

# Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning

LASA Team, Weiwen Xu<sup>♦</sup>, Hou Pong Chan<sup>♦</sup>, Long Li<sup>♦</sup>, Mahani Aljunied, Ruiyuan Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang<sup>♦</sup>, Yu Rong<sup>♦</sup>

DAMO Academy, Alibaba Group

<sup>♦</sup>Equal Contribution, <sup>♣</sup>Project Lead

Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities in understanding common visual elements such as landscapes, household objects, and public events, largely due to their large-scale datasets and advanced training strategies. However, their effectiveness in medical applications remains limited due to the inherent discrepancies between data and tasks in medical scenarios and those in the general domain. Concretely, existing medical MLLMs face the following critical limitations: (1) limited coverage of medical knowledge beyond imaging, (2) heightened susceptibility to hallucinations due to suboptimal data curation processes, and (3) lack of reasoning capabilities tailored for complex medical scenarios. To address these challenges, we first propose a comprehensive data curation procedure that (1) efficiently acquires rich medical knowledge data not only from medical imaging but also from extensive medical texts and general-domain data; and (2) synthesizes accurate medical captions, visual question answering (VQA), and reasoning samples. As a result, we build a multimodal dataset enriched with extensive medical knowledge. Building on the curated data, we introduce our medical-specialized MLLM: LINGSHU. LINGSHU undergoes multi-stage training to embed medical expertise and enhance its task-solving capabilities progressively. Besides, we preliminarily explore the potential of applying reinforcement learning with verifiable rewards paradigm to enhance LINGSHU's medical reasoning ability. Additionally, we develop MEDEVALKIT, a unified evaluation framework that consolidates leading multimodal and textual medical benchmarks for standardized, fair, and efficient model assessment. We evaluate the performance of LINGSHU on three fundamental medical tasks, multimodal QA, text-based QA, and medical report generation. The results show that LINGSHU consistently outperforms the existing open-source multimodal models on most tasks. Furthermore, we conduct five case studies closely aligned with real-world scenarios, demonstrating LINGSHU's potential for practical applications in medical contexts.

**Date:** June 12, 2025

**Correspondence:** Hao Zhang and Yu Rong, {hz.hhea2e, royrong.ry}@alibaba-inc.com

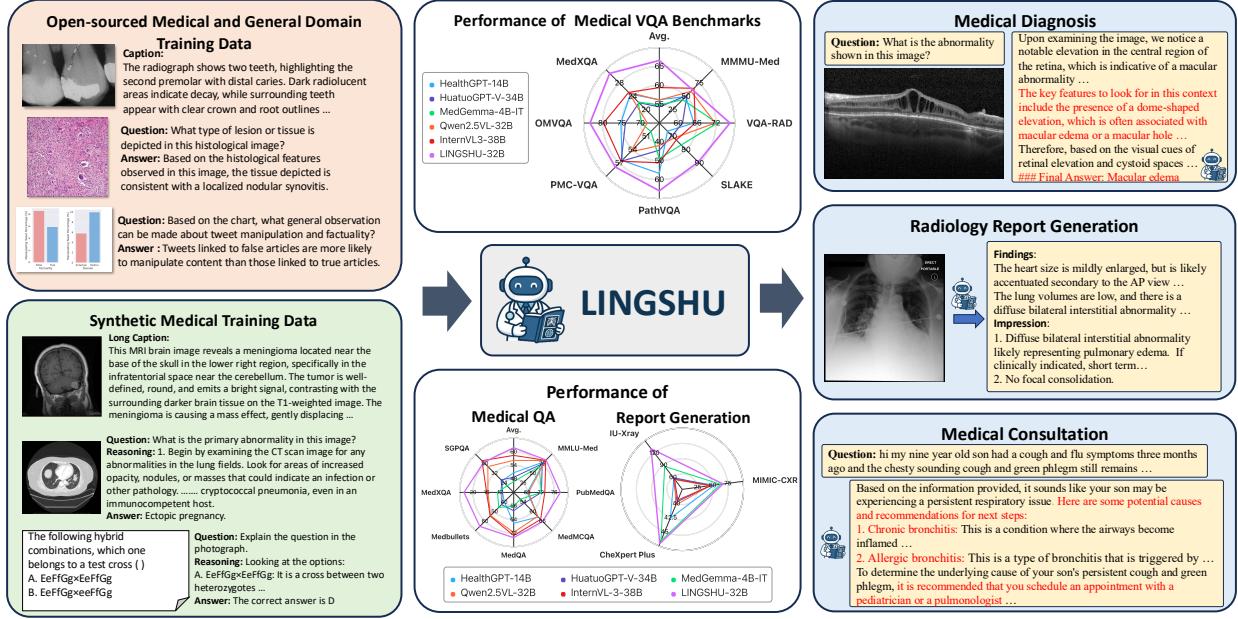
**Homepage:** <https://huggingface.co/lingshu-medical-mllm>



## 1 Introduction

Recent advancements in Multimodal Large Language Models (MLLMs) have significantly propelled progress in multimodal domains (OpenAI, 2024a; DeepMind, 2025a; Chen et al., 2025b; Kimi et al., 2025; Zhu et al., 2025), demonstrating capabilities that approach expert-level proficiency in areas like image captioning, visual question answering, and complex reasoning. Despite these successes, the performance of MLLMs in medical applications remains notably constrained (Yan et al., 2023; Lee et al., 2023; Nori et al., 2023). Biomedical image-text pairs differ fundamentally from typical web-based content, presenting unique challenges that general visual assistants are not well-equipped to handle (Li et al., 2025b). Lacking domain-specific medical visual knowledge, these models often respond to biomedical inquiries either with uncertainty or, worse, with incorrect information and hallucinations (Li et al., 2023a; Chen et al., 2024c).

To address these gaps, recent studies have concentrated on adapting models specifically for medical contexts by integrating general-purpose LLMs or MLLMs with medical multimodal data to construct specialized



**Figure 1** LINGSHU is trained on our meticulously curated datasets, encompassing rich and diverse open-sourced medical data across different modalities, general domain data, and high-quality synthetic medical data. Our results demonstrate that LINGSHU achieves promising performance across various medical benchmarks and modalities. Furthermore, case studies in downstream tasks highlight its practical utility and significant potential in real-world healthcare settings.

models (Wu et al., 2023; Hyland et al., 2024; Zhang et al., 2024a,b; Seyfioglu et al., 2025). These models demonstrate encouraging results in various tasks, including medical VQA, medical report generation, and diagnostic support, underscoring the considerable potential of MLLMs in healthcare scenarios. Pioneering models, such as LLaVA-Med (Li et al., 2023a), directly employ PubMed-derived datasets for vision-language alignment and supervised fine-tuning. However, their performance remains limited due to low data quality and inherent noise within PubMed. Subsequent research (Chen et al., 2024c; Li et al., 2025b; Lin et al., 2025) addresses the challenges with refined training recipes as well as larger and higher-quality medical multimodal datasets. Concurrently, advances in reasoning models, exemplified by OpenAI's o-series (OpenAI, 2024b, 2025b) and DeepSeek-R1 (DeepSeek-AI, 2025), set new benchmarks through advanced post-training techniques. Inspired by reinforcement learning with verifiable rewards (RLVR; Shao et al. (2024); Yu et al. (2025)), some studies (Lai et al., 2025; Pan et al., 2025) attempt to integrate this technique into the training of medical models to improve reasoning capabilities and enhance performance in clinical applications.

Although significant progress has been achieved, existing medical MLLMs typically focus on collecting image-text pairs by distilling knowledge from advanced models (OpenAI, 2024a,b, 2025a,b; DeepMind, 2025a,b). This solution is easy to implement and is scalable, but has notable limitations. First, medical tasks inherently demand extensive and diverse domain-specific knowledge. Nevertheless, the current distilled data are generated by prompting advanced models with medical imaging. These imaging prompts present challenges in extracting essential knowledge related to pharmacology, public health, and clinical contexts from the advanced models. Second, many distillation processes lack additional labels or supervision, relying exclusively on the generative capability of large models, thereby increasing hallucination risks. Third, the existing distilled data fall short in addressing complex medical cases that extend beyond basic image-text correlations and necessitate sequential decision-making.

In this work, we address the challenges of medical data curation via (1) efficiently acquiring rich medical knowledge data by collecting not only multimodal medical data but also extensive medical texts and general-domain multimodal/text data; and (2) synthesizing accurate medical data through a robust pipeline for generating captions, QA pairs, and chain-of-thoughts (CoTs) reasoning samples. Leveraging this data, we develop LINGSHU<sup>1</sup>, which is specifically tailored for the medical domain. LINGSHU models undergo a multi-

<sup>1</sup> LINGSHU originates from a chapter title in the *Huangdi Neijing* ([https://en.wikipedia.org/wiki/Huangdi\\_Neijing](https://en.wikipedia.org/wiki/Huangdi_Neijing)), one

---

stage training paradigm to progressively infuse medical knowledge and enhance their problem-solving abilities. Besides, we investigate the role of RLVR in improving LINGSHU’s multimodal medical reasoning and develop LINGSHU-RL. In addition to the data and model contributions, we find that current models are often evaluated in isolated, non-standardized environments. The lack of accessible, unified evaluation frameworks further restricts the development of the medical field. Therefore, we introduce MEDEVALKIT, a unified framework that consolidates major multimodal and textual medical benchmarks for efficient, standardized model assessment.

Extensive experiments demonstrate that LINGSHU consistently achieve state-of-the-art (SOTA) performance across most multimodal and textual medical VQA tasks, as well as in report generation, for both 7B and 32B configurations. Notably, in medical VQA tasks, LINGSHU-32B outperforms the second-best model by an average of 7.2 accuracy points across seven tasks, even surpassing proprietary models like GPT-4.1 and Claude Sonnet 4. Through comprehensive ablation studies, we further confirm the importance of data quality and medical knowledge coverage on overall performance, validating the effectiveness of our data curation framework. Finally, our case study underscores the great potential of LINGSHU across numerous medical applications, including medical report generation, medical support assistance, and clinical surgery assistance.

To sum up, our contributions are fourfold:

- **Data:** We introduce a novel data curation pipeline that efficiently gathers medical knowledge from various resources and synthesizes high-quality medical captions, QA pairs, and CoT reasoning samples.
- **Training:** We propose our medical-specific MLLMs, LINGSHU in 7B and 32B parameter sizes. LINGSHU models are trained with our tailored, multi-stage training paradigm that progressively infuses extensive medical knowledge into MLLMs and enhances problem-solving abilities. We also explore the role of RLVR in improving the multimodal medical reasoning and develop LINGSHU-RL.
- **Evaluation:** We propose MEDEVALKIT, a unified evaluation framework that consolidates major benchmarks for multimodal and textual medical tasks, streamlining model assessment and advancing standardized performance evaluation in the medical field.
- **Performance:** Through rigorous experimental validations, LINGSHU demonstrates state-of-the-art performance across multiple multimodal and textual medical VQA tasks, as well as in report generation.

**Paper organization.** We first describe our data curation process in §2, followed by the introduction of detailed training methodology (§3). We then elaborate our unified medical evaluation framework, MEDEVALKIT in §4. The experimental results are reported in §5, followed by illustrative case studies in §6. Subsequently, we briefly review related literature on medical MLLMs in §7. Finally, we discuss limitations and outline directions for future research in §8.

## 2 Data Curation

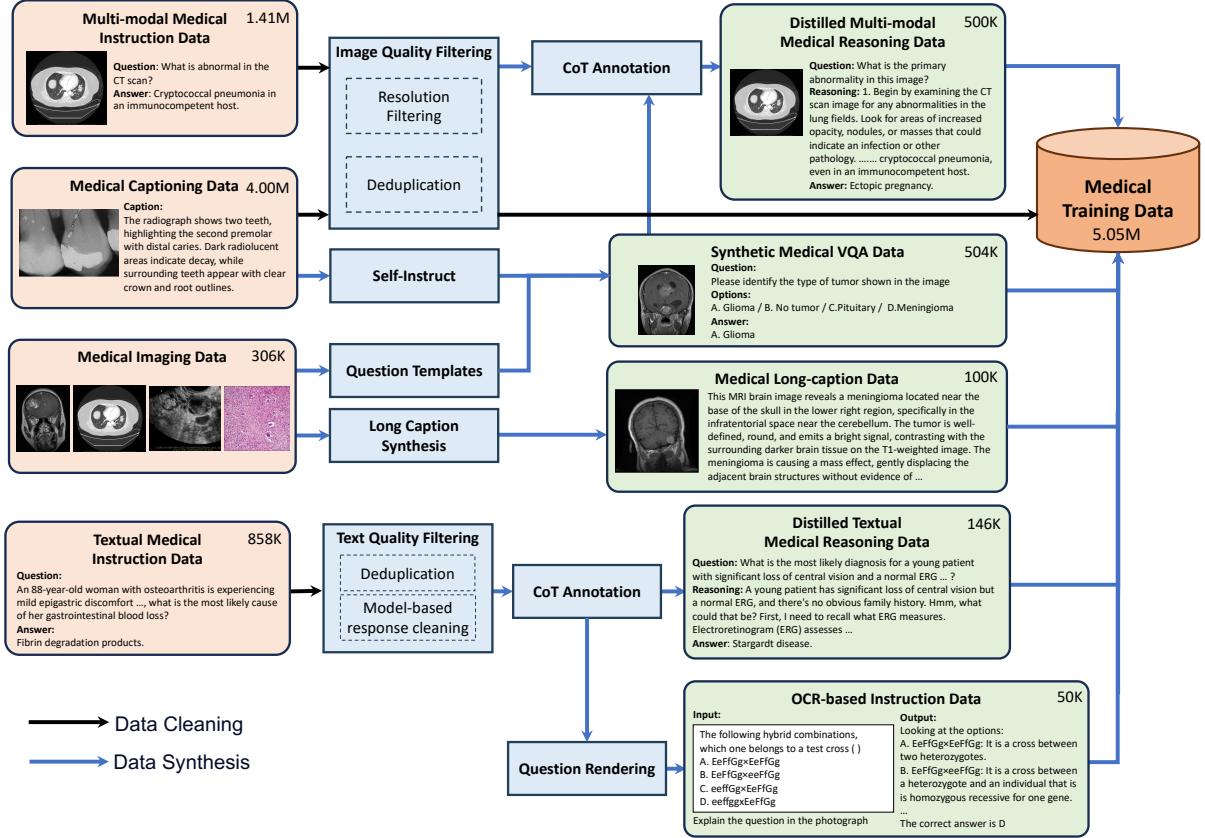
We address the intricate challenges of medical data curation via two primary strategies: (1) comprehensive collection of diverse medical knowledge sources, including medical multimodal datasets, extensive medical texts, and general-domain multimodal/textual data; and (2) synthesis of high-quality medical data through a robust pipeline capable of generating detailed captions, OCR-based samples, QA pairs, and chain-of-thought reasoning examples. The overall medical data curation pipeline is illustrated in Figure 2.

For *data collection*, we curate a wide range of open-source datasets from the web, categorized as follows: (1) **multimodal medical data** for enhancing LINGSHU’s comprehension of medical images and downstream task performance; (2) **unimodal medical data**, including standalone medical images and texts, to enrich domain knowledge and support synthetic data generation; (3) **general-domain data** to improve general visual and language understanding. All data undergo strict quality filtering prior to use. Each dataset is individually preprocessed and cleaned to ensure the quality and relevance.

For *medical data synthesis*, we generate targeted synthetic data to strengthen specific model capabilities: (1) **long-form captions** to improve diagnostic detail extraction from images; (2) **OCR-based instruction data** to enhance text recognition in medical imagery; (3) **VQA samples** for tasks such as diagnosis, anatomical

---

of the earliest classical texts of traditional Chinese medicine.



**Figure 2** The overall data curation pipeline of medical multimodal and textual data.

recognition, and modality classification; (4) **distilled reasoning trajectories** to promote advanced medical reasoning. All synthesized data are subject to a rigorous quality control process to ensure accuracy and reliability.

## 2.1 Data Collection

### 2.1.1 Medical Multimodal Data Collection

We incorporate two primary categories of existing multimodal medical data into our training set: (1) **medical caption data**, which pairs one or more medical images with descriptive text to facilitate semantic alignment between visual and textual modalities; and (2) **medical multimodal instruction data**, which includes image-instruction pairs. The latter comprises two main types: *medical VQA data*, featuring open-ended or multiple-choice questions about medical images, and *medical report data*, consisting of radiology reports that detail clinical findings and impressions. These datasets support MLLM in learning a broad range of medical multimodal tasks. A complete list of the collected datasets is provided in Table 1.

### 2.1.2 Medical Unimodal Data Collection

**Medical text instruction data.** Chen et al. (2023) has demonstrated that supervised fine-tuning (SFT) on medical instruction data is effective in injecting medical knowledge into LLMs. Building on this insight, we incorporate medical text instruction data into our training corpus to enhance the domain knowledge of LINGSHU. The incorporated data spans four primary categories: (1) *medical factoid QA*, including both free-form and multiple-choice questions on medical knowledge; (2) *distilled reasoning data*, which supplements answers with step-by-step reasoning path annotated by large reasoning models; (3) *patient-doctor dialogues*, comprising single-turn exchanges with patient queries and corresponding doctor responses; and (4) *general*

**Table 1** List of collected medical multimodal datasets. For MIMIC-CXR (Johnson et al., 2019) in medical captioning data, we only use its “Findings” part as the image captions.

Type of Data	Collected Datasets
Medical Captioning Data	LLaVA-Med Alignment (Li et al., 2023a), PubMedVision (Chen et al., 2024c), Quilt-LLaVA-Pretrain (Seyfioglu et al., 2025), PMC-OA (Lin et al., 2023), ROCO (Pelka et al., 2018), ROCov2 (Rückert et al., 2024), MIMIC-CXR (Johnson et al., 2019), MediCaT (Subramanian et al., 2020), FairVLMed (Luo et al., 2024), and MedPix-2.0 (Siragusa et al., 2025)
Medical Instruction Data	PathVQA (He et al., 2020), PMC-VQA (Zhang et al., 2024b), SLAKE (Liu et al., 2021), Quilt-LLaVA Instruct (Seyfioglu et al., 2025), VQA-Med-2019 (Ben Abacha et al., 2019), PubMedVision (Chen et al., 2024c), LLaVA-Med (Li et al., 2023a), LLaVA-Med-zh (BUAADreamer, 2024), VQA-RAD (Lau et al., 2018), MIMIC-Ext-MIMIC-CXR-VQA (Bae et al., 2024), CheXpert Plus (Chambon et al., 2024), MIMIC-CXR (Johnson et al., 2019), and IU-Xray (Demner-Fushman et al., 2016)

**Table 2** List of collected medical textual instruction datasets.

Type of Data	Collected Datasets
Medical Factoid QA	MedQA (Jin et al., 2021), MedQuAD (Ben Abacha and Demner-Fushman, 2019), ApolloCorpus (Wang et al., 2024b), and PMC-LLaMA (Wu et al., 2024)
Distilled Textual Medical Reasoning Data	MedThoughts-8K (hw hwei, 2025), MedReason (Wu et al., 2025), medical-o1-reasoning-SFT (Chen et al., 2024b), medical-o1-verifiable-problem (Chen et al., 2024b), and Medical-R1-Distill-Data (Chen et al., 2024b)
Patient-doctor Dialogues	HealthCareMagic-100k (Li et al., 2023b), icliniq-10k (Li et al., 2023b), and HuatuoGPT-II-GPT4-SFT-140K (Chen et al., 2023)
General Medical Instruction Data	AlpaCare-MedInstruct-52k (Zhang et al., 2025c)

*medical instruction data*, covering diverse tasks such as medication summarization and medical text translation. A complete list of the datasets is provided in Table 2.

**Medical imaging data.** It comprises raw images paired with human-annotated metadata, including the depicted organ, diagnostic labels, and imaging modality. This metadata enables the synthesis of additional captioning and VQA samples to improve medical image understanding and downstream task performance. We primarily use these datasets for data synthesis, which is detailed in §2.2. Table 3 summarizes the collected datasets, which span a wide range of medical modalities.

### 2.1.3 General Domain Data Collection

To enhance the general visual understanding, text generation, and multimodal instruction-following capabilities of our medical MLLM, we incorporate diverse general-domain datasets. These include image captioning, textual, and multimodal instruction-following data, which collectively enhance the model’s generalization to a wide range of vision-language tasks. Specifically, we utilize LLaVA-1.5 captions (Liu et al., 2024) and PixMo (Deitke et al., 2024) for image captioning; LLaVA-1.5 (textual instruction subset) and OpenHermes-2.5 (Teknium, 2023) for textual instruction-following; and LLaVA-1.5 (multimodal instruction subset) along with ALLaVA (Chen et al., 2024a) for multimodal instruction-following.

### 2.1.4 Data Cleaning Process

**Data cleaning for medical multimodal data.** Many open-source multimodal medical datasets are extracted automatically from scientific papers and thus often contain noise and redundancy. To ensure data quality, we implement a three-stage cleaning pipeline. First, **image filtering**, where images with dimensions smaller than 64 pixels are removed to eliminate low-resolution or uninformative content. Second, **image deduplication**, where we apply perceptual hashing with a strict Hamming distance threshold of zero to detect and remove exact duplicate images, retaining only one high-quality instance per set. To accelerate this process, we use a chunk-based deduplication technique.<sup>2</sup> Third, **text filtering**, where image caption samples with fewer than 10

<sup>2</sup><https://github.com/xuehuachunsheng/DupImageDetection>

---

**Table 3** List of collected medical imaging datasets of different medical modalities.

Modality	Collected Datasets
X-ray	COVID19-Radiography ( <a href="#">Chowdhury et al., 2020</a> ), NIH-Chest-X-Ray ( <a href="#">Wang et al., 2017</a> ), CheXpert ( <a href="#">Irvin et al., 2019</a> ), and Mendeley Digital Knee X-ray ( <a href="#">Gornale and Patravali, 2020</a> )
CT	KIPA22 ( <a href="#">He et al., 2021</a> ), DeepLesion ( <a href="#">Yan et al., 2017</a> ), and chest-ctscan ( <a href="#">Yi, 2021</a> )
MRI	Brain-Tumor-MRI ( <a href="#">Nickparvar, 2021</a> ), BraTS2024 ( <a href="#">de Verdier et al., 2024</a> ), LLD-MMRI ( <a href="#">Lou et al., 2025</a> ), and MAMA-MIA ( <a href="#">Garrucho et al., 2025</a> )
Ultrasound	AbdomenUS ( <a href="#">Vitale et al., 2020</a> ), Breast Ultrasound Images ( <a href="#">Al-Dhabayani et al., 2020</a> ), Annotated Ultrasound Liver images ( <a href="#">Xu et al., 2022</a> ), and RadIMageNet (subset of ultrasound images) ( <a href="#">Mei et al., 2022</a> )
Dermoscopy	HAM10000 ( <a href="#">Philipp et al., 2018</a> ), ISIC challenge 2020 ( <a href="#">Kurtansky et al., 2024</a> ), and Fitzpatrick17k ( <a href="#">Groh et al., 2021</a> )
Fundus	BRESET ( <a href="#">Nakayama et al., 2024</a> ), PAPILA ( <a href="#">Kovalyk et al., 2022</a> ), and 2015&2019 Diabetic Retinopathy Detection ( <a href="#">Lin, 2022</a> )
Histopathology	CPD ( <a href="#">Wagner et al., 2023</a> ), Breast-Cancer ( <a href="#">Ding et al., 2023</a> ), PanNuke ( <a href="#">Gamper et al., 2020</a> ), and EBHI-Seg ( <a href="#">Shi et al., 2023</a> )
Microscopy	Blood Cell Image Gan ( <a href="#">2018</a> ), BioMediTech RPE ( <a href="#">Nanni et al., 2016</a> ), and mhsma-dataset ( <a href="#">Javadi and Mirroshandel, 2019</a> )

or more than 1024 tokens are excluded to filter out lengthy and potentially noisy examples.

