Zhang *et al*. *Intell Robot* 2024;4(4):457-78
DOI: 10.20517/ir.2024.27

Intelligence & Robotics

Review

Open Access

A survey of datasets in medicine for large language models

Deshiwei Zhang[1], Xiaojuan Xue[2], Peng Gao[3], Zhijuan Jin[4], Menghan Hu[2], Yue Wu[5], Xiayang Ying[6]

[1]School of Civil Engineering, Southeast University, Nanjing 210096, Jiangsu, China.
[2]Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Nanjing, Shanghai 200241, China.
[3]Department of Ophthalmology, Shanghai Tenth People's Hospital of Tongji University, Tongji University School of Medicine, Shanghai 200072, China.
[4]Department of Developmental and Behavioral Pediatrics, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China.
[5]Department of Ophthalmology, Ninth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200011, China.
[6]Department of General Surgery, Pancreatic Disease Center, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200001, China.

Correspondence to: Dr. Xiayang Ying, Department of General Surgery, Pancreatic Disease Center, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, No. 197 Ruijin 2nd Road, Huangpu District, Shanghai 200001, China. E-mail: yingxiayang@hotmail.com; Dr. Menghan Hu, Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, 500 Dongchuan Rd., Shanghai 200241, China. E-mail: mhhu@ce.ecnu.edu.cn

How to cite this article: Zhang D, Xue X, Gao P, Jin Z, Hu M, Wu Y ,Ying X. A survey of datasets in medicine for large language models. *Intell Robot* 2024;4(4):457-78. http://dx.doi.org/10.20517/ir.2024.27

Received: 22 Apr 2024 **First Decision:** 27 Sep 2024 **Revised:** 25 Nov 2024 **Accepted:** 28 Nov 2024 **Published:** 25 Dec 2024

Academic Editor: Simon Yang **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

With the advent of models such as ChatGPT and other models, large language models (LLMs) have demonstrated unprecedented capabilities in understanding and generating natural language, presenting novel opportunities and challenges within the medicine domain. While there have been many studies focusing on the employment of LLMs in medicine, comprehensive reviews of the datasets utilized in this field remain scarce. This survey seeks to address this gap by providing a comprehensive overview of the datasets in medicine fueling LLMs, highlighting their unique characteristics and the critical roles they play at different stages of LLMs' development: pre-training, fine-tuning, and evaluation. Ultimately, this survey aims to underline the significance of datasets in realizing the full potential of LLMs to innovate and improve healthcare outcomes.

Keywords: Large language models (LLMs), NLP, dataset in medicine, Q&A system in medicine

1. INTRODUCTION

Medicine stands as a critical field intricately connected to human well-being, where the integration of advanced technologies such as large language models (LLMs) has shown promising potential [1]. Since the introduction of ChatGPT [2], numerous studies have leveraged such models for various medical applications, demonstrating their adeptness at tasks ranging from biological information extraction [3], medical advice consultation, mental health-related applications and clinical report generation. Furthermore, LLMs have demonstrated their potential to improve patient care [4,5]. The utilization of LLMs in medicine is often facilitated by crafting specialized prompts or instructions, enabling these models to navigate the complexities of medical data effectively.

Existing LLMs can be classified into three types: encoder-only, encoder-decoder, and decoder-only LLMs. Encoder-only LLMs (e.g., BERT [6]) are generally used for tasks that involve understanding text, such as classification and sentiment analysis. Encoder-decoder LLMs (e.g., ChatGLM [7]) are useful for tasks that involve both understanding and generating text, such as summarization. Decoder-only LLMs (e.g., GPT-4 [8]) excel at generative tasks such as sentence completion and open-ended generation. LLMs in medicine are developed through a two-stage process: pre-training and fine-tuning. To pre-train LLMs, two common tasks are employed: language modeling and denoising autoencoding. Language modeling involves predicting the next word in a sequence, helping the model to learn language patterns and semantic relationships effectively [9]. Denoising autoencoding, on the other hand, requires the model to recover the replaced parts of the text, which aids in understanding and generating language outputs precisely [10,11]. The pre-training phase involves training a language model on a large corpus of structured and unstructured text data. For LLMs in medicine, the corpus may include electronic health records (EHR) [12], clinical notes [13], and medical literature [14]. Pre-training lays the foundation for the LLMs, enabling them to grasp the broad nuances of language and acquire generation skills [9,15], preparing LLMs for more specialized tasks in subsequent training stages. It is important to note that some LLMs are pre-trained on general data and fine-tuned on medical data, while others are trained on medical datasets from scratch. For instance, models such as PubMedBERT [16] are specifically pre-trained on biomedical corpora, leading to improved performance in healthcare-specific tasks compared to models that are fine-tuned on medical data after general pre-training.

Having established a solid foundation through pre-training, the fine-tuning phase focuses on domain-specific adaptation. This stage involves diverse medical corpora, such as dialogue datasets, question-answer (QA) pairs, and instructional texts, ensuring the model excels in specialized tasks [17]. Researchers have proposed some fine-tuning methods [18–20] to develop effective medical LLMs. This phase ensures that the model becomes proficient in handling healthcare-specific language, thereby enhancing its accuracy and efficiency. The choice of datasets for training LLMs in medicine depends on the specific task and the type of data required. Common sources of data include EHR, scientific literature, web data, and public knowledge bases. These datasets provide valuable information for training LLMs and enable them to understand and generate medical text, making them versatile tools capable of providing accurate clinical decision support [21].

Lastly, the evaluation of LLMs employs datasets as benchmarks to rigorously assess their performance, such as text classification [22], semantic understanding [23], question answering (QA) [24], and trustworthiness [25]. The application of LLMs in the medical field has gained significant attention in recent years. Medical LLMs have been evaluated and utilized in a range of medical applications, including medical queries [26], medical examinations [27], and medical assistants [28]. By evaluating the model performance on different datasets, researchers can identify areas where the model excels and areas where it needs improvement [29]. This feedback helps refine the model architecture and train methods. In Figure 1, we describe the construction process of the medical Q&A LLMs.

While Q&A systems are a key application of LLMs in healthcare, they represent only a small part of their broader potential. Our survey provides an overview of datasets used for both pre-training and fine-tuning

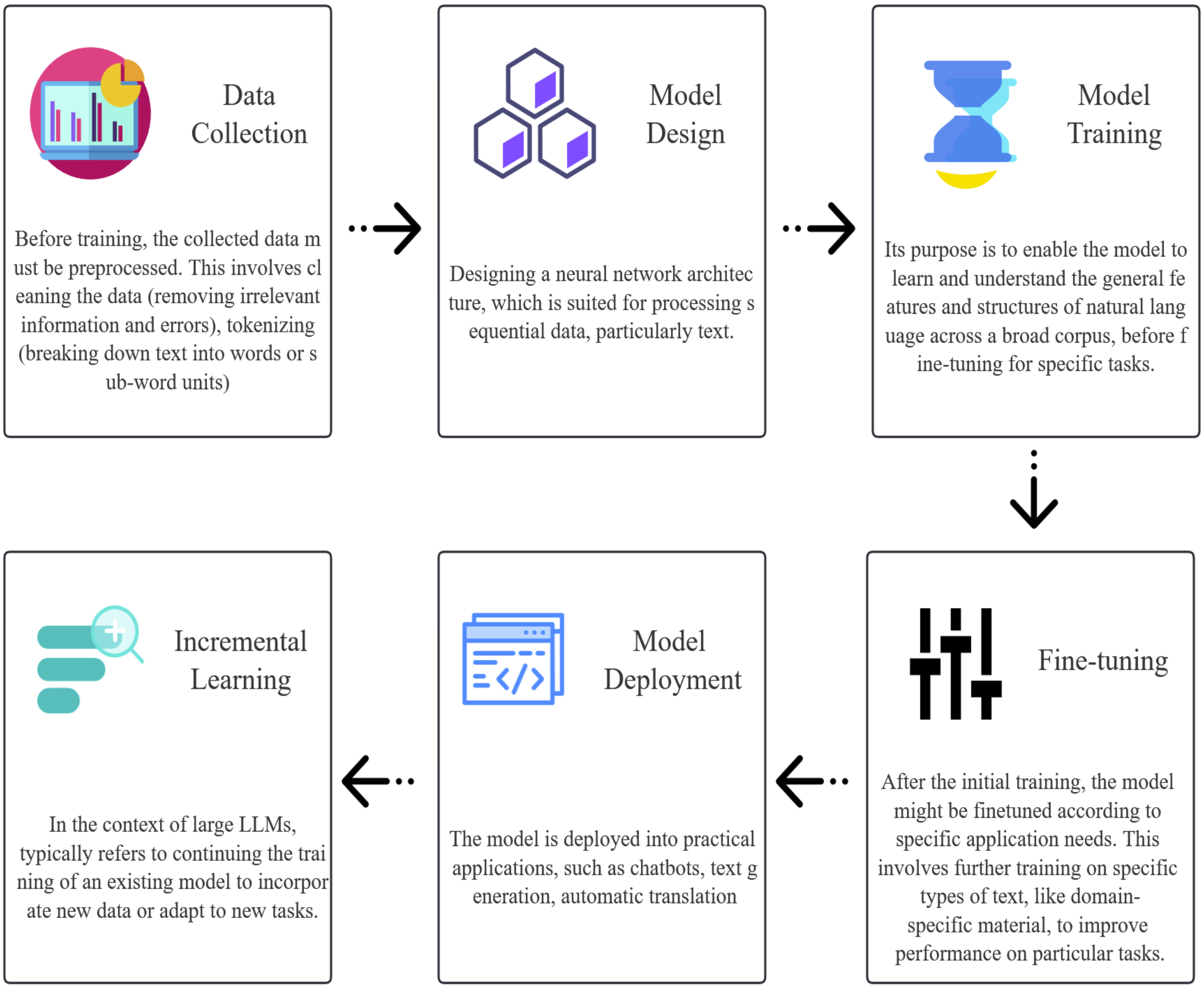


Figure 1. Construction process of the medical Q&A LLMs.

LLMs across a variety of medical tasks. We focus on offering concise summaries and links to these datasets, which can support LLM development across diverse medical applications. Compared with the recent work by Wu *et al*., which focuses on the accessibility and characteristics of publicly available clinical text datasets, our survey encompasses a broader scope by including multimodal datasets[30]. This provides valuable resources for researchers developing LLMs for a wider range of tasks.

To further emphasize the significance of these datasets, it is crucial to recognize that comprehensive overviews of the available data are still scarce. Figure 2 illustrates a timeline of dataset development, spanning dialogue, QA, EHR, summarization, and multimodal data, which have collectively driven the advancement of medical LLMs. By offering a detailed analysis of these datasets and their applications, our survey aims to address this gap in the literature. We hope that this work not only underscores the critical role these datasets play in advancing LLM technology but also encourages further research to unlock the full potential of LLMs in medicine.

