of the field of medical AI and its applications as well as challenges and future works for these technologies.

Author Agreement

We the undersigned declare that this manuscript entitled "Reviewing Clinical Knowledge in Medical Large Language Models: Training and Beyond" is original, has not been published before and is not currently being considered for publication elsewhere. An earlier version of this work was posted on arXiv (arXiv: 2502.20988v1, 28 Feb 2025). The submitted version includes significant revisions as detailed in the Cover Letter. We agree to transfer copyright to Knowledge-Based Systems upon acceptance. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We understand that the Corresponding Author is the sole contact for the Editorial process. He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

CRediT authorship contribution statement

Qiyuan Li: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Haijiang Liu:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis; **Caicai Guo:** Writing – original draft, Visualization, Investigation, Formal analysis, Conceptualization; **Chao Gao:** Visualization, Validation, Data curation; **Deyu Chen:** Investigation, Formal analysis; **Meng Wang:** Validation, Formal analysis; **Feng Gao:** Writing – review & editing, Validation, Supervision; **Frank van Harmelen:** Writing – review & editing; **Jinguang Gu:** Writing – review & editing, Funding acquisition, Conceptualization.

Data availability

I have shared the link to my data at the Attach File step.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grants 2022YFC3300801.

Appendix A. Summary of clinical datasets

We present the summarization and statistics of clinical datasets, which include both pre-training datasets and fine-tuning datasets, as shown in Tables A.1 and A.2, respectively.

Table A.1
A summary of the clinical datasets discussed in Section 2.2, which include both pre-training datasets (text pre-training datasets and multimodal pre-training datasets) and fine-tuning datasets (text fine-tuning datasets and multimodal fine-tuning datasets), in terms of data type, authenticity, diversity, coverage, update frequency, usability, reliability, and language. "Y" means "yes", "N" means "no", "M" means "mixed datasets (i.e., both real and generated data)", "CA" means "conditional access", and "P" means "partially reliable".

	DataType	Authenticity	Diversity						Medical Coverage			Update Frequency	Usability	Reliability	Language
			Text	Figure	Radiology	Pathology	Time series	Numerical Value	Fundamentals	Specialty Knowledge	Time Span				
PubMed	Medical Literature	Y	Y	–	–	–	–	Y	Y	–	–	Once a year, around December	Y	Y	English
MedDialog	doctor-Patient Dialogue	Y	Y	–	–	–	–	–	–	96/172 diseases	2008-2020/ 2010-2020	–	Y	Y	English /Chinese
EHRs	Patient Clinical Records	Y	Y	–	–	–	–	–	–	126 clinical departments	2011–2021	–	N	Y	English
ChiMed-CPT	Combined Dataset	Y	Y	–	–	–	–	–	Y	–	–	–	Y	Y	English /Chinese
GAP-REPLAY	Combined Dataset	Y	Y	–	–	–	–	Y	Y	–	–	–	Y	Y	English
The Pile	Combined Dataset	Y	Y	–	–	–	–	Y	Y	–	–	–	Y	Y	English
MTB	Medical Textbooks	Y	Y	Y	–	–	–	–	Y	–	–	–	N	Y	English
PMC-OA	Literature Image-Text	Y	Y	Y	–	–	–	–	Y	–	–	–	Y	Y	English
MIMIC-CXR	Chest X-ray Image-Text	Y	Y	–	Y	–	–	–	–	14 types of chest findings	2011–2016	–	CA	Y	English
MIMIC-III	Patient Health Data	Y	Y	–	Y	–	Y	Y	Y	–	2001–2012	–	CA	Y	English
MIMIC-IV	Patient Health Data	Y	Y	–	Y	–	Y	Y	Y	–	2008–2019	–	CA	Y	English
MIMIC-CXR-JPG	Chest X-ray Images	Y	–	–	Y	–	–	–	–	14 types of chest findings	2011–2016	–	CA	Y	English
MIMIC-IV-Note	Patient Clinical Records	Y	Y	–	Y	–	Y	Y	Y	–	Within one year after the visit	–	CA	Y	English
CMExam	Medical Licensing Exam QA	Y	Y	–	–	–	–	–	–	26 diseases, 35 clinical departments	–	–	Y	Y	Chinese
MedQA	Medical Board Exam QA	Y	Y	–	–	–	–	–	Y	–	–	–	Y	Y	Simplified Chinese/ Traditional Chinese/ English
PubMedQA	Medical Literature QA	M	Y	–	–	–	–	Y	Y	–	–	–	Y	P	English

Table A.1
continue.

	DataType	Authenticity	Diversity						Medical Coverage			Update Frequency	Usability	Reliability	Language
			Text	Figure	Radiology	Pathology	Time series	Numerical Value	Fundamentals	Specialty Knowledge	Time Span				
cMedQA2	Chinese Community Medicine QA	Y	Y	–	–	–	–	–	Y	–	–	–	Y	Y	Chinese
MedQuAD	Medical QA Pairs	Y	Y	–	–	–	–	–	Y	–	–	–	Y	Y	English
CMD.	Medical Dialogue	Y	Y	–	–	–	–	–	–	6 types of clinical departments	–	–	Y	Y	Chinese
MedDialog-CN	doctor-Patient Dialogue	Y	Y	–	–	–	–	–	–	29 broad categories of specialties & 172 fine-grained specialties	2010–2020	–	Y	Y	Chinese
Multi-MedQA	Combined Dataset	Y	Y	–	–	–	–	–	Y	–	–	–	Y	Y	English
MMedBench	Combined Dataset	N	Y	–	–	–	–	–	Y	21 medical topic categories	–	–	Y	Y	English/ Chinese/ Spanish/ Japanese/ French/ Russian
PathVQA	Pathology Image QA	Y	Y	–	–	Y	–	–	–	Pathology	–	–	Y	Y	English
VQA-RAD	Radiology QA	Y	Y	–	Y	–	–	–	–	Radiology	–	–	Y	Y	English
VQA-med-2018	Medical Image QA	Y	Y	–	Y	–	–	–	–	Radiology	–	–	Y	Y	English
VQA-med-2019	Medical Image QA	Y	Y	–	Y	–	–	–	–	Radiology	–	–	Y	Y	English
VQA-med-2020	Medical Image QA	Y	Y	–	Y	–	–	–	–	Radiology	–	–	Y	Y	English
VQA-med-2021	Medical Image QA	Y	Y	–	Y	–	–	–	–	Radiology	–	–	Y	Y	English
SLAKE	Medical Image QA	Y	Y	–	Y	–	–	–	–	12 diseases, 39 organs	–	–	Y	Y	English/ Chinese
PMC-15M	Biomedical Image-Text Pairs	Y	Y	Y	–	–	–	–	Y	–	–	–	N	Y	English
ChiMed-VL	Combined Dataset	Y	Y	Y	Y	–	–	–	Y	–	–	–	Y	P	English/ Chinese
MultiMed-Bench	Combined Dataset	Y	Y	–	Y	Y	–	–	Y	Pathology, radiology, dermatology, genomics	–	–	CA	Y	English

Table A.2
A statistic of the clinical datasets discussed in Section 2.2, which include both pre-training datasets (text pre-training datasets and multimodal pre-training datasets) and fine-tuning datasets (text fine-tuning datasets and multimodal fine-tuning datasets). "Q" means "question", "A" means "answer".