The full pipeline is applied to all medical image captioning datasets. For instruction data, we omit the text filtering step due to the validity of concise responses, *e.g.*, “yes/no”, word phrases, or short options. Image deduplication in instruction datasets specifically targets PubMedVision ([Chen et al., 2024c](#)) and LLaVA-Med ([Li et al., 2023a](#)), which exhibit a high degree of redundancy. However, cross-dataset deduplication is avoided in this case, as individual images may be paired with diverse instructions. In contrast, image deduplication across all datasets is conducted for medical captioning data.

**Data cleaning for medical text data.** To ensure the quality and compliance of collected medical text instruction data, we implement two key cleaning procedures. First, we apply **model-based response cleaning** to open-source patient-doctor dialogue datasets, such as HealthCareMagic-100k ([Li et al., 2023b](#)) and iCliniq-10k ([Li et al., 2023b](#)), which are typically scraped from online healthcare platforms. These responses often contain sensitive identity information or provide explicit diagnoses and prescriptions, raising privacy and legal concerns. Specifically, we employ LLaMA-3.1-70B ([Dubey et al., 2024](#)) to remove identity-related content and revise responses to avoid direct medical recommendations, as detailed in Appendix A.1. Second, to eliminate redundancy, we perform **text deduplication** across multiple instruction datasets using min-hash locality-sensitive hashing (LSH), retaining only the highest-quality version of each near-duplicate based on the reliability of its source.

## 2.2 Medical Data Synthesis

This section details our methodologies for synthesizing four types of medical multimodal data—long-form captions, OCR-based instruction samples, VQA instances, and distilled reasoning examples—alongside the quality control strategy employed to ensure their accuracy and reliability.

### 2.2.1 Medical Long-form Caption Synthesis

Given the brevity and limited descriptive depth of existing medical caption datasets, we aim to construct enriched captions that capture key visual features in medical images. To this end, we utilize data from medical image segmentation and classification tasks, which include human-annotated disease regions, diagnostic labels, and metadata such as patient gender and X-ray view position. This structured factual knowledge enables the synthesis of detailed and authentic medical captions. To ensure modality diversity and improve the model’s visual perception across imaging types, we source data from multiple modalities and balance their volume in the constructed caption dataset.

---

Specifically, we collect AbdomenUS (Vitale et al., 2020) for **Ultrasound**; PAD-UFES-20 (Pacheco et al., 2020) for **Dermoscopy**; CheXpert (Irvin et al., 2019), NIH-Chest-X-Ray (Rajpurkar et al., 2017), and Mendeley Digital Knee X-ray (Gornale and Patravali, 2020) for **X-ray**; KIPA22 (He et al., 2021) and DeepLesion (Yan et al., 2017) for **CT**; BraTS2024 (de Verdier et al., 2024), Brain-Tumor-MRI (Nickparvar, 2021), LLD-MMRI (Lou et al., 2025), and MAMA-MIA (Garrucho et al., 2025) for **MRI**; and CPD (Wagner et al., 2023), Breast-Cancer (Ding et al., 2023), PanNuke (Gamper et al., 2020), and EBHI-Seg (Shi et al., 2023) for **Histopathology**. For each image, we adopt GPT-4o (OpenAI, 2024a), whose prompt is shown in Appendix A.4, to generate a detailed long-form caption that captures fine-grained visual features. The captioning process follows a five-stage pipeline: (1) metadata preparation, (2) region-of-interest (RoI) identification, (3) annotation with factual knowledge, (4) annotation with doctor preference, and (5) summarization. This procedure yields approximately 100K high-quality medical image captions.

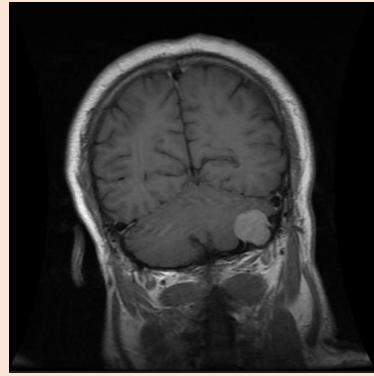
**Stage 1: metadata preparation.** This stage leverages metadata to provide essential contextual information for a given image, including modality, disease type, and, when applicable, camera view and observable patient characteristics. Following Xie et al. (2025), we begin by extracting metadata from each dataset, applying dataset-specific rules to convert this information into short captions. For example, in CPD (Wagner et al., 2023), we generate short captions using a staining method and nuclei score metadata with the template: “*A WSI image of carcinogenic DNA damage caused by ultraviolet (UV) radiation. The image is stained for {staining}. The relative amount of damaged nuclei (bounded in [0,1]) is {nuclei score}.*” To minimize hallucination risk, we exclude metadata not visually inferable from the image, e.g., patient age, based on expert consultation. All personal identifiers are removed to prevent racial or demographic biases. Additionally, we manually retrieve medical knowledge from the web to enhance the annotation process, particularly for models like GPT-4o that may lack domain-specific expertise. Thus, this stage involves metadata preparation from both the dataset and external medical references.

**Stage 2: RoI identification.** Building upon the textual data obtained in Stage 1, this stage incorporates complementary visual information. For image segmentation datasets, it provides disease or abnormality localization through RoIs, which are critical for reducing hallucinations and must be clearly marked in the images. RoIs are available as either segmentation masks or bounding boxes. Bounding boxes are directly overlaid on the original images, while segmentation masks are converted into bounding boxes by computing the minimal enclosing rectangles. For 3D datasets such as MAMA-MIA (Garrucho et al., 2025) and KIPA22 (He et al., 2021), we extract 2D slices and their corresponding masks along the z-axis, and apply the same bounding box rendering procedure to generate 2D images.

**Stage 3: annotation with factual knowledge.** This stage focuses on synthesizing image descriptions by integrating textual and visual information from the previous two stages. For datasets with annotated RoIs, we input bounding-box-rendered images, short captions, and retrieved medical knowledge into GPT-4o to generate descriptive text. For datasets lacking RoI annotations, such as image classification tasks, GPT-4o is prompted with the original image and metadata from Stage 1. Since visual and textual inputs from the last two stages are both manually annotated, the resulting descriptions are considered highly reliable.

**Stage 4: annotation with doctor preference.** Although the generated factual descriptions are highly reliable, they are often constrained by explicitly available metadata and may overlook critical visual details beyond RoIs. To mitigate this, we consult medical professionals to identify key factors they consider when interpreting medical images. Their insights are distilled into task-specific instructions. For instance, in MRI analysis, clinicians focus on sequence type, image orientation, anatomical structures, and visible abnormalities. These instructions are used to guide the annotation model, which, given the original image without bounding box overlays, generates alternative descriptions aligned with expert diagnostic perspectives.

**Stage 5: summarization.** We use GPT-4o to combine the outputs from stages 3 and 4 to consolidate the final caption for each medical image. To avoid conflict, the description from stage 3 is prioritized due to its higher reliability. Figure 3 presents a representative example of a synthesized long-form caption for an MRI image.



This MRI brain image reveals a meningioma located near the base of the skull in the lower right region, specifically in the infratentorial space near the cerebellum. The tumor is well-defined, round, and emits a bright signal, contrasting with the surrounding darker brain tissue on the T1-weighted image. The meningioma is causing a mass effect, gently displacing the adjacent brain structures without evidence of significant invasion. Anatomical structures visible include the frontal and parietal lobes, as well as the cerebellum and brainstem. The gyri and sulci of the cerebral cortex appear normal, with no other abnormalities noted. This imaging highlights the utility of MRI in diagnosing and understanding the spatial relations of intracranial lesions like meningiomas, which can exert pressure on nearby brain tissues and structures.

**Figure 3** An example of the synthesized medical long-form caption for the MRI image.

**Table 4** An example of the synthetic OCR-based instruction sample.

Input	Output
In the following hybrid combinations, which one belongs to a test cross ( ) A. EeFfGg × EeFfGg B. EeFfGg × eeFfGg C. eeffGg × EeFfGg D. eeffgg × EeFfGg	Looking at the options: A. EeFfGg × EeFfGg: It is a cross between two heterozygotes. B. EeFfGg × eeFfGg: It is a cross between a heterozygote and an individual that is homozygous recessive for one gene. C. eeffGg × EeFfGg: It is a cross between an individual that is homozygous recessive for two genes and a heterozygote. D. eeffgg × EeFfGg: It is a cross between an individual that is homozygous recessive for all three genes and a heterozygote. The correct answer is D
Explain the question in the photograph	

### 2.2.2 Medical OCR-based Instruction Data Synthesis

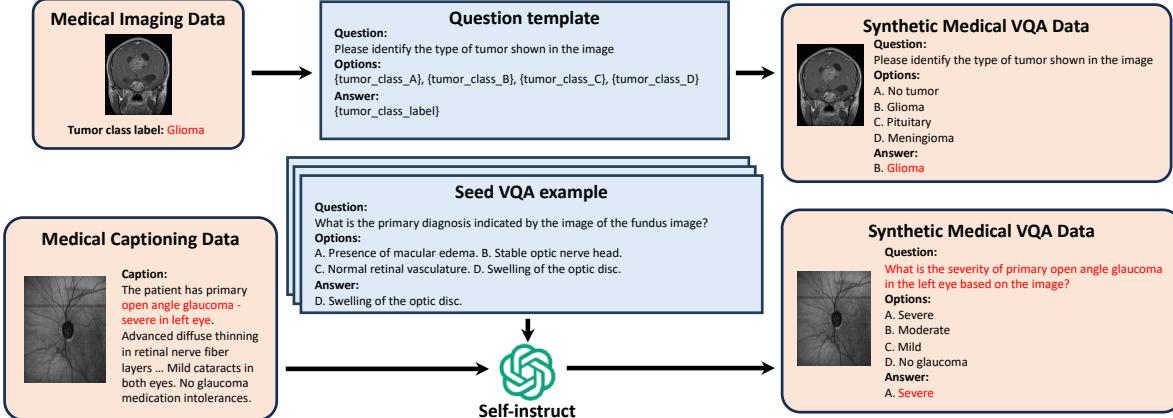
OCR-based instruction datasets enhance the model’s ability to interpret textual information embedded within medical images, a common scenario involving elements such as patient identifiers, age, and biological metrics. To construct these datasets, we collect biology and chemistry exam questions with ground-truth answers from online sources, as these subjects are closely related to the medical domain and offer rich resources. Each question is annotated with detailed reasoning steps using Gemini-2.0-Flash-Thinking (DeepMind, 2025a), and only samples with answers that exactly match the ground-truth are retained. The validated questions are then rendered as images to form multimodal instruction data, where the question image serves as input and the textual answer as output (see Table 4). This process yields a total of 50K high-quality OCR-based samples.

### 2.2.3 Medical VQA Data Synthesis

Medical diagnosis, anatomy identification, and modality classification are core competencies required for MLLMs to perform diverse medical multimodal tasks. To strengthen LINGSHU’s capabilities in these areas, we synthesize additional medical VQA data using two complementary strategies: a template-based method and a self-instruct-based method.

**Template-based method.** This method constructs VQA samples using human-annotated meta-information, *e.g.*, abnormality labels, from public medical imaging datasets. Specifically, we extract datasets containing labels related to anatomical structures, abnormalities, or imaging modalities. For each label type, we manually design question templates. Ground-truth answers are derived from the original annotations, while distractor options are sampled from the label space of the respective dataset. This approach enables a structured generation of high-quality VQA data. An example of this process is illustrated in Figure 4.

**Self-instruct-based method.** To enhance question diversity and linguistic variation, we leverage GPT-4o to generate VQA samples from (image, caption) pairs sourced from medical captioning datasets. GPT-4o is prompted using a few-shot format, where seed examples—drawn from open-source medical VQA



**Figure 4** Illustration of the data synthesis process for synthetic medical VQA data.

datasets—demonstrate questions focused on diagnosis and anatomy. It then produces a question, options, and the correct answer based on the provided caption. This method enables more varied and natural question formulations, as shown in Figure 4. The prompt design is described in Appendix A.2.

#### 2.2.4 Medical Reasoning Data Distillation

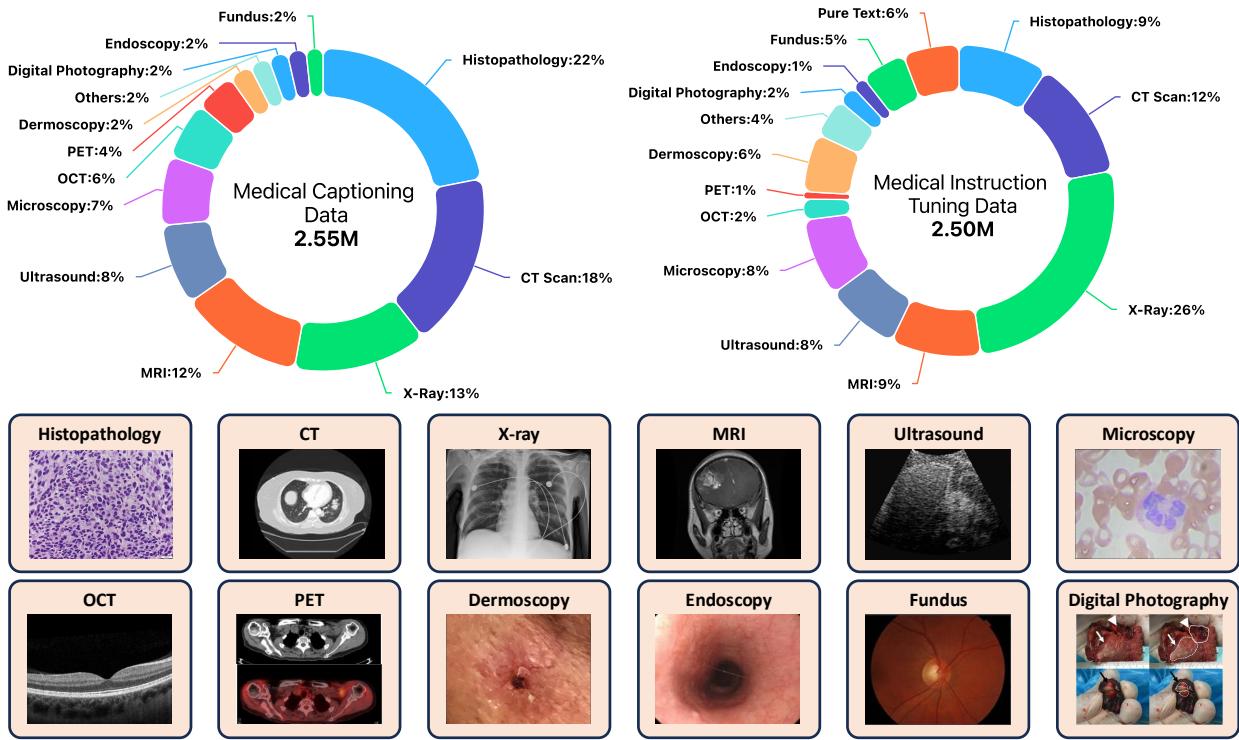
Most existing open-source and our synthesized medical instruction data, primarily composed of short-answer and multiple-choice questions, lack explicit annotations of the underlying reasoning process. To strengthen the model’s medical reasoning capabilities, we employ GPT-4o to generate CoT reasoning trajectories for a subset of both multimodal and textual instruction data. Specifically, for each sample, we provide GPT-4o with the question, ground-truth answer, and, if applicable, answer options and the corresponding medical image. GPT-4o is instructed to produce a step-by-step reasoning path without relying on or explicitly referencing the ground-truth answer. To ensure quality, we implement an LLM-based validation process wherein GPT-4o evaluates the consistency between the reasoning trace and the ground-truth answer. Samples deemed inconsistent are excluded from the final dataset. The prompts are detailed in Appendix A.3.

### 2.3 Dataset Summary

The data collection and synthesis pipelines yield 3.75M high-quality open-source medical samples and 1.30M high-quality synthetic medical samples, respectively. We also conduct rigorous quality checks on all samples to ensure the overall data quality. **To prevent data contamination, we perform strict image and text deduplication by excluding any images and samples overlapping with the evaluation benchmarks in MedEvalKit.** Moreover, we examine the distribution of medical image modalities within the training set. As many multimodal datasets lack explicit modality annotations, we train a modality classifier based on BiomedCLIP (Zhang et al., 2025b) to infer the modalities of unlabeled samples. Figure 5 presents the resulting modality distribution. Our dataset covers over 12 medical imaging modalities, substantially enhancing the model’s versatility and applicability on medical downstream tasks across a wide spectrum of modalities.

## 3 Model Training

LINGSHU is built upon the Qwen2.5-VL model architecture (Bai et al., 2025), which consists of three key components: a large language model (LLM), a vision encoder, and an MLP-based projector. We employ two parameter-scaled variants of Qwen2.5-VL, *i.e.*, 7B-Instruct and 32B-Instruct, as the foundational models. The choice of instruct versions is motivated by two primary considerations. First, these models require significantly less alignment data during training, owing to their inherent image understanding and instruction-following capabilities. Second, they support the direct use of instruction-formatted data during the alignment phase. Consequently, LINGSHU can adopt instruction-style data formatting from the initial training stage, which



**Figure 5** The distribution of medical image modality in our training data (medical subset), where the category “Pure Text” indicates text-only medical data.

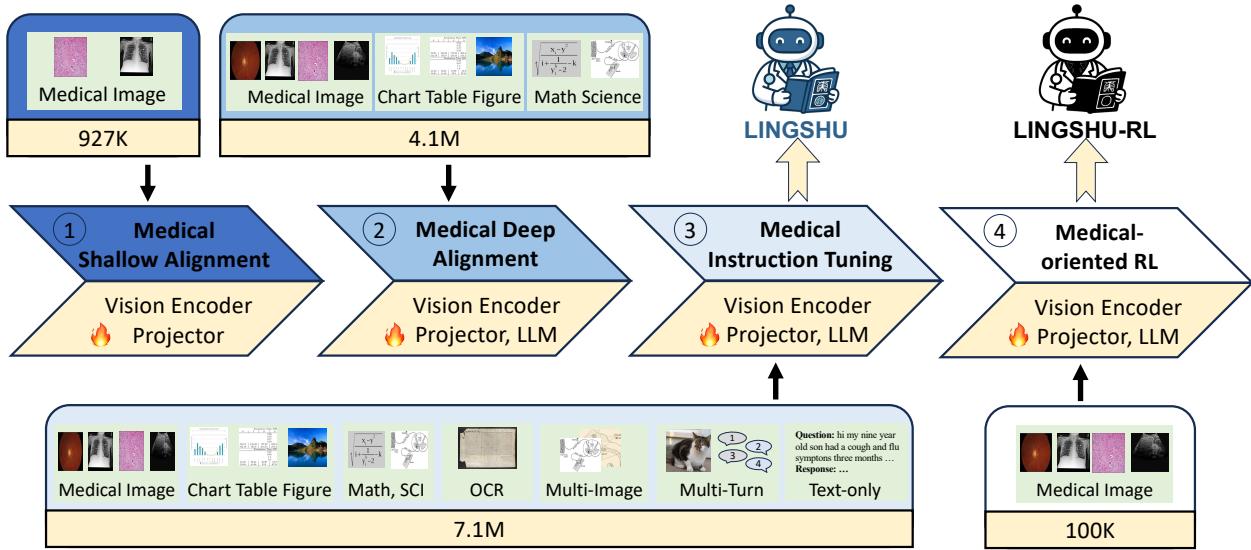
promotes consistency across the multi-stage training pipeline, reduces the overhead of additional instruction alignment, simplifies the training process, and enhances overall training stability.

### 3.1 Training Recipe

Building on prior advances in Multimodal Large Language Models (MLLMs; [Bai et al. \(2025\)](#); [Zhu et al. \(2025\)](#)), we develop a multi-stage training framework to progressively adapt the backbone model to the medical domain. This framework follows a “shallow-to-deep” progression and comprises four sequential stages: (1) **Medical Shallow Alignment**, (2) **Medical Deep Alignment**, (3) **Medical Instruction Tuning**, and (4) **Medical-oriented Reinforcement Learning**. The overall training pipeline is illustrated in Figure 6. Note that our LINGSHU is trained after the first three stages. We further explore the effectiveness of RL in medical reasoning, leading to LINGSHU-RL version.

The initial two stages focus on constructing a robust vision-language foundation tailored to medical scenarios. Medical Shallow Alignment fine-tunes the model using a small set of medical image-text pairs, enabling it to encode medical images and generate corresponding descriptions accurately. This stage facilitates the model’s preliminary understanding of visual medical content. Medical Deep Alignment builds upon this by introducing larger, higher-quality, and semantically richer medical image-text pair datasets. This stage deepens the model’s domain knowledge and achieves finer-grained vision-language alignment.

The subsequent stages aim to enhance the model’s utility in practical medical settings. Medical Instruction Tuning improves the model’s ability to comprehend and execute task-specific instructions across various medical use cases, thereby enhancing its generalizability to downstream tasks. Finally, Medical-oriented Reinforcement Learning incorporates the Reinforcement Learning with Verifiable Rewards (RLVR; [Shao et al. \(2024\)](#); [Lai et al. \(2025\)](#); [Pan et al. \(2025\)](#)) paradigm to strengthen the model’s medical reasoning, problem-solving capacity, and interpretability.



**Figure 6** The training pipeline of LINGSHU, which consists of four stages: medical shallow alignment, medical deep alignment, medical instruction tuning, and medical-oriented reinforcement learning. Note that our LINGSHU is trained after the first three SFT stages. We additionally apply RL on top of it, leading to LINGSHU-RL version.

### 3.1.1 Medical Shallow Alignment

The goal of the medical shallow alignment stage is to establish effective alignment between diverse medical imaging modalities and their corresponding textual descriptions. This alignment enhances the model’s ability to comprehend and interpret medical images, thereby laying a solid foundation for subsequent medical knowledge integration. During this phase, the LLM is kept frozen, while only the vision encoder and the projector are fine-tuned using coarsely annotated medical image-caption data.

