Review Article | Artificial Intelligence

eISSN 2005-8330
https://doi.org/10.3348/kjr.2025.0599
Korean J Radiol 2025

Korean Journal of Radiology
KJR

Multimodal Large Language Models in Medical Imaging: Current State and Future Directions

Yoojin Nam[1,2]*, Dong Yeong Kim[1,3]*, Sunggu Kyung[1,4], Jinyoung Seo[1,5], Jeong Min Song[1,3], Jimin Kwon[1,6], Jihyun Kim[1,5], Wooyoung Jo[1,5], Hyungbin Park[1,5], Jimin Sung[1,7], Sangah Park[1,7], Heeyeon Kwon[1,4], Taehee Kwon[1,5], Kanghyun Kim[1,5], Namkug Kim[1,3,4]

[1]Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea
[2]Department of Radiology, Samsung Changwon Hospital, Sungkyunkwan University School of Medicine, Changwon, Republic of Korea
[3]Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea
[4]Department of Biomedical Engineering, Brain Korea 21 Project, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea
[5]University of Ulsan College of Medicine, Seoul, Republic of Korea
[6]Department of Radiology, Dankook University Hospital, Dankook University College of Medicine, Cheonan, Republic of Korea
[7]Chosun University College of Medicine, Gwangju, Republic of Korea

Multimodal large language models (MLLMs) are emerging as powerful tools in medicine, particularly in radiology, with the potential to serve as trusted artificial intelligence (AI) partners for clinicians. In radiology, these models integrate large language models (LLMs) with diverse multimodal data sources by combining clinical information and text with radiologic images of various modalities, ranging from 2D chest X-rays to 3D CT/MRI. Methods for achieving this multimodal integration are rapidly evolving, and the high performance of freely available LLMs may further accelerate MLLM development. Current applications of MLLMs now span automatic generation of preliminary radiology report, visual question answering, and interactive diagnostic support. Despite these promising capabilities, several significant challenges hinder widespread clinical adoption. MLLMs require access to large-scale, high-quality multimodal datasets, which are scarce in the medical domain. Risks of hallucinated findings, lack of transparency in decision-making processes, and high computational demands further complicate implementation. This review summarizes the current capabilities and limitations of MLLMs in medicine—particularly in radiology—and outlines key directions for future research. Critical areas include incorporating region-grounded reasoning to link model outputs to specific image regions, developing robust foundation models pre-trained on large-scale medical datasets, and establishing strategies for the safe and effective integration of MLLMs into clinical practice.

Keywords: Artificial intelligence; Large language model; Medical imaging; Multimodal large language model

INTRODUCTION

Artificial intelligence (AI), especially deep learning, has significantly impacted clinical medicine, particularly radiology [1]. Convolutional neural networks have enhanced image recognition and segmentation, boosting diagnostic accuracy for specific tasks and altering components of the radiological workflow [2]. However, these early AI applications primarily focused on analyzing images in isolation [3]. This unimodal approach contrasts sharply with real-world clinical radiology, where practitioners routinely combine imaging findings with patient information from electronic health records (EHRs), including clinical notes, laboratory results, and patient history [3-5]. Given the inherently multimodal nature of radiological practice, the limitations of image-centric AI have prompted the

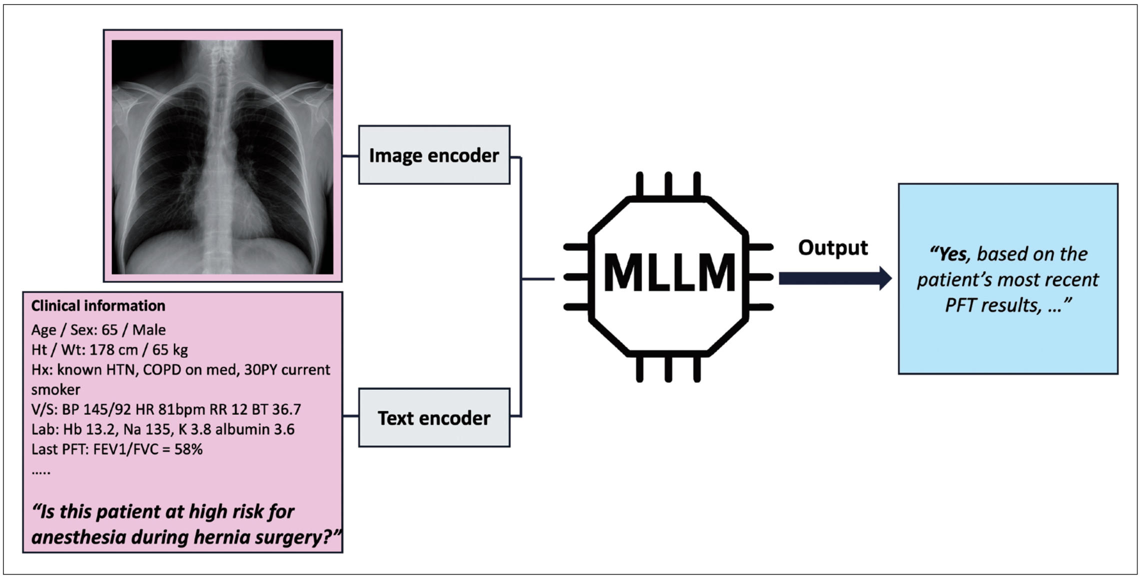
Received: May 14, 2025 **Revised:** July 3, 2025 **Accepted:** July 8, 2025
*These authors contributed equally to this work.
Corresponding author: Namkug Kim, PhD, Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea
• E-mail: namkugkim@gmail.com

emergence of multimodal large language models (MLLMs) designed to integrate diverse clinical and imaging data [3-5].

MLLMs, also known as large multimodal models (LMMs), represent a significant evolution in medical AI [6]. Their core capability is the concurrent processing and integration of heterogeneous data modalities [7]. In clinical settings, this includes various imaging types (e.g., radiologic imaging such as CT, MRI, and X-ray, endoscopy, digital pathology, and various clinical photos) alongside textual data such as radiology reports, clinical notes, and structured EHR data (Fig. 1) [8]. MLLMs build upon large language models (LLMs)—sophisticated AI trained on vast text datasets using transformer-based architectures to understand and generate human-like language [9]. MLLMs extend these capabilities by incorporating advanced computer vision modules and multimodal learning techniques [8]. The defining feature is their ability to integrate and align information across modalities, often mapping them into a shared representational space [8]. This synergy allows for a more comprehensive understanding than unimodal approaches permit [10]. Consequently, MLLMs can tackle complex cross-modal tasks such as radiology report generation (RRG) from images and visual question answering (VQA) that incorporates both imaging and clinical context [11].

The rapid development of MLLMs reflects several converging technological advancements. First, the evolution of LLMs, powered by the transformer architecture and its self-attention mechanism, allows the capture of long-range textual dependencies, with extensive pre-training on web-scale corpora providing broad linguistic and domain-specific knowledge (Fig. 2) [12,13]. Second, parallel innovations in multimodal architectures, particularly vision transformers (ViTs) [14], provide high-capacity encoders adaptable to medical imaging modalities such as CT and MRI. Third, multimodal learning strategies, including contrastive pre-training (Fig. 3A) [15] and instruction-tuned fusion architectures (Fig. 3B) [16], permit the seamless integration of image and text representations, enabling natural-language prompts to drive complex visual reasoning. Finally, the availability of high-performance computing infrastructure, including graphics processing units (GPUs) and tensor processing units (TPUs), provides the computational throughput required to train and deploy these parameter-intensive systems at clinically relevant scales [17]. Together, these advancements support the foundation model (FM) approach, where large, broadly pre-trained models are adapted for specific tasks, potentially accelerating development in specialized domains like medical imaging [18].

This review provides a comprehensive overview of MLLMs in medical imaging, with a particular focus on radiology. Aimed at medical professionals, the discussion highlights clinical relevance, potential applications, and key implementation challenges, based on recent research.