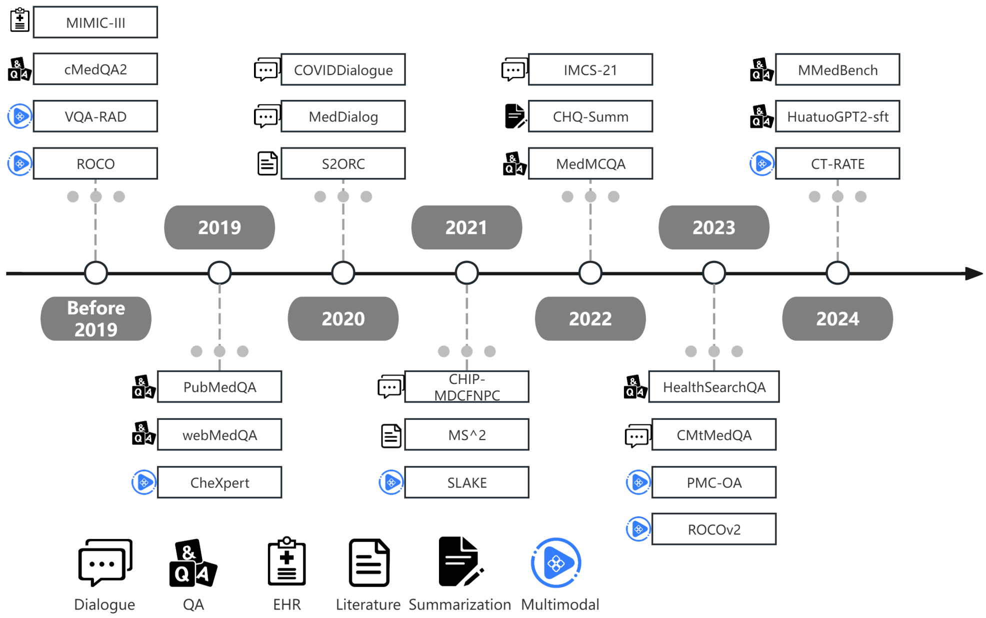


Figure 2. Timeline diagram of the development of medical datasets.

2. OVERVIEW OF DATASETS

2.1. Collection

Datasets for medical LLMs originate from diverse sources. Open-source platforms such as Hugging Face and GitHub provide immediate access to pre-curated datasets, while literature reviews and Google searches uncover additional resources. Together, these strategies ensure a broad and comprehensive collection covering a wide range of medical applications. A summary of these datasets is presented in Table 1.

2.2. Sources

In the realm of medical data analysis, especially for training LLMs tailored for healthcare applications, the diversity and specificity of the datasets are paramount. Unlike earlier pretrained language models (PLMs), contemporary LLMs, with their vast array of parameters, necessitate extensive training data encompassing a comprehensive spectrum of medical knowledge. To meet this requirement, a variety of specialized medical datasets have become increasingly available for research purposes. We categorize these corpora into four groups based on their sources: EHR, Scientific Literature, Web Data, and Public Knowledge Bases.

2.2.1 EHR

EHRs contain comprehensive information about patients' medical history, diagnoses, treatments, medication and allergies. They are widely used in medical research and analysis. In this category, MIMIC-III [31], MIMIC-IV [32] and CPRD [33] are three commonly used datasets for LLM fine-tuning.

MIMIC-III [31] is an openly available dataset featuring de-identified health data from over 40,000 patients who were admitted to the intensive care units at Beth Israel Deaconess Medical Center from 2001 to 2012. This extensive dataset includes records from 58,976 hospital admissions across 38,597 patients, positioning it as a crucial resource for in-depth healthcare research. It is renowned for its substantial inclusion of 2,083,180 de-identified notes, filled with detailed patient histories and clinician observations. MIMIC-III provides a wide array of data, encompassing patient demographics, hourly vital sign measurements, lab test results, medical procedures, medication records, caregiver notes, imaging reports, and mortality data, both during hospital

Table 1. An overview of commonly used datasets in medicine for LLMs

Dataset	Type	Language	Scale	Highlight
MIMIC-III [31]	EHR	English	58 K hospital admissions	Comprising over 58,000 hospital admissions for 46,520 patients (38,645 adults and 7,875 neonates)
MIMIC-IV [32]	EHR	English	504 K admissions	Covering a decade of admissions between 2008 and 2019 and establishing a modular organization
CPRD [33]	EHR	English	2 K primary care practices	Containing over 2,000 primary care practices and including 60 million patients
PubMed [34]	Scientific literature	English	36 M citations	Comprising over 36 M citations and abstracts of biomedical literature
PMC [35]	Scientific literature	English	8 M articles	Consisting of over 8 million available full-text article records
RCT [36]	Scientific literature	English	4 K abstracts	Comprising 4,528 systematic reviews composed by members of the Cochrane collaboration
MSˆ2 [37]	Scientific literature	English	470 K abstracts	The first large-scale, publicly available multi-document summarization dataset of over 470 K documents and 20 K summaries
CDSR [38]	Scientific literature	English	6 K abstracts	Containing a training set of 5,195 abstract pairs, a validation set of 500 abstract pairs, and a test set of 1,000 abstract pairs
SumPubMed [39]	Scientific literature	English	33 K abstracts	Consisting of 33,772 abstracts from PubMed biomedical research paper
The Pile [40]	Scientific literature	English	825 GB text	Containing 825 GB of data from multiple sources, including books, websites, scientific papers, and social media platforms
S2ORC [41]	Scientific literature	English	81.1 M papers	Containing 81.1 M English-language academic papers covering many academic disciplines
CORD-19 [42]	Scientific literature	English	1 M papers	Containing over 1 M papers on COVID-19 and related historical coronavirus research
COMETA [43]	Web data	English	20 K entities	Containing 20,015 English biomedical concept mentions from Reddit annotated with links to SNOMED CT
WebText [44]	Web data	English	40 GB text	Containing highly upvoted links from Reddit
OpenWebText [45]	Web data	English	38 GB text	An accessible open-source alternative to WebText
C4 [11]	Web data	English	750 GB text	Containing about 750 GB clean English text scraped from the web
UMLS [46]	Knowledge base	English	2 M entities	Containing over 2 million names for 900 K concepts from over 60 families of biomedical vocabularies
cMeKG [47]	Knowledge base	Chinese	10 K diseases	Covering more than 10,000 diseases and nearly 20,000 drugs
DrugBank [48]	Knowledge base	English	16 K drug entries	Containing 16,581 drug entries with 5,293 non-redundant protein sequences linked to these drug entries
cMedQA2 [49]	QA	Chinese	108 K QA pairs	Containing 108,000 questions and 203,569 answers
webMedQA [50]	QA	Chinese	63 K QA pairs	Consisting of 63,284 questions, covering most of the clinical departments of common diseases and health problems
Huatuo-26M [51]	QA	Chinese	26 M QA pairs	Containing over 26 million QA pairs, covering diseases, symptoms, treatment methods, and drug information
ChatMed-Dataset [52]	QA	Chinese	110 K QA pairs	Comprising 110,113 medical query-response pairs generated by OpenAI's GPT-3.5 engine
PubMedQA [53]	QA	English	273 K QA pairs	Containing 1 K expert-annotated, 61.2 K unlabeled and 211.3 K artificially generated QA instances
CMCQA [54]	QA	Chinese	1.3 M QA pairs	Consisting of 1.3 million QA pairs, covering 45 departments, such as andrology, stomatology, gynecology and obstetrics
Medical Flashcards [19]	QA	English	33 K instances	Containing 33,955 instances, covering subjects such as anatomy, physiology, pathology and pharmacology
Wikidoc [19]	QA	English	10 K instances	Comprising 10,000 instances from a collaborative platform with up-to-date medical knowledge
WPI. [19]	QA	English	5 K instances	Consisting of 5,942 QA pairs generated from paragraph headings and associated text content
medical [55]	QA	Chinese	2.4 M QA pairs	Consisting of 2.4 million QA pairs, including pre-training, instruction fine-tuning and reward datasets
HuatuoGPT-sft [56]	QA	Chinese	220 K QA pairs	Containing 220,000 QA pairs including distilled data from ChatGPT and real-world data from Doctors
HuatuoGPT2-sft [57]	QA	Chinese	50 K QA pairs	Containing 50,000 QA pairs from diverse sources including encyclopedias, books, academic literature, and web content
HEQ. [51]	QA	Chinese	364 K QA pairs	Consisting of 364,420 QA pairs sourced from medical encyclopedias and medical articles
CMD. [58]	QA	Chinese	792 K instances	Containing 792,099 instances, covering andrology, internal medicine, obstetrics and gynecology, oncology, pediatrics and surgery
HealthSearchQA [59]	QA	English	3 K questions	Consisting of 3,173 commonly searched consumer medical questions
MedDialog-EN [60]	Dialogue	English	260 K dialogues	Comprising 260 K conversations between patients and doctors, covering 96 specialties
MedDialog-CN [60]	Dialogue	Chinese	1.1 M dialogues	Comprising 1.1 million conversations between patients and doctors, covering 172 specialties
IMCS-21 [61]	Dialogue	Chinese	4 K annotated samples	Containing a total of 4,116 annotated samples with 164,731 utterances, covering 10 pediatric diseases
CovidDialog [62]	Dialogue	English&Chinese	1 K consultations	Containing an English dataset containing 603 consultations and a Chinese dataset containing 1,393 consultations
MMMLU [19]	Dialogue	English	3 K instances	Consisting of 3,787 instances from measuring massive multitask language understanding
Pubmed Causal [63]	Dialogue	English	2 K instances	Consisting of 2,446 annotated sentences
HealthCareMagic-100k [64]	Dialogue	English	100 K conversations	Containing 100 K patient-doctor conversations from an online medical consultation website
iCliniq [64]	Dialogue	English	10 K conversations	Comprising 10,000 patient-doctor conversations from a separate source
GenMedGPT-5k [64]	Dialogue	English	5 K conversations	Consisting of 5 K generated conversations between patients and doctors from ChatGPT
MedDG [65]	Dialogue	Chinese	17 K dialogues	Containing 17,864 dialogues sourced from the gastroenterology department of a Chinese medical consultation website
CHIP-MDCFNPC [66]	Dialogue	Chinese	8 K dialogues	Consisting of 8,000 annotated dialogues
DISC-Med-SFT [67]	Dialogue	Chinese	470 K dialogues	Comprising over 470 K distinct examples from existing medical datasets
CMtMedQA [68]	Dialogue	Chinese	70 K dialogues	Containing 70,000 real instances from 14 medical departments, including many proactive doctor inquiries
Alpaca-EN-AN [69]	Instructions	English	52 K instructions	Containing 52 K instruction-following data based on the self-instruct method
Alpaca-CH-AN [70]	Instructions	Chinese	52 K instructions	Containing 52 K instruction-following data with Alpaca prompts translated into Chinese by ChatGPT
sft-20k [71]	Instructions	Chinese	20 K instructions	Containing 20 K instructions and applying specific question templates to semi-structured data
ShenNong-TCM [72]	Instructions	Chinese	110 K instructions	Consisting of over 110,000 pieces of instructional data
MeQSum [73]	Summarization	English	1 K instances	Comprising 1 K consumer health questions and their summaries based on this definition
CHQ-Summ [74]	Summarization	English	1 K instances	Containing 1,507 question-summary pairs with annotations about question focus and question type
MEDIQA-AnS [75]	Summarization	English	156 instances	Containing 156 consumer health questions, corresponding answers, and expert-created summaries
VQA-RAD [76]	Multimodal	English	3 K QA pairs	Consisting of 3,515 visual questions of 11 types and 315 corresponding radiological images
SLAKE [77]	Multimodal	English&Chinese	14 K QA pairs	Containing 642 images with 14,028 QA pairs and 5,232 medical knowledge triplets
PathVQA [78]	Multimodal	English	32 K QA pairs	Consisting of 4,998 pathology images and 32,799 QA pairs
U-Xray [79]	Multimodal	English	3 K reports and 7 K images	Containing 3,955 reports and 7,470 DICOM images
ROCO [80]	Multimodal	English	81 K image-caption pairs	Consisting of 81,000 radiology images and corresponding captions
ROCOv2 [80]	Multimodal	English	79 K image-caption pairs	An updated version of ROCO, including 35,705 new images and manually curated medical concepts
MedICaT [81]	Multimodal	English	217 K images	Containing 217,060 images including subcaption and subfigure annotations
PMC-OA [82]	Multimodal	English	1.6 M image-caption pairs	Consisting of 1.6 M image-caption pairs from PMC's OpenAccess subset, covering diverse modalities or diseases
CheXpert [83]	Multimodal	English	224 K radiographs	Comprising 224,316 chest radiographs of 65,240 patients with associated reports
PadChest [84]	Multimodal	English	160 K images with related text	Consisting of over 160,000 images from 67,000 patients with related text
MIMIC-CXR	Multimodal	English	227 K imaging studies	Containing 227,835 imaging studies with 377,110 images for 64,588 patients
PMC-15M [85]	Multimodal	English	15 M Figure-caption pairs	Consisting of 15 million Figure-caption pairs from 4.4 million scientific articles
CT-RATE [86]	Multimodal	English	5 K chest CT volumes	Comprising 50,188 non-contrast 3D chest CT volumes from 25,692 patients along with corresponding radiology text reports
OpenPath [87]	Multimodal	English	208 K pathology images	Comprising 208,414 pathology images with related descriptions