	Size			Total
	train	dev/val	test	
PubMed	–	–	–	37M biomedical literatures
MedDialog	–	–	–	MedDialog-EN: 0.26M doctor-patient dialogue, MedDialog-CN: 3.4M doctor-patient dialogue
EHRs	–	–	–	290,482,002 clinical notes
ChiMed-CPT	–	–	–	1219K QA, 8K plain text, 406K KG, 3918K dialogue
GAP-REPLAY	–	–	–	43.2K clinical guidelines, 16.1M paper abstract, 5M medical papers, 494K experience replay
The Pile	–	–	–	PubMed Central: 5M publications, PubMed Abstracts: 30M publications
MTB	–	–	–	4721 textbooks
PMC-OA	–	–	–	1.6M medical image-text
MIMIC-CXR	–	–	–	377K medical image-text
MIMIC-III	–	–	–	53.4K EHRs
MIMIC-IV	–	–	–	40K EHRs
MIMIC-CXR-JPG	–	–	–	377K medical images
MIMIC-IV-Note	–	–	–	2.65M medical texts
CMExam	54,497 Q	6811 Q	6811 Q	68,119 Q
MedQA	USMLE: 10,178 Q, MCMLE: 27,400 Q, TWMLE: 11,298 Q	USMLE: 1,272 Q, MCMLE: 3,425 Q, TWMLE: 1,412 Q	USMLE: 1,273 Q, MCMLE: 3,426 Q, TWMLE: 1,413 Q	USMLE: 12,723 Q, MCMLE: 34,251 Q,TWMLE: 14,123 Q
PubMedQA	–	–	–	PQA-L(abeled): 1K QA pairs, PQA-U(nlabeled): 61.2K QA pairs, PQA-A(rtificial): 211.3K QA pairs
cMedQA2	100,000 Q, 188,490 A	4000 Q, 7527 A	4000 Q, 7552 A	108,000 Q, 203,569 A
MedQuAD	–	–	–	47,457 QA pairs
CMD.	Andriatria: 94,596 QA pairs, IM: 220,606 QA pairs, OAGD: 183,751 QA pairs, Oncology: 75,553 QA pairs, Pediatric: 101,602 QA pairs, Surgical: 115,991 QA pairs	–	–	792,099 QA pairs
MedDialog-CN	–	–	–	3.4M doctor-patient dialogue
MultiMedQA	–	MedQA(USMLE): 11,450 QA pairs MedMCQA: 187K QA pairs PubMedQA: 500 QA pairs MMLU: 123 QA pairs LiveQA(TREC-2017): 634 QA pairs - HealthSearchQA: 3,375 Q	MedQA(USMLE): 1,273 QA pairs MedMCQA: 6.1K QA pairs PubMedQA: 500 QA pairs MMLU: 1,089 QA pairs LiveQA(TREC-2017): 104 QA pairs Medication QA: 674 QA pairs	MedQA(USMLE): 12,723 QA pairs MedMCQA: 193.1K QA pairs PubMedQA: 1,000 QA pairs MMLU: 1,212 QA pairs LiveQA(TREC-2017): 738 QA pairs Medication QA: 674 QA pairs HealthSearchQA: 3,375 Q
MMedBench	45,048 QA pairs	–	8518 QA pairs	53,566 QA pairs
PathVQA	17,325 QA pairs, 2499 images	9462 QA pairs, 1499 images	6012 QA pairs, 1000 images	32,799 QA pairs, 4998 images
VQA-RAD	–	–	–	3515 QA pairs, 315 images
VQA-med-2018	5K images with QA pairs	0.5K images with QA pairs	0.5K images with Q only	–
VQA-med-2019	12,792 QA pairs, 3200 images	2000 QA pairs,500 images	500 QA pairs,500 images	–
VQA-med-2020	VQA task: 4,000 QA pairs,4,000 images, VQG task: 2,156 QA pairs,780 images	VQA task: 500 QA pairs,500 images, VQG task: 164 QA pairs,141 images	VQA task: 500 QA pairs,500 images, VQG task: 80 images	–
VQA-med-2021	4000 QA pairs,4000 images	500 images with Q/A about Abnormality	500 images with related Q about Abnormality	-
SLAKE	–	–	–	14,028 Q, 642 images
PMC-15M	–	–	–	15M image-caption pairs
ChiMed-VL	–	–	–	ChiMed-VL-Alignment: 580K image-text pairs,ChiMed-VL-Instruction: 460K QA pairs
MultiMedBench	10,178 Q	–	1273 Q	MedQA(USMLE): 11,451 Q
	182,822 Q	–	4183 Q	MedMCQA: 187,005 Q
	–	–	500 Q	PubMedQA: 500 Q
	58,405 reports	7413 reports	13,057 reports	MIMIC-III: 78,875 reports
	1,797 QA pairs (only free-form and paraphrased Q)	–	451 QA pairs (not filtered)	VQA-RAD: 3,515 QA pairs
	9849 QA pairs	2109 QA pairs	2070 QA pairs	Slake-VQA: 14,028 QA pairs
	19,755 QA pairs	6279 QA pairs	6761 QA pairs	Path-VQA: 32,799 QA pairs
	353,542 samples	2866 samples	4834 samples	MIMIC-CXR(RG task): 361,242 image-text pairs
	–	–	–	MIMIC-CXR(MIC task): 78,875 images
	1838 images	–	460 images	PAD-UFES-20: 2,298 images
	16,000 images	–	4000 images	VinDr-Mammo: 20,000 images
	1,318 images(mass), 1544 images(calcification)	–	378 images(mass), 326 images(calcification)	CBIS-DDSM: 2,620 images
	197,038 candidate variants	–	13,030 candidate variants	PrecisionFDA: 210,068 genomic variants

Appendix B. Industrial text medical LLMs

We present the summarization of industrial medical LLMs building from text corpora, as shown in Table B.1.

Table B.1
Application for plain-text industrial medical LLMs. These models often were built on vast real-life clinical data collected by cooperation's, which reflects the patient condition more comprehensively.

Application Name	Anchor Model	Data Scale	Enterprise Name	Function
UniGPT-Med	UniGPT [a]	2,000B tokens	Unisound AI Technology Co., Ltd.	AI-assisted D&T Medical Record Generation
–	EyeGPT [b]	–	Wenzhou International Optometry Innovation Centre	Clinical Medical Assistance
DaJing TCM	QiHuangWenDao[c]	11M TCM-KG 1,500 books&literatures 100,000 real medical cases 100,000 P&T&M&A[d] 2M diagnosis	Nanjing Dajing TCM IT Co., Ltd.	Intelligent Assisted D&T Intelligent Health Conditioning
DingDangKuaiYao[e]	HealthGPT	–	DingDangKuaiYao Technology Group Co., Ltd.	Quick Medicine Service Online Consultation Chronic and Health Management
Gushengtang TCM Gushengtang-doctor	GushengtangTCM	-	Gushengtang TCM Chain Management Group	AI-assisted D&T Recommended TCM Formulas
Xunfei Xiaoyi	Xinghuo[f]	–	iFLYTEK Co., Ltd.	Report Interpretation TCM Syndromes Identification
Zhiyun Health[g]	ClouD GPT	–	Hangzhou Comms IT Co., Ltd.	AI-assisted Diagnosis AI Drug Device Development
WiNEX[h]	WiNGPT[i]	30B/65B tokens	Winning Health Technology Group Co.,Ltd	Health Care Assistant
Tongyi Renxin[j]	Tongyi Qwen[k]	>3,000B tokens	Alibaba Group	Intelligent Inquiry Report Interpretation Summary Abstract

[a] http://shanhai.unisound.com/
[b] http://eyegpt.com.cn/#/
[c] http://www.dajingtcm.com/dajinggpt
[d] Pulse&Tongue&Meridian&Acupuncture
[e] https://www.ddky.com/
[f] https://xinghuo.xfyun.cn/
[g] https://www.zyhealth.com
[h] https://www.winning.com.cn/WiNEX/
[i] https://github.com/winninghealth/WiNGPT2
[j] https://tongyi.aliyun.com/renxin
[k] https://tongyi.aliyun.com/qianwen

Appendix C. Industrial multimodal medical LLMs

We present the summarization of industrial medical LLMs building from multimodal data, as shown in Table C.1.

Table C.1
Application for multimodal industrial medical LLMs. The integration of numerous clinical data gives the model a better understanding over the fine-grained medical knowledge.

Application Name	Anchor Model	Data Scale	Enterprise Name	Function
StoneNeedle [141]	–	–	AthenaEyesCo., LTD.	AI-assisted Diagnosis AI-assisted Reading Physiological Prediction Sleep Monitoring
Tencent MedLLM	Hunyuan[a]	2.85M medical entities 12.5M medical relations Medical KG&literatures covering 98 % 30M dialogues 360,000 expert's annotation	Tencent Cloud Computing (Beijing) Co., Ltd.	Content Generation Medical Record Structuralization AI-assisted Diagnosis AI Rational Drug Use AI-assisted Reading
PanGu	Pangu-Drug [142]	>3,000B tokens	Huawei Technologies Co., Ltd.	AI-assisted drug R&D
Medical Sense	SenseChat[b]	>20B tokens	Shanghai SenseTime IT Co., Ltd.	AI-assisted diagnosis Clinic Interpreter Robot Medical Record Structuralization
MedLinker[c]	MedGPT	>2B medical texts SFT 8M diagnosis >100 doctors RLHF	Chengdu MedCloud Technology Co., Ltd.	Intelligent Health Inquiry
WeiMai WeiMai-doctor	CareGPT	>1B medical texts Millions of medical and health KB SFT >100 doctors RLHF	Weima Technology Co., Ltd.	Personalized Matching and Recommendation Aiding Decision-making
01Bot[d]	Baidu Wenxin[e]	Hundred billion tokens	Baidu Online Network Technology (Beijing) Co., Ltd.	AI-assisted Diagnosis Medical Record Generation

[a] https://hunyuan.tencent.com/
[b] https://chat.sensetime.com/
[c] https://www.medlinker.com/
[d] https://01.baidu.com/bot.HTML
[e] https://wenxin.baidu.com/

Appendix D. Data Ddistribution of academic LLMs across different medical capabilities

We systematically summarize the distribution of academic LLMs across various medical capabilities, facilitating a better assessment of the model's performance, as shown in Fig. D.1.