Clinical information
Age / Sex: 65 / Male
Ht / Wt: 178 cm / 65 kg
Hx: known HTN, COPD on med, 30PY current smoker
V/S: BP 145/92 HR 81bpm RR 12 BT 36.7
Lab: Hb 13.2, Na 135, K 3.8 albumin 3.6
Last PFT: FEV1/FVC = 58%
.....

"Is this patient at high risk for anesthesia during hernia surgery?"

*"**Yes**, based on the patient's most recent PFT results, ..."*

Fig. 1. Example of a MLLM. A patient's chest X-ray and corresponding clinical information, including a key clinical question, are encoded separately by an image encoder and a text encoder, respectively. By integrating these multimodal inputs, the MLLM generates an appropriate answer to the user's question. MLLM = multimodal large language model

Brief Technical Principles and Concepts of MLLMs

MLLMs extend traditional text-only LLMs by incorporating diverse input modalities such as images, audio, and video. These models learn cross-modal connections, allowing them to process, reason about, and generate information across multiple data types [19]. A typical MLLM architecture

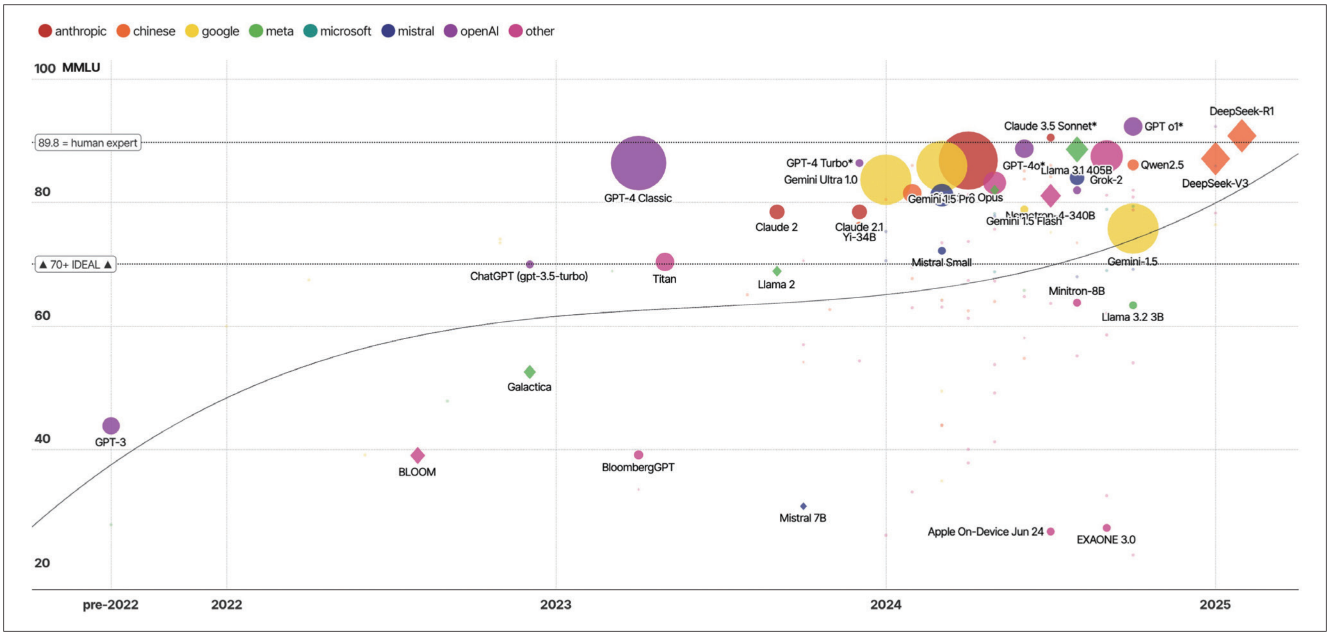


Fig. 2. Global landscape of LLMs and their performance. Bubble chart displaying over 100 public LLM releases from 2020 to 2025, ranked by their MMLU score (y-axis; higher values indicate better performance) and release date (x-axis). Bubble size corresponds to the reported number of training parameters (in billions), while bubble color indicates the developer (e.g., OpenAI, Google, Anthropic). Dashed horizontal lines represent two reference benchmarks: the human-expert level (89.8%) and the "ideal clinical threshold" (70%) referenced in the text. Reprinted under open access from David et al. [13]. *Parameter undisclosed. LLM = large-language model, MMLU = massive multitask language understanding

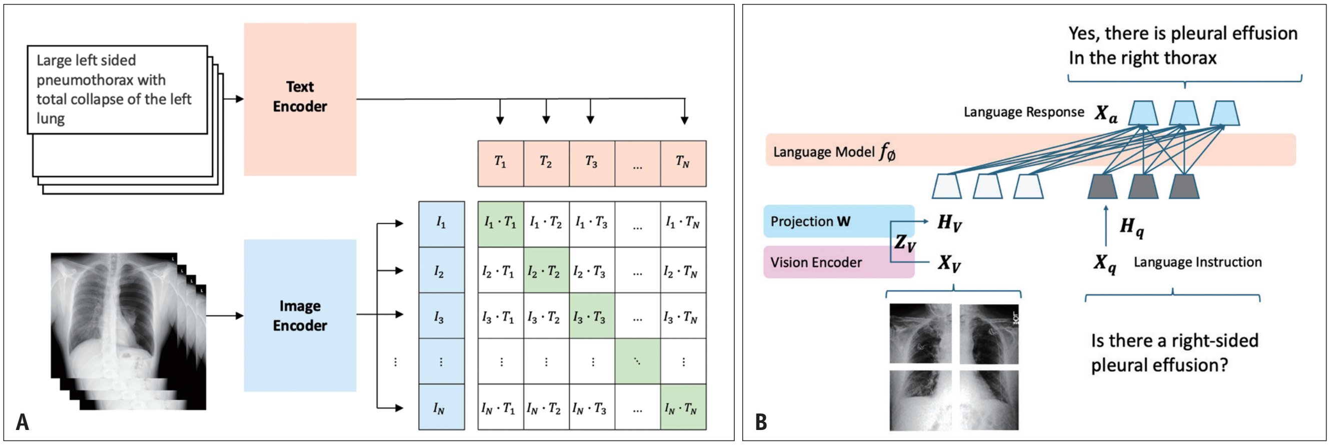


Fig. 3. Multimodal pipelines linking chest radiographs with text. **A:** Contrastive pre-training. Separate image and report encoders are trained to learn a shared embedding space by pulling true image–report pairs together and pushing mismatches apart, enabling zero-shot study retrieval, automatic report drafting, and label-efficient classification. Adapted from Radford et al., *Proc Mach Learn Res* 2021;139:8748-8763, originally published under a CC BY license [15]. **B:** Instruction-tuned fusion. Visual features extracted by a frozen vision encoder are projected and integrated into the token stream of a large language model that has been fine-tuned with radiology-specific instructions. This setup enables the model to process natural-language prompts (e.g., "Is there right-sided effusion?") and generate clinically relevant outputs such as key findings, differential diagnoses, or follow-up recommendations. Adapted from Liu et al., a preprint published under a CC BY license [16]. I = image embedding, Hq = image-aware hidden states after cross attention, Hv = linearly projected visual tokens, T = text embedding, Xa = autoregressively generated answer tokens, Xq = prompt (query) text tokens, Xv = input radiograph, Zv = vision-encoder feature map

comprises pre-trained encoders for different data types, a pre-trained LLM, and a multimodal connector that aligns the representations across components (Fig. 4). Some architectures also include a generative module capable of creating images or videos when needed. In this architecture, the LLM functions as a 'cognitive engine,' maintaining its text-centric pre-trained state to provide powerful reasoning capabilities without requiring additional fine-tuning for multimodal inputs [7]. This design supports a wide range of applications, including RRG, VQA, and content-based retrieval across modalities. Given the substantial computational demands of large-scale MLLM models, optimization strategies developed for LLMs—such as compression, quantization, and knowledge distillation—are critical for reducing inference costs and facilitating deployment in real-world clinical settings [20-22].