Freezing the LLM is a deliberate choice, as the coarse caption data typically consists of short and information-sparse text. Exposing the LLM to such limited textual content could potentially impair its language generation capabilities. In contrast, training the vision encoder and projector enables the model to better map medical visual features into the LLM’s representational space. The use of coarsely annotated data is motivated by its relatively low semantic complexity, which facilitates rapid learning of the general characteristics of medical imaging modalities. This not only aids in effectively adjusting the vision encoder’s outputs for better alignment with the LLM but also contributes to faster training and improved convergence stability. As shown in Table 5, the coarsely annotated medical image-caption data comprises two datasets with relatively short and concise captions, *i.e.*, PMC-OA (Lin et al., 2023) and ROCO (Pelka et al., 2018).

### 3.1.2 Medical Deep Alignment

The goal of the medical deep alignment stage is to comprehensively integrate medical knowledge into MLLM, enhancing its capabilities to understand diverse medical concepts and adapt to various clinical contexts. To achieve this, all model parameters, including those of the LLM, vision encoder, and projector, are unfrozen to allow end-to-end fine-tuning. This process is designed to improve the model’s ability in multimodal knowledge integration and its effectiveness in interpreting complex medical visual data.

During this stage, training is conducted on a substantially more diverse and enriched set of medical image-text pairs. Compared to the preceding shallow alignment phase, the dataset is significantly expanded in terms of modality diversity, linguistic complexity, and structural completeness. It includes images from less common medical imaging modalities, longer and more structured captions, as well as synthetic image-caption pairs generated from medical image classification and segmentation tasks.

Moreover, high-quality general-domain multimodal image-text data is incorporated into the training process alongside medical data. This joint training not only helps preserve the model’s general multimodal capabilities but also introduces it to a wider variety of visual formats, such as charts, tables, graphs, mathematics, and

**Table 5** The overview of data mixture for the four training stages. Note all datasets listed below have been rigorously processed through our data pipeline (detailed in §2), resulting in cleaned and validated versions rather than raw data.

Stage	Training Data Composition	Amount
Medical Shallow Alignment	PMC-OA and ROCO	~927K
Medical Deep Alignment	<b>1. Medical Image Captions:</b> ROCOv2, LLaVA-Med, Quilt-LLaVA, PubMedVision, MIMIC-CXR, FairVLMed, MediCaT, MedPix-2.0, and <i>synthesized long-form medical caption data</i> <b>2. General Image Captions:</b> LLaVA-1.5 Caption and PixMo	~4.1M
Medical Instruction Tuning	<b>1. Medical Multimodal:</b> PathVQA, PMC-VQA, SLAKE, Quilt-LLaVA, VQA-Med-2019, MIMIC-Ext-MIMIC-CXR-VQA, PubMedVision, LLaVA-Med (en & zh), VQA-RAD, CheXpert Plus, MIMIC-CXR, ROCOv2, IU-Xray, and <i>synthesized QA, OCR-based, and CoT data in the medical domain</i> <b>2. General Multimodal:</b> LLaVA-1.5 instruction (multimodal instruct subset only) and ALLaVA <b>3. Medical Text:</b> MedQuAD, medical-o1-verifiable-problem, medical-o1-reasoning-SFT, ApolloCorpus, Medical-R1-Distill-Data, AlpaCare-MedInstruct-52k, HealthCareMagic-100k, icliniq-10k, PMC-LLaMA, HuatuoGPT2-GPT4-SFT-140K, MedReason, MedThoughts-8K, MedQA, and <i>synthesized medical QA data</i> <b>4. General Text:</b> LLaVA-1.5 instructions (textual instruct subset only) and OpenHermes-2.5	~7.1M
Medical-oriented RL	Data curated from MIMIC-Ext-MIMIC-CXR-VQA, PMC-VQA, Kvasir-VQA, Path-VQA, SLAKE, VQA-Med-2019, VQA-RAD and <i>synthesized QA datasets followed by rigorous post-processing</i>	~100K

scientific illustrations. Exposure to these structured visual elements enhances the model’s ability to perform structured data interpretation and cross-modal reasoning.

This is particularly beneficial in medical contexts where diagnostic reports often include laboratory tables, time-series physiological curves, and pathological maps—visual formats that differ significantly from natural images in both distribution and semantics. By learning from these general-domain examples, the model acquires transferable skills in chart interpretation and symbolic reasoning, enabling more precise analysis of abnormal physiological metrics, lesion measurements, and their temporal dynamics in clinical scenarios. The overall medical and general image caption dataset usage is summarized in Table 5.

### 3.1.3 Medical Instruction Tuning

In line with the prevailing practices in MLLMs (Li et al., 2023a; Chen et al., 2024c; Zhu et al., 2025; Chen et al., 2025c), the medical instruction-tuning stage is designed to refine MLLM’s instruction following capability with a broad spectrum of task directives, ensuring that its outputs are accurately aligned with user intentions. To accommodate a wide breadth of clinical use cases, we unlock all parameters and perform large-scale, end-to-end optimization of the entire model.

The training corpus transcends conventional instruction formats, such as image descriptions, question-answer pairs, and multiple-choice questions, by integrating an extensive suite of scenario-oriented queries collected and synthesized from multiple sources. These queries span diagnosis, clinical examination, medical knowledge retrieval, clinical report generation, and anatomical structure localization. This breadth markedly augments the model’s domain competency. In addition, we enrich the corpus with carefully vetted high-quality general-domain data and medical text data, both of which have proven effective in enhancing performance on downstream medical tasks. Incorporating such data is capable of counterbalancing the inherent image-centric information bias of image-text pair data, expanding the model’s conceptual scope beyond visual content, and fostering a more holistic grasp of medical knowledge.

Moreover, our training data comprises advanced and complex data formats, such as multi-image reasoning tasks, multi-turn dialogues, and queries that require detailed reasoning processes, which demand deeper analytical acuity. Collectively, this diversified and sophisticated training data equips the MLLM to navigate complex clinical scenarios with enhanced precision and analytical capabilities.

---

The datasets used in this stage are summarized in Table 5, which are categorized based on their domain (*i.e.*, medical or general) and modality (*i.e.*, multimodal or textual):

- In the **medical multimodal** subset, we initially employ multiple public medical instruction datasets, but found that they often emphasize specific image regions, which may introduce local view bias during training. Thus, we supplement the training with high-quality caption datasets—selected from ROCov2 (Rückert et al., 2024), PubMedVision (Chen et al., 2024c), and MIMIC-CXR (Johnson et al., 2019)—to promote holistic visual understanding. We further incorporate clinical report generation datasets, CheXpert Plus (Chambon et al., 2024) and IU-Xray (Demner-Fushman et al., 2016). Lastly, we synthesize additional medical data, as detailed in §2.2, including OCR-based, QA, and CoT data.
- For the **general multimodal** subset, we select the publicly available instruction-tuned version of LLaVA-1.5 (Liu et al., 2024) and ALLaVA (Chen et al., 2024a).
- In the **medical text** subset, we curate data from diverse public sources and generate instruction data via an automated synthesis pipeline, followed by strict quality validation. We further incorporate open-source long-form medical reasoning datasets, primarily distilled from OpenAI-o1 (OpenAI, 2024b) and DeepSeek-R1 (DeepSeek-AI, 2025), which offer comprehensive analyses of medical problems and effectively enhance the model’s medical knowledge.
- For the **general text** subset, we leverage two public datasets, *i.e.*, the text-only part of LLaVA-1.5 (Liu et al., 2024), and OpenHermes-2.5 (Teknium, 2023).

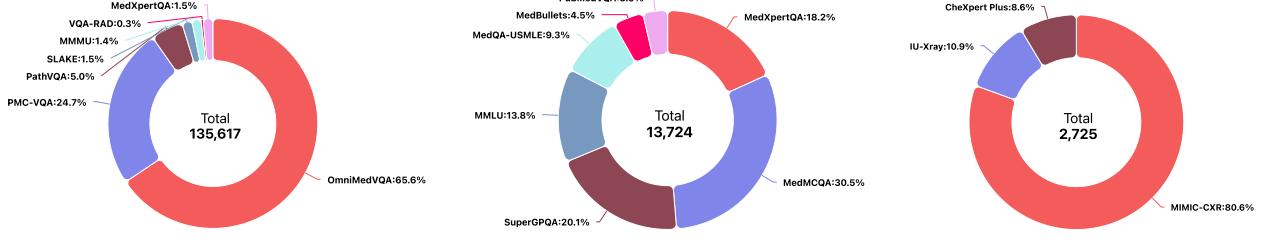
### 3.1.4 Medical-oriented Reinforcement Learning

Recent advancements in reasoning models, exemplified by OpenAI’s o-series (OpenAI, 2024b, 2025b) and DeepSeek-R1 (DeepSeek-AI, 2025), have set new benchmarks in complex tasks through advanced post-training strategies. Central to these improvements is reinforcement learning with verifiable rewards (RLVR), with Group Relative Policy Optimization (GRPO; Shao et al. (2024)) emerging as a prominent method due to its efficiency and effectiveness. It has been increasingly adopted for training MLLMs to enhance reasoning capabilities (Meng et al., 2025; Wang et al., 2025a; Tan et al., 2025). Building on RLVR’s success, some work applies it to medical domains for improving generalizability, interpretability, and reliability (Lai et al., 2025; Pan et al., 2025). Conventional SFT encounters key limitations: 1) over-reliance on answer supervision, which can cause overfitting and shortcut learning, particularly detrimental in medical scenarios; and 2) limited promotion of deliberate reasoning skills. RLVR addresses these issues by encouraging models to autonomously discover reasoning pathways through reward signals, rather than relying on answer memorization or teacher-guided CoT imitation (Chu et al., 2025). Accordingly, we adopt GRPO to train LINGSHU, leveraging a carefully curated medical verifiable dataset.

The medical verifiable dataset is derived from multiple medical sources (Table 5) and undergoes rigorous post-processing. Since most collected samples are in multiple-choice format, we reformulate those with word or phrase answers as open-ended questions to enhance difficulty while maintaining verifiability. This process also leads to a data balance between MCQA and open-ended QA. Additionally, binary-answer questions (*e.g.*, “Yes”/“No”) are downsampled to approximately 5% of the dataset to mitigate potential bias from their over-representation. Samples are then selected based on their modality and query format, aiming to ensure the data balance. A total of 100K examples are curated to form the training set for the RL stage. For reward design, we follow common practices (DeepSeek-AI, 2025; Yeo et al., 2025) by using a strict format reward and an accuracy reward, with respective weights of 0.5 and 1.

## 3.2 Implementation Details

As previously outlined, LINGSHU builds upon two variants of the Qwen2.5-VL-Instruct model, with parameter sizes of 7B and 32B, and is further optimized via continual training. The standard training process is organized into three primary stages, *i.e.*, Medical Shallow Alignment, Medical Deep Alignment, and Medical Instruction Tuning. This leads to our LINGSHU models. Across all stages, we adopt the AdamW optimizer with the cosine learning rate scheduler and a warm-up step of 100. The maximum sequence length is set to 8,192 tokens, the per-device training batch size is configured as 1, and the gradient accumulation step is set to 8.



**Figure 7** An overview of the data distribution across benchmarks for different types of medical tasks.

In the Medical Shallow Alignment stage, LLM remains frozen, only the vision encoder and the projection layer are fine-tuned for *one* epoch, with learning rates set to  $2e-6$  and  $1e-5$ , respectively. During the Medical Deep Alignment and Medical Instruction Tuning stages, we unfreeze the LLM and fine-tune it with a learning rate of  $1e-5$ , while maintaining the same learning rates for the vision encoder and projector. These stages are trained for *one* and *two* epochs, respectively.

To enhance training efficiency, we employ data packing during the Medical Instruction Tuning stage. Preliminary experiments suggest that data packing at this stage does not adversely affect model performance. However, similar attempts in the Medical Shallow and Deep Alignment stages have led to a marked decline in model performance. This decline is likely attributable to the high prevalence of short textual samples within the medical image-caption dataset.<sup>3</sup> When packed into longer sequences, these short entries contribute sparse gradients during normalization, undermining the training dynamics. Additionally, data packing substantially reduces the number of training steps, which may hinder sufficient convergence. Extending the number of epochs to compensate could result in overfitting. Consequently, we opt not to use data packing during the Medical Shallow and Deep Alignment stages.

For the Medical-oriented RL stage, we initialize training from LINGSHU checkpoint and apply GRPO for one epoch using the AdamW optimizer. The maximum sequence length is set to 4096,<sup>4</sup> with a rollout batch size of 512 and a global batch size of 128. The learning rate is configured at  $1e-6$ , the sampling temperature at 1.0, the KL divergence loss coefficient at  $1e-3$ , and 16 responses are sampled for each prompt. The resultant model is named as LINGSHU-RL.

## 4 MedEvalKit: A Unified Medical Evaluation Framework

The rapid progress of medical MLLMs has yielded impressive performance across various tasks. However, the absence of a unified evaluation framework has led to inconsistencies in performance comparison across models. Additionally, reproducing or deploying existing models in a standardized environment often requires substantial time and computational resources. A systematic and rigorous evaluation across diverse medical tasks is essential for accurately assessing model capabilities, identifying strengths and limitations, and providing actionable feedback for further development. Reproducible, quantitative evaluation is thus vital to advancing the optimization of medical MLLMs.

An effective evaluation of medical MLLMs requires a multiple-dimensional assessment, including medical knowledge, visual understanding, and reasoning ability. While existing frameworks, such as LMMs-Eval (Zhang et al., 2025a), Eval-Harness (Gao et al., 2024), and VLMEvalKit (Duan et al., 2024), can be applied to evaluate medical MLLMs, they are primarily designed for general multimodal tasks. As a result, they exhibit limitations in terms of medical data coverage and model adaptation, making them difficult to apply directly. This not only reduces evaluation efficiency but may also introduce potential biases.

To comprehensively and efficiently evaluate the performance of medical MLLMs, we develop a systematic evaluation framework, termed MED EVALKIT, which integrates mainstream medical benchmarks and task

<sup>3</sup>Although we synthesize a certain amount of long-form captions, it constitutes only a small proportion of the overall dataset.

<sup>4</sup>Although a maximum sequence length of 8192 is used in previous stages, preliminary RL experiments show that most medical samples generate significantly shorter outputs. Thus, we reduce the maximum sequence length to 4096 for this stage.

<b>Prompt for Multiple-Choice QA</b> <image> Question: <input question> Options: A. <option A> B. <option B> C. <option C> D. <option D> Answer with the option's letter from the given choices directly.	<b>Prompt for Open-ended QA</b> <image> Question: <input question> Answer the question concisely.	<b>Prompt for Report generation</b> <image>  You are a helpful assistant. Please generate a report for the given images, including both findings and impressions. Return the report in the following format: Findings: {} Impression: {}.
<b>Prompt for Close-ended QA</b> <image> Question: <input question> Answer the question using a single word or phrase.		

**Figure 8** The prompt formats of different medical question types.

types, supporting a variety of question formats, including multiple-choice questions, closed-ended questions, open-ended questions, and medical report generation. It covers representative tasks in medical multimodal and textual understanding, offering broad applicability. MEDEVALKIT accommodates both multimodal and text-only inputs, enabling unified evaluation of both medical MLLMs and medical LLMs. To enhance standardization and reproducibility, we standardized data preprocessing formats and post-processing protocols. A consistent model deployment and inference interface is provided, supporting rapid integration and one-click evaluation. For result assessment, MEDEVALKIT incorporates a dual verification mechanism that combines rule-based evaluation with an LLM-as-a-Judge strategy, integrating both objective and subjective assessment to improve evaluation stability and reliability. Additionally, the framework supports inference acceleration via vLLM (Kwon et al., 2023), enabling high-throughput and parallel evaluation, with strong scalability and engineering usability.

#### 4.1 A Large-scale Medical Evaluation Benchmark

To enable a more comprehensive evaluation of medical MLLMs, we curated a collection of mainstream medical benchmark datasets covering multimodal question answering, text-only question answering, and report generation, comprising a total of 152,066 evaluation samples with 121,622 distinct medical images spanning 16 benchmark datasets. The overview of the medical evaluation benchmark is depicted in Figure 7. More detailed information is presented in Table 25.

**Multimodal QA:** We include the test sets of VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), PathVQA (He et al., 2020), PMC-VQA (v2; Zhang et al. (2024b)), OmniMedVQA (Hu et al., 2024), MMMU (Yue et al., 2024), and MedXpertQA (Zuo et al., 2025). Specifically, we select the Health & Medical subset of MMMU, the multimodal portion of MedXpertQA, and the open-access part of OmniMedVQA. By incorporating these datasets, we encompass a diverse range of medical imaging modalities, including X-ray, CT scans, MRI, PET, ultrasound, microscopy, pathology, OCT, dermoscopy, gastrointestinal (GI) examinations, endoscopy, fundus, and medical-related charts, tables, and figures.

**Text-only QA:** We collect multiple medical text evaluation benchmarks, which comprise the test sets of MMLU (Hendrycks et al., 2021), PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022), MedQA-USMLE (Jin et al., 2021), MedBullets (Chen et al., 2025a), MedXpertQA (Zuo et al., 2025), and SuperGPQA (Team et al., 2025). Specifically, we select the text portion of MedXpertQA, the medicine-related questions of MMLU by following the split in Wang et al. (2024b), and the test set of SuperGPQA by the official split criteria.

**Report Generation:** We utilize the official test splits of three established benchmarks: MIMIC-CXR (Johnson et al., 2019), IU-Xray (Demner-Fushman et al., 2016), and CheXpert Plus (Chambon et al., 2024). These datasets primarily consist of chest X-rays paired with corresponding radiology reports. We follow Zhang et al. (2024c) to filter out samples where both findings and impressions are empty.

Moreover, we standardize the input format for all questions based on the question types, while adhering to the officially recommended chat template of the candidate medical MLLMs to be evaluated. The input prompt format of each question type is illustrated in Figure 8.

## 4.2 Evaluation Metrics

As the evaluation set comprises multiple task types, each requiring its own evaluation criteria, we implement a range of task-specific evaluation protocols accordingly.

For the QA task, **Accuracy** is used as the primary evaluation metric. For *multiple-choice questions*, a two-stage evaluation strategy is adopted: rule-based matching is first applied to assess alignment with answer options; if insufficient, we further employ the official MMMU codebase<sup>5</sup> to compute similarity scores and select the option with the highest score as the final answer. For *open-ended questions*, we directly leverage GPT-4.1 (2025-04-14 version; OpenAI (2025a)) to assess consistency between model outputs and reference answers, with the prompting strategy detailed in Appendix B.

For the report generation task, we develop a multi-metric evaluation approach that integrates semantic and model-based metrics, specifically tailored to address the challenges posed by the lengthy and highly open-ended nature of the outputs. This design enables a more comprehensive and granular assessment of model performance. For the semantic evaluation metrics, we choose **Rouge-L** (Lin, 2004) and **CIDEr** (Vedantam et al., 2015) to measure the quality of model-generated reports with respect to the reference answers. For the model-based evaluation metrics, we follow ReXrank (Zhang et al., 2024c) to utilize **SembScore** (Smit et al., 2020) and **RateScore** (Zhao et al., 2024b) to measure the candidate reports. We also include a composite metric **RadCliQ-v1** (Yu et al., 2023). The detailed descriptions of these metrics are presented in Appendix B.

## 5 Experiments

### 5.1 Experimental Settings

To comprehensively evaluate the performance of our model, LINGSHU, on various medical benchmarks, we compare it against a diverse set of baseline models, encompassing both proprietary models and open-source MLLMs from both general and medical domains. The open-source MLLMs are further categorized based on the parameter size. Specifically, our evaluation includes the following models:

- **Proprietary Models:** GPT-4.1 (2025-04-14 version; OpenAI (2025a)), Claude Sonnet 4 (Anthropic, 2025), and Gemini-2.5-Flash (Preview 05-20 version; DeepMind (2025a)), which are the most representative proprietary models available.
- **Medical MLLMs:** BiomedGPT (Zhang et al., 2024a), Med-R1 (Lai et al., 2025), MedVLM-R1 (Pan et al., 2025), MedGemma (Google, 2025), LLaVA-Med (Li et al., 2023a), HuatuoGPT-V (Chen et al., 2024c), BioMediX2 (Mullappilly et al., 2024), HealthGPT (Lin et al., 2025), and MedDr (He et al., 2024).
- **General-purpose MLLMs:** Qwen2.5-VL-Instruct (Bai et al., 2025), InternVL2.5 (Chen et al., 2025c), and InternVL3 (Zhu et al., 2025).<sup>6</sup>

To ensure a fair and consistent comparison, all models are evaluated using our MEDEVALKIT (detailed in §4) within a standardized evaluation environment.

### 5.2 Performance Comparison on Medical Multimodal Benchmarks

Table 6 presents a detailed comparison between LINGSHU and a diverse set of both proprietary and open-source MLLMs across seven medical multimodal benchmarks. Proprietary models such as GPT-4.1, Claude Sonnet 4, and Gemini-2.5-Flash, demonstrate consistently strong performance, with Gemini-2.5-Flash attaining the highest average score (65.1) among them. Among open-source models with fewer than 10B parameters, LINGSHU-7B achieves the highest average score (61.8), exceeding the best-performing baseline in this category by a substantial margin (+4.5). It ranks first on five out of seven benchmarks—SLAKE (83.1), PathVQA (61.9), PMC-VQA (56.3), OmniMedVQA (82.9), and MedXpertQA (26.7)—and secures second place on the remaining two, i.e., MMMU-Med and VQA-RAD. Notably, on PathVQA, LINGSHU-7B outperforms the next best open-source model, MedGemma-4B-IT, by a wide margin (61.9 versus 48.8). Furthermore,

<sup>5</sup><https://github.com/MMMU-Benchmark/MMMU>

<sup>6</sup>Although categorized as general-purpose MLLMs, the InternVL series also incorporates massive medical multimodal data into its training corpus (Li et al., 2025b; Chen et al., 2025c; Zhu et al., 2025).

**Table 6** Performance comparison of LINGSHU with other MLLMs on medical multimodal benchmarks, where **bold** and underline scores indicate the best and the second-best method, respectively. Note that OMVQA and MedXQA indicate OmniMedVQA and MedXpertQA-Multimodal benchmarks, respectively.  $\diamond$ Med-R1 is trained on part of the OmniMedVQA test set, making its results on OmniMedVQA meaningless.  $\heartsuit$ BiomedGPT and MedDr do not support multi-image inputs, which are required in MedXpertQA-Multimodal. Models colored in light gray, gray and green denote the medical MLLMs, general-purpose MLLMs and our LINGSHU, respectively.