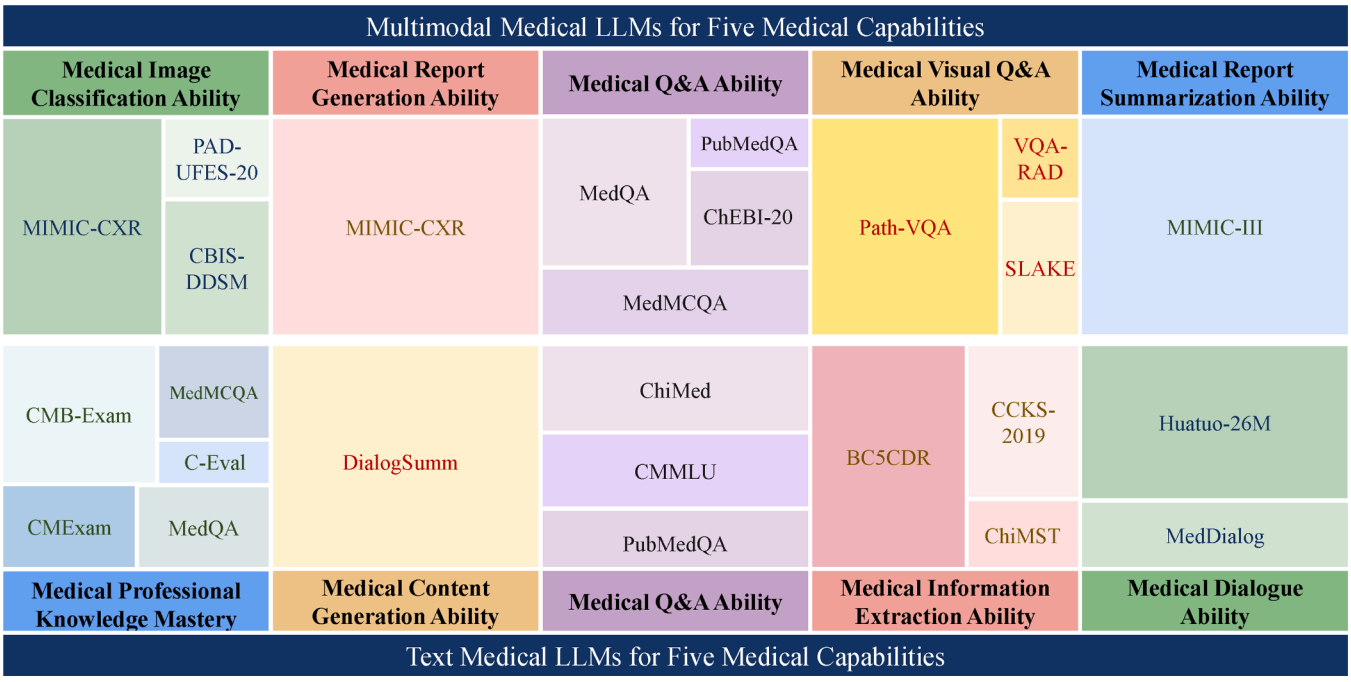


Fig. D.1. Data distribution of different capabilities of medical LLMs. Above and below are the five medical capabilities of the multimodal and text medical LLMs and their corresponding assessment datasets, respectively.

References

[1] S. Zhou, Z. Xu, M. Zhang, C. Xu, Y. Guo, Z. Zhan, S. Ding, J. Wang, K. Xu, Y. Fang, L. Xia, J. Yeung, D. Zha, G.B. Melton, M. Lin, R. Zhang, Large language models for disease diagnosis: a scoping review, CoRR abs/2409.00097 (2024). https://doi.org/10.48550/arXiv.2409.00097.

[2] G.K. Gupta, P. Pande, LLMs in disease diagnosis: a comparative study of deepseek-r1 and O3 mini across chronic health conditions, CoRR abs/2503.10486 (2025). https://doi.org/10.48550/arXiv.2503.10486.

[3] W.H.K. Chiu, W.S.K. Ko, W.C.S. Cho, S.Y.J. Hui, W.C.L. Chan, M.D. Kuo, Evaluating the diagnostic performance of large language models on complex multimodal medical cases, J. Med. Internet Res. 26 (2024). https://www.jmir.org/2024/1/e53724.

[4] D. Wang, S. Zhang, Large language models in medical and healthcare fields: applications, advances, and challenges, Artif. Intell. Rev. 57 (299) (2024). https://doi.org/10.1007/s10462-024-10921-0.

[5] Q. Xie, Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, X. Peng, J. Huang, J. Zhang, V.K. Keloth, X. Zhou, L. Qian, H. He, D. Shung, L. Ohno-Machado, Y. Wu, H. Xu, J. Bian, Medical foundation large language models for comprehensive text analysis and beyond, NPJ Dig. Med. 8 (1) (2025). https://doi.org/10.1038/s41746-025-01533-1.

[6] F. Busch, L. Hoffmann, C. Rueger, E.H.C. van Dijk, R. Kader, E. Ortiz-Prado, M.R. Makowski, L. Saba, M. Hadamitzky, J.N. Kather, D. Truhn, R. Cuocolo, L.C. Adams, K.K. Bressem, Current applications and challenges in large language models for patient care: a systematic review, Commun. Med. 5 (26) (2025). https://doi.org/10.1038/s43856-024-00717-2.

[7] F. Gaber, M. Shaik, F. Allega, A.J. Bilecz, F. Busch, K. Goon, V. Franke, A. Akalin, Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis, NPJ Dig. Med. 8 (263) (2025). https://doi.org/10.1038/s41746-025-01684-1.

[8] Y. Kim, C. Park, H. Jeong, Y.S. Chan, X. Xu, D. McDuff, H. Lee, M. Ghassemi, C. Breazeal, H.W. Park, MDAGents: an adaptive collaboration of LMs for medical decision-making, in: Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems, 2522, 2024, pp. 79410–79452. https://dl.acm.org/doi/10.5555/3737916.3740438.

[9] J.W. Kim, A. Podlasek, K. Shidara, F. Liu, A. Alaa, D. Bernardo, Limitations of large language models in clinical problem-solving arising from inflexible reasoning, CoRR abs/2502.04381 (2025). https://doi.org/10.48550/arXiv.2502.04381.

[10] P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis, D. Rueckert, Evaluation and mitigation of the limitations of large language models in clinical decision-making, Nat. Med. 30 (2613–2622) (2024). https://doi.org/10.1038/s41591-024-03097-1.

[11] J. Ye, H. Tang, Multimodal large language models for medicine: a comprehensive survey, CoRR abs/2504.21051 (2025). https://doi.org/10.48550/arXiv.2504.21051.

[12] J.C.L. Ong, S.Y.-H. Chang, W. William, A.J. Butte, N.H. Shah, L.S.T. Chew, N. Liu, F. Doshi-Velez, W. Lu, J. Savulescu, D.S.W. Ting, Medical ethics of large language models in medicine, NEJM AI 1 (7) (2024) AIra2400038. https://ai.nejm.org/doi/full/10.1056/AIra2400038.

[13] S. Sandmann, S. Hegselmann, M. Fujarski, L. Bickmann, B. Wild, R. Eils, J. Varghese, Benchmark evaluation of deepseek large language models in clinical decision-making, Nat. Med. (2025). https://doi.org/10.1038/s41591-025-03727-2.

[14] T. Hirosawa, T. Suzuki, T. Shiraishi, A. Hayashi, Y. Fujii, T. Harada, T. Shimizu, Adapting artificial intelligence concepts to enhance clinical decision-making: a hybrid intelligence framework, Int. J. Gen. Med. 17 (5417–5422) (2024). https://doi.org/10.2147/IJGM.S497753.

[15] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S.W. Ting, Large language models in medicine, Nat. Med. 29 (8) (2023) 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.

[16] M. Wornow, Y. Xu, R. Thapa, B. Patel, E. Steinberg, S. Fleming, M.A. Pfeffer, J. Fries, N.H. Shah, The shaky foundations of large language models and foundation models for electronic health records, NPJ Dig. Med. 6 (2023) 135. https://doi.org/10.1038/s41746-023-00879-8.

[17] C.W. Safranek, A.E. Sidamon-Eristoff, A. Gilson, D. Chartash, The role of large language models in medical education: applications and implications, JMIR Med. Educ. 9 (2023) e50945. https://doi.org/10.1038/s41746-023-00879-8.