Architectures

Modality-specific encoders transform complex data types—such as images, audio, and video—into simpler, meaningful representations. These encoders extract essential features from each modality, converting high-dimensional inputs into streamlined formats suitable for downstream processing. Instead of training new encoders from scratch (i.e., using randomly initialized weights without prior knowledge), researchers typically employ existing pre-trained models [7]. A popular choice is contrastive language-image pre-training (CLIP) [15], which aligns visual data with corresponding textual descriptions through extensive training (Fig. 3A).

To bridge the modality gap between non-text data and natural language, a multimodal connector is introduced as a learnable interface. Since LLMs operate solely on textual input, this connector maps diverse data types to their corresponding textual representations within a shared semantic space. This approach enables the outputs of specialized encoders to be translated into formats interpretable by the LLM, thereby eliminating the need to train a multimodal model from scratch [7]. Multimodal connectors can be categorized into four main types (Fig. 4).

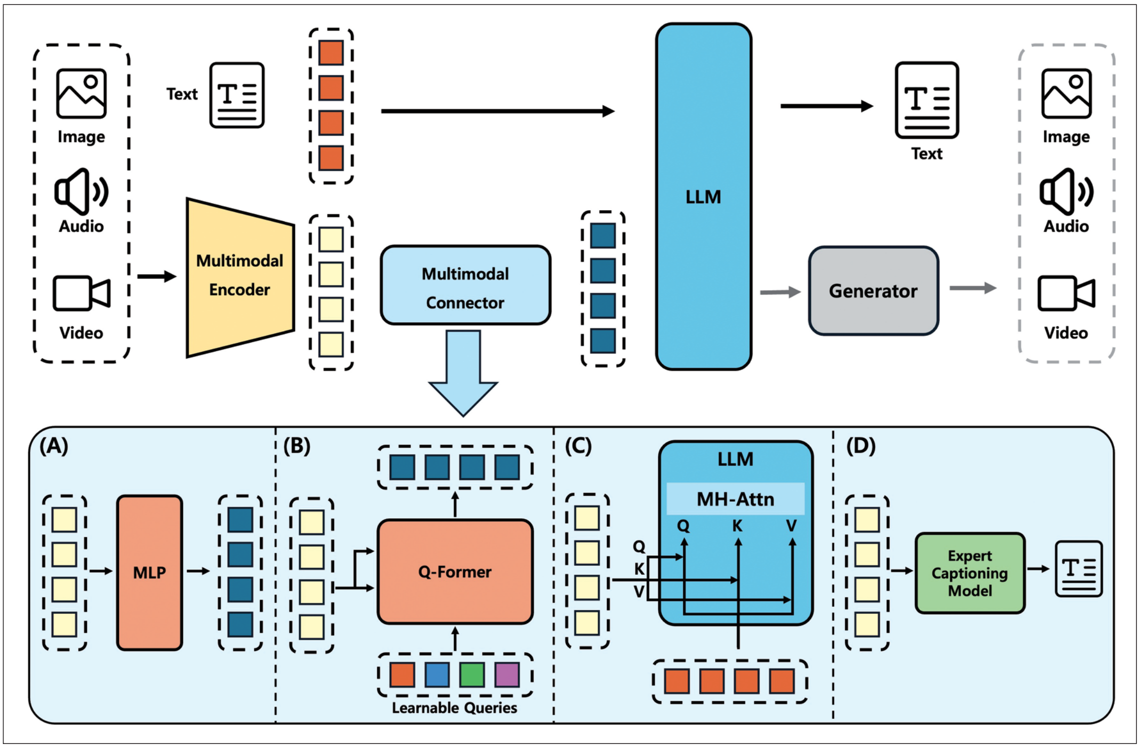


Fig. 4. Schematic illustration of a typical multimodal LLM architecture. The architecture generally comprises three main components: an encoder, a connector, and a LLM. An optional generator can be attached to the LLM to generate additional modalities beyond text. The encoder processes input data—such as images, audio, or video—and extracts modality-specific features. These features are then transformed by the connector to enhance the LLM's interpretability. Connectors are broadly categorized into four types: **(A)** projection-based, **(B)** query-based, **(C)** fusion-based, and **(D)** expert-driven language transformation connectors. Adapted from Yin et al., *Natl Sci Rev* 2024;11:nwae403, originally published under a CC BY license [7]. LLM = large language model, MLP = multi-layer perceptron, Q-Former = query transformer, MH-Attn = multi-head attention

Projection-based connectors employ a multi-layer perceptron (MLP), a type of neural network analogous to neural communication in the human brain (Fig. 4A) [16]. The MLP transforms visual data into a representation that aligns closer to language, making it easier for the LLM to understand and process [23]. Query-based connectors utilize specialized trainable 'query tokens' to extract salient visual details from images [24]. These tokens guide the model in efficiently identifying and retrieving relevant visual information (Fig. 4B) [25]. Both projection- and query-based connectors convert features into token representations, which are then passed to the LLM alongside text tokens, enabling effective multimodal information integration within the LLM's processing pipeline. Fusion-based connectors facilitate feature-level integration within the LLM architecture (Fig. 4C) [26]. Through a cross-attention mechanism, the model establishes direct interactions between pairs of visual and language representations, enabling effective multimodal information integration [27]. Specifically, this mechanism allows language representations to selectively focus on and incorporate relevant details from visual inputs by forming pairwise relationships [19]. As a result, feature-level fusion allows for richer multimodal interactions throughout the LLM's processing stages [7]. Expert-driven language transformations convert non-linguistic data directly into text, similar to image captioning (Fig. 4D) [25]. This approach uses specialized models to translate multimodal inputs directly into language that LLMs can process without additional training [7,28]. While straightforward to implement, this method often results in information loss when complex data like videos are reduced to text descriptions that cannot fully preserve spatial-temporal relationships [7,29].

Pre-trained LLMs form the cognitive backbone of modern multimodal systems, providing a significantly more efficient alternative to building models from scratch. Their extensive training on large-scale text corpora enables broad reasoning and contextual understanding, which can be leveraged for multimodal tasks [7,30]. These models inherently support capabilities such as zero-shot generalization and few-shot learning, chain-of-thought reasoning, and instruction following [31]. Empirical studies indicate that larger models improve accuracy, contextual understanding, fluency, and problem-solving, while demonstrating emergent capabilities like cross-lingual understanding [7,30].

Training Strategies

MLLMs are typically developed through three sequential stages: pre-training, instruction tuning, and alignment tuning. Each stage uses different data types and learning objectives to progressively improve the model's cross-modal understanding and reasoning capabilities [7]. In the pre-training stage, a multimodal connector learns to align visual and textual representations, often using autoregressive captioning on image-text pairs [32]. Research indicates that selectively fine-tuning components of the vision encoder enables more precise alignment between modalities [33]. The training data at this stage includes both large-scale web-collected materials and refined content from human annotation or high-performance MLLMs [34].

During instruction tuning, the model is fine-tuned using datasets containing diverse natural language instructions and multimodal inputs, teaching it to follow complex directives reliably (Fig. 3B). The model is trained to generate appropriate responses to various inputs, including images and text. This process involves fine-tuning the LLM using low-rank adaptation (LoRA) while simultaneously training the multimodal connector to process heterogeneous input modalities effectively [35]. The vision encoder may be selectively fine-tuned, depending on performance needs. Several strategies are employed during this stage, including converting question-answer pairs into instruction formats or generating multimodal instructions using advanced LLMs [36]. Research shows that incorporating both multimodal with text-only data during instruction tuning significantly enhances the model's conversational quality and instruction-following capabilities, thereby improving its adaptability across various tasks [37].

The final stage, alignment tuning, optimizes the model's outputs to better reflect human preferences, thereby improving response quality and reliability. This is typically achieved through reinforcement learning from human feedback [38]. A reward model is first trained on human preference data, after which the policy model is fine-tuned to maximize reward scores. This stage relies on small-scale, high-quality comparison responses, which help reduce hallucination risks and better reflect human preferences.