stays and post-discharge.

Evolving from MIMIC-III, MIMIC-IV [32] is a relational database detailing actual patient admissions at a tertiary academic medical center in Boston, MA, USA. This updated version includes modernized data spanning from 2008 to 2019, capturing a wide array of medical information such as laboratory measurements, administered medications, and recorded vital signs. Designed to support a broad spectrum of healthcare research, MIMIC-IV is instrumental for investigations in clinical decision-making, patient care optimization, and epidemiological studies, providing a rich foundation for advancing medical knowledge and improving healthcare outcomes.

In contrast to the U.S.-based datasets, CPRD[33] contains coded and anonymized EHR data from a network of over 2000 practices in the UK. This dataset is linked to secondary care and other health and administrative databases, providing a representative sample of the population by age, sex, and ethnicity. It includes information on demographic characteristics, diagnoses and symptoms, drug exposures, and vaccination history. Covering 60 million patients, among whom nearly 18 million are currently registered, CPRD notably provides a significant longitudinal perspective on health outcomes and trends, with 25% of these patients having been followed for at least 20 years.

2.2.2 Scientific literature

Scientific literature datasets, such as PubMed[34], provide access to a vast collection of research papers, articles, and abstracts related to life sciences and biomedical topics. These datasets are valuable for training healthcare language models as they contain high-quality academic and professional text.

PubMed is a freely accessible database that supports access to several National Library of Medicine (NLM) literature resources, which contains citations and abstracts related to biomedical topics and life sciences. It provides more than 36 million citations and abstracts of biomedical literature, including content from MEDLINE, PubMed Central (PMC)[35], and online books. These citations may include links to full-text content from other sources such as PMC or the publisher's website. Launched online in 1996, PubMed is maintained and upgraded by the National Center for Biotechnology Information (NCBI). The dataset contains high-quality text, making it particularly suitable for training medical LLMs.

Closely related to PubMed, PMC serves as a significant repository that archives open access full-text articles published in biomedical and life sciences journals. PMC contains over eight million full-text article records covering biomedical and life science research from the late 1700s to the present, including articles formally published in a scholarly journal, author manuscripts accepted for publication in a journal, and preprint versions of articles. It is a crucial component of the NLM collection, complementing its extensive print and licensed electronic journal holdings.

Turning to datasets tailored for systematic reviews, RCT[36] comprises 4,528 reviews conducted by the Cochrane collaboration's members. These systematic reviews sourced from PubMed of all trials are relevant to specific clinical questions. RCT is constructed from the abstracts of systematic reviews and the clinical trials' titles and abstracts that are summarized by these reviews. Similarly, CDSR[38] is a dataset of high-quality systematic reviews in various medical domains. It contains a training set of 5,195 source-target pairs, a validation set of 500 abstract pairs, and a test set of 1,000 abstract pairs.

For summarization tasks in the biomedical field, MSˆ2 (multi-document summarization of medical studies)[37] is a multi-document summarization dataset comprising over 470 K documents and 20 K summaries in the biomedical domain. It is constructed from papers in the Semantic Scholar literature corpus, containing a large amount of related markup.

Broadening the scope to multidisciplinary research, S2ORC[41] is a large corpus consisting of 81.1 million English-language academic papers covering many academic disciplines. It represents a significant stride in academic literature aggregation. S2ORC is meticulously constructed using data from the Semantic Scholar literature corpus, which integrates papers from a variety of sources including publishers, archives such as arXiv or PubMed and resources such as MAG.

For highly specialized research, CORD-19[42] is a dataset containing over 1M papers on COVID-19 and related historical coronavirus research, including full text content for nearly 370 K papers. This dataset is valuable for pandemic research, emphasizing the essential role of domain-specific resources in meeting urgent biomedical needs.

2.2.3 Web data

Web data includes a broad spectrum of text that can be sourced from the internet, embodying a vast array of information types and formats. Among these, social media content stands out as one of the most prevalent and rich data sources. Reddit is a popular online platform where users can submit various types of content, including links, text posts, images, and videos. These submissions can be endorsed or disapproved by others through "upvotes" or "downvotes". Content that garners a significant number of upvotes is typically regarded as valuable and can serve as a rich source for creating high-quality datasets.

WebText[44] is a well-known corpus, which is compiled from highly upvoted links on Reddit, but it is not publicly available. In response to the limited availability of WebText, OpenWebText[45], an open-source alternative, is released.

COMETA[43] is an entity linking dataset of medical terminology. It contains 20,015 English biomedical concept mentions from Reddit expert-annotated with their corresponding SNOMED-CT links, covering a wide range of concepts such as symptoms, diseases, anatomical expressions and procedures across a range of conditions.

Colossal Clean Crawled Corpus (C4)[11] is a dataset containing about 750 GB clean English text scraped from the Common Crawl web dump.

2.2.4 public knowledge bases

There exist many public knowledge bases in medicine, such as UMLS[46], CMeKG[47] and DrugBank[48].

UMLS[46] is one of the most popular repository of biomedical vocabularies, which is developed by the US NLM. It has over two million names representing around 900,000 concepts sourced from over 60 families of biomedical vocabularies, as well as 12 million relations among these concepts.

CMeKG[47] is a Chinese medical knowledge graph, which is a structured description of professional medical knowledge. It is constructed by referencing authoritative international medical standards and a wide range of sources such as clinical guidelines, industry standards and medical textbooks. This knowledge graph lays a foundation for a medical QA system and serves as a comprehensive resource for medical information.

DrugBank[48] is a comprehensive freely available database containing detailed drug, drug-target, drug action and drug interaction information. The most recent version (5.1.12) has 16,581 drug entries including 2,769 approved small molecule drugs, 1,620 approved biologics, 135 nutraceuticals and over 6,723 experimental drugs. Additionally, 5,291 non-redundant protein sequences are linked to these drug entries. DrugBank's data is highly structured and accessible in various formats, including SMILES, SDF, MOL, PDB, InChI, and InChIKey for chemical structures, FASTA for sequence data, and XML and JSON for textual data. These standardized formats make DrugBank's data easily usable for training LLMs in tasks such as drug discovery and pharmacological research. Known for its high data quality, DrugBank curates information from peer-reviewed scientific literature, patents, and reputable databases, offering extensive and comprehensive data on drug targets. Regular updates ensure the dataset remains accurate and reliable. DrugBank's structured data on drug-target interactions makes it highly suitable for LLM tasks such as QA on drug-related topics. Additionally, its detailed pharmacological data supports LLMs in generating accurate summaries of drug mechanisms, and its information on interactions and metabolic pathways can be leveraged for dialogue generation in systems assisting healthcare professionals in drug-related decision-making.

It is important to note that, while datasets such as MIMIC-III form the foundation for many LLMs, they do have certain limitations. For example, MIMIC-III comes from a single institution - Beth Israel Deaconess Medical Center in Boston, USA, which makes it less representative of healthcare practices and patient demographics in other regions or countries, limiting the generalizability of models trained on this data. The dataset spans from 2001 to 2012, and medical practices, technologies, and treatment protocols have evolved significantly since then. This makes LLMs trained on MIMIC-III potentially less effective when applied to current healthcare scenarios. Moreover, MIMIC-III primarily contains data from critical care units, focusing on patients with severe conditions. As a result, it lacks coverage of broader healthcare settings, such as outpatient care or chronic disease management, which limits the scope of models trained on this dataset. Even MIMIC-III, one of the largest clinical text datasets available, contains only 0.5 billion tokens of clinical text from a single hospital - far less than the tens of billions used in LLM training[88]. Similarly, while CPRD provides a robust source of UK population health data, it has certain limitations. The dataset primarily reflects the ethnic and disease distributions specific to the UK, which may introduce regional and population biases that affect the model's generalizability. Additionally, CPRD data are primarily derived from primary care practices, making it more representative of outpatient care. Although CPRD includes linkages to hospital care and other health-related datasets, its emphasis on primary care may limit the depth of information available for studies that require detailed inpatient or intensive care data. This highlights the pressing need for more extensive and diverse clinical text datasets to advance clinical LLMs.