[18] H. Zhou, F. Liu, B. Gu, X. Zou, J. Huang, J. Wu, Y. Li, S.S. Chen, P. Zhou, J. Liu, Y. Hua, C. Mao, C. You, X. Wu, Y. Zheng, L. Clifton, Z. Li, J. Luo, D.A. Clifton, A survey of large language models in medicine: progress, application, and challenge, CoRR abs/2311.05112 (2024). https://doi.org/10.48550/arXiv.2311.05112.

[19] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, E. Cambria, A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics, Inform. Fusion 118 (2025) 102963. https://doi.org/10.1016/j.inffus.2025.102963.

[20] Z. Ali, Y. Huang, I. Ullah, J. Feng, C. Deng, N. Thierry, A. Khan, A.U. Jan, X. Shen, W. Rui, G. Qi, Deep learning for medication recommendation: a systematic survey, Data Intell. 5 (2) (2023) 303–354. https://doi.org/10.1162/dint_a_00197.

[21] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, G. Wang, Instruction tuning for large language models: a survey, CoRR abs/2308.10792 (2023). https://doi.org/10.48550/arXiv.2308.10792.

[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLama: open and efficient foundation language models, CoRR abs/2302.13971 (2023). https://doi.org/10.48550/arXiv.2302.13971.

[23] T.L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilic, D. Hesslow, et al., Bloom: a 176b-parameter open-access multilingual language model, CoRR abs/2211.05100 (2022). https://doi.org/10.48550/arXiv.2211.05100.

[24] Y. Gu, R. Tinn, H. Cheng, M.R. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthcare (HEALTH) 3 (1) (2022) 2:1–2:23. https://doi.org/10.1145/3458754.

[25] G. Zeng, W. Yang, Z. Ju, Y. Yang, S. Wang, R. Zhang, M. Zhou, J. Zeng, X. Dong, R. Zhang, H. Fang, P. Zhu, S. Chen, P. Xie, Meddialog: large-scale medical dialogue datasets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020, pp. 9241–9250. https://doi.org/10.18653/v1/2020.emnlp-main.743.

[26] Z. Deng, W. Gao, C. Chen, Z. Niu, Z. Gong, R. Zhang, Z. Cao, F. Li, Z. Ma, W. Wei, L. Ma, OphGLM: an ophthalmology large language-and-vision assistant, Artif. Intell. Med. 157 (2024) 103001. https://doi.org/10.1016/j.artmed.2024.103001.

[27] X. Yang, A. Chen, N. PourNejatian, H.C. Shin, K.E. Smith, C. Parisien, C. Compas, C. Martin, A.B. Costa, M.G. Flores, Y. Zhang, T. Magoc, C.A. Harle, G. Lipori, D.A. Mitchell, W.R. Hogan, E.A. Shenkman, J. Bian, Y. Wu, A large language model for electronic health records, NPJ Digital Med. 5 (1) (2022) 194. https://doi.org/10.1038/s41746-022-00742-2.

[28] Q. Ye, J. Liu, D. Chong, P. Zhou, Y. Hua, A. Liu, Qilin-MED: multi-stage knowledge injection advanced medical large language model, CoRR abs/2310.09089 (2023). https://doi.org/10.48550/arXiv.2310.09089.

[29] Z. Chen, A. Hernández-Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M.-A. Hartley, M. Jaggi, A. Bosselut, Meditron-70b: scaling medical pretraining for large language models, CoRR abs/2311.16079 (2023). https://doi.org/10.48550/arXiv.2311.16079.

[30] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, The pile: an 800GB dataset of diverse text for language modeling, CoRR abs/2101.00027 (2021). https://arxiv.org/abs/2101.00027.

[31] E. Bolton, A. Venigalla, M. Yasunaga, D. Hall, B. Xiong, T. Lee, R. Daneshjou, J. Frankle, P. Liang, M. Carbin, C.D. Manning, BiomedLM: a 2.7b parameter language model trained on biomedical text, CoRR abs/2403.18421 (2024). https://doi.org/10.48550/arXiv.2403.18421.

[32] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E.P. Reis, P. Rajpurkar, Med-flamingo: a multimodal medical few-shot learner, in: S. Hegselmann, A. Parziale, D. Shanmugam, S. Tang, M.N. Asiedu, S. Chang, T. Hartvigsen, H. Singh (Eds.), Proceedings of the 3rd Machine Learning for Health Symposium, volume 225 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 353–367. https://proceedings.mlr.press/v225/moor23a.html.

[33] W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, W. Xie, PMC-CLIP: contrastive language-image pre-training using biomedical documents, in: Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, volume 14227 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 525–536. https://doi.org/10.1007/978-3-031-43993-3_51.

[34] A.E.W. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, C. Deng, R.G. Mark, S. Horng, Mimic-CXR: a large publicly available database of labeled chest radiographs, CoRR abs/1901.07042 (2019). http://arxiv.org/abs/1901.07042.

[35] Z. Wang, Z. Wu, D. Agarwal, J. Sun, MedCLIP: contrastive learning from unpaired medical images and text, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2022, pp. 3876–3887. https://doi.org/10.18653/v1/2022.emnlp-main.256.

[36] A.E.W. Johnson, T.J. Pollard, L. Shen, L.-W.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, Mimic-III, a freely accessible critical care database, Sci. Data 3 (160035) (2016). https://doi.org/10.1038/sdata.2016.35.

[37] E. Alsentzer, J.R. Murphy, W. Boag, W. Weng, D. Jin, T. Naumann, M.B.A. McDermott, Publicly available clinical BERT embeddings, CoRR abs/1904.03323 (2019). http://arxiv.org/abs/1904.03323.

[38] A.E.W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T.J. Pollard, S. Hao, B. Moody, B. Gow, L.-w.H. Lehman, L.A. Celi, R.G. Mark, Mimic-IV, a freely accessible electronic health record dataset, Sci. Data 10 (1) (2023). https://doi.org/10.1038/s41597-022-01899-x.

[39] A.E.W. Johnson, T.J. Pollard, N.R. Greenbaum, M.P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R.G. Mark, S.J. Berkowitz, S. Horng, Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019 . https://arxiv.org/abs/1901.07042.

[40] A. Johnson, T. Pollard, S. Horng, L.A. Celi, R. Mark, Mimic-IV-note: Deidentified free-text clinical notes, 2023 . https://doi.org/10.13026/1n74-ne17.

[41] J. Fehr, Llama-care, a multimodal medical large language model for hospital discharge instruction generation, 2023 . https://jf-11.github.io/pages/llama_care.html.

[42] J. Liu, P. Zhou, Y. Hua, D. Chong, Z. Tian, A. Liu, H. Wang, C. You, Z. Guo, L. Zhu, M.L. Li, Benchmarking large language models on CMExam - a comprehensive Chinese medical exam dataset, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, 2283, 2023, pp. 52430–52452. https://dl.acm.org/doi/10.5555/3666122.3668405.

[43] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, P. Szolovits, What disease does this patient have? A large-scale open domain question answering dataset from medical exams, Appl. Sci. 11 (14) (2021) 6421. https://doi.org/10.3390/app11146421.

[44] Q. Jin, B. Dhingra, Z. Liu, W.W. Cohen, X. Lu, PubmedQA: a dataset for biomedical research question answering, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2019, pp. 2567–2577. https://doi.org/10.18653/v1/D19-1259.