Representative General-Purpose MLLMs

Table 1 presents prominent MLLM models that have significantly influenced architectural and training paradigms. Flamingo [26], developed by DeepMind, was an early model that effectively used feature fusion to combine visual

Table 1. Summary of representative general-purpose multimodal LLMs

Model (year)	Connector type	Vision encoder	LLM backbone	Pre-training data	Model size	Instruction tuning	Key features
Flamingo (2022) [26]	Fusion-based	NFNet image encoder with Perceiver resampler	Chinchilla (70B)	ALIGN and LTIP datasets	3.2–80B	No (few-shot only)	Few-shot multimodal learning achieves SOTA performance on 16 tasks with a few, without additional fine-tuning
BLIP2 (2023) [25]	Query-based	ViT from CLIP	OPT (2.7B, 6.7B) or FlanT5 (3B, 11B)	129M image–text pairs	3.1–12.1B	No	Strong zeroshot captioning, outperforming Flamingo by 8.7% on zero-shot VQA
LLaVA (2023) [16]	Projection-based	ViT from CLIP	Vicuna (13B)	595K image-text pairs	UN	Yes	Achieves 85% of GPT-4's performance on a multimodal instruction-following dataset
GPT4 (2023) [39]	UN	UN	Native multimodal	UN	UN	Yes	Advanced VQA, reduced hallucination, robust multimodal reasoning
Gemini (2023) [40]	UN	UN	Native multimodal	UN	UN	Yes	Strong reasoning across modalities; tool integration
Claude 3 (2024) [41]	UN	UN	Native multimodal	UN	UN	Yes	Strong OCR, structured data understanding, extended context window

LLM = large language model, ALIGN = a large-scale ImaGe and noisy text embedding, LTIP = long text and image pairs, SOTA = state-of-the-art, BLIP = bootstrapped language-image pre-training, ViT = vision transformer, CLIP = contrastive language-image pre-training, OPT = open pre-trained transformer, T5 = text-to-text transfer transformer, VQA = vision question answering, LLaVA = large language and vision assistant, GPT = generative pre-trained transformer, UN = undisclosed, OCR = optical character recognition

information with language processing (Fig. 4C). It inserted visual features into the middle layers of an LLM using cross-attention, allowing the language components to attend selectively to relevant visual elements, promoting deeper multimodal integration. Subsequent query-based approaches improved efficiency by simplifying visual-language connections. Bootstrapped language-image pre-training (BLIP)-2 [25] advanced this paradigm by introducing a two-stage query transformer (Q-Former) that connects an image encoder to an LLM. The Q-Former uses learnable query tokens to extract salient visual features and convert them into LLM-compatible format, enabling end-to-end vision-to-language generation without complete retraining (Fig. 4B). Remarkably, BLIP-2 achieved performance comparable to Flamingo while using 54 times fewer trainable parameters. Further simplifying the architecture, large language and vision assistant (LLaVA) [16] adopted projection-based connectors (Fig. 4A), requiring only 595000 image-text pairs for initial alignment—significantly fewer than BLIP-2's 100 million samples.

Recent models, such as GPT, Gemini, and Claude, increasingly use extensive instruction tuning and human feedback alignment, applying text-only reinforcement learning techniques to visual domains [39-41]. While larger models and more training data generally improve multimodal capabilities, as seen in the evolution from Flamingo to more advanced models, strategic architectural design choices, such as the use of lightweight adapters, have enabled smaller models to achieve impressive results in specific areas [42,43]. These developments highlight the critical balance between model scale and design efficiency in advancing multimodal AI.

2D Medical Imaging MLLMs

Early research on MLLM in medicine primarily focused on 2D radiological images, particularly chest X-rays (CXRs) [44]. This emphasis was driven by two principal factors: 1) the relative maturity and cross-domain transferability of 2D vision encoders pre-trained on large-scale natural image datasets [14], and 2) the availability of extensive, publicly accessible datasets that pair 2D medical images with corresponding radiology reports [45].

Typical Datasets for 2D MLLM Research

Several high-quality public radiology datasets have significantly advanced the development and evaluation of 2D medical multimodal models. Among them, MIMIC CXR [46], which contains hundreds of thousands of CXRs paired with corresponding radiology reports, serves as a primary resource for training models on RRG tasks. Building on this, the Chest ImaGenome [47] introduces detailed annotations, including anatomical structure bounding boxes, which support more granular analysis of model attention and enable more rigorous evaluation methodologies.

For VQA in medical imaging, benchmarks like VQA-RAD [48] and SLAKE [49] provide carefully selected CXRs and CT images, accompanied by clinician-generated questions and expert-verified answers. These datasets are crucial for rigorously evaluating a model's ability to interpret medical images and extract clinically relevant information. A more recent resource, PMC-VQA [50], is created by taking radiology figures and their captions from PubMed Central articles and converting them into 227 thousands of diverse question-answer pairs using automated methods. This large-scale dataset expands the range of clinical scenarios and supports the fine-tuning of MLLMs to enhance their radiology reasoning capabilities.

Collectively, these datasets—whether offering comprehensive report collections (MIMIC-CXR), focused Q&A benchmarks (VQA-RAD, SLAKE), or large-scale synthetic VQA pairs (PMC-VQA)—have played a central role in shaping both the main research tasks and standard evaluation protocols used across 2D medical MLLM studies. Typically, models are pre-trained on the extensive MIMIC-CXR dataset and subsequently evaluated on VQA-RAD, SLAKE, or PMC-VQA to assess their performance in real-world radiology applications [51].

Typical 2D MLLMs and Architectures

Contrastive learning aligns radiologic images and their corresponding reports by pulling matched pairs together in the embedding space while pushing mismatched pairs further apart, thereby significantly reducing the need for manual annotations [15]. ConVIRT [52] first applied a bidirectional contrastive loss to CXR–report pairs,

Table 2. Summary of representative 2D multimodal LLMs in radiology trained by contrastive learning

Model	Base architecture (vision + LLM)	Key technique(s)	Primary task(s)	Dataset(s) used	Key strength/ contribution
ConVIRT [52]	ResNet50 + ClinicalBERT	Bidirectional imagetext contrastive pretraining, largebatch unsupervised learning	Zeroshot classification & retrieval	MIMICCXR v2 (227K) + internal musculoskeletal set (48K pairs)	First medical imagetext contrastive framework
MedCLIP [53]	ViT + BioClinicalBERT	Decoupled contrastive learning, semantic matching loss (using medical knowledge)	Zero-shot classification, supervised classification, image-text retrieval	Unpaired images/ text (e.g., CheXpert, MIMIC-CXR)	High data efficiency, addresses false negatives, strong zero-shot performance
BioViL-T [54]	Hybrid CNN-transformer multi-image encoder + CXR-BERT	Temporal vision-language pretraining, contrastive learning	Progression classification, phrase grounding, RRG	MIMIC-CXR (longitudinal pairs)	First model with temporal awareness, SOTA on temporal tasks
BioMedCLIP [56]	ViT + PubMedBERT	Large-scale contrastive pre-training	Cross modal retrieval, zero-shot/few-shot/ full-shot image classification, VQA	PMC-15M (15 million diverse biomedical image-text pairs)	Domain-specific adaptations, positive transfer learning demonstrated

LLM = large language model, ConVIRT = contrastive learning of medical visual representations, ResNet = residual network, BERT = bidirectional encoder representation from Transformers, MIMIC-CXR = medical information mart for intensive care chest X-ray dataset, CXR = chest X-ray, MedCLIP = medical contrastive language-image pre-training, ViT = vision transformer, BioClinicalBERT = clinical BERT pretrained on biomedical notes, CheXpert = chest X-ray expert-labeled dataset from Stanford, BioViL-T = biomedical vision-language model with temporal modeling, CNN = convolutional neural network, CXR-BERT = BERT variant trained on chest X-ray reports, RRG = radiology report generation, SOTA = state-of-the-art, BioMedCLIP = biomedical CLIP-style pretraining using ViT and PubMedBERT, PubMedBERT = BERT pretrained on PubMed abstracts, VQA = visual question answering, PMC = PubMed Central establishing a strong radiology-specific pre-training baseline. MedCLIP [53] improved data efficiency and enabled zero-shot transfer with limited paired examples by introducing separate image and text encoders along with a terminology-aware semantic loss. BioViL-T [54] extended this approach by integrating CXR-BERT [55] text encoders with multi-image transformers to model temporal disease progression, achieving state-of-the-art (SOTA) performance in phrase grounding and progression tasks. BioMedCLIP [56] further demonstrated that pre-training on large, diverse biomedical image–text corpora can outperform radiology-specific models, highlighting the benefit of cross-domain knowledge transfer. Table 2 summarizes these methods.