2.3. Datasets structure

Datasets for medical LLMs can be classified into two broad categories based on their structure: conventional text data and multimodal data.

2.3.1 Conventional text data

This category comprises datasets primarily centered on text, which are essential for training models to understand and generate human language. Within this category, QA and dialogue datasets stand out as the most widely utilized.

QA datasets

QA datasets are designed to train models capable of answering human-posed questions. These datasets typically consist of question-answer pairs, which are used to enable models to comprehend the question context and generate accurate responses based on the information available in the dataset or linked knowledge bases.

cMedQA2[49], the extension and amendment of version 1.0, is a dataset designed specifically for Chinese community medical QA. It is collected from an online Chinese medical QA forum, where users post their queries and receive answers from qualified doctors. It contains 108,000 questions and 203,569 answers, doubling the number of questions and answers compared to version v1.0, and performs some data cleaning preprocessing steps, such as eliminating greeting words and replacing English punctuation with Chinese punctuation.

webMedQA[50] is a real-world Chinese medical QA dataset collected from professional health-related consultancy websites. The dataset is collected through some steps including data preprocessing and removing questions with more than one best-adopted reply. It consists of 63,284 questions, covering most of the clinical departments of common diseases and health problems.

Huatuo-26M[51], named after the ancient Chinese physician Hua Tuo, stands as the largest Chinese medical QA dataset available today. It is collected from multiple sources through text cleaning and data deduplication methods, including an online medical consultation website, medical encyclopedias, and medical knowledge bases. It contains over 26 million QA pairs, covering various aspects such as diseases, symptoms, treatment methods, and drug information. Huatuo-26M significantly expands the scale of existing medical QA datasets and offers an unprecedented resource in the Chinese medical domain.

PubMedQA[53] is a novel biomedical QA dataset collected from PubMed abstracts. It aims to answer research questions with yes/no/maybe, using the corresponding abstracts. PubMedQA has 1K expert-annotated, 61.2 K unlabeled and 211.3 K artificially generated QA instances. It is the first QA dataset where reasoning over the contexts, especially their quantitative contents, is required to answer the questions.

HealthSearchQA[59] is a new dataset of 3,173 commonly searched consumer medical questions. It is curated using seed medical conditions and their associated symptoms. The dataset diverges from other medical text QA datasets in three significant ways including question only, free text response and open domain.

Dialogue

Dialogue datasets record conversational exchanges, mirroring real human-to-human interactions. They are crucial for training models to understand conversational nuances, and provide accurate and contextually appropriate responses.

MedDialog-CN[60] is a Chinese dataset containing conversations between doctors and patients. It contains 1.1 million dialogues, covering 29 broad categories of specialties and 172 fine-grained specialties. The data is collected from an online consultation website. MedDialog-EN[60] is an English dataset with 0.26 million dialogues, covering 51 categories of communities and 96 specialties. The data is collected from two online platforms of healthcare services.

IMCS-21[61] is a dialogue dataset that contains a total of 4,116 annotated samples with 164,731 utterances, covering ten pediatric diseases: bronchitis, fever, diarrhea, upper respiratory infection, dyspepsia, cold, cough, jaundice, constipation and bronchopneumonia.

Pubmed Causal[63] is a dataset for causal statements in science publications, containing 2,446 annotated sentences.

Instructions

Instruction datasets comprise step-by-step directives or guidelines intended to train models to perform specific tasks or understand procedural language, which is particularly useful for instructional AI applications.

Alpaca[69] is a dataset based on the self-instruct[89] method. This dataset employs the text-davinci-003 model on the 175 human-crafted instruction-output pairs from Self-Instruct to generate 52,000 new instructions along with inputs and outputs. Moreover, around 40% of the examples have an input in the final dataset.

sft-20k[71] is a dataset originating from the QiZhen medical knowledge base, which includes real medical QA data between patients and doctors, and drug text knowledge. It constructs an instructional dataset by applying specific question templates to semi-structured data. Qizhen Medical Knowledge Base collects QA data on various topics such as diseases, medications, diagnostic tests, surgeries, prognoses, and dietary information, summing up to 560,000 instructional entries.

ShenNong-TCM-Dataset[72] is constructed on the foundation of TCM-neo4j, an open-source medical knowledge graph. It employs an innovative entity-centric self-instruction method, leveraging ChatGPT to generate over 110,000 pieces of instructional data centered around TCM.

Summarization

Summarization is a concise description that captures the salient details of information. In the medical domain, summarization can be useful for helping people easily understand and address the diverse nature of questions and answers.

MeQSum[73] is a dataset comprising 1,000 consumer health questions and their expert-crafted summaries. The questions are carefully chosen from a collection provided by the U.S. NLM, ensuring a diverse and representative selection of consumer health questions. The dataset is particularly noteworthy for its method of summarization, meticulously carried out by three medical experts adhering to stringent guidelines to ensure the quality and utility of the summaries. The summarization process followed by these experts is governed by two critical principles. First, the summary must allow the retrieval of correct and complete answers to the original questions. Second, the summary cannot be shortened further without meeting the first condition.

CHQ-Summ[74] is a CHQ summarization dataset consisting of 1,507 consumer health questions and corresponding summaries. It is created from the Yahoo community QA forum that has a diverse set of users' questions. It contains additional annotations about question focus and question type of the original question, which are all annotated by domain experts.

MEDIQA-AnS[75] is a dataset designed for question-driven, consumer-focused summarization. It contains 156 consumer health questions, corresponding answers to these questions, and manually generated summaries of these answers.

In Table 2, we compare some medical QA datasets. A notable distinction among the datasets is their source. Most datasets, such as cMedQA2, webMedQA, and Huatuo-26M, are sourced from community-driven medical consultation platforms such as Xywy Community, Baidu Doctor, and Qianwen Health. These platforms are primarily Chinese-language websites, which means the datasets predominantly represent Chinese patients and healthcare practitioners, resulting in limited ethnic and geographical diversity. Additionally, the range of diseases covered may not be fully comprehensive, and the quality and reliability of the answers can vary due to differences in expertise among the responding doctors. Despite the limitations, these datasets provide a rich repository of user-generated medical inquiries and professional responses, making them particularly useful for developing consumer-facing medical QA systems. On the other hand, PubMedQA and MeQSum leverage more formal medical literature and research databases such as PubMed and the U.S. NLM. While these datasets do include some international content, they are primarily based on U.S. sources, thus reflecting healthcare practices, patient demographics, and disease prevalence patterns that are more representative of the American population. This may limit their generalizability to other ethnic groups and geographical regions. A major limitation of datasets sourced from online communities is the variability in the accuracy and reliability of the information. These datasets may have limited coverage of diseases, and the quality of responses can vary significantly based on the expertise of the contributing doctors. Conversely, datasets such as PubMedQA offer a higher degree of reliability but may lack the diversity of inquiries that arise from everyday medical concerns.

A critical factor in the utility of these datasets is the inclusion of real versus generative data. Datasets such as Huatuo-26M, Chatmed, and ShenNong incorporate both real and synthetic data. The inclusion of generative data, when carefully modeled, can significantly enhance the coverage of edge cases and underrepresented medical conditions. It enables QA systems to handle rare or hypothetical medical inquiries, which may not often arise in real-world consultations. However, generative data can also introduce noise into the training process, particularly if the synthetic data is of lower quality or inaccurately generated. This underscores the need for careful validation of generated content to prevent the dissemination of incorrect or harmful medical advice.

Table 2. Overview of medical question answering datasets

Dataset	Format	Size	Release time	Field	Source	Contain real data	Contain generative data	Language
cMedQA2	Q+A	108k+203k	2017	General medical field	Xywy Community	Yes	No	Chinese
webMedQA	Q+A	316k+63k	2019	General medical field	Baidu Doctor & 120ask	Yes	No	Chinese
Huatuo-26M	Q+A	26m+26m	2023	General medical field	Wiki, Qianwen Health	Yes	Yes	Chinese
Chatmed	Q+A	500k+500k	2023	Medical Question & Answering	Search Engines	No	Yes	Chinese
PubMedQA	Q+context+A	500+500	2019	General medical field	PQA-Labeled, PQA-Unlabeled, PQA-Artificial	Yes	Yes	English
HealthSearchQA	Q	3175	2023	General medical field	Search Engines	Yes	No	English
CovidDialog	Q+A	15k	2020	COVID-19 and pneumonia	Haodf	Yes	No	Chinese
IMCS-21	Q+context+A	164k	2022	Pediatric medicine	Muzhi Doctor	Yes	Yes	Chinese
sft-20k	Q+A	560k+180k+298k	2023	General medical field	Qizhen	Yes	Yes	Chinese
ShenNong	Q+A	110k	2023	Traditional Chinese Medicine	Open-source medical knowledge graph	No	Yes	Chinese
MeQSum	Q+A	1k+4k+3k	2019	General medical field	U.S. National Library of Medicine, Quora	Yes	No	English
CHQ-Summ	Q+A	1560	2022	Healthcare	Yahoo! Answers	Yes	No	English

Each dataset is suited for specific applications in medical QA. For instance, IMCS-21 is tailored to pediatric medicine, making it highly specialized for QA systems focusing on children's health. The detailed, context-specific nature of this dataset makes it ideal for systems that require deep, domain-specific knowledge. Similarly, CovidDialog, released in 2020 in response to the COVID-19 pandemic, offers targeted information on COVID-19 and other pneumonia-related conditions, making it an invaluable resource for QA applications focused on respiratory diseases. The release dates of these datasets are indicative of their relevance to current healthcare challenges. For example, datasets such as Huatuo-26M and Chatmed, released in 2023, reflect the latest developments in large-scale QA system training and integrate recent developments in medical knowledge, making them well-suited for modern medical applications.

In summary, each dataset within this comparative analysis offers unique strengths depending on the target application, the need for real or synthetic data, and the specific medical domain. While large datasets such as Huatuo-26M and sft-20k are indispensable for building comprehensive, large-scale QA models, smaller datasets such as PubMedQA and MeQSum remain critical for high-accuracy tasks that demand evidence-based answers. Thus, the choice of dataset should be guided by the specific needs of the QA system - whether for broad coverage, specialized knowledge, or linguistic precision.

2.3.2 Multimodal data

This category includes datasets that involve multiple modalities, such as text, images, and time series data. In the medical domain, multimodal language models offer a promising direction for further research. In Figure 3, we present a partial display of the content from four multimodal datasets.