[45] S. Zhang, X. Zhang, H. Wang, L. Guo, S. Liu, Multi-scale attentive interaction networks for chinese medical question answer selection, IEEE Access 6 (2018) 74061–74071. https://doi.org/10.1109/ACCESS.2018.2883637.
[46] A.B. Abacha, D. Demner-Fushman, A question-entailment approach to question answering, BMC Bioinformatics 20 (511) (2019). https://doi.org/10.1186/s12859-019-3119-4.
[47] K. Singhal, S. Azizi, T. Tu, S.S. Mahdavi, J. Wei, H.W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B.A.y. Arcas, D. Webster, G.S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, V. Natarajan, Large language models encode clinical knowledge, Nature 620 (2023) 172–180. https://doi.org/10.1038/s41586-023-06291-2.
[48] P. Qiu, C. Wu, X. Zhang, W. Lin, H. Wang, Y. Zhang, Y. Wang, W. Xie, Towards building multilingual language model for medicine, Nat. Commun. 15 (8384) (2024). https://doi.org/10.1038/s41467-024-52417-z.
[49] X. He, Y. Zhang, L. Mou, E.P. Xing, P. Xie, PathVQA: 30000+ questions for medical visual question answering, CoRR abs/2003.10286 (2020). https://arxiv.org/abs/2003.10286.
[50] J.J. Lau, S. Gayen, A.B. Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, Sci. Data 5 (180251) (2018). https://doi.org/10.1038/sdata.2018.251.
[51] S.A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, M.P. Lungren, Overview of imageCLEF 2018 medical domain visual question answering task, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, volume 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. https://ceur-ws.org/Vol-2125/paper_212.pdf.
[52] A.B. Abacha, S.A. Hasan, V.V. Datla, J. Liu, D. Demner-Fushman, H. Müller, VQA-MED: overview of the medical visual question answering task at imageclef 2019, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. https://ceur-ws.org/Vol-2380/paper_272.pdf.
[53] A.B. Abacha, V.V. Datla, S.A. Hasan, D. Demner-Fushman, H. Müller, Overview of the VQA-Med task at imageCLEF 2020: Visual question answering and generation in the medical domain, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. https://ceur-ws.org/Vol-2696/paper_106.pdf.
[54] A.B. Abacha, M. Sarrouti, D. Demner-Fushman, S.A. Hasan, H. Müller, Overview of the VQA-Med task at imageCLEF 2021: Visual question answering and generation in the medical domain, in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 1081–1088. https://ceur-ws.org/Vol-2936/paper-87.pdf.
[55] B. Liu, L. Zhan, L. Xu, L. Ma, Y. Yang, X. Wu, Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering, in: 18th IEEE International Symposium on Biomedical Imaging, IEEE, 2021, pp. 1650–1654. https://doi.org/10.1109/ISBI48211.2021.9434010.
[56] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, M.P. Lungren, T. Naumann, H. Poon, Large-scale domain-specific pretraining for biomedical vision-language processing, CoRR abs/2303.00915 (2023). https://doi.org/10.48550/arXiv.2303.00915.
[57] J. Liu, Z. Wang, Q. Ye, D. Chong, P. Zhou, Y. Hua, Qilin-MED-VL: towards Chinese large vision-language model for general healthcare, CoRR abs/2310.17956 (2023). https://doi.org/10.48550/arXiv.2310.17956.
[58] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, A. Palepu, B. Mustafa, A. Chowdhery, Y. Liu, S. Kornblith, D. Fleet, P. Mansfield, S. Prakash, R. Wong, S. Virmani, C. Semturs, S.S. Mahdavi, B. Green, E. Dominowska, B.A.y. Arcas, J. Barral, D. Webster, G.S. Corrado, Y. Matias, K. Singhal, P. Florence, A. Karthikesalingam, V. Natarajan, Towards generalist biomedical AI, NEJM AI 1 (3) (2024) AIoa2300138. https://ai.nejm.org/doi/full/10.1056/AIoa2300138.
[59] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N.A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: aligning language models with self-generated instructions, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2023, pp. 13484–13508. https://doi.org/10.18653/v1/2023.acl-long.754.
[60] Y. Ji, Y. Deng, Y. Gong, Y. Peng, Q. Niu, L. Zhang, B. Ma, X. Li, Exploring the impact of instruction data scaling on large language models: an empirical study on real-world use cases, CoRR abs/2303.14742 (2023). https://doi.org/10.48550/arXiv.2303.14742.
[61] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, OpenAI Blog (2018).
[62] Y. Li, Z. Li, K. Zhang, R. Dan, Y. Zhang, Chatdoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge, Cureus 15 (6) (2023). https://doi.org/10.7759/cureus.40895.
[63] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, T. Liu, Huatuo: tuning LLaMA model with Chinese medical knowledge, CoRR abs/2304.06975 (2023). https://doi.org/10.48550/arXiv.2304.06975.
[64] H. Xie, Y. Chen, X. Xing, J. Lin, X. Xu, PsyDT: Using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2025, pp. 1081–1115. https://aclanthology.org/2025.acl-long.55/.
[65] H. Qiu, H. He, S. Zhang, A. Li, Z. Lan, SMILE: single-turn to multi-turn inclusive language expansion via chatgpt for mental health support, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, 2024, pp. 615–636. https://doi.org/10.18653/v1/2024.findings-emnlp.34.
[66] Y. Chen, X. Xing, J. Lin, H. Zheng, Z. Wang, Q. Liu, X. Xu, Soulchat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, 2023, pp. 1170–1183. https://doi.org/10.18653/v1/2023.findings-emnlp.83.
[67] C. Zhang, R. Li, M. Tan, M. Yang, J. Zhu, D. Yang, J. Zhao, G. Ye, C. Li, X. Hu, CPsyCoun: a report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling, in: Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, 2024, pp. 13947–13966. https://doi.org/10.18653/v1/2024.findings-acl.830.
[68] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems, 159, 2020, pp. 1877–1901. https://dl.acm.org/doi/abs/10.5555/3495724.3495883.
[69] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W.L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, Z. Liu, P. Zhang, Y. Dong, J. Tang, GLM-130B: an open bilingual pre-trained model, in: The Eleventh International Conference on Learning Representations, OpenReview.net, 2023. https://openreview.net/pdf?id=-Aw0rrrPUF.
[70] Y. Zhu, R. Kiros, R.S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: towards story-like visual explanations by watching movies and reading books, in: 2015 IEEE International Conference on Computer Vision, IEEE Computer Society, 2015, pp. 19–27. https://doi.org/10.1109/ICCV.2015.11.
[71] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, Adv. Neural Inf. Process. Syst. 35 (2022) 27730–27744. https://dl.acm.org/doi/10.5555/3600270.3602281.
[72] C. Peng, X. Yang, A. Chen, K.E. Smith, N.M. Pournejatian, A.B. Costa, C. Martin, M.G. Flores, Y. Zhang, T. Magoc, G.P. Lipori, D.A. Mitchell, N.S. Ospina, M.M. Ahmed, W.R. Hogan, E.A. Shenkman, Y. Guo, J. Bian, Y. Wu, A study of generative large language model for medical research and healthcare, NPJ Digital Med. 6 (2023) 1–26. https://doi.org/10.1038/s41746-023-00958-w.
[73] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.
[74] N. Deng, H. Fu, X. Chen, Named entity recognition of traditional chinese medicine patents based on biLSTM-CRF, Wireless Commun. Mobile Comput. 2021(6696205) (2021) 1–12. https://doi.org/10.1155/2021/6696205.
[75] Z. Hu, Z. Ni, J. Shi, S. Xu, B. Xu, A knowledge-enhanced two-stage generative framework for medical dialogue information extraction, Mach. Intell. Res. 21 (1) (2024) 153–168. https://doi.org/10.1007/s11633-023-1461-5.
[76] Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, L. Sun, H. Wu, Unified structure generation for universal information extraction, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2022, pp. 5755–5772. https://doi.org/10.18653/v1/2022.acl-long.395.
[77] H. Zhang, J. Chen, F. Jiang, F. Yu, Z. Chen, G. Chen, J. Li, X. Wu, Z. Zhang, Q. Xiao, X. Wan, B. Wang, H. Li, HuatuoGPT, towards taming language model tobe a doctor, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, 2023, pp. 10859–10885. https://doi.org/10.18653/v1/2023.findings-emnlp.725.
[78] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S.R. Pfohl, H. Cole-Lewis, D. Neal, Q.M. Rashid, M. Schaekermann, A. Wang, D. Dash, J.H. Chen, N.H. Shah, S. Lachgar, P.A. Mansfield, S. Prakash, B. Green, E. Dominowska, B.A.y. Arcas, N. Tomašev, Y. Liu, R. Wong, C. Semturs, S.S. Mahdavi, J.K. Barral, D.R. Webster, G.S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, V. Natarajan, Towards expert-level medical question answering with large language models, Nat. Med. 31 (2025) 943–950. https://doi.org/10.1038/s41591-024-03423-7.
[79] Y. Chen, Z. Wang, X. Xing, H. Zheng, Z. Xu, K. Fang, J. Wang, S. Li, J. Wu, Q. Liu, X. Xu, BianQue: balancing the questioning and suggestion ability of health LLMs with multi-turn health conversations polished by chatGPT, CoRR abs/2310.15896 (2023). https://doi.org/10.48550/arXiv.2310.15896.
[80] H. Xiong, S. Wang, Y. Zhu, Z. Zhao, Y. Liu, L. Huang, Q. Wang, D. Shen, DoctorGLM: fine-tuning your chinese doctor is not a herculean task, CoRR abs/2304.01097 (2023). https://doi.org/10.48550/arXiv.2304.01097.
[81] C. Wu, X. Zhang, Y. Zhang, Y. Wang, W. Xie, PMC-LLAMA: further finetuning llama on medical papers, CoRR abs/2304.14454 (2023). https://doi.org/10.48550/arXiv.2304.14454.
[82] W. Zhu, X. Wang, Chatmed: a chinese medical large language model, 2023, (https://github.com/michael-wzhu/ChatMed).
[83] W.Y. Wei Zhu, X. Wang, Shennong-TCM: A traditional chinese medicine large language model, 2023, (https://github.com/michael-wzhu/ShenNong-TCM-LLM).
[84] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H.W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A.M. Dai, T.S. Pillai, M. Pellat,

A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, PaLM: scaling language modeling with pathways, J. Mach. Learn. Res. 24 (2023) 240:1–240:113. http://jmlr.org/papers/v24/22-1144.HTML.