Research then shifted toward MLLMs, where general architectures are adapted through instruction tuning or fine-tuning to enhance reasoning and generation capabilities. Several MLLMs have shown notable progress in 2D radiological image analysis (Table 3). LLaVA-Med [57] pairs a CLIP-based ViT with a Vicuna/Llama LLM. After aligning 15 million PMC image–caption pairs and a brief GPT-4-guided tuning stage, it achieves expert-level performance in VQA and medical dialogue. Med-PaLM M [58] adopts a generalist design by integrating PaLM-E [59] with a ViT, enabling a single set of parameters to handle multiple biomedical modalities. Fine-tuning on MultiMedBench [58] enabled the model to attain SOTA performance across all evaluated tasks, and its automatically generated chest radiograph reports were preferred over those of human radiologists in approximately 40% of blinded comparisons. Med-Flamingo [60] adapts the few-shot OpenFlamingo-9B framework to the medical domain. Without weight updates, it can answer image-based exam questions and generate free-form explanations deemed accurate by experts. Targeting thoracic imaging, X-rayGPT [61] maps a MedCLIP [53] encoder into Vicuna-7B using a single linear projector and 217000 annotated summaries, supporting concise impression generation, abnormality description, and interactive VQA.

In summary, the current landscape of 2D radiology MLLMs is shaped by two complementary approaches: contrastive pre-training for efficient visual representation learning and instruction-tuned generalist models for advanced reasoning and generation capabilities.

Target Tasks for 2D MLLMs

Integrating visual and language processing in MLLMs has enabled several capabilities that directly enhance radiology workflows. First, RRG leverages MLLMs to translate complex image features into coherent narrative text, automatically

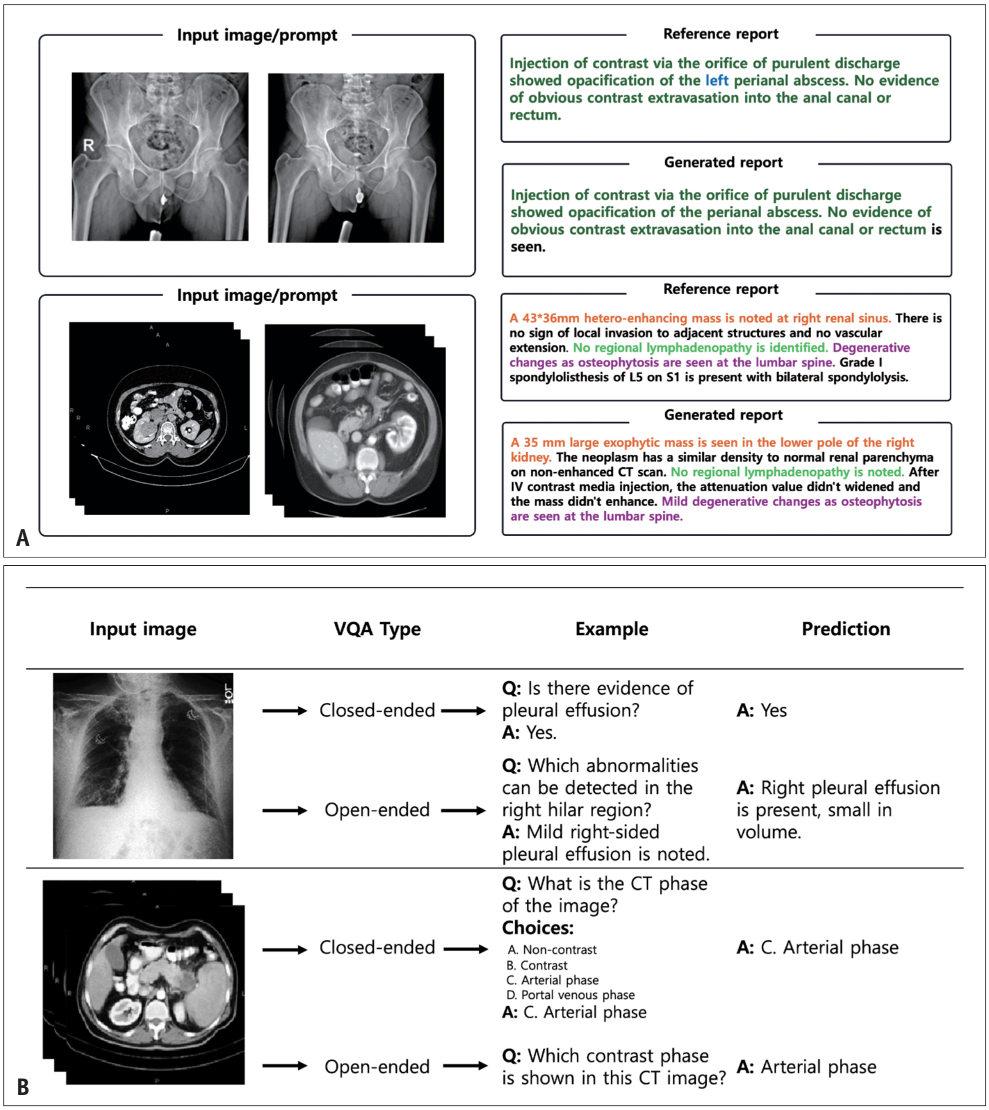
Table 3. Summary of representative 2D MLLMs in radiology trained via instruction tuning or fine-tuning

Model	Base architecture (vision + LLM + interface)	Key technique(s)	Primary task(s)	Dataset(s) used	Key strength/contribution
LLaVA-Med [57]	CLIP/ViT or BioMedCLIP/ViT + Vicuna/Llama + Linear Projection	Instruction tuning, curriculum learning	Visual conversation, VQA	PMC-15M (captions), GPT-4 generated instruction-following data	Efficient adaptation of general MLLM for biomedical conversation
Med-PaLM M [58]	ViT-e, ViT-22B + PaLM-E 8B, 62B, 540B + Linear projection	End-to-end fine-tuning, instruction prompting, one-shot exemplar	VQA, RRG, classification, genomics etc.	MultiMedBench (diverse medical tasks/modalities)	Generalist model with SOTA performance on many tasks, strong reasoning
Med-Flamingo [60]	CLIP ViT-L/14 + Llama-7B + Perceiver Resampler & Cross-Attention	Continued pre-training, few-shot in-context learning	Generative VQA, Rationale generation	MTB, publications (PMC-OA)	First medical MLLM with few-shot learning capability for VQA and reasoning
X-rayGPT [61]	MedClip ViT + Vicuna-7B + Linear Projection	Medical visual-text, alignment fine-tuning	Imageconditioned CXR summary generation, interactive VQA	Generated summaries (217K) from MIMIC-CXR, OpenI reports	Specialized conversational model for CXRs

MLLM = multimodal large language model, LLM = large language model, LLaVA = large language and vision assistant, CLIP = contrastive language-image pretraining, ViT = vision transformer, VQA = visual question answering, PMC = PubMed Central, GPT = generative pre-trained transformer, Med-PaLM = medical pathways language model, PaLM = pathways language model, RRG = radiology report generation, MultiMedBench = benchmark suite for diverse medical tasks, SOTA = state-of-the-art, MTB = medical textbooks, PMC-OA = PubMed Central open access subset, X-rayGPT = instruction-tuned conversational MLLM for chest radiographs, MedCLIP = medical contrastive language-image pretraining, CXR = chest X-ray generating the “Findings” and “Impression” sections of a report (Fig. 5A) [4]. This automation can reduce radiologists’ workload, accelerate report delivery times, and improve consistency across examinations by standardizing language and minimizing repetitive dictation. Second, VQA allows clinicians—and potentially patients—to interactively query medical images and receive accurate, contextual responses (Fig. 5B) [50,62]. This interactive approach supports quick clinical decision-making and serves as an intuitive educational tool for trainees. Finally, text-to-image retrieval applications allow radiologists to search large imaging archives using natural language queries (e.g., “find all CXRs suggesting tuberculosis”) or retrieve relevant reports for a specific image [7]. These systems can streamline research cohort selection, support case reviews in team meetings, and enhance quality control by efficiently identifying similar studies.