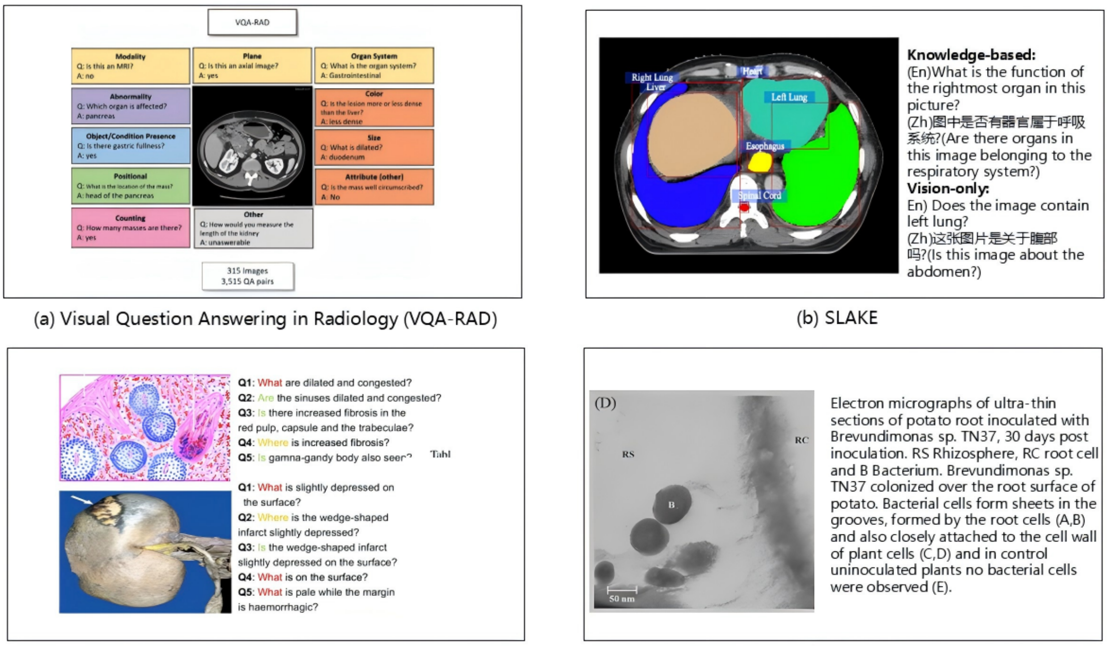
VQA Datasets

Medical visual question answering (Med-VQA) has tremendous potential in medicine, particularly in fields such as radiology and pathology. These two fields are rich in both imaging data and textual reports, making them prime candidates for VQA applications.

VQA-RAD [76] is a manually-crafted dataset in radiology where questions and answers are given by clinicians. It contains 3,515 visual questions of 11 types and 315 corresponding radiological images.

SLAKE [77] is a large bilingual dataset with comprehensive semantic labels annotated by experienced physicians and an extendable knowledge base for Med-VQA. It contains 642 radiology images including 12 diseases and 39 organs of the whole body, with 14,028 QA pairs and 5232 medical knowledge triplets.

PathVQA [78] is a pathology VQA dataset containing 32,799 QA pairs of eight categories, generated from 4,998



(a) Visual Question Answering in Radiology (VQA-RAD) (b) SLAKE

(c) PathVQA (d) PMC-OA

Figure 3. Partial content display of multimodal dataset.

images. The majority of questions in PathVQA are open-ended, and the other half are "yes/no" questions.

ROCO[80] is a multimodal image dataset, containing over 81K radiology images with several medical imaging modalities. It is constructed by retrieving all image-caption pairs from PMC. All images have corresponding captions, keywords extracted from the image caption, UMLS Concept Unique Identifiers and Semantic Type.

MedICaT[81] is a dataset that encompasses medical figures, captions, subfigures and subcaptions, and inline references that enable the study of these figures in context. The dataset's content is meticulously extracted from open-access articles available in PMC, ensuring figures and captions. Additionally, the corresponding reference texts are sourced from the S2ORC[41]. It contains 217,060 figures collected from 131,410 open-access scientific papers. Moreover, the dataset includes inline references for approximately 25,000 figures from the ROCO[80] dataset.

Among these datasets, VQA-RAD is a pioneering dataset for radiology VQA, consisting of radiological images and related medical questions. It focuses on enhancing the understanding of diagnostic images through QA, making it valuable for medical decision support systems. SLAKE extends the concept by incorporating a broader set of modalities (CT, MRI, X-rays) and both visual and textual inputs. This allows for more complex, multimodal reasoning tasks. PathVQA is specific to pathology images, such as histopathological slides, and is valuable for disease diagnosis in pathology using VQA techniques.

U-Xray, CheXpert, PadChest, and MIMIC-CXR focus on chest X-ray classification and segmentation. These datasets are widely used in developing models for automatic disease detection (e.g., pneumonia, pneumothorax) and diagnosis from X-rays. CheXpert, with over 224,000 labeled images, provides precise annotations for several lung diseases, making it a benchmark for chest disease classification. PadChest contains over 160,000 labeled images and expands its scope with Spanish-language reports, contributing to multilingual model training. ROCO and ROCOv2 are aimed at report generation from medical images. They cover a variety of medical imaging modalities, such as X-rays, CT scans, and MRIs, along with their corresponding textual descriptions.

These datasets enable models to generate human-readable medical reports, which can assist radiologists in producing structured reports efficiently.

MIMIC-CXR, with over 370,000 chest X-rays and associated clinical reports, is one of the largest and most diverse datasets available for medical imaging research. It is widely used for tasks such as disease classification, report generation, and clinical decision support. Similarly, CheXpert and PadChest offer large volumes of annotated images but focus more specifically on X-ray data, whereas MIMIC-CXR includes a broader range of accompanying clinical information, making it suitable for more complex multimodal tasks. U-Xray, though smaller in size, remains valuable for tasks such as detecting lung diseases in X-ray images. Its specialized focus on specific modalities and conditions makes it ideal for benchmarking models on chest X-ray classification tasks.

SLAKE, ROCO, MedlCaT, and PMC-15M offer multimodal data, which combines medical images with textual annotations or descriptions. This allows models to tackle tasks that require understanding both modalities simultaneously, such as VQA, report generation, or image-text matching. PMC-15M, with its combination of images and full-text articles, supports complex natural language processing (NLP) tasks in medical research, while ROCO enables report generation tasks that can directly influence clinical workflow efficiency.

Multi-omics datasets

Multi-omics datasets, which integrate various types of biological data such as genomics, proteomics, and metabolomics, provide a rich and multidimensional foundation for developing LLMs.

Cancer multi-omics datasets [90] include ten datasets that contain multi-omics data from different cancer types, with each dataset corresponding to a specific cancer. These datasets typically include three key omics layers: gene expression, DNA methylation, and miRNA expression. The number of patients ranges from 170 for acute myeloid leukemia (AML) to 621 for breast invasive carcinoma (BIC), offering a diverse range of data for training LLMs in cancer research.

Each dataset offers unique benefits depending on the intended use case and the specific tasks in medical image processing, report generation, or VQA. CheXpert, PadChest, and MIMIC-CXR stand out in disease detection and classification from chest X-rays due to their size and rich annotations. VQA-RAD, SLAKE, and PathVQA are essential for advancing VQA tasks in radiology and pathology. Datasets such as PMC-15M and MedlCaT provide invaluable resources for multimodal tasks that combine medical images with textual data, enabling more sophisticated models for clinical decision support and medical research. What is more, the integration of multi-omics data serves as both a guide for biomedical researchers in identifying suitable deep learning-based fusion methods and an indication of promising directions for improving multi-omics data fusion techniques.

In summary, the diversity and richness of data sources lay a solid foundation for the development of medical LLMs, facilitating significant advancements in understanding medical language, processing medical data, and providing medical decision support. Each dataset's unique structure and characteristics cater to specific aspects of healthcare AI technology, ranging from basic QA systems to complex diagnostic tools and patient management systems. By leveraging these resources, LLMs demonstrate substantial potential to enhance patient care quality and accelerate medical research.

3. DATASET APPLICATION

Datasets are fundamental to the deployment of LLMs in medicine. They are mainly employed in three aspects: pre-training, fine-tuning, and evaluation. The application of datasets has been presented in Figure 4.

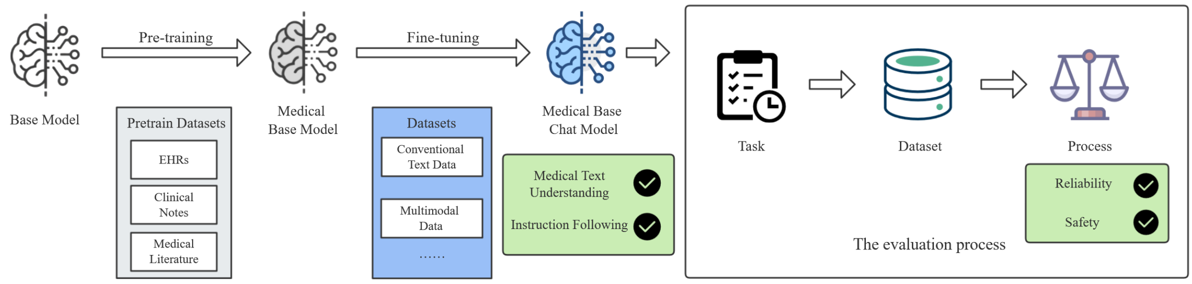


Figure 4. The application of datasets in developing medical Q&A LLMs.

3.1. Pre-training

In the pre-training phase, a large corpus of text data including both structured and unstructured text is used to train the LLMs. The corpus typically consists of various sources such as EHRs, clinical notes, and medical literature[91]. Some of the most commonly used medical datasets for pre-training medical LLMs include PubMed[34], MIMIC-III clinical notes[31], and PMC literature[35]. High-quality datasets form the backbone of LLM pre-training, with MIMIC-III and PubMed serving as pivotal resources. MIMIC-III provides valuable and extensive data for research, which has facilitated the development of several LLMs, such as ClinicalBert[14], GatorTron[92], and BlueBERT[93]. PubMed, known for its extensive biomedical literature, has been the foundation for training models such as PubMedBERT[16], BioBERT[94] and GatorTron. These models benefit from PubMed's extensive collection of research articles and studies, which enables them to capture a wide range of biomedical knowledge essential for engaging with scientific texts.

A notable example is GatorTronGPT, which is trained on a massive corpus of 82 billion words of de-identified clinical text[13] and 195 billion words of general English text from the Pile dataset[40]. GatorTronGPT is trained from scratch using the GPT-3[95] architecture with five billion and 20 billion parameters. By leveraging such diverse datasets, models such as GatorTronGPT gain the ability to understand both general English and domain-specific medical language, enhancing their utility in clinical applications. These datasets can be employed in combination to enrich the pre-training phase. BlueBERT combines both PubMed and MIMIC-III for pre-training; BioBERT is pre-trained on both PubMed and PMC; MEDITRON[96] is pre-trained on the GAP-REPLAY data mixture that contains papers from PubMed and PMC. Through pre-training on these medical corpora, LLMs are equipped with rich medical knowledge and to tackle various healthcare-related tasks.