Models	MMMU-Med	VQA-RAD	SLAKE	PathVQA	PMC-VQA	OMVQA	MedXQA	Avg.
Proprietary Models								
GPT-4.1	75.2	65.0	72.2	55.5	55.2	75.5	45.2	63.4
Claude Sonnet 4	74.6	67.6	70.6	54.2	54.4	65.5	43.3	61.5
Gemini-2.5-Flash	76.9	68.5	75.8	55.4	55.4	71.0	52.8	65.1
Open-source Models (<10B)								
BiomedGPT $\heartsuit$	24.9	16.6	13.6	11.3	27.6	27.9	-	-
Med-R1-2B $\diamond$	34.8	39.0	54.5	15.3	47.4	-	21.1	-
MedVLM-R1-2B	35.2	48.6	56.0	32.5	47.6	77.7	20.4	45.4
MedGemma-4B-IT	43.7	<b>72.5</b>	<u>76.4</u>	48.8	49.9	69.8	22.3	54.8
LLaVA-Med-7B	29.3	53.7	48.0	38.8	30.5	44.3	20.3	37.8
HuatuoGPT-V-7B	47.3	67.0	67.8	48.0	53.3	74.2	21.6	54.2
BioMediX2-8B	39.8	49.2	57.7	37.0	43.5	63.3	21.8	44.6
Qwen2.5VL-7B	50.6	64.5	67.2	44.1	51.9	63.6	22.3	52.0
InternVL2.5-8B	53.5	59.4	69.0	42.1	51.3	<u>81.3</u>	21.7	54.0
InternVL3-8B	<b>59.2</b>	65.4	72.8	48.6	<u>53.8</u>	79.1	<u>22.4</u>	<u>57.3</u>
LINGSHU-7B	<u>54.0</u>	<u>67.9</u>	<b>83.1</b>	<b>61.9</b>	<b>56.3</b>	<b>82.9</b>	<b>26.7</b>	<b>61.8</b>
Open-source Models (>10B)								
HealthGPT-14B	49.6	65.0	66.1	<u>56.7</u>	56.4	75.2	24.7	56.2
HuatuoGPT-V-34B	51.8	61.4	69.5	44.4	56.6	74.0	22.1	54.3
MedDr-40B $\heartsuit$	49.3	65.2	66.4	53.5	13.9	64.3	-	-
InternVL3-14B	<u>63.1</u>	66.3	<u>72.8</u>	48.0	54.1	78.9	23.1	58.0
Qwen2.5V-32B	59.6	<u>71.8</u>	71.2	41.9	54.5	68.2	25.2	56.1
InternVL2.5-38B	61.6	61.4	70.3	46.9	<u>57.2</u>	79.9	24.4	57.4
InternVL3-38B	<b>65.2</b>	65.4	72.7	51.0	56.6	79.8	<u>25.2</u>	<u>59.4</u>
LINGSHU-32B	62.3	<b>76.5</b>	<b>89.2</b>	<b>65.9</b>	<b>57.9</b>	<b>83.4</b>	<b>30.9</b>	<b>66.6</b>

LINGSHU-7B achieves an average performance comparable to GPT-4.1 and Claude Sonnet 4, highlighting its competitiveness despite significantly smaller parameters. With increased model capacity, LINGSHU-32B **attains a state-of-the-art average score of 66.6 across all benchmarks, outperforming both proprietary and open-source counterparts**. These results validate the effectiveness of our approach and demonstrate its strong potential for broad application across diverse medical multimodal tasks.

### 5.3 Performance Comparison on Medical Textual Benchmarks

Although primarily designed as multimodal models, LINGSHU exhibit remarkable proficiency in handling medical text-based tasks, underscoring its generalization capabilities across data modalities. As reported in Table 7, LINGSHU-7B achieves the highest average accuracy among all open-source models with fewer than 10B parameters. It leads on four key benchmarks—PubMedQA (76.6), MedQA-USMLE (63.3), Medbullets (56.2), and MedXpertQA (16.5)—and ranks a close second on the remaining three benchmarks.

For larger-scale models, LINGSHU-32B delivers even more pronounced improvements. It outperforms the second-best open-source competitor, InternVL3-38B, by an average of 3.4 percentage points in accuracy and secures top performance on six of the seven evaluated benchmarks, which highlights the scalability of LINGSHU. Nevertheless, a performance gap persists between LINGSHU and leading proprietary systems such as GPT-4.1, Claude Sonnet 4, and Gemini-2.5-Flash, particularly in tasks requiring deep clinical reasoning or access to extensive medical knowledge. Closing this gap is a primary objective of our ongoing research, which aims to

**Table 7** Performance of LINGSHU and other MLLMs on medical text benchmarks. Note that MedQA, MedXQA, and SGPQA denote MedQA-USMLE, MedXpertQA-Text, and SuperGPQA-Medical benchmarks, respectively. Models colored in light gray, gray and green denote the medical MLLMs, general-purpose MLLMs and our LINGSHU, respectively.

Models	MMLU-Med	PubMedQA	MedMCQA	MedQA	Medbullets	MedXQA	SGPQA	Avg.
Proprietary Models								
GPT-4.1	89.6	75.6	77.7	89.1	77.0	30.9	49.9	70.0
Claude Sonnet 4	91.3	78.6	79.3	92.1	80.2	33.6	56.3	73.1
Gemini-2.5-Flash	84.2	73.8	73.6	91.2	77.6	35.6	53.3	69.9
Open-source Models (<10B)								
Med-R1-2B	51.5	66.2	39.1	39.9	33.6	11.2	17.9	37.0
MedVLM-R1-2B	51.8	66.4	39.7	42.3	33.8	11.8	19.1	37.8
MedGemma-4B-IT	66.7	72.2	52.2	56.2	45.6	12.8	21.6	46.8
LLaVA-Med-7B	50.6	26.4	39.4	42.0	34.4	9.9	16.1	31.3
HuatuoGPT-V-7B	69.3	72.8	51.2	52.9	40.9	10.1	21.9	45.6
BioMediX2-8B	68.6	75.2	52.9	58.9	45.9	13.4	25.2	48.6
Qwen2.5VL-7B	73.4	<u>76.4</u>	52.6	57.3	42.1	12.8	26.3	48.7
InternVL2.5-8B	74.2	76.4	52.4	53.7	42.4	11.6	26.1	48.1
InternVL3-8B	<b>77.5</b>	75.4	<b>57.7</b>	<u>62.1</u>	<u>48.5</u>	<u>13.1</u>	<b>31.2</b>	<u>52.2</u>
LINGSHU-7B	<u>74.5</u>	<b>76.6</b>	<u>55.9</u>	<b>63.3</b>	<b>56.2</b>	<b>16.5</b>	<u>26.3</u>	<b>52.8</b>
Open-source Models (>10B)								
HealthGPT-14B	80.2	68.0	63.4	66.2	39.8	11.3	25.7	50.7
HuatuoGPT-V-34B	74.7	72.2	54.7	58.8	42.7	11.4	26.5	48.7
MedDr-40B	65.2	77.4	38.4	59.2	44.3	12.0	24.0	45.8
InternVL3-14B	81.7	<u>77.2</u>	62.0	70.1	49.5	14.1	37.9	56.1
Qwen2.5VL-32B	83.2	68.4	63.0	71.6	54.2	15.6	37.6	56.2
InternVL2.5-38B	<u>84.6</u>	74.2	<u>65.9</u>	<u>74.4</u>	<u>55.0</u>	14.7	39.9	58.4
InternVL3-38B	83.8	73.2	64.9	73.5	54.6	<u>16.0</u>	<b>42.5</b>	58.4
LINGSHU-32B	<b>84.7</b>	<b>77.8</b>	<b>66.1</b>	<b>74.7</b>	<b>65.4</b>	<b>22.7</b>	<u>41.1</u>	<b>61.8</b>

further enhance domain adaptation, instruction-following capabilities, and medical reasoning proficiency in future iterations of LINGSHU.

#### 5.4 Performance Comparison on Report Generation Benchmarks

In addition to the conventional QA benchmarks, we further evaluate the model’s performance on a practical task of high clinical relevance: automatic medical report generation. The results, summarized in Table 8, cover three widely adopted benchmarks: MIMIC-CXR, CheXpert Plus, and IU-Xray. A case study about report generation can be found in §6.5. For models with fewer than 10B parameters, LINGSHU-7B consistently surpasses all competing models on all three datasets, attaining either the best or second-best score for every evaluation metric. In the larger model size (>10B), performance differences among open-source models become more pronounced. LINGSHU-32B delivers the strongest overall results, leading in almost all metrics. Notably, its IU-Xray score is nearly double that of the compared baselines. These quantitative gains are accompanied by two noteworthy qualitative observations: (1) On MIMIC-CXR and CheXpert Plus, larger models occasionally achieve lower performance compared to their smaller counterparts. We posit that this counterintuitive result may be because larger models produce more elaborate and stylistically varied descriptions, nuances that current automatic metrics fail to reward adequately. This finding highlights an urgent need for more sophisticated evaluation protocols for free-text medical report generation tasks. (2) Although the InternVL series achieves commendable scores on multimodal and textual QA benchmarks (see Tables 6 and 7), it underperforms markedly in report generation. This disparity reveals a substantial disconnect between benchmark-centric evaluations and the demands of real-world medical applications, underscoring the importance of task-specific assessment when deploying medical AI systems.

**Table 8** Comprehensive evaluation of medical report generation on MIMIC-CXR, CheXpert Plus, and IU-Xray. Models colored in light gray, gray and green denote the medical MLLMs, general-purpose MLLMs and our LINGSHU, respectively. Metrics included are two semantic-based: ROUGE-L and CIDEr, two model-based: RaTE and SembScore (Semb), and one composite metric: RadCliQ-v1 (RadCliQ). As RadCliQ-v1 indicates better performance when the score is lower, we follow Zhang et al. (2024c) to take its reciprocal, *i.e.*,  $\text{RadCliQ}^{-1}$ . All scores are scaled by a factor of 100 to enhance clarity and comprehension.

Models	MIMIC-CXR					CheXpert Plus					IU-Xray				
	ROUGE-L	CIDEr	RaTE	Semb	RadCliQ <sup>-1</sup>	ROUGE-L	CIDEr	RaTE	Semb	RadCliQ <sup>-1</sup>	ROUGE-L	CIDEr	RaTE	Semb	RadCliQ <sup>-1</sup>
Proprietary Models															
GPT-4.1	9.0	82.8	51.3	23.9	57.1	24.5	78.8	45.5	23.2	45.5	30.2	124.6	51.3	47.5	80.3
Claude Sonnet 4	20.0	56.6	45.6	19.7	53.4	22.0	59.5	43.5	18.9	43.3	25.4	88.3	55.4	41.0	72.1
Gemini-2.5-Flash	25.4	80.7	50.3	29.7	59.4	23.6	72.2	44.3	27.4	44.0	33.5	129.3	55.6	50.9	91.6
Open-source Models (<10B)															
Med-R1-2B	19.3	35.4	40.6	14.8	42.4	18.6	37.1	38.5	17.8	37.6	16.1	38.3	41.4	12.5	43.6
MedVLM-R1-2B	20.3	40.1	41.6	14.2	48.3	20.9	43.5	38.9	15.5	40.9	22.7	61.1	46.1	22.7	54.3
MedGemma-4B-IT	25.6	81.0	52.4	29.2	62.9	27.1	79.0	47.2	29.3	46.6	30.8	103.6	57.0	46.8	86.7
LLaVA-Med-7B	15.0	43.4	12.8	18.3	52.9	18.4	45.5	38.8	23.5	44.0	18.8	68.2	40.9	16.0	58.1
HuatuoGPT-V-7B	23.4	69.5	48.9	20.0	48.2	21.3	64.7	44.2	19.3	39.4	29.6	104.3	52.9	40.7	63.6
BioMediX2-8B	20.0	52.8	44.4	17.7	53.0	18.1	47.9	40.8	21.6	43.3	19.6	58.8	40.1	11.6	53.8
Qwen2.5VL-7B	24.1	63.7	47.0	18.4	55.1	22.2	62.0	41.0	17.2	43.1	26.5	78.1	48.4	36.3	66.1
InternVL2.5-8B	23.2	61.8	47.0	21.0	56.2	20.6	58.5	43.1	19.7	42.7	24.8	75.4	51.1	36.7	67.0
InternVL3-8B	22.9	66.2	48.2	21.5	55.1	20.9	65.4	44.3	25.2	43.7	22.9	76.2	51.2	31.3	59.9
LINGSHU-7B	<b>30.8</b>	<b>109.4</b>	<b>52.1</b>	<b>30.0</b>	<b>69.2</b>	<b>26.5</b>	<b>79.0</b>	<b>45.4</b>	<b>26.8</b>	<b>47.3</b>	<b>41.2</b>	<b>180.7</b>	<b>57.6</b>	<b>48.4</b>	<b>108.1</b>
Open-source Models (>10B)															
HealthGPT-14B	21.4	64.7	48.4	16.5	52.7	20.6	66.2	44.4	22.7	42.6	22.9	81.9	50.8	16.6	56.9
HuatuoGPT-V-34B	23.5	68.5	48.5	23.0	47.1	22.5	62.8	42.9	22.1	39.7	28.2	108.3	54.4	42.2	59.3
MedDr-40B	15.7	62.3	45.2	12.2	47.0	24.1	66.1	44.7	24.2	44.7	19.4	62.9	40.3	7.3	48.9
InternVL3-14B	22.0	63.7	48.6	17.4	46.5	20.4	60.2	44.1	20.7	39.4	24.8	93.7	55.0	38.7	55.0
Qwen2.5VL-32B	15.7	50.2	47.5	17.1	45.2	15.2	54.8	43.4	18.5	40.3	18.9	73.3	51.3	38.1	54.0
InternVL2.5-38B	22.7	61.4	47.5	18.2	54.9	21.6	60.6	42.6	20.3	45.4	28.9	96.5	53.5	38.5	69.7
InternVL3-38B	22.8	64.6	47.9	18.1	47.2	20.5	62.7	43.8	20.2	39.4	25.5	90.7	53.5	33.1	55.2
LINGSHU-32B	<b>28.8</b>	<b>96.4</b>	<b>50.8</b>	<b>30.1</b>	<b>67.1</b>	<b>25.3</b>	<b>75.9</b>	<b>43.4</b>	<b>24.2</b>	<b>47.1</b>	<b>42.8</b>	<b>189.2</b>	<b>63.5</b>	<b>54.6</b>	<b>130.4</b>

**Table 9** Performance comparison of LINGSHU and its variant after medical-oriented RL, *i.e.*, LINGSHU-RL, on medical multimodal benchmarks. Note that OMVQA and MedXQA indicate OmniMedVQA and MedXpertQA-Multimodal benchmarks, respectively.

Models	MMMU-Med	VQA-RAD	SLAKE	PathVQA	PMC-VQA	OMVQA	MedXQA	Avg.
LINGSHU-7B	54.0	67.9	<b>83.1</b>	<b>61.9</b>	56.3	<b>82.9</b>	<b>26.7</b>	<b>61.8</b>
LINGSHU-RL-7B	<b>54.7</b>	<b>68.3</b>	82.5	61.1	<b>57.0</b>	81.3	<b>26.7</b>	61.5

## 5.5 Performance of Medical-oriented RL

We further conduct the initial exploration of RLVR to boost the medical multimodal reasoning capabilities of LINGSHU. As reported in Table 9, while the RL-trained variant (LINGSHU-RL-7B) achieves marginal improvements on MMMU-Med (+0.7%), PMC-VQA (+0.7%), and VQA-RAD (+0.4%), it shows performance degradation on SLAKE (-0.6%), PathVQA (-0.8%), and OMVQA (-1.6%). The average performance across all datasets remains relatively stable, indicating that the current RLVR training has not achieved the same improvements as those observed in the math/code domains. We identify two potential causes: (1) Reward Design: Our implementation adopted the prevailing approach of accuracy-oriented, rule-based rewards. However, medical reasoning is knowledge-driven, and acceptable answers, especially for open-ended questions, often exhibit substantial linguistic variability. Simple heuristics reward, therefore, may provide an unreliable signal of correctness, which can misdirect optimization. (2) Data Quality: A significant portion of medical multimodal problems do not necessitate complex reasoning processes, such as straightforward classification problems like identifying organ types. These data fail to effectively contribute to enhancing the model’s reasoning capabilities while introducing unnecessary noise to the training process. In this case, our current approach lacks criteria for selecting such reasoning-focused medical problems.

**Table 10** Ablation study examining data composition during the medical instruction tuning stage.  $\Delta|\text{Data}|$  denotes the reduced amount of training data. The symbol  $\downarrow$  indicates performance that is significantly lower compared to LINGSHU, while  $\Downarrow$  denotes an even more substantial decrease in performance.

Models	$\Delta \text{Data} $	MMMU-Med	VQA-RAD	SLAKE	PathVQA	PMC-VQA	OMVQA	MedXQA
Qwen2.5VL-7B	-	50.6	64.5	67.2	44.1	51.9	63.6	22.3
LINGSHU-7B	-	54.0	67.9	83.1	61.9	56.3	82.9	26.7
<i>Overall Data Composition</i>								
- Medical Multimodal	2.7M	54.1	66.5 $\downarrow$	69.0 $\Downarrow$	45.2 $\Downarrow$	56.2	78.0 $\Downarrow$	25.6 $\downarrow$
- General Multimodal	1.2M	51.4 $\Downarrow$	66.7 $\downarrow$	83.6	62.1	54.5 $\downarrow$	81.3 $\downarrow$	26.0
- General Text	1M	53.3 $\downarrow$	65.2 $\downarrow$	83.6	61.4	55.5 $\downarrow$	81.3 $\downarrow$	25.9 $\downarrow$
- Medical Text	173K	50.9 $\Downarrow$	66.7 $\downarrow$	82.4	60.2 $\downarrow$	54.8 $\downarrow$	82.0	24.0 $\Downarrow$
<i>Effect of Selected High-value Data</i>								
- Medical Caption (stage 3)	2M	52.9 $\downarrow$	66.3 $\downarrow$	82.7	61.7	54.1 $\downarrow$	81.4 $\downarrow$	26.8
- Medical Caption (stage 1&2)	4.8M	53.8	67.4	83.3	60.6 $\downarrow$	54.9 $\downarrow$	81.5 $\downarrow$	26.7
- Synthetic Multimodal	1.1M	52.1 $\downarrow$	67.4	81.6 $\downarrow$	60.1 $\downarrow$	56.5	71.6 $\Downarrow$	26.1
- Medical CoT-Text	162K	52.6 $\downarrow$	66.3 $\downarrow$	83.6	61.1	53.6 $\Downarrow$	81.6 $\downarrow$	25.3 $\downarrow$

## 5.6 Ablation Study on Data Composition

To gain insights into the impact of different data types during the medical instruction tuning stage, we conducted an ablation study by selectively removing specific categories of training data, as shown in Table 10.

Overall, all four categories of training data outlined in §3.1.3, namely, {Medical, General}  $\times$  {Multimodal, Text}, collectively contribute to LINGSHU’s medical multimodal task-solving ability, even when some data sources are not explicitly tailored for medical domain. Notably, medical textual data emerges as the most critical component; the removal of merely 173K samples leads to substantial performance degradation across five of the seven evaluated tasks.

Each data type contributes differently depending on the task. For instance, the exclusion of medical multimodal data (- *Medical Multimodal*) results in pronounced performance declines on benchmarks such as SLAKE, PathVQA, and OmniMedVQA. This likely reflects the heavy reliance of these datasets on histopathology and X-ray imagery, which are well represented in the medical multimodal training corpus. In contrast, removing general multimodal data (- *General Multimodal*) has a more adverse effect on tasks like MMMU-Med and PMC-VQA. These tasks appear to require a nuanced understanding of medical content presented in complex formats such as public health charts—knowledge that is underrepresented in current medical multimodal datasets.

We hypothesize that such specialized knowledge can be effectively supplemented through the inclusion of general-domain multimodal and textual data, thereby compensating for existing limitations in domain-specific resources. These observations reinforce our hypothesis: *acquiring comprehensive medical knowledge for multimodal problem solving should not rely exclusively on medical multimodal data*. Rather, **it necessitates the integration of medical text and general-domain data to develop a more robust and generalizable model**.

Beyond the overall data composition, we examine the contributions of several high-value data components:

- **Medical Caption (Stage 3):** As mentioned in §3.1.3, we retain a subset of high-quality medical caption data during Stage 3 to mitigate local view bias. Experimental results indicate substantial performance degradation on benchmarks such as MMMU-Med, VQA-RAD, PMC-VQA, and OmniMedVQA upon its removal. This underscores the continued importance of incorporating high-quality caption data in the final stages of instruction tuning.
- **Medical Caption (Stage 1&2):** To assess the role of early-stage training, we bypass Stages 1 and 2 and directly fine-tune the model using Stage 3 instruction data. This variant leads to moderate performance drops on PathVQA, PMC-VQA, and OmniMedVQA, suggesting that early-stage alignment remains a critical step in the training pipeline.
- **Synthetic Multimodal Medical data:** Excluding all synthetic multimodal data, as defined in §2.2, yields notable declines in performance across MMMU-Med, SLAKE, PathVQA, and especially OmniMedVQA,

which includes a broad spectrum of medical imaging modalities. These findings highlight the essential role of synthetic data in enriching modality diversity and compensating for gaps in existing datasets through the infusion of additional medical knowledge.

- **Distilled Textual Medical Reasoning Data:** Building on prior evidence of the significance of textual data, we further isolate the impact of distilled medical reasoning texts. Results reveal that removing only this subset produces performance drops on five out of seven tasks, nearly equivalent to removing all medical textual data. We attribute this effect to the dense medical knowledge encoded within chain-of-thought (CoT) reasoning traces. This finding suggests a promising direction for future research: developing diverse, high-quality corpora of medical reasoning texts to systematically enhance medical expertise acquisition.

## 5.7 Results per Modality on OmniMedVQA

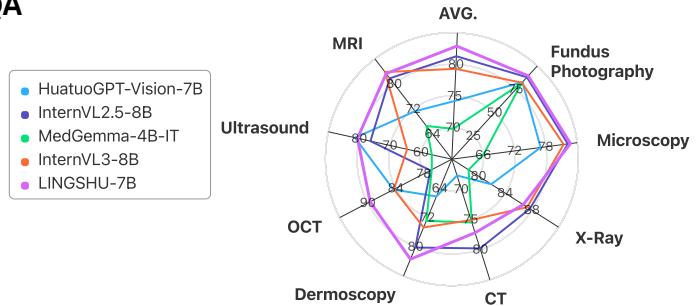
The analysis in §5.2 demonstrates the effectiveness of LINGSHU in diverse medical tasks. Here, we investigate how these MLLMs perform in terms of different medical modalities. Figure 9 benchmarks LINGSHU-7B against the four strongest baselines based on their average performance in medical multimodal tasks, *i.e.*, InternVL-3-8B, InternVL-2.5-8B, HuatuoGPT-V-7B, and MedGemma-4B-IT, on OmniMedVQA.

Across the eight medical imaging modalities evaluated, LINGSHU either matches or exceeds the competitors on most fronts. It achieves leading performance in microscopy, MRI, dermoscopy, and OCT modalities that demand precise recognition of micro-textural and meso-structural cues; this suggests that LINGSHU is particularly effective at capturing both high-frequency detail and anatomical regularities. LINGSHU also delivers promising performance on ultrasound and fundus photography, underscoring its robustness in domains characterized by speckle noise (ultrasound) and fine vascular patterns (fundus). In CT and X-ray, LINGSHU trails the best baseline by a modest margin. Overall, LINGSHU establishes itself as a modality-agnostic yet high-resolution specialist, outperforming larger competitors despite its comparatively compact parameter budget. The gaps identified in modalities like CT and X-ray offer targeted opportunities for improving LINGSHU in the future.