Progress in 3D Medical Imaging MLLMs

The transition to 3D medical imaging in MLLMs is driven by the clinical need for detailed spatial information inherent in volumetric modalities like CT and MRI, which allows for superior pathology localization, disease staging, and surgical



Reference report

Injection of contrast via the orifice of purulent discharge showed opacification of the left perianal abscess. No evidence of obvious contrast extravasation into the anal canal or rectum.

Generated report

Injection of contrast via the orifice of purulent discharge showed opacification of the perianal abscess. No evidence of obvious contrast extravasation into the anal canal or rectum is seen.

Reference report

A 43*36mm hetero-enhancing mass is noted at right renal sinus. There is no sign of local invasion to adjacent structures and no vascular extension. No regional lymphadenopathy is identified. Degenerative changes as osteophytosis are seen at the lumbar spine. Grade I spondylolisthesis of L5 on S1 is present with bilateral spondylolysis.

Generated report

A 35 mm large exophytic mass is seen in the lower pole of the right kidney. The neoplasm has a similar density to normal renal parenchyma on non-enhanced CT scan. No regional lymphadenopathy is noted. After IV contrast media injection, the attenuation value didn't widened and the mass didn't enhance. Mild degenerative changes as osteophytosis are seen at the lumbar spine.

Input image	VQA Type	Example	Prediction
	Closed-ended	**Q:** Is there evidence of pleural effusion? **A:** Yes.	**A:** Yes
	Open-ended	**Q:** Which abnormalities can be detected in the right hilar region? **A:** Mild right-sided pleural effusion is noted.	**A:** Right pleural effusion is present, small in volume.
	Closed-ended	**Q:** What is the CT phase of the image? **Choices:** A. Non-contrast B. Contrast C. Arterial phase D. Portal venous phase **A:** C. Arterial phase	**A:** C. Arterial phase
	Open-ended	**Q:** Which contrast phase is shown in this CT image?	**A:** Arterial phase

Fig. 5. Representative vision-language tasks in radiology. **A:** Automated report generation using RadFM. The upper and lower panels illustrate generated reports compared to the corresponding reference reports. Keywords that are correctly matched or missed are highlighted using color coding to enhance interpretability. **B:** VQA. Examples include both closed-ended (choice-based) and open-ended (free-text) responses. Adapted from Wu et al., a preprint published under a CC BY license [71]. RadFM = radiology foundation model, VQA = visual question answering

planning compared to 2D imaging [63]. In contrast, applying 2D networks to individual slices is both computationally inefficient and fails to capture inter-slice contextual information. Similarly, selecting only a few representative slices risks overlooking critical spatial relationships [64].

However, developing true 3D MLLMs presents several significant challenges. These include the lack of large, annotated 3D datasets; the substantial computational demands from processing many voxels, leading to token explosion; difficulties in adapting pre-trained 2D vision models to 3D structures; and the need for new evaluation methods that can assess spatial features in a volumetric context [44,45,65,66]. Despite these obstacles, research in 3D medical MLLM has gained momentum, driven by pressing clinical needs and supported by the emergence of large-scale 3D medical image-text datasets [44,45].

New Datasets Fueling 3D MLLM Development

Several new datasets are accelerating 3D MLLM development by providing well-annotated volumetric studies with matching textual reports.

CT-RATE [67] includes over 25000 non-contrast chest CT scans from about 21000 patients, reconstructed into nearly 50000 volumes, each paired with the dictated radiology report. RadGenome-Chest CT [68] builds directly upon CT-RATE by incorporating organ- and lesion-level masks—spanning about 200 anatomical classes—generated via automated segmentation methods [69]. These anatomical annotations are aligned at the sentence level with the associated report text, enabling models to learn explicit voxel-to-language correspondences. This fine-grained alignment is an essential prerequisite for clinically meaningful, location-specific reasoning.

M3D-Data [45] extends the scope further by providing 120000 publicly available 3D studies paired with free-text descriptions, along with 662000 instruction–response exemplars covering key volumetric tasks, from RRG and VQA to slice-level localization and segmentation. To address privacy concerns, all data are sourced exclusively from open-access repositories.

MedErr-CT [70] augments approximately 3000 CT-RATE studies with 41000 question-answer pairs covering six clinically salient error categories. It provides a robust training and evaluation resource for developing error-aware models and represents the first benchmark designed to systematically assess a 3D MLLM's ability to classify, localize, and correct inaccuracies in radiology reports.

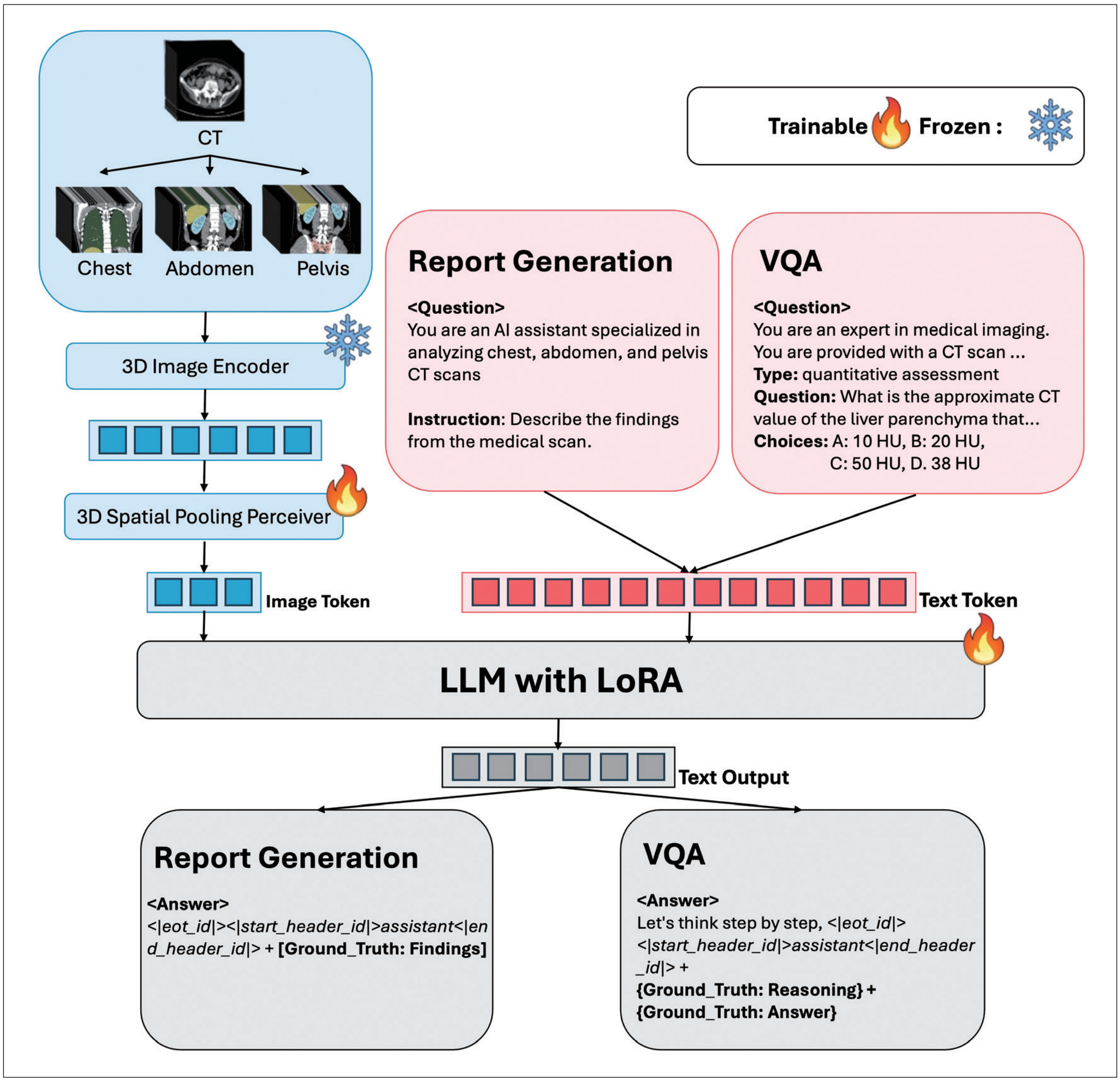
Collectively, these resources advance the field beyond basic image-report matching toward datasets that embed detailed anatomical context. Through voxel-wise or region-wise alignment of images and language, 3D MLLMs are now capable of generating targeted findings (e.g., "tiny cavity in the right upper lobe"), answering spatially grounded queries, and supporting nuanced clinical decision-making, capabilities that conventional paired datasets alone could not deliver.