3.2. Fine-tuning

After pre-training, LLMs require further fine-tuning to enhance their abilities. This phase leverages domain-specific datasets, such as dialogue data, QA pairs, and instructional texts, enabling the models to develop a nuanced understanding of both natural language and medical terminology. It tailors the model for specific medical tasks, allowing it to interpret and generate medical texts effectively.

For LLMs in medicine, supervised fine-tuning (SFT) and instruction fine-tuning (IFT) are two commonly used methods. These methods involve training the model on specific datasets to adapt it to the desired domain or task. SFT utilizes high-quality medical corpus such as physician-patient conversations, medical QA, and knowledge graphs. IFT constructs instruction-based training datasets, typically comprising instruction-input-output triples, to enhance the ability of instruction following. An example of IFT is the training process of Med-PaLM 2[18], where the base LLM is PaLM 2[97]. Med-PaLM 2 is fine-tuned using the datasets including MedQA, MedMCQA, HealthSearchQA, LiveQA, and MedicationQA. The fine-tuning followed the protocol used by Chung *et al.*, resulting in improved performance on medical Q&A benchmarks[98].

In addition, building upon the CMeKG[47], BenTsao[99] utilizes diverse instructional data for its instruction tuning process. MedAlpaca[19], built upon the LLaMA[100], is fine-tuned using over 160,000 medical QA pairs sourced from Medical Meadow[19]. Similarly, ChatDoctor[64] is obtained by fine-tuning the LLaMA model[100] on HealthCareMagic-100k[64] following the Stanford Alpaca[69] training method. It is first fine-tuned with Alpaca's data and further refined on HealthCareMagic-100k to improve its medical knowledge accuracy. Other models such as Qilin-Med[101] and Zhongjing[68] are obtained by incorporating the knowledge graph from ChiMed[101] and CMtMedQA[68] to perform fine-tuning on the Baichuan[102] and LLaMA[100] respectively to enhance their medical reasoning capabilities.

3.3. Evaluation

Evaluation is critical to the success of LLMs, which helps us better understand their strengths and weaknesses. Evaluating LLMs for medical applications typically involves using Q&A benchmarks where the models answer questions from a dataset, and their responses are scored based on predefined metrics, such as accuracy, precision, and recall, which are highly relevant to clinical applications. For text generation tasks, BLEU and ROUGE metrics are commonly applied to assess how closely the generated output matches the ground truth. For example, MultiMedQA[59] is designed to evaluate the capabilities of LLMs in answering medical questions across various formats, including multiple-choice and long-form answers. This benchmark compiles datasets from diverse sources, such as professional medical exams, medical research, and consumer health inquiries, providing a more thorough assessment of LLM performance beyond traditional multiple-choice accuracy or standard natural language generation metrics such as BLEU. The evaluation process tests LLMs not only on their factual accuracy but also on their medical reasoning capabilities and ability to handle both open-domain and closed-domain questions.

In addition, metrics such as Recall@K for retrieval tasks and AUC for classification tasks are frequently employed. For instance, PMC-CLIP[82], pretrained on the PMC-OA dataset, is evaluated using Recall@K for image-text retrieval on ROCO[80] and AUC and accuracy metrics for image classification, where it showed strong capabilities in these tasks. Clinical prediction with LLMs (CPLLM)[103] is evaluated on four prediction tasks: patient hospital readmission prediction, along with three specific diagnosis predictions for Chronic Kidney Disease, Acute and Unspecified Renal Failure, and Adult Respiratory Failure. The first two diagnoses are derived from the MIMIC-IV dataset, while the last diagnosis is derived from the eICU-CRD dataset[104]. MIMIC-IV provides the start time for admission and discharge times, and eICU-CRD associates each diagnosis with a timestamp, making them similarly applicable for patient readmission prediction tasks.

Models such as Med-PaLM[59] and Med-PaLM 2[18] are tested on MultiMedQA. USMLE[105], PubMedQA[53], and MedMCQA[106] are three popular datasets to evaluate their effectiveness. Codex-Med[107], PMC-LLaMA[108], Galactica[109], GatorTronGPT[12] and Med-PaLM 2[18] are evaluated on these three datasets. USMLE is also used to evaluate the performance of MedAlpaca in a zero-shot setting. Additionally, iCliniq[64] is used to test ChatDoctor's performance for a quantitative evaluation. HuatuoGPT[56] undergoes evaluation using three Chinese QA datasets: cMedQA2[49], webMedQA[50], and Huatuo26M[51] with GPT-4 and doctors comparing the responses from HuatuoGPT and making evaluations. The cMedQA2 dataset is also used to evaluate ClinicalGPT[110], which is conducted using automated evaluation metrics, with GPT-4 serving as the reference model. Other models such as LLaVA-Med[111] and Med-Flamingo[112] are evaluated on VQA datasets. VQA-RAD[76], SLAKE[77] and Path-VQA[78] are used to evaluate LLaVA-Med. VQA-RAD, Path-VQA and Visual USMLE are used to evaluate Med-Flamingo to measure their performance in medical-related tasks. Recently, MMedBench, covering 21 medical fields, has been designed to assess the accuracy of multiple-choice QA tasks and the ability to generate rationales across multiple languages[113]. The evaluation of eleven LLMs shows that MMed-Llama 3, built on the foundation of LLaMA 3, demonstrates strong performance compared with models such as LLaMA, ChatDoctor, and MedAlpaca. The linguistic diversity of datasets used in evaluations provides deeper insights into the models' capabilities, not only across various medical domains but also in their adaptability to multilingual healthcare contexts.

4. CHALLENGES

Several important factors such as data availability, data curation, quality, and deserve careful consideration.

4.1. Availability and open access

Accessibility of datasets is a key factor in determining their usability for external researchers. They can generally be categorized into three groups based on their accessibility [114,115]. Open access datasets are easily available to external researchers, often requiring only simple registration or an email request. These datasets provide valuable resources without the need for complex approvals. In contrast, regulated access datasets require formal agreements, such as institutional approvals, ethical clearance, or payments, due to the sensitive nature of the data. While these safeguards ensure compliance with regulations, they also create barriers that can slow down access. Lastly, inaccessible datasets are publicly listed as available but are often difficult to obtain due to issues such as non-responsiveness or outdated access links. Many medical datasets fall into the regulated access category, requiring formal approvals to protect sensitive information, while some may be inaccessible despite being listed as available [30]. These barriers highlight the challenges researchers face when trying to access medical data for LLM development.

4.2. Data curation and quality

Data curation tasks, including discovering, extracting, transforming, cleaning, and integrating data, remain critical yet resource-intensive efforts for organizations [116]. Data scientists often spend over 80% of their time on these tasks [117], as generic tools are rarely sufficient for the diverse and domain-specific requirements encountered in practice. The quality of available datasets also varies significantly, as many datasets are poorly curated, with incomplete or unlabeled data, making it difficult to train effective models. The process of human labeling is resource-intensive and requires domain expertise, while unsupervised learning methods face challenges due to the need for high accuracy [118]. Improved data curation practices, such as better structuring and labeling, are essential to enhance the quality and usability of medical datasets for LLM training.

4.3. Data scarcity and fragmentation

Data scarcity and fragmentation remain significant obstacles in medical research, as medical data is frequently siloed across different institutions and stored in various formats. Multimodal biomedical data fusion has become essential in modern healthcare research, integrating diverse data sources such as medical images, biomarkers, and physiological signals to provide a more comprehensive understanding of biological systems [119]. This approach enhances decision-making in key areas such as disease diagnosis, treatment planning, and patient monitoring by leveraging the strengths of each data modality.

4.4. Ethical considerations

Ethical concerns about using LLMs in the medical domain are significant, particularly around patient privacy, safety, and sensitive data use. A key issue is the collection and potential exposure of protected health information input into LLM application programming interfaces (APIs), which could be accessed by unauthorized parties. The lack of transparency from companies on how they store and use this data raises ethical questions about submitting sensitive information. Thus, strict controls for de-identification and informed consent must be implemented when handling protected health information in LLM APIs. Another major issue is the leakage of personally identifiable information (PII) [120,121]. LLMs trained on large datasets may inadvertently expose sensitive PII, such as email addresses or other confidential details, through vulnerabilities such as prompt injection attacks. So, it is essential to implement rigorous safeguards and data protection measures when deploying LLMs in clinical settings.

Moreover, data that disproportionately represents certain populations may result in biased models that perform poorly for underrepresented groups, exacerbating health disparities. Researchers must consider these ethical factors and, where possible, implement bias mitigation strategies and uphold the highest standards of data privacy and protection. There are also concerns about bias in medical datasets, which can result in LLMs that benefit certain populations while marginalizing others. In particular, the datasets used to train these models typically come from well-funded institutions in high-income, English-speaking countries. This leads to a significant under-representation of perspectives from other regions of the world, causing LLMs to adopt views that are biased toward the healthcare processes of high-income countries[122]. As a result, additional training could be integrated into the model development process to ensure that LLMs serve diverse patient populations equitably, and global datasets should be incorporated to reduce geographical and socioeconomic biases.

4.5. Limitations and future directions

Current medical datasets are relatively smaller than those used for general LLMs, covering a limited portion of the medical knowledge domain[59]. LLMs trained on these datasets may perform well on benchmarks, but they often struggle with real-world tasks such as differential diagnosis and personalized treatment planning[18]. While generating high-quality synthetic datasets for training could help broaden the model's knowledge, it risks causing LLMs to experience forgetting[123]. Further research is necessary to validate the effectiveness of synthetic data for medical LLMs and to develop techniques that mitigate such risks.

For evaluation, existing medical Q&A benchmarks often rely on metrics such as classification accuracy or natural language generation scores (e.g., BLEU[124]), which may not cover the full breadth of clinical scenarios and decision-making processes that occur in real-world medical practice. Multiple-choice tasks, often featured in these benchmarks, are much easier than real-world medical decisions that require synthesizing patient information and formulating individualized treatment plans, as they are grounded by experts. Although the MultiMedQA benchmark addresses some of these gaps by offering a diverse set of questions from medical exams, research, and consumer health queries, it is not exhaustive enough. It currently lacks coverage across all medical and scientific domains and is limited to English-language datasets, which restricts its applicability in global healthcare settings. To effectively evaluate LLMs, it is crucial to expand these datasets to include multilingual evaluations and more comprehensive clinical tasks, such as open-ended assessments that mirror actual clinical workflows. This expansion will enable models to be tested on their ability to reason through medical complexities, and provide accurate responses that are essential in real-world clinical environments.