[85] R. Anil, A.M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, et al., PaLM 2 technical report, CoRR abs/2305.10403 (2023). https://doi.org/10.48550/arXiv.2305.10403.

[86] G. Wang, G. Yang, Z. Du, L. Fan, X. Li, ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation, CoRR abs/2306.09968 (2023). https://doi.org/10.48550/arXiv.2306.09968.

[87] J. Chen, X. Wang, A. Gao, F. Jiang, S. Chen, H. Zhang, D. Song, W. Xie, C. Kong, J. Li, X. Wan, H. Li, B. Wang, HuatuoGPT-II, one-stage training for medical adaption of LLMs, CoRR abs/2311.09774 (2023). https://doi.org/10.48550/arXiv.2311.09774.

[88] Y. Chen, Z. Wang, X. Xing, Z. Xu, K. Fang, S. Li, J. Wang, X. Xu, Bianque-1.0: improving the "question" ability of medical chat model through finetuning with hybrid instructions and multi-turn doctor QA datasets (2023). https://github.com/scutcyr/BianQue.

[89] Y. Tian, R. Gan, Y. Song, J. Zhang, Y. Zhang, Chimed-GPT: a Chinese medical large language model with full training regime and better alignment to human preferences (2024) 7156–7173. https://doi.org/10.18653/v1/2024.acl-long.386.

[90] R. Gan, Z. Wu, R. Sun, J. Lu, X. Wu, D. Zhang, K. Pan, P. Yang, Q. Yang, J. Zhang, Y. Song, Ziya2: data-centric learning is all LLMs need, CoRR abs/2311.03301 (2023). https://doi.org/10.48550/arXiv.2311.03301.

[91] Y. Luo, J. Zhang, S. Fan, K. Yang, M. Hong, Y. Wu, M. Qiao, Z. Nie, BiomedGPT: an open multimodal large language model for biomedicine, IEEE J. Biomed. Health Inform. (2024). https://doi.org/10.1109/JBHI.2024.3505955.

[92] K. Lo, L.L. Wang, M. Neumann, R. Kinney, D.S. Weld, S2ORC: the semantic scholar open research corpus, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447.

[93] K. Zhang, R. Zhou, E. Adhikarla, Z. Yan, Y. Liu, J. Yu, Z. Liu, X. Chen, B.D. Davison, H. Ren, J. Huang, C. Chen, Y. Zhou, S. Fu, W. Liu, T. Liu, X. Li, Y. Chen, L. He, J. Zou, Q. Li, H. Liu, L. Sun, A generalist vision-language foundation model for diverse biomedical tasks, Nat. Med. 30 (2024) 3129–3141. https://doi.org/10.1038/s41591-024-03185-2.

[94] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, in: Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, 2022, pp. 23318–23340. https://proceedings.mlr.press/v162/wang22al.html.

[95] D. Driess, F. Xia, M.S.M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, P. Florence, PaLM-E: an embodied multimodal language model, in: International Conference on Machine learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 8469–8488. https://proceedings.mlr.press/v202/driess23a.html.

[96] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, Llava-med: training a large language-and-vision assistant for biomedicine in one day, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, 1240, 2023, pp. 28541–28564. https://dl.acm.org/doi/10.5555/3666122.3667362.

[97] OpenAI, Gpt-4 technical report, CoRR abs/2303.08774 (2023). https://doi.org/10.48550/arXiv.2303.08774.

[98] R. Wang, Y. Duan, J. Li, P. Pang, T. Tan, XrayGLM: the first Chinese medical multimodal model that chest radiographs summarization, 2023, (https://github.com/WangRongsheng/XrayGLM).

[99] C. Shu, B. Chen, F. Liu, Z. Fu, E. Shareghi, N. Collier, Visual med-alpaca: A parameter-efficient biomedical LLM with visual capabilities, 2023, (https://github.com/cambridgeltl/visual-med-alpaca).

[100] Q. Li, C. Qiu, H. Liu, J. Gu, D. Luo, Decoupled contrastive learning for multilingual multimodal medical pre-trained model, Neurocomputing 633 (2025) 129809. https://doi.org/10.1016/j.neucom.2025.129809.

[101] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Qin, Y. Zheng, X. Qiu, X. Huang, Q. Zhang, T. Gui, The rise and potential of large language model based agents: a survey, Sci. China Inform. Sci. 68 (121101) (2025). https://doi.org/10.1007/s11432-024-4222-0.

[102] W. Zhou, Y.E. Jiang, L. Li, J. Wu, T. Wang, S. Qiu, J. Zhang, J. Chen, R. Wu, S. Wang, S. Zhu, J. Chen, W. Zhang, N. Zhang, H. Chen, P. Cui, M. Sachan, Agents: an open-source framework for autonomous language agents, CoRR abs/2309.07870 (2023). https://doi.org/10.48550/arXiv.2309.07870.

[103] C.H. Song, B.M. Sadler, J. Wu, W. Chao, C. Washington, Y. Su, LLM-planner: few-shot grounded planning for embodied agents with large language models, in: IEEE/CVF International Conference on Computer Vision, IEEE, 2023, pp. 2986–2997. https://doi.org/10.1109/ICCV51070.2023.00280.

[104] Y.M. Cho, S. Rai, L.H. Ungar, J. Sedoc, S.C. Guntuku, An "integrative survey on mental health conversational agents to bridge computer science and medical perspectives", in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023, pp. 11346–11369. https://doi.org/10.18653/v1/2023.emnlp-main.698.

[105] M.R. Ali, S.Z. Razavi, R. Langevin, A.A. Mamun, B. Kane, R. Rawassizadeh, L.K. Schubert, E. Hoque, A virtual conversational agent for teens with autism spectrum disorder: experimental results and design lessons, in: IVA '20: ACM International Conference on Intelligent Virtual Agents, 2, ACM, 2020, pp. 1–8. https://doi.org/10.1145/3383652.3423900.

[106] M. Abbasian, I. Azimi, A.M. Rahmani, R.C. Jain, Conversational health agents: a personalized LLM-powered agent framework, CoRR abs/2310.02374 (2023). https://doi.org/10.48550/arXiv.2310.02374.

[107] K. Chen, X. Li, T. Yang, H. Wang, W. Dong, Y. Gao, Mdteamgpt: a self-evolving llm-based multi-agent framework for multi-disciplinary team medical consultation, CoRR abs/2503.13856 (2025). https://doi.org/10.48550/arXiv.2503.13856.

[108] K. Lan, B. Jin, Z. Zhu, S. Chen, S. Zhang, K.Q. Zhu, M. Wu, Depression diagnosis dialogue simulation: self-improving psychiatrist with tertiary memory, CoRR abs/2409.15084 (2024). https://doi.org/10.48550/arXiv.2409.15084.

[109] Q. Jin, Y. Yang, Q. Chen, Z. Lu, GeneGPT: augmenting large language models with domain tools for improved access to biomedical information, Bioinformatics 40 (2) (2024). https://doi.org/10.1093/bioinformatics/btae075.

[110] A.J. Goodell, S.N. Chu, D. Rouholiman, L.F. Chu, Augmentation of chatGPT with clinician-informed tools improves performance on medical calculation tasks, MedRxiv (2023). https://www.medrxiv.org/content/early/2023/12/15/2023.12.13.23299881.

[111] X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan, M. Gerstein, Medagents: large language models as collaborators for zero-shot medical reasoning (2024) 599–621. https://doi.org/10.18653/v1/2024.findings-acl.33.

[112] L. Yue, S. Xing, J. Chen, T. Fu, ClinicalAgent: clinical trial multi-agent system with large language model-based reasoning, in: Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 11, ACM, 2024, pp. 1–10. https://doi.org/10.1145/3698587.3701359.

[113] B. Li, T. Yan, Y. Pan, J. Luo, R. Ji, J. Ding, Z. Xu, S. Liu, H. Dong, Z. Lin, Y. Wang, Mmedagent: learning to use medical tools with multi-modal agent, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, 2024, pp. 8745–8760. https://aclanthology.org/2024.findings-emnlp.510.

[114] S. Gao, R. Zhu, Z. Kong, A. Noori, X. Su, C. Ginder, T. Tsiligkaridis, M. Zitnik, TXagent: an AI agent for therapeutic reasoning across a universe of tools, CoRR abs/2503.10970 (2025). https://doi.org/10.48550/arXiv.2503.10970.