Typical 3D MLLMs and Architectures

Recent architectural innovations aim to overcome the major computational and representational challenges of applying MLLMs to 3D medical images (Fig. 6). Table 4 summarizes several notable designs created for this purpose.

Radiology foundation model (RadFM) [71] integrates a 3D ViT with a perceiver that distills whole-volume features into a fixed 32-token sequence, enabling RRG and VQA on high-end GPUs, albeit with lower disease-specific accuracy compared to specialist models. MedBLIP [72] targets Alzheimer's MRI, pairing images with EHR data. It employs a learnable patch layer to adapt a frozen 2D ViT for 3D input, and uses the medical query transformer (MedQFormer) to filter task-relevant features, enabling strong zero-shot reporting. CT2Rep [63] tokenizes chest CT volumes for input into a 3D autoregressive transformer with relational memory. Cross-attention to prior scans mimics routine longitudinal comparisons, aligning automated CT reporting more closely to clinical workflows. M3D-LaMed [45] combines a 3D ViT with an aggressively pooled perceiver to process full-resolution volumes on a standard workstation. A downstream LLM performs RRG, VQA, retrieval, and segmentation, functioning as a multitask assistant. Med-2E3 [65] fuses global 3D and slice-level 2D encoders, enhancing RRG and VQA performance, particularly for subtle focal lesions often missed using purely volumetric compression.

Across these systems, performance gains rely less on scaling LLMs and more on radiology-specific intermediate modules. These include perceiver-based compression [45,71], query-driven feature filtering [72], sequential patch encoding [63], and dual-stream fusion [65], which collectively transform gigabyte-scale inputs into concise, diagnostically meaningful representations. Further gains are likely to stem from further refinement of such cross-modal connectors and the integration of prior-study context, rather than from simply scaling language parameters.



Report Generation

<Question>
You are an AI assistant specialized in analyzing chest, abdomen, and pelvis CT scans

Instruction: Describe the findings from the medical scan.

VQA

<Question>
You are an expert in medical imaging.
You are provided with a CT scan ...
Type: quantitative assessment
Question: What is the approximate CT value of the liver parenchyma that...
Choices: A: 10 HU, B: 20 HU, C: 50 HU, D. 38 HU

Report Generation

<Answer>
<|*eot_id*|><|*start_header_id*|>*assistant*<|*end_header_id*|> + **[Ground_Truth: Findings]**

VQA

<Answer>
Let's think step by step, <|*eot_id*|><|*start_header_id*|>*assistant*<|*end_header_id*|> +
{Ground_Truth: Reasoning} +
{Ground_Truth: Answer}

Fig. 6. Representative architecture for integrating 3D medical imaging with LLMs. To address the challenges of processing volumetric data, recent designs segment CT scans by anatomical region (e.g., chest, abdomen, pelvis) and employ frozen 3D encoders in combination with trainable fusion modules. Visual and textual inputs are integrated using LoRA-adapted LLMs to support tasks such as report generation and VQA. This figure illustrates a model developed in our laboratory based on the M3D-LaMed framework. Adapted from Bai et al., a preprint published under a CC BY license [45]. LLM = large language model, LoRA = low-rank adaptation, VQA = visual question answering

Advancing MLLMs for Medicine

MLLMs for medicine are rapidly advancing by addressing existing limitations and exploring new paradigms to enhance clinical utility. Key developments include improved spatial granularity by incorporating region-level visual details [16,45,63,73,74] and the use of large-scale pre-trained FMs for broad generalization [75,76]. Additionally, emerging paradigms, such as Vision-Language-Action (VLA) frameworks, further extend capabilities by integrating medical image interpretation with language-guided clinical actions [77,78]. Additionally, collaborative multi-agent approaches enable knowledge sharing among specialized AI agents, facilitating more comprehensive medical reasoning [79,80]. Collectively, these innovations lay the groundwork for next-generation medical MLLMs that can operate within real-world resource constraints while maintaining diagnostic fidelity.

Region-Focused MLLMs

While general research on MLLMs has shown that adding local visual details to LLMs significantly enhances their

Table 4. Summary of typical 3D multimodal LLMs in radiology

Model	Key architectural feature(s)	Primary task(s)	Training data highlights	Key strength/contribution
RadFM [71]	3D ViT + Perceiver Module	RRG, VQA (multimodal 2D/3D)	Pre-train: MedMD (16M scans); Fine-tune: RadMD (3M image-report pairs)	Foundation model concept for radiology, handles multiple modalities
MedBLIP [72]	Learnable 3D Patch Embedding + Frozen 2D Encoder + MedQFormer + Frozen LLM (LoRA)	Zero-shot Classification, VQA (3D + EHR)	AD datasets (30K MRI + EHR)	Lightweight, integrates 3D images with EHR text using query mechanism
CT2Rep [63]	3D Autoregressive Causal Transformer + Hierarchical Memory Decoder	3D CT RRG (chest)	CT-RATE	First dedicated 3D RRG model, sequential processing, longitudinal data integration
M3D-LaMed [45]	3D ViT + 3D Spatial Pooling Perceiver + LLM	RRG, VQA, retrieval, positioning, segmentation (3D)	M3D-Data (120K pairs, 662K instructions)	Efficient 3D token compression via perceiver, generalist across multiple 3D tasks
Med-2E3 [65]	Integrated 3D Encoder + 2D Encoder + Text-Guided Inter-Slice Scoring	RRG, VQA (3D)	Benchmarked on public 3D datasets (M3D-data)	Novel dual-encoder approach mimicking radiologist workflow, task-specific attention

LLM = large language model, RadFM = radiology foundation model, ViT = vision transformer, RRG = radiology report generation, VQA = visual question answering, MedMD = medical multimodal dataset, RadMD = radiology multimodal dataset, MedBLIP = medical bootstrapping language-image pre-training from 3D medical images and texts, MedQFormer = medical query transformer, LoRA = low-rank adaptation, EHR = electronic health record, AD = Alzheimer's disease, M3D-LaMed = a versatile multi-modal large language model for 3D medical image analysis, M3D-Data = 3D multi-modal medical dataset, Med-2E3 = 2D-enhanced 3D medical multimodal large language model

ability to reason about specific regions [73,74], similar work in medical imaging remains in its early stages. Most medical MLLMs still process entire medical images as single global units, limiting their ability to generate text descriptions about specific, clinically relevant areas. This limitation is especially problematic in radiology, where reports typically describe findings across multiple anatomical locations and disease processes [81]. To address these challenges, several region-focused medical MLLMs have been developed (Table 5).