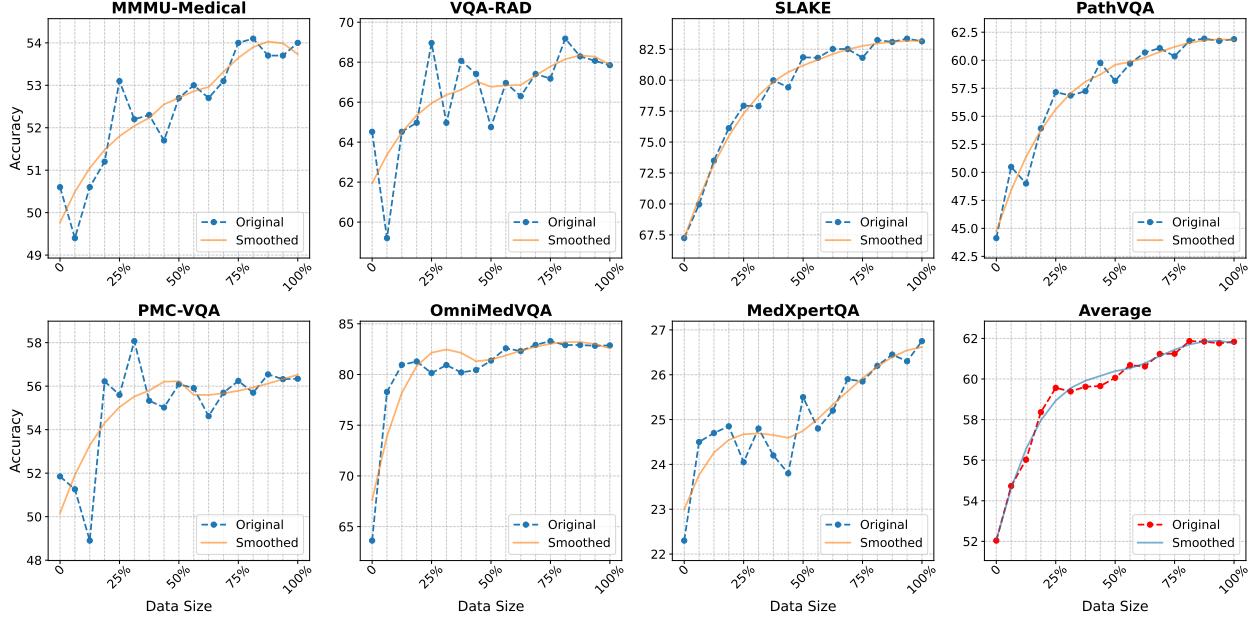
## 5.8 Analysis of Data Volume Scaling

Figure 10 traces accuracy as the number of training samples is progressively expanded from 0% to 100% of its full size. A common pattern emerges across all seven benchmarks: accuracy rises sharply with the first quarter of additional data, then continues to climb—albeit with diminishing marginal gains—until it converges near the 75–100% mark. The smoothed curves reveal a near-monotonic trajectory that is partially obscured by the volatility of the raw scores, confirming that the overall trend is robust.

Two further observations merit emphasis. First, tasks that are intrinsically image-centric and less text-intensive (*e.g.*, SLAKE and PathVQA) saturate earlier than concept-heavy challenges such as MedXpertQA, which still registers steady improvements at the upper end of the data scale. Second, the aggregate plot mirrors the individual tasks, climbing from roughly 52% to 62% accuracy, thereby quantifying a clear data scaling for training medical MLLMs. Taken together, these findings highlight the decisive role of large-scale data acquisition while also signaling an eventual plateau, after which further progress will likely depend on qualitative data enrichment or architectural innovation rather than sheer volume alone.



**Figure 9** Performance of different models across different medical modalities on OmniMedVQA.



**Figure 10** The impact on downstream task performance when increasing the scale of training data.

## 6 Case Studies

This section presents illustrative case studies that showcase LINGSHU in various clinical applications, such as medical diagnosis and report generation. All examples are produced with the LINGSHU-32B model and, when feasible, compared with equivalently sized baseline models. The resulting comparisons elucidate the tangible advantages LINGSHU can offer to both healthcare professionals and patients.

### 6.1 Visual Question Answering across Different Medical Imaging Modalities

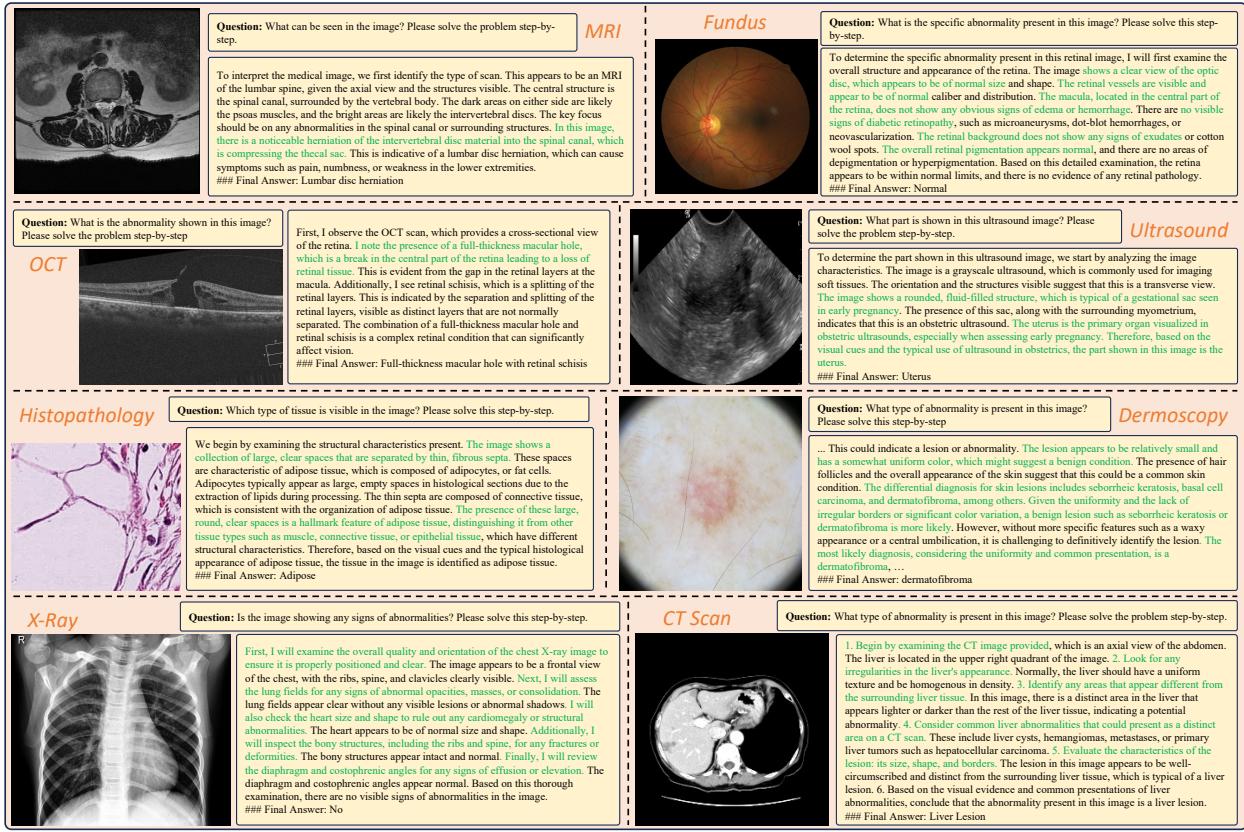
Figure 11 presents the medical VQA cases of LINGSHU across eight distinct medical modalities. The results highlight LINGSHU is capable of transparently reasoning its process to derive a reliable decision.

**Reasoning trajectory.** LINGSHU traces a diagnostic path that mirrors standard clinical workflows. In the chest-radiograph example, for instance, it inspects anatomical regions in the canonical order: first the “*lungs*,” then the “*heart*,” followed by the “*bony structures*,” and finally the “*diaphragm*” and “*costophrenic angles*.” This systematic progression ensures comprehensive coverage of relevant findings.

**Decision refinement.** Rather than issuing a final diagnosis immediately, the model incrementally constrains the differential. In the dermoscopy case, the initial hypothesis set included “*seborrheic keratosis*,” “*basal-cell carcinoma*,” and “*dermatofibroma*.” Continued analysis reduced the options to the first and third, before converging on “*dermatofibroma*” after incorporating additional visual cues. This step-wise narrowing underscores the reliability and clinical plausibility of LINGSHU’s decision-making process.

### 6.2 Medical Diagnosis

Figure 12 compares the diagnostic reasoning of LINGSHU with that of the second-best model, InternVL3, on a medical diagnosis case. Guided by the dome-shaped retinal elevation observed in the image, LINGSHU initially considers two differential diagnoses, “*macular edema*” and “*macular hole*,” and subsequently converges on “*macular edema*” after incorporating additional clinical cues such as subretinal fluid accumulation and cystoid spaces. InternVL3, by contrast, issues an incorrect verdict with unwarranted confidence, neither assessing alternative hypotheses nor integrating ancillary visual evidence. This discrepancy aligns with the broader modality analysis in §6.1, highlighting LINGSHU’s more clinically grounded and cautious diagnostic strategy, resulting in more dependable outcomes.



**Figure 11** Case studies on LINGSHU in visual question answering across various medical imaging modalities.

### 6.3 Medical Knowledge

Beyond its proficiency in clinical image interpretation, LINGSHU exhibits sustained mastery of specialised biomedical knowledge. As depicted in Figure 13, LINGSHU clearly interprets the intricate processes occurring within synapses. Notably, it precisely identifies the synapse type as noradrenergic, based on a thorough understanding of neurotransmitter and receptor interactions. This is in stark contrast to the original Qwen2.5-VL model, which, while adept at describing synaptic processes, fails to accurately classify them, a crucial piece of information. Moreover, the Qwen2.5-VL produces a great amount of incoherent or irrelevant content, a sign of unreliability in its understanding and application of medical knowledge.

### 6.4 Public Health

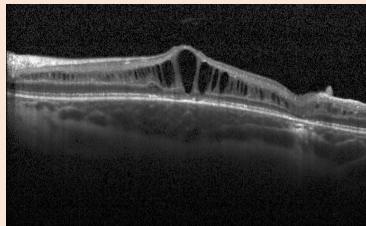
As shown in Figure 14, a public health case, LINGSHU precisely extracts the relevant elements from the cohort-study table, identifying incidence rates of 58 per 100,000 in the smoking (exposed) cohort and 5 per 100,000 among non-smokers. It then applies the standard formula, subtracting the latter from the former to obtain an attributable risk of 53 per 100,000. This example underscores LINGSHU’s capacity to interpret tabular data accurately, implement appropriate public-health calculations, and articulate the population-level implications of those results.

### 6.5 Report Generation

Figure 15 presents an X-ray report produced by LINGSHU. It rigorously follows the conventional radiological structure, *i.e.*, distinct “Findings” followed by an “Impression.” Within the Findings, it correctly delineates suspected abnormalities involving both the cardiac silhouette and pulmonary parenchyma. Furthermore, LINGSHU exhibits clinical insight by appending actionable recommendations, such as obtaining a follow-up radiograph after diuretic therapy. These results indicate that the model not only detects radiographic

## Diagnosis

**Question:** What is the abnormality shown in this image?



### LINGSHU:

Upon examining the image, we notice a notable elevation in the central region of the retina, which is indicative of a macular abnormality. The macula is the central part of the retina responsible for sharp vision, and any elevation or distortion here can suggest pathology.

The key features to look for in this context include the presence of a dome-shaped elevation, which is often associated with macular edema or a macular hole. Macular edema is characterized by fluid accumulation in the macula, leading to swelling and elevation. In this image, the central elevation and the appearance of the retinal layers suggest fluid accumulation beneath the retina, which is consistent with macular edema. This is further supported by the presence of cystoid spaces within the retina, which are typical of fluid-filled cysts seen in cystoid macular edema.

Therefore, based on the visual cues of retinal elevation and cystoid spaces, the specific abnormality shown in this image is macular edema.

### Final Answer: Macular edema

### InternVL3:

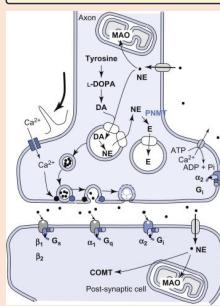
To analyze the specific abnormality shown in this image, let's break it down step by step:

1. \*\*Analyze the Layers\*\*: - The retina has several distinct layers, and each layer has a specific appearance in an OCT scan. The layers include the retinal nerve fiber layer, ganglion cell layer, inner plexiform layer, inner nuclear layer, outer plexiform layer, outer nuclear layer, and the retinal pigment epithelium (RPE).
2. \*\*Look for Abnormalities\*\*: - In this image, there is a noticeable disruption in the normal architecture of the retina. Specifically, there is a large, dark area in the central part of the scan, which suggests a significant structural change.
3. \*\*Interpret the Findings\*\*: - The dark area in the center is indicative of a full-thickness defect in the retina. This is characteristic of a macular hole, which is a small break in the macula, an area in the retina responsible for central vision.
4. \*\*Conclusion\*\*: - The specific abnormality shown in this image is a macular hole. This condition can cause blurred or distorted central vision and is often associated with aging, but it can also be caused by trauma or other eye conditions.

**Figure 12** The comparison between LINGSHU and InternVL3 (the second-best MLLM in Table 6) on medical diagnosis. The green parts are some important reasoning during the generation of LINGSHU. The red part indicates the thoughtless reasoning of InternVL3.

## Medical Knowledge

**Question:** What is this?



### LINGSHU:

To determine what is depicted in the image, I will analyze the visual elements and the context provided. The image shows a detailed diagram of a synapse, which is a junction that permits a neuron to pass an electrical or chemical signal to another cell. Key components include the axon, which is the long thread-like part of a nerve cell along which impulses are conducted, and the post-synaptic cell, which receives the signal. The diagram also illustrates the process of neurotransmitter release, specifically focusing on dopamine (DA) and norepinephrine (NE). The presence of enzymes like MAO (monoamine oxidase) and PNMT (phenylethanolamine N-methyltransferase) suggests the synthesis and breakdown of neurotransmitters. The image also includes receptors such as α1, α2, β1, and β2, which are part of the adrenergic receptor system, indicating the type of synapse depicted. Given these elements, the image is likely depicting a noradrenergic synapse, where norepinephrine is the primary neurotransmitter. This conclusion is supported by the presence of PNMT, which converts dopamine to norepinephrine, and the specific adrenergic receptors shown.

### Final Answer: noradrenergic synapse

### Qwen2.5-VL:

This diagram illustrates the process of neurotransmitter synthesis, storage, release, and degradation in neurons, specifically focusing on catecholamine neurotransmitters such as dopamine (DA) and norepinephrine (NE). Here's a breakdown of what each part represents:

1. Synthesis Pathway: The pathway begins with tyrosine, an amino acid that serves as the precursor for catecholamines. Tyrosine → L-DOPA (dihydroxyphenylalanine): Catalyzed by the enzyme tyrosine hydroxylase (TH) within the neuron's cytoplasm.
2. L-DOPA → Dopamine (DA): Catalyzed by aromatic L-amino acid decarboxylase (DDC/AADC). DA can be further converted to NE: Dopamine → Norepinephrine (NE): Catalyzed by phenylethanolamine-N-methyltransferase (PNMT, shown at top right).

...  
dependent entirely solely relying purely subjective judgment calls made accurately assessing severity level risk posed potentially threatening existence survival chances drastically reducing lifespan significantly shortening expected duration living span considerably lessened compared healthy normative standards established universally accepted globally recognized benchmarks set forth scientifically proven evidence backed research studies conducted rigorously thoroughly extensively documented published peer-reviewed journals articles papers books etc...

**Figure 13** Case study of LINGSHU in explaining medical knowledge, where the red part is the incoherent or irrelevant content generated by Qwen2.5-VL.

anomalies with precision but also furnishes contextually appropriate management advice, reflecting the holistic practices of experienced radiologists.

## 6.6 Patient-Doctor Dialogue

Figure 16 showcases LINGSHU’s text-understanding ability prowess within a real-world consultation scenario. The model interprets the patient’s concerns, proposes plausible differential diagnoses, directs the individual to the relevant clinical specialty, and specifies suitable diagnostic tests. Such performance underscores LINGSHU’s promise as an effective assistant for medical consultations and broader healthcare support.

## 7 Related work

The rapid development of multimodal large language models (MLLMs) has achieved tremendous improvements across various domains, which has sparked widespread interest in their applications within the medical field and driven significant progress in this area, highlighting their enormous potential within healthcare scenarios (Tian et al., 2023; AlSaad et al., 2024). Early endeavors attempt to combine LLMs with specialized medical visual encoders through a linear transformation layer to establish vision-language alignment for medical image understanding and analysis (Li et al., 2023a; Moor et al., 2023; Liu et al., 2023; Zhang et al., 2024a). Building upon this foundation, subsequent research has adopted similar architectural paradigms for developing medical MLLMs. These efforts incorporate various strategies, such as constructing more comprehensive training datasets (Ikezogwo et al., 2023; Chen et al., 2024c; Li et al., 2025b; Hamamci et al.,

## Public Health

	case	Non-case	total	incidence (/ 100 thousands)
exposure group	58	99942	100000	58
non-exposed group	5	99995	100000	5

**Question:** The results of a prospective cohort study of smoking and lung cancer were as follows. The attributable risk(AR) was

To calculate the attributable risk (AR) from the given data, we need to follow these steps:

1. **Identify the incidence rates in both groups\*\*:**

- The incidence rate in the exposure group (smokers) is 58 per 100,000.
- The incidence rate in the non-exposure group (non-smokers) is 5 per 100,000.

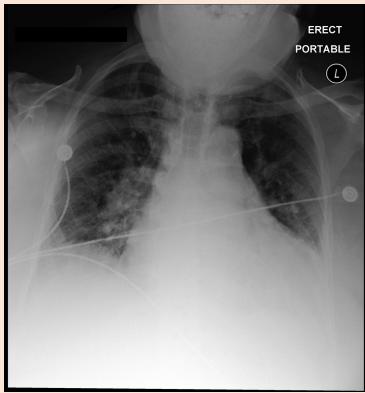
2. **Calculate the attributable risk (AR)\*\*:**

- AR is calculated as the difference between the incidence rate in the exposed group and the incidence rate in the non-exposed group.
- AR = Incidence rate in exposed group - Incidence rate in non-exposed group
- AR = 58 - 5
- AR = 53

Therefore, the attributable risk (AR) is 53 per 100,000. This means that for every 100,000 people, there are 53 more cases of lung cancer in smokers compared to non-smokers.

**Figure 14** Case study of LINGSHU in dealing with a public health problem.

## Report generation



Please generate a medical report for this image in the format of Findings and Impression. Make it as detail as possible.

**Findings:**

The heart size is mildly enlarged, but is likely accentuated secondary to the AP view. The mediastinal and hilar contours are unremarkable.

The lung volumes are low, and there is a diffuse bilateral interstitial abnormality, likely representing pulmonary edema.

There is no large pleural effusion or pneumothorax.

**Impression:**

1. Diffuse bilateral interstitial abnormality likely representing pulmonary edema. If clinically indicated, short term follow-up radiographs following diuresis can be performed to document resolution.

2. No focal consolidation.

**Figure 15** Case study of LINGSHU in generating medical report.

2025), designing sophisticated training recipes (Nath et al., 2025; Wang et al., 2024a), employing efficient fine-tuning techniques (Lin et al., 2025), integrating mixture-of-experts mechanisms (He et al., 2024), and leveraging reinforcement learning (Lai et al., 2025; Pan et al., 2025), to improve performance across diverse medical tasks. In parallel, widely used proprietary models have integrated medical knowledge into their platforms, such as Med-Gemini (DeepMind, 2024b) and Med-PaLM (Singhal et al., 2023a,b), and achieve strong performance on diverse medical tasks (Xie et al., 2024; Saab et al., 2024; Yang et al., 2024; Aydin and Karaarslan, 2025; Arora et al., 2025). Beyond general medical MLLMs, many studies also explore specialized MLLMs tailored to particular diseases or clinical contexts to support better targeted downstream applications, such as models developed specifically for pathology (Lu et al., 2024; Wang et al., 2024c; Zhao et al., 2024a; Seyfioglu et al., 2025), radiology (Hyland et al., 2024; Christensen et al., 2024; Shui et al., 2025; Zambrano Chaves et al., 2025; Pai et al., 2025; Tanno et al., 2025), ophthalmology (Deng et al., 2024), pan-cancer analysis (Keyl et al., 2025), specific cancer types (Sammut et al., 2022; Pai et al., 2024; Niu et al., 2025), etc. Furthermore, building on the capabilities of medical MLLMs, researchers have initiated the development of medical agents that extend model functionality through integration with external medical tools, as illustrated by systems like AgentClinic (Schmidgall et al., 2025), Agent Hospital (Li et al., 2025a), AIME (DeepMind, 2024a), and MedAgent-Pro (Wang et al., 2025b). In contrast, our work focuses on developing a multimodal foundation model that accommodates diverse medical modalities and supports tasks such as question answering, diagnosis, report generation, and diverse clinical applications.

## Medical Support

**Question:** hi my nine year old son had a cough and flu symptoms three months ago and the chesty sounding cough and green phlegm still remains. it did seem to get better but never totally went and has now picked up again...he has only ever had antibiotics once in his live which suggests he his generally fitand well and active...never short of breath or weezy...so why would this be....

Based on the information provided, it sounds like your son may be experiencing a persistent respiratory issue. [Here are some potential causes and recommendations for next steps](#):

1. **Chronic bronchitis:** This is a condition where the airways become inflamed and produce excess mucus, leading to a persistent cough and green phlegm. It is often caused by repeated respiratory infections or exposure to irritants like cigarette smoke.
  2. **Allergic bronchitis:** This is a type of bronchitis that is triggered by allergies. It can cause a persistent cough and green phlegm, along with other allergy symptoms like sneezing and itchy eyes.
  3. **Asthma:** While asthma typically causes wheezing and shortness of breath, it can also cause a persistent cough and green phlegm in some cases.
  4. **Sinusitis:** This is an inflammation of the sinuses that can cause a persistent cough and green phlegm. It is often caused by a bacterial or viral infection.
  5. **Bronchiectasis:** This is a condition where the airways become damaged and scarred, leading to a persistent cough and green phlegm. It is often caused by repeated respiratory infections or exposure to irritants.
- To determine the underlying cause of your son's persistent cough and green phlegm, [it is recommended that you schedule an appointment with a pediatrician or a pulmonologist](#). They can perform a physical examination, review your son's medical history, and order appropriate tests (such as a chest X-ray or a lung function test) to help determine the underlying cause. Additionally, they can provide recommendations for treatment and management of the condition.