[115] Y. Jiang, K.C. Black, G. Geng, D. Park, A.Y. Ng, J.H. Chen, MedAgentbench: dataset for benchmarking LLMs as agents in medical applications, CoRR abs/2501.14654 (2025). https://doi.org/10.48550/arXiv.2501.14654.

[116] F. Xu, P. Cheng, F. Gao, Y. Jin, S. Yan, Q. Huang, Y. Wang, X. Ren, J. Gu, An agent-based adaptive medical dialogue service for personalized healthcare, Intern. J. Web Serv. Res. (IJWSR) 22 (1) (2025) 1–28. http://dx.doi.org/10.4018/IJWSR.371758.

[117] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: a roadmap, IEEE Trans. Knowl. Data Eng. 36 (7) (2024) 3580–3599. https://doi.org/10.1109/TKDE.2024.3352100.

[118] A. Mavridis, S. Tegos, C. Anastasiou, M. Papoutsoglou, G. Meditskos, Large language models for intelligent RDF knowledge graph construction: results from medical ontology mapping, Front. Artif. Intell. 8 (2025). https://doi.org/10.3389/frai.2025.1546179.

[119] H.B. Giglou, J. D'Souza, F. Engel, S. Auer, LLMs4Om: matching ontologies with large language models, in: The Semantic Web: ESWC 2024 Satellite Events, volume 15344 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 25–35. https://doi.org/10.1007/978-3-031-78952-6_3.

[120] X. Yang, N. PourNejatian, H.C. Shin, K.E. Smith, C. Parisien, C. Compas, C. Martin, M.G. Flores, Y. Zhang, T. Magoc, C.A. Harle, G. Lipori, D.A. Mitchell, W.R. Hogan, E.A. Shenkman, J. Bian, Y. Wu, GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records, medRxiv (2022). https://www.medrxiv.org/content/early/2022/02/28/2022.02.27.22271257.

[121] J. Zhang, Graph-toolformer: to empower LLMs with graph reasoning ability via prompt augmented by chatGPT, CoRR abs/2304.11116 (2023). https://doi.org/10.48550/arXiv.2304.11116.

[122] C. Feng, X. Zhang, Z. Fei, Knowledge solver: teaching LLMs to search for domain knowledge from knowledge graphs, CoRR abs/2309.03118 (2023). https://doi.org/10.48550/arXiv.2309.03118.

[123] V. Arsenyan, S. Bughdaryan, F. Shaya, K.W. Small, D. Shahnazaryan, Large language models for biomedical knowledge graph construction: information extraction from EMR notes, in: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, 2024, pp. 295–317. https://aclanthology.org/2024.bionlp-1.23.

[124] K. Zuo, Y. Jiang, F. Mo, P. Lio, Kg4diagnosis: a hierarchical multi-agent LLM framework with knowledge graph enhancement for medical diagnosis, CoRR abs/2412.16833 (2024). https://doi.org/10.48550/arXiv.2412.16833.

[125] B. Mo, K. Yu, J. Kazdan, P. Mpala, L. Yu, C. Cundy, C.I. Kanatsoulis, S. Koyejo, KGGeN: extracting knowledge graphs from plain text with language models, CoRR abs/2502.09956 (2025). https://doi.org/10.48550/arXiv.2502.09956.

[126] H. Chen, X. Shen, Q. Lv, J. Wang, X. Ni, J. Ye, SAC-KG: exploiting large language models as skilled automatic constructors for domain knowledge graph, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2024, pp. 4345–4360. https://doi.org/10.18653/v1/2024.acl-long.238.

[127] J. Wu, J. Zhu, Y. Qi, Medical graph RAG: towards safe medical large language model via graph retrieval-augmented generation, CoRR abs/2408.04187 (2024). https://doi.org/10.48550/arXiv.2408.04187.

[128] P. Jiang, C. Xiao, A. Cross, J. Sun, Graphcare: enhancing healthcare predictions with personalized knowledge graphs, in: The Twelfth International Conference on Learning Representations, OpenReview.net, 2024. https://openreview.net/forum?id=tVTN7Zs0ml.