5. CONCLUSIONS

This survey presents a comprehensive overview of the datasets in medicine and their pivotal role in developing LLMs. Datasets serve not only as a foundation for training LLMs but also as benchmarks for evaluating their performance. Each stage of datasets' application is critical for ensuring that the models are practically effective in real-world medical settings. Looking forward, the continued expansion and refinement of these datasets will be essential. Future research should focus on enhancing dataset transparency and quality, addressing privacy concerns, and integrating multimodal data to enrich model training. The development of medical datasets is a dynamic and evolving field that holds the key to unlocking the full potential of LLMs in medicine.

DECLARATIONS

Acknowledgments

We thank the Editor-in-Chief and all reviewers for their comments.

Authors' contributions
Made substantial contributions to the research, reviewed and summarized the literature, wrote and edited the original draft: Zhang D, Xue X, Hu M, Ying X
Investigation, data analysis and interpretation: Zhang D, Xue X, Gao P
Supervision: Hu M, Jin Z, Wu Y, Ying X

Availability of data and materials
Not applicable.

Financial support and sponsorship
None.

Conflicts of interest
Hu M is an Editorial Board Member of the journal *Intelligence & Robotics*, while the other authors declare that there are no conflicts of interest.

Ethical approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

REFERENCES

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29:1930–40. DOI
2. Introducing ChatGPT; 2022. Available from: https://openai.com/blog/chatgpt. [Last accessed on 12 Dec 2024]
3. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: Goldberg Y, Kozareva Z, Zhang Y, editors. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. pp. 1998–2022. DOI
4. Ji S, Zhang T, Yang K, et al. Domain-specific continued pretraining of language models for capturing long context in mental health; 2023. Available from: https://arxiv.org/abs/2304.10447. [Last accessed on 12 Dec 2024]
5. Sheng B, Guan Z, Lim LL, et al. Large language models for diabetes care: potentials and prospects. *Sci Bull* 2024;69:583–88. DOI
6. Devlin J, Chang MW, Lee K, Toutanova K.BERT: pre-training of deep bidirectional transformers for language understanding; 2019. Available from: https://arxiv.org/abs/1810.04805. [Last accessed on 12 Dec 2024]
7. Du Z, Qian Y, Liu X, et al. GLM: General language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics; 2022. pp. 320–35. DOI
8. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report; 2024. Available from: https://arxiv.org/abs/2303.08774. [Last accessed on 12 Dec 2024]
9. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems. vol. 33. Curran Associates, Inc.; 2020. pp. 1877–901. Available from: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf. [Last accessed on 12 Dec 2024]
10. Lewis M, Liu Y, Goyal N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. pp. 7871–80. DOI
11. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 2020;21:1–67. Available from: http://jmlr.org/papers/v21/20-074.html.
12. Peng C, Yang X, Chen A, et al. study of generative large language model for medical research and healthcare. *NPJ Digit Med* 2023;6:210. DOI
13. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med* 2022;5:194. DOI

14. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical bert embeddings; 2019. Available from: https://arxiv.org/abs/1904.03323. [Last accessed on 12 Dec 2024]
15. Chowdhery A, Narang S, Devlin J, et al. PaLM: Scaling language modeling with pathways; 2022. Available from: https://arxiv.org/abs/2204.02311. [Last accessed on 12 Dec 2024]
16. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. New York, NY, USA: Association for Computing Machinery; 2021. DOI
17. Minaee S, Mikolov T, Nikzad N, et al. Large language models: a survey; 2024. Available from: https://arxiv.org/abs/2402.06196. [Last accessed on 12 Dec 2024]
18. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models; 2023. Available from: https://arxiv.org/abs/2305.09617. [Last accessed on 12 Dec 2024]
19. Han T, Adams LC, Papaioannou JM, et al. MedAlpaca – An open-source collection of medical conversational AI models and training data; 2023. Available from: https://arxiv.org/abs/2304.08247. [Last accessed on 12 Dec 2024]
20. Toma A, Lawler PR, Ba J, et al. Clinical camel: an open expert-level medical language model with dialogue-based knowledge encoding; 2023. Available from: https://arxiv.org/abs/2305.12031. [Last accessed on 12 Dec 2024]
21. He K, Mao R, Lin Q, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics; 2024. Available from: https://arxiv.org/abs/2310.05694. [Last accessed on 12 Dec 2024]
22. Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models; 2023. Available from: https://arxiv.org/abs/2211.09110. [Last accessed on 12 Dec 2024]
23. Tao Z, Jin Z, Bai X, et al. EvEval: A comprehensive evaluation of event semantics for large language models; 2023. Available from: https://arxiv.org/abs/2305.15268. [Last accessed on 12 Dec 2024]
24. Bai Y, Ying J, Cao Y, et al. Benchmarking foundation models with language-model-as-an-examiner. In: Oh A, Naumann T, Globerson A, et al., editors. Advances in Neural Information Processing Systems. vol. 36. Curran Associates, Inc.; 2023. pp. 78142–67. Available from: https://proceedings.neurips.cc/paper_files/paper/2023/file/f64e55d03e2fe61aa4114e49cb654acb-Paper-Datasets_and_Benchmarks.pdf. [Last accessed on 12 Dec 2024]
25. Wang B, Chen W, Pei H, et al. DecodingTrust: a comprehensive assessment of trustworthiness in GPT models; 2024. Available from: https://arxiv.org/abs/2306.11698. [Last accessed on 12 Dec 2024]
26. Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. *Fertil Steril* 2023;120:575–83. DOI
27. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312. DOI
28. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst* 2023;47:33. DOI
29. Chang Y, Wang X, Wang J, et al. A Survey on Evaluation of Large Language Models. *ACM Trans Intell Syst Technol* 2024;15. DOI
30. Wu J, Liu X, Li M, et al. Clinical Text Datasets for Medical Artificial Intelligence and Large Language Models — A Systematic Review. *NEJM AI* 2024;1:AIra2400012. DOI
31. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:1–9. DOI
32. Johnson AE, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 2023;10:1. DOI
33. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: clinical practice research datalink (CPRD). *Int J Epidemiol* 2015;44:827–36. DOI
34. PubMed data; 2023. Available from: https://pubmed.ncbi.nlm.nih.gov/download/. [Last accessed on 12 Dec 2024]
35. PMC; 2024. Available from: https://www.ncbi.nlm.nih.gov/pmc/. [Last accessed on 12 Dec 2024]
36. Wallace BC, Saha S, Soboczenski F, Marshall IJ. Generating (factual?) narrative summaries of rcts: experiments with neural multi-document summarization. *AMIA Jt Summits Transl Sci Proc* 2021;2021:605-14. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8378607/. [Last accessed on 12 Dec 2024]
37. DeYoung J, Beltagy I, van Zuylen M, Kuehl B, Wang LL. MS^2: multi-document summarization of medical studies. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. pp. 7494–513. Available from: https://aclanthology.org/2021.emnlp-main.594. [Last accessed on 12 Dec 2024]
38. Guo Y, Qiu W, Wang Y, Cohen T. Automated lay language summarization of biomedical scientific reviews; 2022. Available from: https://arxiv.org/abs/2012.12573. [Last accessed on 12 Dec 2024]
39. Gupta V, Bharti P, Nokhiz P, Karnick H. SumPubMed: Summarization dataset of PubMed scientific articles. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop; 2021. pp. 292–303. DOI
40. Gao L, Biderman S, Black S, et al. The pile: an 800GB dataset of diverse text for language modeling; 2020. Available from: https://arxiv.org/abs/2101.00027. [Last accessed on 12 Dec 2024]
41. Lo K, Wang LL, Neumann M, Kinney R, Weld D. S2ORC: The semantic scholar open research corpus. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. pp. 4969–83. DOI
42. Wang LL, Lo K, Chandrasekhar Y, et al. CORD-19: The COVID-19 Open Research Dataset. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. Online: Association for Computational Linguistics; 2020. Available from: https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1. [Last accessed on 12 Dec 2024]

43. Basaldella M, Liu F, Shareghi E, Collier N. COMETA: A corpus for medical entity linking in the social media. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. pp. 3122–37. Available from: https://www.aclweb.org/anthology/2020.emnlp-main.253. [Last accessed on 12 Dec 2024]
44. Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners; 2019. Available from: https://api.semanticscholar.org/CorpusID:160025533. [Last accessed on 12 Dec 2024]
45. Gokaslan A, Cohen V. OpenWebText Corpus; 2019. Available from: http://Skylion007.github.io/OpenWebTextCorpus. [Last accessed on 12 Dec 2024]
46. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32:D267–70. DOI
47. Liu S, Yang H, Li J, Kolmanič S. Preliminary study on the knowledge graph construction of Chinese ancient history and culture. *Information* 2020;11:186. DOI
48. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;46:D1074–82. DOI
49. Zhang S, Zhang X, Wang H, Guo L, Liu S. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access* 2018;6:74061–71. DOI
50. He J, Fu M, Tu M. Applying deep matching networks to Chinese medical question answering: A study and a dataset. *BMC Med Inform Decis Mak* 2019;19:52. DOI
51. Li J, Wang X, Wu X, et al. Huatuo-26M, a large-scale Chinese medical QA dataset; 2023. Available from: https://arxiv.org/abs/2305.01526. [Last accessed on 12 Dec 2024]
52. Zhu W. ChatMed-Dataset: An GPT generated medical query-response datasets for medcial large language models. GitHub; 2023. Available from: https://github.com/michael-wzhu/ChatMed. [Last accessed on 12 Dec 2024]
53. Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. 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 (EMNLP-IJCNLP); 2019. pp. 2567–77. DOI
54. Xia F, Li B, Weng Y, et al. MedConQA: Medical Conversational Question Answering System based on Knowledge Graphs. In: Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Abu Dhabi, UAE: Association for Computational Linguistics; 2022. pp. 148–58. Available from: https://aclanthology.org/2022.emnlp-demos.15. [Last accessed on 12 Dec 2024]
55. Xu M. MedicalGPT: training medical GPT model; 2023. Available from: https://github.com/shibing624/MedicalGPT. [Last accessed on 12 Dec 2024]
56. Zhang H, Chen J, Jiang F, et al. HuatuoGPT, towards taming language model to be a doctor; 2023. Available from: https://arxiv.org/abs/2305.15075. [Last accessed on 12 Dec 2024]
57. Chen J, Wang X, Ji K, et al. HuatuoGPT-II, One-stage training for medical adaption of LLMs; 2024. Available from: https://arxiv.org/abs/2311.09774. [Last accessed on 12 Dec 2024]
58. Xiong H, Wang S, Zhu Y, et al. DoctorGLM: Fine-tuning your Chinese doctor is not a herculean task; 2023. Available from: https://arxiv.org/abs/2304.01097. [Last accessed on 12 Dec 2024]
59. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023;620:172–80. DOI
60. He X, Chen S, Ju Z, et al. MedDialog: two large-scale medical dialogue datasets; 2020. Available from: https://arxiv.org/abs/2004.03329. [Last accessed on 12 Dec 2024]
61. Chen W, Li Z, Fang H, et al. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics* 2022;39:btac817. DOI
62. Yang W, Zeng G, Tan B, et al. On the generation of medical dialogues for COVID-19; 2020. Available from: https://arxiv.org/abs/2005.05442. [Last accessed on 12 Dec 2024]
63. Yu B, Li Y, Wang J. Detecting causal language use in science findings. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. pp. 4664–74. DOI
64. Li Y, Li Z, Zhang K, et al. ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus* 2023;15. DOI
65. Liu W, Tang J, Cheng Y, et al. MedDG: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. In: CCF International Conference on Natural Language Processing and Chinese Computing. Springer; 2022. pp. 447–59. DOI
66. Zhang N, Chen M, Bi Z, et al. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics; 2022. pp. 7888–915. Available from: https://aclanthology.org/2022.acl-long.544. [Last accessed on 12 Dec 2024]
67. Bao Z, Chen W, Xiao S, et al. DISC-MedLLM: bridging general large language models and real-world medical consultation; 2023. Available from: https://arxiv.org/abs/2308.14346. [Last accessed on 12 Dec 2024]
68. Yang S, Zhao H, Zhu S, et al. Zhongjing: enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue; 2023. Available from: https://arxiv.org/abs/2308.03549. [Last accessed on 12 Dec 2024]
69. Taori R, Gulrajani I, Zhang T, et al. Stanford alpaca: an instruction-following LLaMA model. GitHub; 2023. Available from: https://github.com/tatsu-lab/stanford_alpaca. [Last accessed on 12 Dec 2024]