**Figure 16** Case study of LINGSHU in a real-life medical support.

## 8 Conclusion, Limitations, and Future Work

### 8.1 Conclusion

In this work, we introduce LINGSHU, a domain-specialized multimodal foundation model designed for medical AI, addressing key challenges in medical data construction, model training, and evaluation. By unifying large-scale, high-quality medical and general-domain corpora with a comprehensive data curation pipeline, encompassing image captions, QA pairs, and CoT annotations, LINGSHU acquires an expansive and diverse knowledge base. Our multi-stage training paradigm progressively enhances the model’s medical understanding and problem-solving skills. Besides, we also explore the potential of incorporating advanced reinforcement learning with verifiable rewards to further boost performance. Beyond model development, we emphasize the importance of evaluation standardization in the medical domain, and introduce MEDEVALKIT, a unified framework that consolidates multiple multimodal and textual benchmarks while enforcing standardized evaluation. This toolkit enables fair, reproducible, and transparent comparisons across models. Extensive experiments confirm the model’s superiority. Across a broad spectrum of medical VQA and report generation tasks, LINGSHU consistently surpasses the open-source baselines and narrows the gaps to proprietary models. Overall, LINGSHU and MEDEVALKIT provide a high-performing model, a robust evaluation toolkit, and empirically grounded guidelines on data curation, staged training, and evaluation. These contributions represent a step forward in aligning MLLMs with real-world medical applications.

### 8.2 Limitations

Despite LINGSHU’s promising performance across various medical tasks, several limitations remain:

**Data Quality and Diversity:** While extensive medical multimodal and textual data are collected and synthesized, yielding notable gains. Dataset quality and diversity remain limited. Open-source medical multimodal data often exhibit low annotation accuracy, poor image resolution, and uneven modality distribution. Additionally, in the absence of full expert medical supervision, many samples are generated using proprietary models, which may introduce severe hallucinations and factual errors. Despite rigorous quality control and manual verification, such issues persist and may hinder the model’s generalizability to downstream tasks.

**Model Performance and Generalization:** Although LINGSHU achieves promising results across several medical benchmarks, particularly in question answering and report generation, it still underperforms compared to state-of-the-art proprietary models. Moreover, its generalization to broader and more diverse medical tasks remains insufficiently explored. Enhancing its capabilities through the inclusion of additional downstream

---

scenarios and modality types is promising but entails substantial challenges in data integration, preprocessing, and model design.

**Training Paradigm and Reinforcement Learning:** While we validate the effectiveness of our data strategy and training paradigm through ablation studies, the optimal data mixture and training configuration remain insufficiently explored. Our exploration of RLVR in medical settings is preliminary; although it offers moderate results in downstream tasks, the overall impact is limited, and a deeper understanding of its effective application in medical contexts is still needed.

### 8.3 Future Work

Building on LINGSHU, we identify several directions for future work to address current limitations and further enhance the model’s capabilities:

**High-Quality Medical Data Construction:** Given the scarcity of high-quality multimodal medical data, it is crucial to increase efforts in constructing and curating more diverse, enriched, and high-quality medical image-text datasets. A key challenge is that medical data quality often requires expert evaluation, which is resource-intensive and infeasible to scale. To address this, it is essential to build robust data generation and quality control systems, including specialized assessment models and fine-grained data synthesis pipelines. Incorporating human-in-the-loop mechanisms for iterative refinement may further improve the efficiency and reliability of the data production process.

**Toward Comprehensive Medical Multimodal Benchmarking:** Current medical multimodal benchmarks fall short of capturing real-world clinical complexity, limiting their ability to evaluate model utility in practical downstream tasks. The recent release of HealthBench (Arora et al., 2025) by OpenAI—a realistic, open-ended evaluation framework for assessing LLMs across professional dimensions—highlights the need for similarly comprehensive and application-aligned benchmarks in the medical multimodal domain.

**Model Expansion:** Exploring advanced model architectures tailored to medical multimodal scenarios is essential. Medical data encompasses not only general formats of image and video sequence but also 3D imaging (*e.g.*, CT, MRI), ultra-high-resolution images (*e.g.*, histopathology WSI), and molecular, protein, and genomic data. Although existing frameworks (DeepMind, 2024b) can be adapted to these modalities through optimized data pipelines, such adaptations may introduce information loss and limit the model’s ability to capture subtle variations in medical modalities. Future work will focus on extending LINGSHU to natively support WSI, 3D imaging, and omics data, enabling seamless integration and comprehensive understanding across diverse medical modalities.

**Post-Training Techniques:** Further exploration of post-training techniques is essential, particularly in adapting reinforcement learning for medical MLLMs to better align outputs with task-specific goals and contextual demands. Unlike the logic-driven nature of mathematical or programming reasoning, medical reasoning is predominantly knowledge-driven, relying on domain expertise and clinical experience. This distinction underscores the necessity for tailored reinforcement learning approaches in medical contexts, including the use of specialized reward functions, custom reward models, and process-level supervision.

**Evaluation:** Currently, MED EVALKIT primarily assesses MLLM capabilities within evaluation metrics from general domains. We have attempted to incorporate certain domain-specific metrics into evaluations for particular tasks, such as report generation. In future work, it is crucial to introduce medical-specific evaluation measures, such as the Concordance Index (c-index), Clinical Efficacy Score, and Decision Curve Analysis, to assess model performance in real medical applications. Additionally, exploring methodologies for integrating human expert evaluation to enhance model credibility and safety indicates another promising direction for future investigation.

**Roadmap of Lingshu:** This work marks our initial effort in the medical domain. Future developments will focus on enhancing LINGSHU through improved and diversified medical datasets and benchmarks, exploring more effective post-training strategies tailored to medical tasks, scaling to larger model sizes, and optimizing performance for specific applications such as report generation, diagnosis, and staging prediction. We also plan to expand language support and develop efficient, task-oriented agent systems based on LINGSHU to address the diverse needs of real-world clinical scenarios.

---

## References

- Walid Al-Dhabyani, Mohammed Gomaa, Hussen Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. <https://www.sciencedirect.com/science/article/pii/S2352340919312181>.
- Rawan AlSaad, Alaa Abd-alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. Multimodal large language models in health care: Applications, challenges, and future outlook. *J Med Internet Res*, 26, Sep 2024. <https://www.jmir.org/2024/1/e59505>.
- Anthropic. Introducing claude 4, May 2025. <https://www.anthropic.com/news/clause-4>.
- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *ArXiv*, abs/2505.08775, 2025. <https://arxiv.org/abs/2505.08775>.
- Omer Aydin and Enis Karaarslan. Openai chatgpt interprets radiological images: Gpt-4 as a medical doctor for a fast check-up. *ArXiv*, abs/2501.06269, 2025. <https://arxiv.org/abs/2501.06269>.
- Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei JI, Eric Chang, Tackeun Kim, et al. Mimic-ext-mimic-cxr-vqa: A complex, diverse, and large-scale visual question answering dataset for chest x-ray images, 2024. <https://physionet.org/content/mimic-ext-mimic-cxr-vqa/1.0.0/>.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025. <https://arxiv.org/abs/2502.13923>.
- Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. <https://aclanthology.org/W05-0909/>.
- Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(511), 2019. <https://doi.org/10.1186/s12859-019-3119-4>.
- Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*, 2019. [https://ceur-ws.org/Vol-2380/paper\\_272.pdf](https://ceur-ws.org/Vol-2380/paper_272.pdf).
- BUAADreamer. llava-med-zh-instruct-60k dataset, 2024. <https://huggingface.co/datasets/BUAADreamer/llava-med-zh-instruct-60k>.
- Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *ArXiv*, abs/2405.19538, 2024. <https://arxiv.org/abs/2405.19538>.
- Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *ArXiv*, abs/2402.11684, 2024a. <https://arxiv.org/abs/2402.11684>.
- Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. <https://aclanthology.org/2025.naacl-long.182/>.
- Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. Huatuogpt-ii, one-stage training for medical adaption of llms. *ArXiv*, abs/2311.09774, 2023. <https://arxiv.org/abs/2311.09774>.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-01, towards medical complex reasoning with llms. *ArXiv*, abs/2412.18925, 2024b. <https://arxiv.org/abs/2412.18925>.

---

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *ArXiv*, abs/2406.19280, 2024c. <https://arxiv.org/abs/2406.19280>.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *ArXiv*, abs/2501.17811, 2025b. <https://arxiv.org/abs/2501.17811>.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huirong Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhui Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *ArXiv*, abs/2412.05271, 2025c. <https://arxiv.org/abs/2412.05271>.

Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, Mamun Bin Ibne Reaz, and Mohammad Tariqul Islam. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020. <https://ieeexplore.ieee.org/document/9144185>.

Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision-language foundation model for echocardiogram interpretation. *Nature Medicine*, 30:1481–1488, May 2024. <https://doi.org/10.1038/s41591-024-02959-y>.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *The Second Conference on Parsimony and Learning (Recent Spotlight Track)*, 2025. <https://openreview.net/forum?id=d3E3LWmTar>.

Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, et al. The 2024 brain tumor segmentation (brats) challenge: glioma segmentation on post-treatment mri. *ArXiv*, abs/2405.18368, 2024. <https://arxiv.org/abs/2405.18368>.

Google DeepMind. Amie: A research ai system for diagnostic medical reasoning and conversations, Jan 2024a. <https://research.google/blog/amie-a-research-ai-system-for-diagnostic-medical-reasoning-and-conversations/>.

Google DeepMind. Advancing medical ai with med-gemini, May 2024b. <https://research.google/blog/advancing-medical-ai-with-med-gemini/>.

Google DeepMind. Gemini: A family of highly capable multimodal models. *ArXiv*, abs/2312.11805, 2025a. <https://arxiv.org/abs/2312.11805>.

Google DeepMind. Gemini 2.5: Our most intelligent ai model, Mar 2025b. <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking>.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*, abs/2501.12948, 2025. <https://arxiv.org/abs/2501.12948>.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *ArXiv*, abs/2409.17146, 2024. <https://arxiv.org/abs/2409.17146>.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. <https://doi.org/10.1093/jamia/ocv080>.

---

Zhuo Deng, Weihao Gao, Chucheng Chen, Zhiyuan Niu, Zheng Gong, Ruiheng Zhang, Zhenjie Cao, Fang Li, Zhaoyi Ma, Wenbin Wei, and Lan Ma. Ophglm: An ophthalmology large language-and-vision assistant. *Artif. Intell. Med.*, 157(C), November 2024. <https://doi.org/10.1016/j.artmed.2024.103001>.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. <https://aclanthology.org/N19-1423/>.

Kexin Ding, Mu Zhou, He Wang, Olivier Gevaert, Dimitris Metaxas, and Shaoting Zhang. A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. *Scientific Data*, 10(1):231, 2023. <https://doi.org/10.1038/s41597-023-02125-y>.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM '24*, page 11198–11201, New York, NY, USA, 2024. Association for Computing Machinery. <https://doi.org/10.1145/3664647.3685520>.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Diewuke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *ArXiv*, abs/2407.21783, 2024. <https://arxiv.org/abs/2407.21783>.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *ArXiv*, abs/2003.10778, 2020. <https://arxiv.org/abs/2003.10778>.

Sheng Gan. Bccd dataset, 2018. [https://github.com/Shenggan/BCCD\\_Dataset](https://github.com/Shenggan/BCCD_Dataset).

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. <https://zenodo.org/records/12608602>.

Lidia Garrucho, Kaisar Kushibar, Claire-Anne Reidel, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier del Riego, Alessandro Catanese, Katarzyna Gwoździewicz, Maria-Laura Cosaka, Pasant M Abo-Elhoda, Sara W Tantawy, Shorouq S Sakrana, Norhan O Shawky-Abdelfatah, Amr Muhammad Abdo Salem, Androniki Kozana, Eugen Divjak, Gordana Ivanac, Katerina Nikiforaki, Michail E Klontzas, Rosa García-Dosdá, Meltem Gulsun-Akpınar, Oğuz Lafçı, Ritse Mann, Carlos Martín-Isla, Fred Prior, Kostas Marias, Martijn P A Starmans, Fredrik Strand, Oliver Díaz, Laura Igual, and Karim Lekadir. A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations. *Scientific Data*, 12(1):453, 2025. <https://doi.org/10.1038/s41597-025-04707-4>.

Google. Medgemma: A gemma 3 variant optimized for medical text and image comprehension, May 2025. <https:////deepmind.google/models/gemma/medgemma>.

Shivanand Gornale and Pooja Patravali. Digital knee x-ray images, 2020. <https://doi.org/10.17632/t9ndx37v5h.1>.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2021. [https://openaccess.thecvf.com/content/CVPR2021W/ISIC/papers/Groh\\_Evaluating\\_Deep\\_Neural\\_Networks\\_Trained\\_on\\_Clinical\\_Images\\_in\\_Dermatology\\_CVPRW\\_2021\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2021W/ISIC/papers/Groh_Evaluating_Deep_Neural_Networks_Trained_on_Clinical_Images_in_Dermatology_CVPRW_2021_paper.pdf).

---

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Doga, Omer Faruk Durugol, Weicheng Dai, Murong Xu, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *ArXiv*, abs/2403.17834, 2025. <https://arxiv.org/abs/2403.17834>.

Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Luyang Luo, Huiling Xiang, Xi Lin, Mingxiang Wu, Yifan Peng, George Shih, Ziyang Xu, Xian Wu, Qiong Wang, Ronald Cheong Kin Chan, Varut Vardhanabuti, Winnie Chiu Wing Chu, Yefeng Zheng, Pranav Rajpurkar, Kang Zhang, and Hao Chen. Gsco: Towards generalizable ai in medicine via generalist-specialist collaboration. *ArXiv*, abs/2404.15127, 2024. <https://arxiv.org/abs/2404.15127>.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *ArXiv*, abs/2003.10286, 2020. <https://arxiv.org/abs/2003.10286>.

Yuting He, Guanyu Yang, Jian Yang, Rongjun Ge, Youyoung Kong, Xiaomei Zhu, Shaobo Zhang, Pengfei Shao, Huazhong Shu, Jean-Louis Dillenseger, Jean-Louis Coatrieux, and Shuo Li. Meta grayscale adaptive network for 3d integrated renal structures segmentation. *Medical Image Analysis*, 71:102055, 2021. <https://www.sciencedirect.com/science/article/pii/S1361841521001018>.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. <https://openreview.net/forum?id=d7KBjmI3GmQ>.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimed-vqa: A new large-scale comprehensive evaluation benchmark for medical lvml. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22170–22183, June 2024. [https://openaccess.thecvf.com/content/CVPR2024/papers/Hu\\_OmniMedVQA\\_A\\_New\\_Large-Scale\\_Comprehensive\\_Evaluation\\_Benchmark\\_for\\_Medical\\_LVLM\\_CVPR\\_2024\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2024/papers/Hu_OmniMedVQA_A_New_Large-Scale_Comprehensive_Evaluation_Benchmark_for_Medical_LVLM_CVPR_2024_paper.pdf).

hw hwei. Medthoughts-8k dataset, 2025. <https://huggingface.co/datasets/hw-hwei/MedThoughts-8K>.

Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. Maira-1: A specialised large multimodal model for radiology report generation. *ArXiv*, abs/2311.13668, 2024. <https://arxiv.org/abs/2311.13668>.

Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. <https://openreview.net/forum?id=OL2JQo00kq>.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Cheexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. <https://doi.org/10.1609/aaai.v33i01.3301590>.

Saadil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. [https://datasets-benchmarks-proceedings.neurips.cc/paper\\_files/paper/2021/file/c8ffe9a587b126f152ed3d89a146b445-Paper-round1.pdf](https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/c8ffe9a587b126f152ed3d89a146b445-Paper-round1.pdf).

Soroush Javadi and Seyed Abolghasem Mirroshandel. A novel deep learning method for automatic assessment of human sperm images. *Computers in Biology and Medicine*, 109:182–194, 2019. <https://www.sciencedirect.com/science/article/pii/S0010482519301386>.

Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021. <https://www.mdpi.com/2076-3417/11/14/6421>.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *ArXiv*, abs/1909.06146, 2019. <https://arxiv.org/abs/1909.06146>.

---

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–10, 2019. <https://doi.org/10.1038/s41597-019-0322-0>.

Julius Keyl, Philipp Keyl, Grégoire Montavon, René Hosch, Alexander Brehmer, Liliana Mochmann, Philipp Jurmeister, Gabriel Dernbach, Moon Kim, Sven Koitka, Sebastian Bauer, Nikolaos Bechrakis, Michael Forsting, Dagmar Führer-Sakel, Martin Glas, Viktor Grünwald, Boris Hadaschik, Johannes Haubold, Ken Herrmann, Stefan Kasper, Rainer Kimmig, Stephan Lang, Tienush Rassaf, Alexander Roesch, Dirk Schadendorf, Jens T. Siveke, Martin Stuschke, Ulrich Sure, Matthias Totzeck, Anja Welt, Marcel Wiesweg, Hideo A. Baba, Felix Nensa, Jan Egger, Klaus-Robert Müller, Martin Schuler, Frederick Klauschen, and Jens Kleesiek. Decoding pan-cancer treatment outcomes using multimodal real-world data and explainable artificial intelligence. *Nature Cancer*, 6:307–322, February 2025. <https://doi.org/10.1038/s43018-024-00891-1>.

Kimi, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, Ziwei Chen, and Zongyu Lin. Kimi-vl technical report. *ArXiv*, abs/2504.07491, 2025. <https://arxiv.org/abs/2504.07491>.

Oleksandr Kovalyk, Juan Morales-Sánchez, Rafael Verdú-Monedero, Inmaculada Sellés-Navarro, Ana Palazón-Cabanes, and José-Luis Sancho-Gómez. Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific Data*, 9(1):291, 2022. <https://doi.org/10.1038/s41597-022-01388-1>.

Nicholas Kurtansky, Veronica Rotemberg, Maura Gillis, Kivanc Kose, Walter Reade, and Ashley Chow. Isic 2024 - skin cancer detection with 3d-tbp, 2024. <https://kaggle.com/competitions/isic-2024-challenge>.

Woosuk Kwon, Zuhuan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP ’23, page 611–626, New York, NY, USA, 2023. Association for Computing Machinery. <https://doi.org/10.1145/3600006.3613165>.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *ArXiv*, abs/2503.13939, 2025. <https://arxiv.org/abs/2503.13939>.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. <https://doi.org/10.1038/sdata.2018.251>.

Peter Lee, Sébastien Bubeck, and Joseph Petro. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239, 2023. <https://www.nejm.org/doi/full/10.1056/NEJMsr2214184>.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023a. <https://openreview.net/forum?id=GSuP99u2kR>.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents. *ArXiv*, abs/2405.02957, 2025a. <https://arxiv.org/abs/2405.02957>.

Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyan Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, Yanjun Li, Pengcheng Chen, Xiaowei Hu, Zhongying Deng, Yuanfeng Ji, Jin Ye, Yu Qiao, and Junjun He. Gmai-vl & gmai-vl-5.5m: A large vision-language model and a comprehensive multimodal dataset towards general medical ai. *ArXiv*, abs/2411.14522, 2025b. <https://arxiv.org/abs/2411.14522>.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023b. <https://doi.org/10.7759/cureus.40895>.

---

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. <https://aclanthology.org/W04-1013/>.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao, Hui Lin, Yuetong Zhuang, and Beng Chin Ooi. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *ArXiv*, abs/2502.09838, 2025. <https://arxiv.org/abs/2502.09838>.

Wei-Ming Lin. Resized 2015-2019 diabetic retinopathy detection, 2022. <https://www.kaggle.com/datasets/c7934597/resized-2015-2019-diabetic-retinopathy-detection>.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 525–536, Cham, 2023. Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-43993-3\\_51](https://doi.org/10.1007/978-3-031-43993-3_51).

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654, 2021. <https://ieeexplore.ieee.org/document/9434010>.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, June 2024. [https://openaccess.thecvf.com/content/CVPR2024/papers/Liu\\_Improved\\_Baselines\\_with\\_Visual\\_Instruction\\_Tuning\\_CVPR\\_2024\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2024/papers/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.pdf).

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *ArXiv*, abs/2310.17956, 2023. <https://arxiv.org/abs/2310.17956>.

Meng Lou, Hanning Ying, Xiaoqing Liu, Hong-Yu Zhou, Yuqin Zhang, and Yizhou Yu. Sdr-former: A siamese dual-resolution transformer for liver lesion classification using 3d multi-phase imaging. *Neural Networks*, 185:107228, 2025. <https://www.sciencedirect.com/science/article/pii/S0893608025001078>.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Kenji Ikamura, Georg Gerber, Ivy Liang, Long Phi Le, Tong Ding, Anil V Parwani, and Faisal Mahmood. A multimodal generative ai copilot for human pathology. *Nature*, 634:466–473, October 2024. <https://doi.org/10.1038/s41586-024-07618-3>.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301, 2024. [https://openaccess.thecvf.com/content/CVPR2024/papers/Luo\\_FairCLIP\\_Harnessing\\_Fairness\\_in\\_Vision-Language\\_Learning\\_CVPR\\_2024\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2024/papers/Luo_FairCLIP_Harnessing_Fairness_in_Vision-Language_Learning_CVPR_2024_paper.pdf).

Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahra A. Fayad, and Yang Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 0(ja):e210315, 2022. <https://doi.org/10.1148/ryai.210315>.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiansheng Han, Botian Shi, Wenhui Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *ArXiv*, abs/2503.07365, 2025. <https://arxiv.org/abs/2503.07365>.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR, 10 Dec 2023. <https://proceedings.mlr.press/v225/moor23a.html>.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alsejari, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. Bimedix2: Bio-medical expert lmm for diverse medical modalities. *ArXiv*, abs/2412.07769, 2024. <https://arxiv.org/abs/2412.07769>.

Luis Filipe Nakayama, David Restrepo, João Matos, Lucas Zago Ribeiro, Fernando Korn Mallerbi, Leo Anthony Celi, and Caio Saito Regatieri. Brset: A brazilian multilabel ophthalmological dataset of retina fundus photos. *PLOS Digital Health*, 2024. <https://doi.org/10.1371/journal.pdig.0000454>.

---

Loris Nanni, Michelangelo Paci, Florentino Luciano Caetano dos Santos, Heli Skottman, Kati Juuti-Uusitalo, and Jari Hyttinen. Texture descriptors ensembles enable image-based classification of maturation of human stem cell-derived retinal pigmented epithelium. *PLoS One*, 11(2), 2016. <https://doi.org/10.1371/journal.pone.0149399>.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yucheng Tang, Pengfei Guo, Can Zhao, Ziyue Xu, Yufan He, Greg Heinrich, Yee Man Law, Benjamin Simon, Stephanie Harmon, Stephen Aylward, Marc Edgar, Michael Zephyr, Song Han, Pavlo Molchanov, Baris Turkbey, Holger Roth, and Daguang Xu. Vila-m3: Enhancing vision-language models with medical expert knowledge. *ArXiv*, abs/2411.12915, 2025. <https://arxiv.org/abs/2411.12915>.