[129] Z. Bao, W. Chen, S. Xiao, K. Ren, J. Wu, C. Zhong, J. Peng, X. Huang, Z. Wei, DiscMedLLM: bridging general large language models and real-world medical consultation, CoRR abs/2308.14346 (2023). https://doi.org/10.48550/arXiv.2308.14346.
[130] Y. Gao, R. Li, E. Croxford, J. Caskey, B.W. Patterson, M. Churpek, T. Miller, D. Dligach, M. Afshar, Leveraging medical knowledge graphs into large language models for diagnosis prediction: design and application study, JMIR AI 4 (2025). https://doi.org/10.2196/58670.
[131] P. Jiang, C. Xiao, M. Jiang, P. Bhatia, T.A. Kass-Hout, J. Sun, J. Han, Reasoning-enhanced healthcare predictions with knowledge graph community retrieval, CoRR abs/2410.04585 (2024). https://doi.org/10.48550/arXiv.2410.04585.
[132] J. Wu, W. Deng, X. Li, S. Liu, T. Mi, Y. Peng, Z. Xu, Y. Liu, H. Cho, C.-I. Choi, Y. Cao, H. Ren, X. Li, X. Li, Y. Zhou, Medreason: eliciting factual medical reasoning steps in LLMs via knowledge graphs, CoRR abs/2504.00993 (2025). https://doi.org/10.48550/arXiv.2504.00993.
[133] H. Chen, X. Shen, J. Wang, Z. Wang, Q. Lv, J. He, R. Wu, F. Wu, J. Ye, Knowledge graph finetuning enhances knowledge manipulation in large language models, in: The Thirteenth International Conference on Learning Representations, OpenReview.net, 2025. https://openreview.net/forum?id=oMFOKjwaRS.
[134] Z. Heyi, W. Xin, H. Lifan, L. Zhao, C. Zirui, C. Zhe, Research on question answering system on joint of knowledge graph and large language models, J. Front. Comput. Sci. Technol. 17 (10) (2023). http://fcst.ceaj.org/EN/10.3778/j.issn.1673-9418.2308070.
[135] K. Soman, P.W. Rose, J.H. Morris, R.E. Akbas, B. Smith, B. Peetoom, C. Villouta-Reyes, G. Cerono, Y. Shi, A. Rizk-Jackson, S. Israni, C.A. Nelson, S. Huang, S.E. Baranzini, Biomedical knowledge graph-optimized prompt generation for large language models, Bioinformatics 40 (2024). https://doi.org/10.1093/bioinformatics/btae560.
[136] M.R. Rezaei, R.S. Fard, J. Parker, R.G. Krishnan, M. Lankarany, Adaptive knowledge graphs enhance medical question answering: bridging the gap between LLMs and evolving medical knowledge, CoRR abs/2502.13010 (2025). https://doi.org/10.48550/arXiv.2502.13010.
[137] R. Xu, H. Cui, Y. Yu, X. Kan, W. Shi, Y. Zhuang, M.D. Wang, W. Jin, J.C. Ho, C. Yang, Knowledge-infused prompting: assessing and advancing clinical text data generation with large language models, in: Findings of the Association for Computational Linguistics, Association for Computational Linguistics, 2024, pp. 15496–15523. https://doi.org/10.18653/v1/2024.findings-acl.916.
[138] D. Li, S. Yang, Z. Tan, J.Y. Baik, S. Yun, J. Lee, A. Chacko, B. Hou, D. Duong-Tran, Y. Ding, H. Liu, L. Shen, T. Chen, DALK: dynamic co-augmentation of LMs and KG to answer Alzheimer's disease questions with scientific literature, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 2187–2205. https://aclanthology.org/2024.findings-emnlp.119/.
[139] X. Jiang, R. Zhang, Y. Xu, R. Qiu, Y. Fang, Z. Wang, J. Tang, H. Ding, X. Chu, J. Zhao, Y. Wang, Hykge: a hypothesis knowledge graph enhanced RAG framework for accurate and reliable medical LMs responses, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2025, pp. 11836–11856. https://aclanthology.org/2025.acl-long.580/.
[140] Z. Zhang, Z. Lin, Y. Zheng, X. Wu, How much medical knowledge do LLMs have? An evaluation of medical knowledge coverage for LLMs, in: Proceedings of the ACM on Web Conference 2025, ACM, 2025, pp. 5330–5341. https://doi.org/10.1145/3696410.3714535.
[141] W. Liu, Y. Zuo, Stone needle: a general multimodal large-scale model framework towards healthcare, CoRR abs/2306.16034 (2023). https://doi.org/10.48550/arXiv.2306.16034.
[142] X. Lin, C. Xu, Z. Xiong, X. Zhang, N. Ni, B. Ni, J. Chang, R. Pan, Z. Wang, F. Yu, Q. Tian, H. Jiang, M. Zheng, N. Qiao, Pangu drug model: learn a molecule like a human, Sci. China Life Sci. 66 (2023). https://doi.org/10.1007/s11427-022-2239-y.
[143] K. Papineni, S. Roukos, T. Ward, W. Zhu, BLEU: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL, 2002, pp. 311–318. https://aclanthology.org/P02-1040/.
[144] C.-Y. Lin, ROUGE: a package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, 2004, pp. 74–81. https://aclanthology.org/W04-1013/.
[145] S. Banerjee, A. Lavie, METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, 2005, pp. 65–72. https://aclanthology.org/W05-0909/.
[146] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, Bertscore: evaluating text generation with BERT, in: 8th International Conference on Learning Representations, OpenReview.net, 2020. https://openreview.net/forum?id=SkeHuCVFDr.
[147] W.-w. Yim, Y. Fu, A.B. Abacha, N. Snider, T. Lin, M. Yetisgen, ACI-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation, Sci. Data 10 (2023) 586. https://doi.org/10.1038/s41597-023-02487-3.
[148] W. Yue, X. Wang, W. Zhu, M. Guan, H. Zheng, P. Wang, C. Sun, X. Ma, TCMBench: a comprehensive benchmark for evaluating large language models in traditional chinese medicine, CoRR abs/2406.01126 (2024). https://doi.org/10.48550/arXiv.2406.01126.
[149] X. Wang, G. Chen, D. Song, Z. Zhang, Z. Chen, Q. Xiao, J. Chen, F. Jiang, J. Li, X. Wan, B. Wang, H. Li, CMB: a comprehensive medical benchmark in Chinese, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2024, pp. 6184–6205. https://doi.org/10.18653/v1/2024.naacl-long.343.
[150] W. Zhu, X. Wang, H. Zheng, M. Chen, B. Tang, PromptCBLUE: a Chinese prompt tuning benchmark for the medical domain, CoRR abs/2310.14151 (2023). https://doi.org/10.48550/arXiv.2310.14151.
[151] S. Lyu, C. Chi, H. Cai, L. Shi, X. Yang, L. Liu, X. Chen, D. Zhao, Z. Zhang, X. Lyu, M. Zhang, F. Li, X. Ma, Y. Shen, J. Gu, W. Xue, Y. Huang, RJUA-QA: a comprehensive qa dataset for urology, CoRR abs/2312.09785 (2023). https://doi.org/10.48550/arXiv.2312.09785.
[152] S. Kweon, J. Kim, H. Kwak, D. Cha, H. Yoon, K. Kim, J. Yang, S. Won, E. Choi, EHRNoteQA: an LLM benchmark for real-world clinical practice using discharge summaries, in: Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems, 3958, 2024, pp. 124575–124611. https://dl.acm.org/doi/10.5555/3737916.3741874.
[153] J.C.L. Ong, S.Y.-H. Chang, W. William, A.J. Butte, N.H. Shah, L.S.T. Chew, N. Liu, F. Doshi-Velez, W. Lu, J. Savulescu, D.S.W. Ting, Medical ethics of large language models in medicine, NEJM AI 1 (7) (2024) AIra2400038. https://ai.nejm.org/doi/abs/10.1056/AIra2400038.
[154] P. Yang, H. Wang, Y. Huang, S. Yang, Y. Zhang, L. Huang, Y. Zhang, G. Wang, S. Yang, L. He, Y. Huang, LMKG: a large-scale and multi-source medical knowledge graph for intelligent medicine applications, Knowl. Based Syst. 284 (2024) 111323. https://doi.org/10.1016/j.knosys.2023.111323.
[155] P. Chandak, K. Huang, M. Zitnik, Building a knowledge graph to enable precision medicine, Sci. Data 10 (67) (2023). https://doi.org/10.1038/s41597-023-01960-3.
[156] P. Xia, K. Zhu, H. Li, W. Shi, S. Wang, L. Zhang, J. Zou, H. Yao, Mmed-RAG: versatile multimodal Rag system for medical vision language models, in: The Thirteenth International Conference on Learning Representations, OpenReview.net, 2025. https://openreview.net/forum?id=s5epFPdIW6.
[157] S. Reddy, Evaluating large language models for use in healthcare: a framework for translational value assessment, Inform. Med. Unlocked 41 (2023) 101304. https://doi.org/10.1016/j.imu.2023.101304.
[158] J. Maharjan, A. Garikipati, N.P. Singh, L. Cyrus, M. Sharma, M. Ciobanu, G. Barnes, R. Thapa, Q. Mao, R. Das, OpenmedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models, Sci. Rep. 14 (14156) (2024). https://doi.org/10.1038/s41598-024-64827-6.
[159] T. Savage, A. Nayak, R. Gallo, E. Rangan, J.H. Chen, Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine, NPJ Digital Med. 7 (1) (2024). https://doi.org/10.1038/s41746-024-01010-1.
[160] L. Tang, Z. Sun, B.R.S. Idnay, J.G. Nestor, A. Soroush, P.A. Elias, Z. Xu, Y. Ding, G. Durrett, J.F. Rousseau, C. Weng, Y. Peng, Evaluating large language models on medical evidence summarization, NPJ Digital Med. 6 (2023). https://doi.org/10.1038/s41746-023-00896-7.
[161] B. Wang, H. Zhao, H. Zhou, L. Song, M. Xu, W. Cheng, X. Zeng, Y. Zhang, Y. Huo, Z. Wang, Z. Zhao, D. Pan, F. Yang, F. Kou, F. Li, F. Chen, G. Dong, H. Liu, H. Zhang, J. He, J. Yang, K. Wu, K. Wu, L. Su, L. Niu, L. Sun, M. Wang, P. Fan, Q. Shen, R. Xin, S. Dang, S. Zhou, W. Chen, W. Luo, X. Chen, X. Men, X. Lin, X. Dong, Y. Zhang, Y. Duan, Y. Zhou, Z. Ma, Z. Wu, Baichuan-ML: pushing the medical capability of large language models, CoRR abs/2502.12671 (2025). https://doi.org/10.48550/arXiv.2502.12671.
[162] Z. Liu, Q. Tu, W. Ye, Y. Xiao, Z. Zhang, H. Cui, Y. Zhu, Q. Ju, S. Li, J. Xie, Exploring the inquiry-diagnosis relationship with advanced patient simulators, CoRR abs/2501.09484 (2025). https://doi.org/10.48550/arXiv.2501.09484.
[163] H. Yu, J. Zhou, L. Li, S. Chen, J. Gallifant, A. Shi, X. Li, W. Hua, M. Jin, G. Chen, Y. Zhou, Z. Li, T. Gupte, M. Chen, Z. Azizi, Y. Zhang, T.L. Assimes, X. Ma, D.S. Bitterman, L. Lu, L. Fan, AIpatient: simulating patients with ehrs and LLM powered agentic workflow, CoRR abs/2409.18924 (2024). https://doi.org/10.48550/arXiv.2409.18924.
[164] L.Y. Jiang, X.C. Liu, N.P. Nejatian, M. Nasir-Moin, D. Wang, A. Abidin, K. Eaton, H.A. Riina, I. Laufer, P. Punjabi, M. Miceli, N.C. Kim, C. Orillac, Z. Schnurman, C. Livia, H. Weiss, D. Kurland, S. Neifert, Y. Dastagirzada, D. Kondziolka, A.T.M. Cheung, G. Yang, M. Cao, M. Flores, A.B. Costa, Y. Aphinyanaphongs, K. Cho, E.K. Oermann, Health system-scale language models are all-purpose prediction engines, Nature (619) (2023) 357–362. https://doi.org/10.1038/s41586-023-06160-y.
[165] Z. Kraljevic, D. Bean, A. Shek, R. Bendayan, H. Hemingway, J.A. Yeung, A. Deng, A. Baston, J. Ross, E. Idowu, J.T. Teo, R.J.B. Dobson, Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study, Lancet Digital Health 6 (4) (2024) e281–e290. https://doi.org/10.1016/S2589-7500(24)00025-6.
[166] X. Wang, N. Chen, J. Chen, Y. Hu, Y. Wang, X. Wu, A. Gao, X. Wan, H. Li, B. Wang, Apollo: an lightweight multilingual medical LLM towards democratizing medical AI to 6b people, CoRR abs/2403.03640 (2024). https://doi.org/10.48550/arXiv.2403.03640.
[167] S. Pieri, S.S. Mullappilly, F.S. Khan, R.M. Anwer, S.H. Khan, T. Baldwin, H. Cholakkal, Bimedix: bilingual medical mixture of experts LLM, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, 2024, pp. 16984–17002. https://aclanthology.org/2024.findings-emnlp.989.