70. Peng B, Li C, He P, Galley M, Gao J. Instruction tuning with GPT-4; 2023. Available from: https://arxiv.org/abs/2304.03277. [Last accessed on 12 Dec 2024]
71. QiZhenGPT: an open source Chinese medical large language model. GitHub; 2023. Available from: https://github.com/CMKRG/QiZhenGPT. [Last accessed on 12 Dec 2024]
72. Wei Zhu WY, Wang X. ShenNong-TCM: a traditional chinese medicine large language model. GitHub; 2023. Available from: https://github.com/michael-wzhu/ShenNong-TCM-LLM. [Last accessed on 12 Dec 2024]
73. Abacha AB, Demner-Fushman D. On the summarization of consumer health questions. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. pp. 2228–34. DOI
74. Yadav S, Gupta D, Demner-Fushman D. CHQ-Summ: a dataset for consumer healthcare question summarization; 2022. Available from: https://arxiv.org/abs/2206.06581. [Last accessed on 12 Dec 2024]
75. Savery M, Abacha AB, Gayen S, Demner-Fushman D. Question-driven summarization of answers to consumer health questions. *Sci Data* 2020;7:322. DOI
76. Lau JJ, Gayen S, Ben Abacha A, Demner-Fushman D. A dataset of clinically generated visual questions and answers about radiology images. *Sci Data* 2018;5:1–10. DOI
77. Liu B, Zhan LM, Xu L, et al. Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE; 2021. pp. 1650–54. DOI
78. He X, Zhang Y, Mou L, Xing E, Xie P. PathVQA: 30000+ questions for medical visual question answering; 2020. Available from: https://arxiv.org/abs/2003.10286. [Last accessed on 12 Dec 2024]
79. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc* 2016;23:304–10. DOI
80. Pelka O, Koitka S, Rückert J, Nensa F, Friedrich CM. Radiology objects in context (roco): a multimodal image dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3. Springer; 2018. pp. 180–89. DOI
81. Sanjay Subramanian SMBBMvZSPSSMG Lucy Lu Wang, Hajishirzi H. MedICaT: a dataset of medical images, captions, and textual references. In: Findings of EMNLP; 2020. . DOI
82. Lin W, Zhao Z, Zhang X, et al. PMC-CLIP: contrastive language-image pre-training using biomedical documents; 2023. Available from: https://arxiv.org/abs/2303.07240. [Last accessed on 12 Dec 2024]
83. Irvin J, Rajpurkar P, Ko M, et al. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33; 2019. pp. 590–97. DOI
84. Bustos A, Pertusa A, Salinas JM, De La Iglesia-Vaya M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Med Image Anal* 2020;66:101797. DOI
85. Zhang S, Xu Y, Usuyama N, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs; 2024. Available from: https://arxiv.org/abs/2303.00915. [Last accessed on 12 Dec 2024]
86. Hamamci IE, Er S, Almas F, et al. Developing generalist foundation models from a multimodal dataset for 3D computed tomography; 2024. Available from: https://arxiv.org/abs/2403.17834. [Last accessed on 12 Dec 2024]
87. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual–language foundation model for pathology image analysis using medical twitter. *Nat Med* 2023;29:2307–16. DOI
88. Kaplan J, McCandlish S, Henighan T, et al.Scaling laws for neural language models; 2020. Available from: https://arxiv.org/abs/2001.08361. [Last accessed on 12 Dec 2024]
89. Wang Y, Kordi Y, Mishra S, et al. Self-Instruct: aligning language models with self-generated instructions; 2023. Available from: https://arxiv.org/abs/2212.10560. [Last accessed on 12 Dec 2024]
90. Leng D, Zheng L, Wen Y, et al. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol* 2022;23:171. DOI
91. Zhou H, Liu F, Gu B, et al. A survey of large language models in medicine: progress, application, and challenge; 2024. Available from: https://arxiv.org/abs/2311.05112. [Last accessed on 12 Dec 2024]
92. Yang X, PourNejatian N, Shin HC, et al. Gatortron: A large language model for clinical natural language processing. *medRxiv* 2022:2022–02. DOI
93. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task; 2019. pp. 58–65. DOI
94. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36:1234–40. DOI
95. Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. *Minds Mach* 2020 Dec;30:681–94. DOI
96. Chen Z, Cano AH, Romanou A, et al. MEDITRON-70B: scaling medical pretraining for large language models; 2023. Available from: https://arxiv.org/abs/2311.16079. [Last accessed on 12 Dec 2024]
97. Anil R, Dai AM, Firat O, et al. PaLM 2 technical report; 2023. Available from: https://arxiv.org/abs/2305.10403. [Last accessed on 12 Dec 2024]
98. Chung HW, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. *J Mach Learn Res* 2024;25:1–53. Available from: http://jmlr.org/papers/v25/23-0870.html.
99. Wang H, Liu C, Xi N, et al. HuaTuo: Tuning LLaMA model with Chinese medical knowledge; 2023. Available from: https://arxiv.org/

abs/2304.06975. [Last accessed on 12 Dec 2024]
100. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient doundation language models; 2023. Available from: https://arxiv.org/abs/2302.13971. [Last accessed on 12 Dec 2024]
101. Ye Q, Liu J, Chong D, et al. Qilin-Med: Multi-stage knowledge injection advanced medical large language model; 2024. Available from: https://arxiv.org/abs/2310.09089. [Last accessed on 12 Dec 2024]
102. Yang A, Xiao B, Wang B, et al. Baichuan 2: open large-scale language models; 2023. Available from: https://arxiv.org/abs/2309.10305. [Last accessed on 12 Dec 2024]
103. Shoham OB, Rappoport N. CPLLM: clinical prediction with large language models; 2024. Available from: https://arxiv.org/abs/2309.11295. [Last accessed on 12 Dec 2024]
104. Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data* 2018;5:180178. DOI
105. Jin D, Pan E, Oufattole N, et al. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Appl Sci* 2021;11:6421. DOI
106. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering; 2022. Available from: https://arxiv.org/abs/2203.14371. [Last accessed on 12 Dec 2024]
107. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions?; 2023. Available from: https://arxiv.org/abs/2207.08143. [Last accessed on 12 Dec 2024]
108. Wu C, Lin W, Zhang X, et al. PMC-LLaMA: towards building open-source language models for medicine; 2023. Available from: https://arxiv.org/abs/2304.14454. [Last accessed on 12 Dec 2024]
109. Taylor R, Kardas M, Cucurull G, et al. Galactica: a large language model for science; 2022. Available from: https://arxiv.org/abs/2211.09085. [Last accessed on 12 Dec 2024]
110. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation; 2023. Available from: https://arxiv.org/abs/2306.09968. [Last accessed on 12 Dec 2024]
111. Li C, Wong C, Zhang S, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day; 2023. Available from: https://arxiv.org/abs/2306.00890. [Last accessed on 12 Dec 2024]
112. Moor M, Huang Q, Wu S, et al. Med-Flamingo: a multimodal medical few-shot learner; 2023. Available from: https://arxiv.org/abs/2307.15189. [Last accessed on 12 Dec 2024]
113. Qiu P, Wu C, Zhang X, et al. Towards building multilingual language model for medicine; 2024. Available from: https://arxiv.org/abs/2402.13963. [Last accessed on 12 Dec 2024]
114. Wen D, Khan SM, Xu AJ, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. *Lancet Digit Health* 2022;4:e64–74. DOI
115. Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health* 2021;3:e51–66. DOI
116. Freitas A, Curry E. In: Cavanillas JM, Curry E, Wahlster W, editors. Big data curation. Cham: Springer International Publishing; 2016. pp. 87–118. DOI
117. Rezig EK, Cao L, Stonebraker M, et al. Data Civilizer 2.0: a holistic framework for data preparation and analytics. *Proc VLDB Endow* 2019 Aug;12:1954–57. DOI
118. Liu F, You C, Wu X, et al. Auto-encoding knowledge graph for unsupervised medical report generation; 2021. Available from: https://arxiv.org/abs/2111.04318. [Last accessed on 12 Dec 2024]
119. Duan J, Xiong J, Li Y, Ding W. Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion* 2024;112:102536. DOI
120. Li H, Guo D, Fan W, et al. Multi-step jailbreaking privacy attacks on ChatGPT; 2023. Available from: https://arxiv.org/abs/2304.05197. [Last accessed on 12 Dec 2024]
121. Shen X, Chen Z, Backes M, Shen Y, Zhang Y. "Do Anything Now": characterizing and evaluating in-the-wild jailbreak prompts on large language models; 2024. Available from: https://arxiv.org/abs/2308.03825. [Last accessed on 12 Dec 2024]
122. Li H, Moon JT, Purkayastha S, et al. Ethics of large language models in medicine and medical research. *Lancet Digit Health* 2023;5:e333–35. DOI Pubmed
123. Shumailov I, Shumaylov Z, Zhao Y, et al. The curse of recursion: training on generated data makes models forget; 2024. Available from: https://arxiv.org/abs/2305.17493. [Last accessed on 12 Dec 2024]
124. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02. USA: Association for Computational Linguistics; 2002. p. 311–318. DOI