Msoud Nickparvar. Brain tumor mri dataset, 2021. <https://www.kaggle.com/dsv/2645886>.

Chuang Niu, Qing Lyu, Christopher D. Carothers, Parisa Kaviani, Josh Tan, Pingkun Yan, Mannudeep K. Kalra, Christopher T. Whitlow, and Ge Wang. Medical multimodal multitask foundation model for lung cancer screening. *Nature Communications*, 16, February 2025. <https://doi.org/10.1038/s41467-025-56822-w>.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *ArXiv*, abs/2303.13375, 2023. <https://arxiv.org/abs/2303.13375>.

OpenAI. Gpt-4o system card. *ArXiv*, abs/2410.21276, 2024a. <https://arxiv.org/abs/2410.21276>.

OpenAI. Openai o1 system card. *ArXiv*, abs/2412.16720, 2024b. <https://arxiv.org/abs/2412.16720>.

OpenAI. Introducing gpt-4.1 in the api, Apr 2025a. <https://openai.com/index/gpt-4-1/>.

OpenAI. Introducing o3 and o4-mini, Apr 2025b. <https://openai.com/index/introducing-o3-and-o4-mini/>.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, Miami, Florida, USA, November 2024. Association for Computational Linguistics. <https://aclanthology.org/2024.findings-emnlp.21/>.

Andre G.C. Pacheco, Gustavo R. Lima, Amanda S. Salomão, Breno Krohling, Igor P. Biral, Gabriel G. de Angelo, Fábio C.R. Alves Jr, José G.M. Esgario, Alana C. Simora, Pedro B.C. Castro, Felipe B. Rodrigues, Patricia H.L. Frasson, Renato A. Krohling, Helder Knidel, Maria C.S. Santos, Rachel B. do Espírito Santo, Telma L.S.G. Macedo, Tania R.P. Canuto, and Luíz F.S. de Barros. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221, 2020. <https://www.sciencedirect.com/science/article/pii/S235234092031115X>.

Suraj Pai, Dennis Bontempi, Ibrahim Hadzic, Vasco Prudente, Mateo Sokač, Tafadzwa L. Chaunzwa, Simon Bernatz, Ahmed Hosny, Raymond H. Mak, Nicolai J. Birkbak, and Hugo J. W. L. Aerts. Foundation model for cancer imaging biomarkers. *Nature Machine Intelligence*, 6:354–367, March 2024. <https://doi.org/10.1038/s42256-024-00807-9>.

Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressem, Benjamin H. Kann, Andriy Fedorov, Raymond H. Mak, and Hugo J. W. L. Aerts. Vision foundation models for computed tomography. *ArXiv*, abs/2501.09001, 2025. <https://arxiv.org/abs/2501.09001>.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, Apr 2022. <https://proceedings.mlr.press/v174/pal22a.html>.

Jiazheng Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *ArXiv*, abs/2502.19634, 2025. <https://arxiv.org/abs/2502.19634>.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. <https://aclanthology.org/P02-1040/>.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. 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*, pages 180–189, Cham, 2018. Springer International Publishing. [https://link.springer.com/chapter/10.1007/978-3-030-01364-6\\_20](https://link.springer.com/chapter/10.1007/978-3-030-01364-6_20).

---

Tschandl Philipp, Rosendahl Cliff, and Kittler Harald. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–10, 2018. <https://doi.org/10.1038/sdata.2018.161>.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *ArXiv*, abs/1711.05225, 2017. <https://arxiv.org/abs/1711.05225>.

Johannes Rückert, Louise Bloch, Raphael Brügel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S. Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, Henning Müller, Peter A. Horn, Felix Nensa, and Christoph M. Friedrich. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1), June 2024. <http://dx.doi.org/10.1038/s41597-024-03496-6>.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Housby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of gemini models in medicine. *ArXiv*, abs/2404.18416, 2024. <https://arxiv.org/abs/2404.18416>.

Stephen-John Sammut, Mireia Crispin-Ortuzar, Suet-Feung Chin, Elena Provenzano, Helen A. Bardwell, Wenxin Ma, Wei Cope, Ali Dariush, Sarah-Jane Dawson, Jean E. Abraham, Janet Dunn, Louise Hiller, Jeremy Thomas, David A. Cameron, John M. S. Bartlett, Larry Hayward, Paul D. Pharoah, Florian Markowetz, Oscar M. Rueda, Helena M. Earl, and Carlos Caldas. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*, 601:623–629, January 2022. <https://doi.org/10.1038/s41586-021-04278-5>.

Samuel Schmidgall, Rojin Ziae, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *ArXiv*, abs/2405.07960, 2025. <https://arxiv.org/abs/2405.07960>.

Mehmet Saygin Seyfoglu, Wisdom O. Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. *ArXiv*, abs/2312.04746, 2025. <https://arxiv.org/abs/2312.04746>.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. <https://arxiv.org/abs/2402.03300>.

Liyu Shi, Xiaoyan Li, Weiming Hu, Haoyuan Chen, Jing Chen, Zizhen Fan, Minghe Gao, Yujie Jing, Guotao Lu, Deguo Ma, et al. Ebhi-seg: A novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. *Frontiers in Medicine*, 10:1114673, 2023. <https://doi.org/10.3389/fmed.2023.1114673>.

Zhongyi Shui, Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and Ling Zhang. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. In *The Thirteenth International Conference on Learning Representations*, 2025. <https://openreview.net/forum?id=nYpPAT4L3D>.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620:172–180, August 2023a. <https://doi.org/10.1038/s41586-023-06291-2>.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950, 2023b. <https://doi.org/10.1038/s41591-024-03423-7>.

---

Irene Siragusa, Salvatore Contino, Massimo La Ciura, Rosario Alicata, and Roberto Pirrone. Medpix 2.0: A comprehensive multimodal biomedical data set for advanced ai applications. *ArXiv*, 2025. <https://arxiv.org/abs/2407.02994>.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online, November 2020. Association for Computational Linguistics. <https://aclanthology.org/2020.emnlp-main.117/>.

Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. MedICaT: A dataset of medical images, captions, and textual references. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2112–2120, Online, November 2020. Association for Computational Linguistics. [https://aclanthology.org/2020.findings-emnlp.191/](https://aclanthology.org/2020.findings-emnlp.191).

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *ArXiv*, abs/2503.20752, 2025. <https://arxiv.org/abs/2503.20752>.

Ryutaro Tanno, David G. T. Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, Karan Singhal, Mike Schaeckermann, Rhys May, Roy Lee, SiWai Man, Sara Mahdavi, Zahra Ahmed, Yossi Matias, Joelle Barral, S. M. Ali Eslami, Danielle Belgrave, Yun Liu, Sreenivasa Raju Kalidindi, Shravya Shetty, Vivek Natarajan, Pushmeet Kohli, Po-Sen Huang, Alan Karthikesalingam, and Ira Ktena. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31, February 2025. <https://doi.org/10.1038/s41591-024-03302-1>.

P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi Li, Yunwen Li, David Ma, Yuansheng Ni, Haoran Que, Qiya Wang, Zhoufutu Wen, Siwei Wu, Tyshawn Hsing, Ming Xu, Zhenzhu Yang, Zekun Moore Wang, Junting Zhou, Yuelin Bai, Xingyuan Bu, Chenglin Cai, Liang Chen, Yifan Chen, Chengtuo Cheng, Tianhao Cheng, Keyi Ding, Siming Huang, Yun Huang, Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma, Tianyang Pang, Zhongyuan Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Shanghaoran Quan, Yizhou Tan, Zili Wang, Chenqing Wang, Hao Wang, Yiya Wang, Yubo Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue, Tianyang Zhan, Chun Zhang, Jinyang Zhang, Xiyue Zhang, Xingjian Zhang, Yue Zhang, Yongchi Zhao, Xiangyu Zheng, Chenghua Zhong, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min Yang, Meng Cao, Xiang Yue, Zhaoxiang Zhang, Wangchunshu Zhou, Jiaheng Liu, Qunshu Lin, Wenhai Huang, and Ge Zhang. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *ArXiv*, abs/2502.14739, 2025. <https://arxiv.org/abs/2502.14739>.

Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. <https://huggingface.co/datasets/teknium/OpenHermes-2.5>.

Dianzhe Tian, Shitao Jiang, Lei Zhang, Xin Lu, and Yiyao Xu. The role of large language models in medical image processing: a narrative review. *Quantitative Imaging in Medicine and Surgery*, 14(1):1108, 2023. <https://doi.org/10.21037/qims-23-892>.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. [https://openaccess.thecvf.com/content\\_cvpr\\_2015/papers/Vedantam\\_CIDEr\\_Conensus-Based\\_Image\\_2015\\_CVPR\\_paper.pdf](https://openaccess.thecvf.com/content_cvpr_2015/papers/Vedantam_CIDEr_Conensus-Based_Image_2015_CVPR_paper.pdf).

Santiago Vitale, José Ignacio Orlando, Emmanuel Iarussi, and Ignacio Larrabide. Improving realism in patient-specific abdominal ultrasound simulation using cyclegans. *International Journal of Computer Assisted Radiology and Surgery*, 15(2):183–192, 2020. <https://doi.org/10.1007/s11548-019-02046-5>.

Patrick Wagner, Maximilian Springenberg, Marius Kröger, Rose KC Moritz, Johannes Schleusener, Martina C Meinke, and Jackie Ma. Semantic modeling of cell damage prediction: a machine learning approach at human-level performance in dermatology. *Scientific Reports*, 13(1):8336, 2023. <https://doi.org/10.1038/s41598-023-35370-7>.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *ArXiv*, abs/2504.08837, 2025a. <https://arxiv.org/abs/2504.08837>.

Sheng Wang, Zihao Zhao, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. Interactive computer-aided

---

diagnosis on medical image using large language models. *Communications Engineering*, 3, September 2024a. <https://doi.org/10.1038/s44172-024-00271-8>.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. [https://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Wang\\_ChestX-ray8\\_Hospital-Scale\\_Chest\\_CVPR\\_2017\\_paper.pdf](https://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf).

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *ArXiv*, abs/2403.03640, 2024b. <https://arxiv.org/abs/2403.03640>.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, Fang Wang, Yulong Peng, Junyou Zhu, Jing Zhang, Christopher R. Jackson, Jun Zhang, Deborah Dillon, Nancy U. Lin, Lynette Sholl, Thomas Denize, David Meredith, Keith L. Ligon, Sabina Signoretti, Shuji Ogino, Jeffrey A. Golden, MacLean P. Nasrallah, Xiao Han, Sen Yang, and Kun-Hsing Yu. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634:970–978, October 2024c. <https://doi.org/10.1038/s41586-024-07894-z>.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaoxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *ArXiv*, abs/2503.18968, 2025b. <https://arxiv.org/abs/2503.18968>.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *ArXiv*, abs/2308.02463, 2023. <https://arxiv.org/abs/2308.02463>.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024. <https://doi.org/10.1093/jamia/ocae045>.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *ArXiv*, abs/2504.00993, 2025. <https://arxiv.org/abs/2504.00993>.

Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang, Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang Xie, and Yuyin Zhou. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *ArXiv*, abs/2409.15277, 2024. <https://arxiv.org/abs/2409.15277>.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. In *The Thirteenth International Conference on Learning Representations*, 2025. <https://openreview.net/forum?id=IwgmgidYPS>.

Yiming Xu, Bowen Zheng, Xiaohong Liu, Tao Wu, Jinxiu Ju, Shijie Wang, Yufan Lian, Hongjun Zhang, Tong Liang, Ye Sang, Rui Jiang, Guangyu Wang, Jie Ren, and Ting Chen. Annotated ultrasound liver images. Zenodo, 2022. <https://doi.org/10.5281/zenodo.7272660>. Data set.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations. *ArXiv*, abs/1710.01766, 2017. <https://arxiv.org/abs/1710.01766>.

Zhilin Yan, Kai Zhang, Rong Zhou, Lifang He, Xiang Li, and Lichao Sun. Multimodal chatgpt for medical applications: an experimental study of gpt-4v. *ArXiv*, abs/2310.19061, 2023. <https://arxiv.org/abs/2310.19061>.

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, Eric Wang, Ellery Wulczyn, Fayaz Jamil, Theo Guidroz, Chuck Lau, Siyuan Qiao, Yun Liu, Akshay Goel, Kendall Park, Arnav Agharwal, Nick George, Yang Wang, Ryutaro Tanno, David G. T. Barrett, Wei-Hung Weng, S. Sara Mahdavi, Khaled Saab, Tao Tu, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, Jorge Cuadros, Gregory Sorensen, Yossi Matias, Katherine Chou, Greg Corrado, Joelle Barral, Shravya Shetty, David Fleet, S. M. Ali Eslami, Daniel Tse, Shruthi Prabhakara, Cory McLean, Dave Steiner, Rory Pilgrim, Christopher Kelly, Shekoofeh Azizi, and Daniel Golden. Advancing multimodal medical capabilities of gemini. *arXiv*, abs/2405.03162, 2024. <https://arxiv.org/abs/2405.03162>.

---

Edward Yeo, Yuxuan Tong, Xinyao Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in LLMs. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2025. <https://openreview.net/forum?id=AgtQlhMQOV>.

Sunne Yi. Chest-ctscan dataset, 2021. <https://tianchi.aliyun.com/dataset/93929>.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023. [https://www.cell.com/patterns/pdf/S2666-3899\(23\)00157-5.pdf](https://www.cell.com/patterns/pdf/S2666-3899(23)00157-5.pdf).

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *ArXiv*, abs/2503.14476, 2025. <https://arxiv.org/abs/2503.14476>.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, June 2024. [https://openaccess.thecvf.com/content/CVPR2024/papers/Yue\\_MMMU\\_A\\_Massive\\_Multi-discipline\\_Multimodal\\_Understanding\\_and\\_Reasoning\\_Benchmark\\_for\\_CVPR\\_2024\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2024/papers/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.pdf).

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhamo Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1), April 2025. <http://dx.doi.org/10.1038/s41467-025-58344-x>.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. A generalist vision-language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, August 2024a. <http://dx.doi.org/10.1038/s41591-024-03185-2>.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. <https://aclanthology.org/2025.findings-naacl.51/>.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *ArXiv*, abs/2303.00915, 2025b. <https://arxiv.org/abs/2303.00915>.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. <https://openreview.net/forum?id=SkeHuCVFDr>.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *ArXiv*, abs/2305.10415, 2024b. <https://arxiv.org/abs/2305.10415>.

Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. Rexrank: A public leaderboard for ai-powered radiology report generation. *ArXiv*, abs/2411.15122, 2024c. <https://arxiv.org/abs/2411.15122>.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. Alpacare: Instruction-tuned large language models for medical application, 2025c. <https://arxiv.org/abs/2310.14558>.

---

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, Brian Piening, Carlo Bifulco, Mu Wei, Hoifung Poon, and Sheng Wang. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, 22(1):166–176, November 2024a. <http://dx.doi.org/10.1038/s41592-024-02499-w>.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. <https://aclanthology.org/2024.emnlp-main.836/>.

Jinguo Zhu, Weiyun Wang, Zhe Chen, ZhaoYang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xinguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhui Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *ArXiv*, abs/2504.10479, 2025. <https://arxiv.org/abs/2504.10479>.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *ArXiv*, abs/2501.18362, 2025. <https://arxiv.org/abs/2501.18362>.

# Appendix

## A More Details about Data Curation

### A.1 Model-based response cleaning for patient-doctor dialogue datasets

We show the prompt for model-based response cleaning for patient-doctor dialogue datasets in Table 11.

**Table 11** The prompt used for model-based response cleaning for patient-doctor dialogue datasets.

Response to patient:  
{response}

Revise the above response in a tone of an emphatic health-care agent. You should offer reliable information about medical conditions, but avoid giving explicit diagnosis and prescription. Add suggestions about which kinds of doctors the patient should consult. Include a disclaimer statement saying that it is not a definitive diagnosis. Begin the revised response with "Revised response:".

### A.2 Medical VQA Data Synthesis

We display the prompt of our self-instruct based method for medical VQA data synthesis in Table 12.

**Table 12** The prompt used for synthesizing medical VQA data in our self-instruct based method.

Given a medical image and its description, generate a clinically relevant multiple-choice question that could be asked to a medical professional.

If the given medical image matches its description, your output should be in JSON format with three keys:

- "question": A medically appropriate question related to the image.
- "answer": Correct answer to the question.
- "wrong answers": A list of wrong but relevant answers that could be given by the medical professional.

There should be 1 to 3 options. The format is ["answer1", "answer2", ...]

If the given medical image does not match its description, simply output "Error: the input description does not match the input image."

The question should be directly answerable from the image.

Below are some examples of questions and answers for other images:

{seed\_example\_1}  
{seed\_example\_2}

Now, use the following image and the image caption to generate the question:

Image caption: {caption}

### A.3 Medical reasoning data distillation

We depict the prompts for distilling the CoT reasoning process from GPT4-o into medical open-ended VQA datasets and medical multiple-choice VQA datasets in Figure 13 and Figure 14 respectively. The prompt used

---

for checking the quality of the generated reasoning traces is displayed in Table 15.

**Table 13** The prompt used for annotating CoT reasoning process for open-ended VQA data.

You are a medical expert specializing in interpreting medical images. The user will provide you with the following inputs:

- A medical image
- A question about the image
- The groundtruth answer

Your task is to simulate a step-by-step diagnostic reasoning process that leads to the ground-truth answer. However, you must assume you do not know the ground-truth answer in advance, and you should not reference it explicitly during your reasoning. Please do not give empty output.

- Carefully analyze both the image and the question.
- Reason step by step, as a medical expert would, using visual and textual cues to reach a logical conclusion.
- The reasoning should be concise, clear, and clinically sound.
- After the reasoning, reveal the correct answer by copying the ground-truth answer.

Output in JSON Format:

```
{  
  "reasoning": "<step-by-step reasoning process here>",  
  "answer": "<groundtruth answer>"  
}
```

Question: {question}

Groundtruth answer: {answer}

#### A.4 Prompts for Medical Long-form Caption Synthesis

We detail the prompts used for medical long-form caption synthesis across various stages. The prompt for Stage 3 (factual knowledge annotation) is presented in Table 16. For Stage 4 (annotation with doctor preference), the prompts tailored for different types of medical images are displayed in Tables 17, 18, 19, and 20. Finally, the summarization prompt for Stage 5 is depicted in Table 21. In this Stage 5 prompt, the variable {doctor\_preferred\_features} should be replaced by the doctor-preferred features specified in Tables 22 and 23, according to the type of the source image.

## B More Details about MedEvalKit

The judgment prompt used for GPT-4.1 (OpenAI, 2025a) in the LLM-as-a-Judge module is presented in Table 24.

For the report generation task, we develop a multi-metric evaluation approach that integrates semantic and model-based metrics. Our MEDEVALKIT supports the following metrics, and in this paper, we selectively present a subset of these.

The semantic evaluation metrics:

- **BLEU** (Papineni et al., 2002): BLEU (Bilingual Evaluation Understudy) is a widely adopted metric in machine translation and text generation tasks. It evaluates the quality of generated text by measuring n-gram precision between candidate and reference texts, with scores ranging from 0 to 1. This metric

**Table 14** The prompt used for annotating CoT reasoning process for multiple-choice VQA data.

You are a medical expert specializing in interpreting medical images. The user will provide you with the following inputs:

- A medical image
- A multiple-choice question related to the image
- A list of answer options
- The groundtruth answer

Your task is to simulate a step-by-step diagnostic reasoning process that leads to the ground-truth answer. However, you must assume you do not know the ground-truth answer in advance, and you should not reference it explicitly during your reasoning. Please do not give empty output.

- Carefully analyze both the image and the question.
- Evaluate each answer option logically and methodically.
- Reason step by step, as a medical expert would, using visual and textual cues to reach a logical conclusion.
- The reasoning should be concise, clear, and clinically sound.
- After the reasoning, reveal the correct answer by copying the ground-truth answer exactly.

Output in JSON Format:

```
{  
  "reasoning": "<step-by-step reasoning process here>",  
  "answer": "<groundtruth answer>"  
}
```

Question:

```
{question}
```

Options:

```
{options}
```

Groundtruth answer:

```
{answer}
```

assesses the overlap of contiguous sequences of n words between the generated report and the reference standard.

- **Rouge** (Lin, 2004): ROUGE (Recall-Oriented Understudy for Gisting Evaluation) comprises a set of metrics designed to automatically assess summary quality by comparing generated summaries against reference human-authored summaries. This metric evaluates the overlap between the generated text and reference summaries, focusing particularly on recall-oriented measurements.
- **METEOR** (Banerjee and Lavie, 2005): METEOR (Metric for Evaluation of Translation with Explicit ORdering) is an automated metric designed for machine translation evaluation, which assesses translation quality based on the concept of flexible uni-gram matching between machine-generated translations and human reference translations. This metric incorporates semantic matching principles beyond simple word overlap.
- **CIDEr** (Vedantam et al., 2015): CIDEr (Consensus-based Image Description Evaluation) is a metric that quantifies the similarity between generated sentences and a collection of human-authored reference sentences. This metric specifically measures how well the generated content aligns with human consensus,

**Table 15** The prompt used for checking the quality of generated CoT reasoning traces.

You are a medical expert. The user will provide you with the following inputs:

- A question related to an image.
- A step-by-step diagnostic reasoning process that leads to a predicted answer.
- The groundtruth answer.

Your task is to check whether the predicted answer in the reasoning process is consistent with the ground-truth answer.

If the predicted answer in the reasoning process is inconsistent with the ground-truth answer. Please first output "Inconsistent." Then give a short explanation.

Otherwise, please first output "Consistent." Then give a short explanation.

Question: {question}

Reasoning process: {reasoning}

Groundtruth answer: {answer}

effectively capturing the semantic similarity between machine-generated and reference descriptions.

The model-based evaluation metrics:

- **BERTScore** (Zhang et al., 2020): BERTScore is a neural-based evaluation metric that leverages pre-trained BERT models (Devlin et al., 2019) to assess textual similarity. This metric computes the cosine similarity between BERT embeddings of model-generated radiology reports and ground-truth reference reports, providing a more semantically nuanced evaluation that captures contextual relationships between texts.
- **SembScore** (Smit et al., 2020): SembScore (CheXbert Labeler Vector Similarity) is a domain-specific metric developed for radiology report evaluation. This metric calculates the cosine similarity between vectors of 14 pathological indicators, which are automatically extracted by the CheXbert labeler from both model-generated and ground-truth radiology reports. This specialized metric focuses specifically on assessing the clinical content alignment between generated and reference reports.
- **RadGraph-F1** (Yu et al., 2023): RadGraph-F1 serves as a specialized metric for radiology report evaluation. It quantifies the overlap of clinical entities and their relationships extracted by RadGraph (Jain et al., 2021) from both candidate and reference reports. This metric provides a structured assessment of how well the generated reports capture the underlying clinical relationships present in the reference reports.
- **RadCliQ-v1** (Yu et al., 2023)<sup>7</sup>: RadCliQ is a comprehensive metric designed for evaluating radiology report generation, integrating multiple complementary measures including BLEU, BERTScore, SembScore, and RadGraph-F1 to provide a holistic assessment of report quality.
- **RaTEScore** (Zhao et al., 2024b)<sup>8</sup>: RaTEScore is an entity-aware metric specifically designed for radiology report evaluation. This metric emphasizes critical medical entities, including diagnostic findings and anatomical details, while demonstrating robustness to complex medical synonyms and sensitivity to negation expressions.
- **GREEN** (Ostmeier et al., 2024)<sup>9</sup>: GREEN (Generated Radiology Report Evaluation and Error Notation) is an LLM-based metric designed for evaluating generated radiology reports. This metric leverages large

<sup>7</sup>We implement it using the official codebase at <https://github.com/rajpurkarlab/CXR-Report-Metric>

<sup>8</sup>We implement it using the official codebase at <https://github.com/MAGIC-AI4Med/RaTEScore>

<sup>9</sup>We implement it using the official codebase at <https://github.com/Stanford-AIMI/GREEN>

---

language models to identify and interpret clinically significant errors in both quantitative and qualitative aspects of candidate reports. The evaluation framework enables a comprehensive assessment of clinical accuracy and relevance in generated reports.

Table 25 summarizes the statistics of the evaluation benchmarks utilized in the MEDEVALKIT framework.

**Table 16** The prompt used for the factual knowledge annotation stage of medical long caption synthesis.

You are provided with a biomedical image from a medical dataset, the disease type (or organ name if there is no disease) of the dataset ('Disease or organ'), the medical Knowledge of the disease ('Knowledge'), a coarse caption ('Coarse Caption') of the image and the medical imaging methods ('Modality'). Your task is to answer the following questions based on the image, bounding box (if provided), caption, disease type and disease knowledge, and condense your answers into caption-styled text.

### question1

Give me a detailed description of the image, including type of the image, organs in the image, approximate location of these organs and relevant locations of these organs and any medical devices (if present) visible in the image as detailedly as possible. Note when answering question1:

1. You can use all the meta information I provided. But never mention that this information has been given. You should NEVER mention anything about the bounding box like the contour itself and its outline. Assuming you are introducing to someone who can only see the original image.
2. If there is some statistical information provided, like the damage area or the relative damage ratio, etc., you should mention it in your response.
3. Not all disease knowledge is relevant to this image; only utilize disease knowledge pertinent to the condition depicted in this image for analysis.
4. The coarse caption may not explicitly describe the image, for example, there may appear multiple organs in the caption. You should utilize your knowledge to figure out the most ONE organ and ONE disease to give your description.
5. Do not explain or emphasize your analysis. Do not suggest or indicate any consequent effects of the disease.

### question2 Specify the specific location of the disease or the organ in the image and its relative position to other reference objects in the image. Describe what is unusual in this area indicating the disease (color, texture, size and other features). Note when answering question2:

1. You can use information from 'Coarse Caption' to locate the disease or the organ. If no, you need to locate the disease or organ by yourself.
2. The image may contain multiple bounding boxes, and the contents of these contours may not necessarily represent the affected areas. Therefore, you need to first answer the questions based on the contents within each bounding box. Afterward, analyze the location of the disease based on your answers.
3. Do not use phrase "bounding box" in your response. The bounding box if provided is only a visual aid for your reference. Do not say something correlates with or matches any of provided information. Assuming you are introducing to someone who cannot see the bounding box and the provided textual information. If you want to mention the bounding box, use the position of the bounding box (e.g. the top-left side of the image). Do not contain phrases "caption", "medical annotation", "medical knowledge".
4. Do not say anything that is not needed in your analysis, like introduction of the disease and medical equipments.
5. Do not explain or emphasize your analysis.

### question3 What may be the relationship between the content in the bounding box and other regions (others being cause of the disease/jointly affected by the diseases/one affect the others/relative positional relationships)? Why and is it possible? Note when answering question3:

1. Utilize external knowledge, if possible, to choose relationships and give necessary analysis.
2. You can only give an explanation to your choice within two sentences.
3. Do not summarize what you've said.
4. Do not emphasize your analysis.

### Integrate Information

Describe your answers in a descriptive sentence, not in a "Question-Answer" style. Combine and slightly shorten your answers to the above three questions into a coherent text, keeping as much information of your answers as possible.

Note when integrating information and outputting your response:

1. Don't respond saying you're unable to assist with requests.
2. You should only output your combined and shortened text in the "caption" field.

**Table 17** The prompts used for the annotation with doctor preference stage of medical long caption synthesis for MRI, X-ray, and CT images.

**MRI instruction:**

Write a description of the MRI image to include the following information, if these information are visually discernible from the image:

1. The sequence used (e.g., T1-weighted, T2-weighted, FLAIR, etc.)
2. Plane/orientation of the image (axial, coronal, sagittal)
3. Describe the anatomical structures visible and their normal or abnormal appearance (e.g., brain lobes, ventricles, midline structures).
4. Identify and describe any abnormal findings, shape, location and effects (e.g. compression)
5. Describe signal intensity: hyperintense, hypointense, or isointense

**X-ray instruction:**

Write a description of the X-ray image to include the following information, if these information are visually discernible from the image:

1. Specific area of the body being imaged (e.g. chest area, lungs, knee, femur, etc.)
2. Body part or organ being examined, either fully or partially
3. Anatomical planes (e.g. Axial, Sagittal, and Coronal), specific X-ray imaging projections (e.g. AP, PA, lateral view, oblique view, etc.), and clearly discernible left vs right side (either clearly labeled in the image, or clearly determined from image).
4. Abnormalities that are clearly observed in the image (e.g. fractures, definite osteophytes, effusions, miliary patterns, vascular markings, diaphragm depression, air bronchograms, thickened interlobular septa, mediastinum shifts, deep sulcus sign, etc.), acute trauma signs (e.g. fractures), implants or other foreign bodies.
5. Gender and age-range of patient if clearly discernible (e.g. pediatric vs. adult anatomy)
6. When 2 (or a series of) images of the same patient taken across time are provided, describe the clearly discernible difference.
7. Identify and describe any abnormal findings, shape, location and effects (e.g. unexpected wires, prostheses, etc.)

**CT instruction:**

Write a description of the CT image to include the following information, if these information are visually discernible from the image:

1. The plane/orientation of the image (axial, coronal, sagittal).
2. Contrast usage: whether contrast material is used (contrast-enhanced) or not (non-contrast).
3. Describe the anatomical structures visible and their normal or abnormal appearance (e.g., organs, bones, blood vessels).
4. Identify and describe any abnormal findings including their location, size, shape, and potential effects on surrounding structures.
4. Assess and describe the density or attenuation of the tissues and structures as hypoattenuating (darker), isoattenuating (similar), or hyperattenuating (brighter) relative to surrounding tissues.

**Table 18** The prompts used for the annotation with doctor preference stage of medical long caption synthesis for histopathology, skin lesion, and knee xray images.

**Histopathology instruction:**

Write a description of the histopathology image to include the following information, if these information are visually discernible from the image:

Tissue Type: Identify the tissue or organ from which the sample is taken (e.g., liver, lung, skin).

Staining Technique: Specify the staining method used (e.g., Hematoxylin and Eosin (H&E), PAS, Trichrome).

Cellular Morphology: Describe the appearance of the cells, including size, shape, and organization (e.g., normal, dysplastic, neoplastic).

Tissue Architecture: Assess the arrangement and structural organization of the tissue (e.g., normal, inflamed, fibrotic).

Abnormal Findings: Identify and describe any pathological changes, such as the presence of tumors, necrosis, inflammation, or infections.

Distribution and Patterns: Note the distribution of abnormal findings within the tissue (e.g., focal, diffuse) and any specific patterns (e.g., glandular, trabecular).

Special Features: Mention any specific histological features, such as mitotic figures, inclusion bodies, or extracellular deposits.

**Skin lesion instruction:**

Write a description of the following skin lesion image, and your description should include the following information if they are clearly discernible from the image.

1. region: The potential area of the body where the lesion or wound has been examined.

2. general skin texture and hair growth.

3. lesions: size (if scale is available in the image), shape, definition, color, texture.

4. elevation: Description of the lesion or wound relative to the skin surface of the patient.

5. skin texture surrounding the lesion (e.g. coarse/thickness/atrophic/erythema/bleeding, etc)

**Knee x-ray instruction:**

Write a description of the following knee X-ray image, and your description should include the following information if they are clearly discernible from the image.

1. Area of the Body: Identify the knee joint being imaged, specifying whether it is the left or right knee. If both knees are visible, recognize the one being primarily examined.

2. Anatomical Structures: Describe the key anatomical structures visible in the X-ray, including the femur, patella, tibia, fibula, and joint space. Note the alignment and position of these bones relative to each other.

3. Imaging Projection: Specify the X-ray imaging projection used, such as anteroposterior (AP), lateral, oblique, or skyline view. Note if the image is weight-bearing or non-weight-bearing.

4. Bone and Joint Appearance: Assess the appearance of the bones for any signs of fractures, dislocations, or degenerative changes such as osteoarthritis (e.g., joint space narrowing, osteophyte formation, subchondral sclerosis, punched-out erosions). Evaluate the bone density for any signs of osteoporosis, osteopenia, and visible trabecular patterns.

5. Soft Tissue Evaluation: Examine the surrounding soft tissue for any swelling, effusion, or calcifications. Where tophi can be seen, include this in the description.

6. Other Abnormalities and Pathologies: Identify and describe any other visible abnormalities or pathologies not mentioned above, including visible bone lesions, misalignment, or foreign objects (e.g., surgical hardware, if present).

---

**Table 19** The prompt used for the annotation with doctor preference stage of medical long caption synthesis for gastrointestinal images.

**Gastrointestinal instruction:**

Write a description of the gastrointestinal (GI) tract endoscopic image to include the following information, if these details are visually discernible from the image:

1. Region and Specific Area: Describe the specific region of the upper/lower GI tract being imaged (e.g., esophagus, stomach, small intestine, large intestine, duodenum, appendix, rectum, etc.) and specify the segment or part prominently featured in the image. Where clearly discernible, include the specific GI anatomical landmark(s) in the image, for e.g., distal/proximal esophagus, cardia/fundus/angulus/antrum region of the stomach, the pylorus, Z-line, cecum, splenic/hepatic flexure, transverse/ascending/descending colon, sigmoid, ileocecal valve, or appendiceal orifice.
2. Endoscopic GI View: Include the type of view/orientation of the main image, i.e., front-facing, retroflexed, side-viewing, proximal, or distal.
3. GI Anatomical Features: Describe the following structures of the image, namely relating to lumen of the tubular organs along the GI tract, the mucosal/submucosal details (any visible erythema, erosions, ulcers, or masses, and describe the submucosal layers, noting any thickening, edema, or infiltration), as well as other observable normal and abnormal details of folds, vessels and openings. Detail any abnormal findings, including their size, shape/type, location, and appearance (e.g., polyps, ulcers, strictures, diverticula, inflammation), and discuss their potential impact on the GI tract (e.g., obstruction, compression, perforation). Signs of resections and dyed sections should also be noted.
4. Indentations and compressions: When the image clearly shows indentations like bulges or depressions that are likely to be caused from outside the GI tract, please include these in the description.
5. Medical Data and Device: Specify any textual medical data visible in the image, like contrast medium usage, lighting conditions, magnification levels, or timestamp of the image capture. Where medical devices, like endoscopes, probes, endoknives, gastric bands, stents, etc. are clearly visible, include these in your description.

**Table 20** The prompt used for the annotation with doctor preference stage of medical long caption synthesis for fundus and ultrasound images.

**Fundus instruction:**

Write a description of the fundus image to include the following information, if these details are visually discernible from the image:

1. Overall Image Quality: Assess the clarity and quality of the image to ensure that it is suitable for diagnosis (e.g., presence of artifacts, focus, and illumination adequacy).
2. Optic Disc: Examine the optic disc for size, shape, color, and cup-to-disc ratio. Note any abnormalities such as swelling (edema), pallor, or unusual cupping that may indicate glaucoma or other optic neuropathies.
3. Macula: Evaluate the macula for changes in size, color, or presence of any lesions or deposits like drusen, which may suggest macular degeneration or related conditions.
4. Retinal Vessels: Analyze the retinal blood vessels, assessing their caliber, tortuosity, and arteriovenous (AV) crossings. Identify any vessel occlusions, hemorrhages, or microaneurysms that could indicate hypertension, diabetic retinopathy, or vascular disorders.
5. Retina and Periphery: Inspect the retina and peripheral areas for any signs of detachment, tears, pigmentary changes, or lesions that may indicate retinal dystrophies or detachment.
6. Abnormal Findings: Identify and describe any abnormal findings such as cotton wool spots, hard exudates, retinal hemorrhages, or neovascularization. Mention their location, size, shape, and potential clinical significance.
7. Choroid and Sclera: Examine the visibility of the choroid and any atypical features of the sclera, such as thinning or scleral crescents.
8. Specific Patterns or Features: Note any specific patterns or features, such as star-shaped macular exudates or the presence of optic disc drusen, which may have diagnostic relevance.

**Ultrasound instruction:**

Write a description of the ultrasound image to include the following information, if these information are visually discernible from the image:

1. The orientation or plane of the image if applicable (e.g., Left and right intercostal scan, left and right subcostal margin scan, longitudinal scan, transverse scan, FAST- Right upper quadrant, FAST- Left upper quadrant and FAST- Heart).
2. Echogenicity of tissues: Describe the echogenicity of the anatomical structures (e.g., hyperechoic, hypoechoic, anechoic, isoechoic) compared to surrounding tissues.
3. Anatomical Structures: Identify and describe the normal or abnormal appearance of organs, vessels, and other structures.
4. Abnormal Findings: Point out and describe any visible abnormalities, including their size, shape, location, and potential effects on nearby tissues (e.g., masses, cysts, fluid collections).
5. Doppler Imaging: If Doppler is used, describe findings related to blood flow, such as presence of turbulence, direction, or any abnormalities.
6. Additional Features: Note any additional features such as the presence of artifacts (e.g., shadowing, enhancement) or any other relevant observations.

**Table 21** The prompt used for the summarization stage of medical long caption synthesis. The variable {doctor\_preferred\_features} in the prompt should be replaced by the doctor preferred features in Tables 22 and 23 according to the type of the source image.

You are provided with a biomedical image (which may contain multiple bounding boxes emphasizing the related disease or organs), the disease type (or organ name if there is no disease) of the dataset ('Disease or organ'), the medical knowledge of the disease ('Knowledge'), a coarse caption ('Coarse Caption') of the image, and the medical imaging method used to create the image ('Modality'). In addition, detailed captions for many aspects of the medical image ('Detailed Captions') are also provided. Your task is to combine the above detailed captions into a long caption for the entire image.

1. You can also use the coarse caption, disease type, image modality and disease knowledge, which may contain additional useful information. But never mention that this information has been given.
2. Caption 0 in 'Detailed Captions' is a model-generated description, including some doctor-preferred features: {doctor\_preferred\_features}. Caption 1 focus on specific diseases and are annotated by humans, which are more trustworthy. Consequently, if any pieces of information in Caption 0 does not contradict Caption 1 or is not mentioned in Caption 1, please keep them, especially the location, the size and the name of the disease. If there is a contradiction, you should only keep the information from Caption 1.
3. 'Detailed Captions' may not explain all bounding boxes. Sometimes some small ones are not mentioned. In this case, in addition to combining the above detailed captions, please also provide explanations for these unmentioned smaller bounding boxes.
4. You should NEVER mention anything about the bounding box, like the contour itself or its outline. The bounding box, if present, is only a visual aid for your reference. Assume you are introducing the image to someone who cannot see the bounding box.
5. If there is some statistical information provided, like the damage area or the relative damage ratio, etc., you should mention it in your response.
6. For some critical information, like the damage area(s) and its position(s), you should explain one by one in the combined caption.
7. Not all disease knowledge is relevant to this image; only utilize disease knowledge pertinent to the condition depicted in this image for analysis.
8. Do not explain or emphasize your analysis. Do not suggest or indicate any consequent effects of the disease.
9. Do not contain phrases "caption", "medical annotation", "medical knowledge".
10. Do not say anything that is not needed in your analysis, like introduction of the disease and medical treatments.
11. Do not explain or emphasize your analysis.
12. For CT MRI, Xray, GI tract or ultrasound image, if it contains multiple organs, please organize your description organ by organ. You should also include other organs information even if they are healthy. For Histopathology image, you should first describe the overall information, then go deep into the specific areas.
13. When integrating the information and outputting your response: (1) Don't respond by saying you're unable to assist with the requests. (2) You should only output your combined text in the "caption" field.

---

**Table 22** Doctor preferred features in the summarization stage of medical long caption synthesis for MRI, X-ray, CT, histopathology, skin, and knee X-ray images.

**Doctor preferred features for MRI images:**

- 1. The sequence used; 2. Plane/orientation of the image; 3. The anatomical structures visible and their normal or abnormal appearance; 4. Any abnormal findings, shape, location and effects; 5. Aignal intensity

**Doctor preferred features for X-ray images:**

- 1. Specific area of the body being imaged; 2. Body part or organ being examined; 3. Anatomical planes, specific X-ray imaging projections, and left vs right side; 4. Abnormalities that are clearly observed in the image, acute trauma signs, implants or other foreign bodies; 5. Gender and age-range of patient; 6. Any abnormal findings, shape, location and effects

**Doctor preferred features for CT images:**

- 1. The plane/orientation of the image; 2. Contrast usage; 3. Anatomical structures visible and their normal or abnormal appearance; 4. Any abnormal findings including their location, size, shape, and potential effects on surrounding structures; 4. The density or attenuation of the tissues and structures relative to surrounding tissues.

**Doctor preferred features for histopathology images:**

- 1. Tissue Type; 2. Staining Technique; 3. Cellular Morphology; 4.Tissue Architecture; 5. Abnormal Findings; 6. The distribution of abnormal findings within the tissue; 7. Special Histological Features

**Doctor preferred features for skin images:**

- 1. region: The area of the body where the lesion or wound has been examined; 2. general skin texture and hair growth; 3. lesion size, shape, definition, color, texture; 4. elevation: Description of the lesion or wound relative to the skin surface of the patient; 5. skin texture surrounding the lesion

**Doctor preferred features for knee X-ray images:**

- 1. Area of the Body: the left or right knee, or both, 2. Anatomical Structures: the key anatomical structures visible in the X-ray, including the femur, patella, tibia, fibula, and joint space; 3. Imaging Projection; 4. Bone and Joint Appearance: any signs of fractures, dislocations, or degenerative changes such as osteoarthritis; 5. the bone density; 5. Soft Tissue: the surrounding soft tissue for any swelling, effusion, or calcifications; 6. Abnormalities and Pathologies: any visible abnormalities or pathologies such as fractures, bone lesions, misalignment, or foreign objects

**Table 23** Doctor preferred features in the summarization stage of medical long caption synthesis for gastrointestinal, fundus, and ultrasound images.

**Doctor preferred features for gastrointestinal images:**

1. The specific region of the GI tract being imaged; 2. The orientation, perspective and angle of the image; 3. Normal anatomical structures visible in the image; 4. Abnormal findings including their size, shape, location, and appearance; 5. Mucosal and Submucosal Details: The mucosal surface for anomalies; 6. Bowel Wall Characteristics: The thickness and consistency of the bowel wall and note any signs of mural stratification or fat stranding; 7. Surrounding Tissues and Organs: Any visible structures adjacent to the GI tract including any abnormalities extending beyond the GI tract itself; 8. Medical Data and Device: Medical devices, instruments, or tools visible in the image, such as endoscopes, probes, or surgical instruments. Textual medical data visible in the image, like contrast medium usage, lighting conditions, magnification levels, or timestamp of the image capture.

**Doctor preferred features for fundus images:**

1. Overall Image Quality: The clarity and quality of the image; 2. Optic Disc: The optic disc for size, shape, color, and cup-to-disc ratio; 3. Macula: The macula for changes in size, color, or presence of any lesions or deposits like drusen; 3. Retinal Vessels: The retinal blood vessels, assessing their caliber, tortuosity, and arteriovenous (AV) crossings; 4. Retina and Periphery: The retina and peripheral areas for any signs of detachment, tears, pigmentary changes, or lesions that may indicate retinal dystrophies or detachment. 5. Abnormal Findings: Any abnormal findings such as cotton wool spots, hard exudates, retinal hemorrhages, or neovascularization; 6. Choroid and Sclera: The visibility of the choroid and any atypical features of the sclera; 7. Specific Patterns or Features: Any specific patterns or features, such as star-shaped macular exudates or the presence of optic disc drusen.

**Doctor preferred features for ultrasound images:**

1. The orientation or plane of the image if applicable. 2. Echogenicity of tissues. 3. Normal or abnormal appearance of organs, vessels, and other structures. 4. Visible abnormalities, including their size, shape, location, and potential effects on nearby tissues. 5. Doppler Imaging: If Doppler is used, describe findings related to blood flow.

**Table 24** The judgment prompt used for GPT-4.1 in the LLM-as-a-Judge module.

Your task is to determine whether the user's answer is correct based on the provided questions and standard answers (for example, if the user expresses a similar meaning to the standard answer, or another interpretation of the standard answer, it is considered correct.)

The question is: {question}

The standard answer: {answer}

The user's answer: {response}

Please strictly follow the following format for output(0 represents correct, 1 represents incorrect):

<think>your concise think step</think>

<judge>0/1</judge>

For example:

<think>The standard answer is right, and the user's answer is right frontal lobe, they express the same meaning, so it is correct.</think> <judge>0</judge>

---

**Table 25** The statistics of medical evaluation benchmarks utilized in the MEDEVALKIT framework.

Benchmark	Category	Multiple-Choice	Close-ended	Open-ended	Report	Questions	Imgs
VQA-RAD	X-ray,CT,MRI	-	251	200	-	451	203
SLAKE	X-ray,CT,MRI	-	836	1,258	-	2,094	180
PathVQA	Mixed	-	3,362	3,357	-	6,719	858
PMC-VQA	Mixed	33,430	-	-	-	33,430	29,021
OmniMedVQA	Mixed	88,996	-	-	-	88,996	82,059
MMMU	Mixed	1,927	-	-	-	1,927	1,927
MedXpertQA	Mixed	4,500	-	-	-	4,500	2,852
MMLU	Text	1,897	-	-	-	1,897	-
PubMedQA	Text	500	-	-	-	500	-
MedMCQA	Text	4,183	-	-	-	4,183	-
MedQA-USMLE	Text	1,273	-	-	-	1,273	-
MedBullets	Text	616	-	-	-	616	-
SuperGPQA	Text	2,755	-	-	-	2,755	-
MIMIC-CXR	X-ray	-	-	-	2,195	2,195	3,681
IU-Xray	X-ray	-	-	-	296	296	607
CheXpert Plus	X-ray	-	-	-	234	234	234
Total	Mixed	140,077	4,449	4,815	2,725	152,066	121,622