Researchers have fine-tuned MLLM using specialized training data that connects textual descriptions to specific image locations, a process known as "Refer-and-Ground" conversations [82-84]. These datasets enable models to localize anatomical structures, detect lesions, and generate reports that reference specific image areas. For example, MAIRA-2 [81] established a benchmark for CXR reporting by linking each clinical finding to its exact location. Other efforts have focused on integrating regions-of-interest (ROI) features directly into language models. One such approach, Region-Guided Radiology Report Generation (RGRG) [85], identifies potential abnormalities, extracts features from these ROIs, and then uses these detailed features to generate reports. More recently, pixel-level guidance via semantic segmentation has become a powerful alternative. MAIRA-SEG [86], for example, uses a specialized system to generate segmentation tokens from outlines of major organs and lesions, enhancing the CXR information fed into the language model (Fig. 7). Similarly, Reg2RG [87] applies this approach to CT scans by providing organ outlines that maintain spatial relationships, while preserving texture details in the visual features, effectively combining detailed and overall information. MedRegion-CT [88] further extends region-focused modeling for chest CT by fusing global and segmentation-guided regional tokens with quantitative organ- and lesion-level attributes, generating organ-specific paragraphs that achieve SOTA accuracy. Spatial grounding will likely shape the next generation of radiology AI models. Systems that combine overall context with region-specific features can create focused, interpretable reports centered on specific findings.

Multimodal Foundation Models: Balancing Specialist and Generalist Approaches

AI in medical imaging is undergoing a significant transformation with the emergence of FMs [89]. These large-scale deep learning systems are trained on vast, diverse datasets, often using self-supervised learning to reduce the need for manual annotations [18,75,90,91]. Unlike traditional approaches that develop specialized models for specific tasks, FMs serve as flexible platforms that can be efficiently fine-tuned or prompted for various clinical applications, while requiring substantially less task-specific labeled data [92]. This evolution is driven by increased access to large, multi-institutional data repositories [45-50,67,68] and a growing recognition that complex clinical challenges require models with broad

Table 5. Summary of typical region-focused multimodal LLMs in radiology

Model	Imaging modality	Key input data	Key method/approach	Primary task(s)
MAIRA-2 [81]	CXR	CXR, text-bounding box pairs	Fine-tuned using "refer-and-ground" conversation corpora (free-text dialogue coupled to bounding-box coordinates). Links each reported finding in CXR reports to an annotated location	Formalized a grounded radiology report generation benchmark
RGRG [85]	CXR	CXR, detected ROI features	Detects or allows radiologists selection of candidate abnormalities. Utilizes features corresponding to these ROIs. Conditions the report generator on these refined, fine-grained embeddings	Report generation conditioned on ROI
MAIRA-SEG [86]	CXR	CXR, pseudo-masks (SEG info)	Leverages the mask-aware extractor of Osprey. Derives "SEG tokens" from pseudo-masks of major organs and lesions to enrich the CXR representation ingested by the LLM	Enhance CXR representation using pixel-level SEG
Reg2RG [87]	CT	CT, organ masks, embeddings	Uses universal SEG to supply organ masks (preserving geometric context). Retains local texture in accompanying visual embeddings. Enables coherent fusion of local (mask + texture) and global cues	Fuse local and global cues in CT imaging
MedRegion-CT [88]	CT	CT, pseudo organ/ lesion masks + patient-specific attributes	Pools global & SEG-guided region tokens via R^2 Token Pooling, adds mask-driven visual extractor and quantitative attribute prompts to condition the LLM, yielding organ-wise paragraphs	Region-grounded CT report generation

LLM = large language model, MAIRA = multimodal AI for radiology application, CXR = chest X-ray, RGRG = region-guided radiology report generation, ROI = region of interest, SEG = segmentation, Reg2RG = region-guided referring and grounding framework for report generation, R^2 = region representative

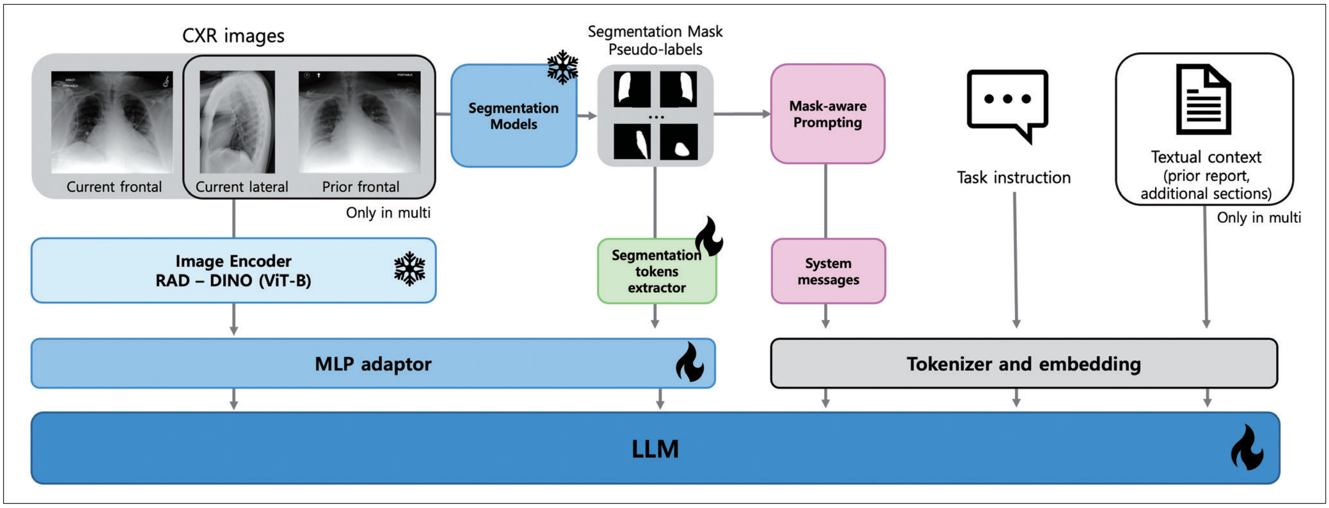


Fig. 7. Representative architecture of region-grounded MLLMs. This figure illustrates the MAIRA-Seg framework, a region-focused MLLM architecture that integrates segmentation-aware spatial tokens with CXR image and text inputs. A frozen vision encoder and segmentation model generate structured representations, which guide a trainable LLM in producing fine-grained, mask-aware radiology reports. Adapted from Sharma et al., *Proc Mach Learn Res* 2025;259:941-960, originally published under a CC BY license [86]. MLLM = multimodal large language model, MAIRA-Seg = Mask-Aware Instruction-tuned Radiology Assistant with Segmentation, CXR = chest X-ray, LLM = large language model, MLP = multi-layer perceptron

visual and semantic knowledge acquired during pre-training [75,93-95].

The heterogeneous nature of medical data and clinical needs has led to the development of various types of FMs. Some models are designed for specific imaging modalities—for example, BrainIAC [96] for MRI and MaCo [97] for CXRs. These models achieve high performance within their respective domains by capturing the unique characteristics of each imaging method, but they typically struggle when applied to other types of medical images [92]. Other FMs are even more narrowly focused, targeting specific anatomical regions or clinical tasks. For instance, MoME [98] specializes in brain lesion segmentation, while MedYOLO [99] is designed for 3D object detection. These specialized models achieve SOTA performance while reducing the annotation work needed for highly specialized clinical applications. More broadly, FMs can be viewed along a continuum from generalist to specialist. Generalist models, such as RadFM [71] and M4oE [100], are designed to operate across multiple imaging modalities, often incorporating diverse data types. New architectural approaches, particularly the Mixture of Experts framework used in models like MoME [98] and M4oE [100], seek to balance these approaches by using specialized expert sub-networks within a more flexible structure.

Perhaps most ambitiously, FMs facilitate multimodal data integration, striving to emulate clinical reasoning by synthesizing information from diverse sources like imaging, EHRs, text, and genomics [101,102]. The development of robust multimodal FMs, alongside comprehensive benchmarks such as CLIMB [103], is pivotal for achieving a holistic patient assessment and advancing precision medicine (Fig. 8) [104].

FMs represent a promising evolution in medical imaging AI, offering more generalizable, adaptable, and data-efficient systems. They speed up AI development through efficient fine-tuning, allowing quick adaptation to new tasks with minimal labeled data, a key advantage in settings involving rare diseases or limited datasets [97,105,106]. While challenges in data access, computational requirements, validation methods, and building clinical trust require significant attention, the potential to improve clinical workflows and patient outcomes provides strong motivation for continued responsible innovation in this field [75,91,107].

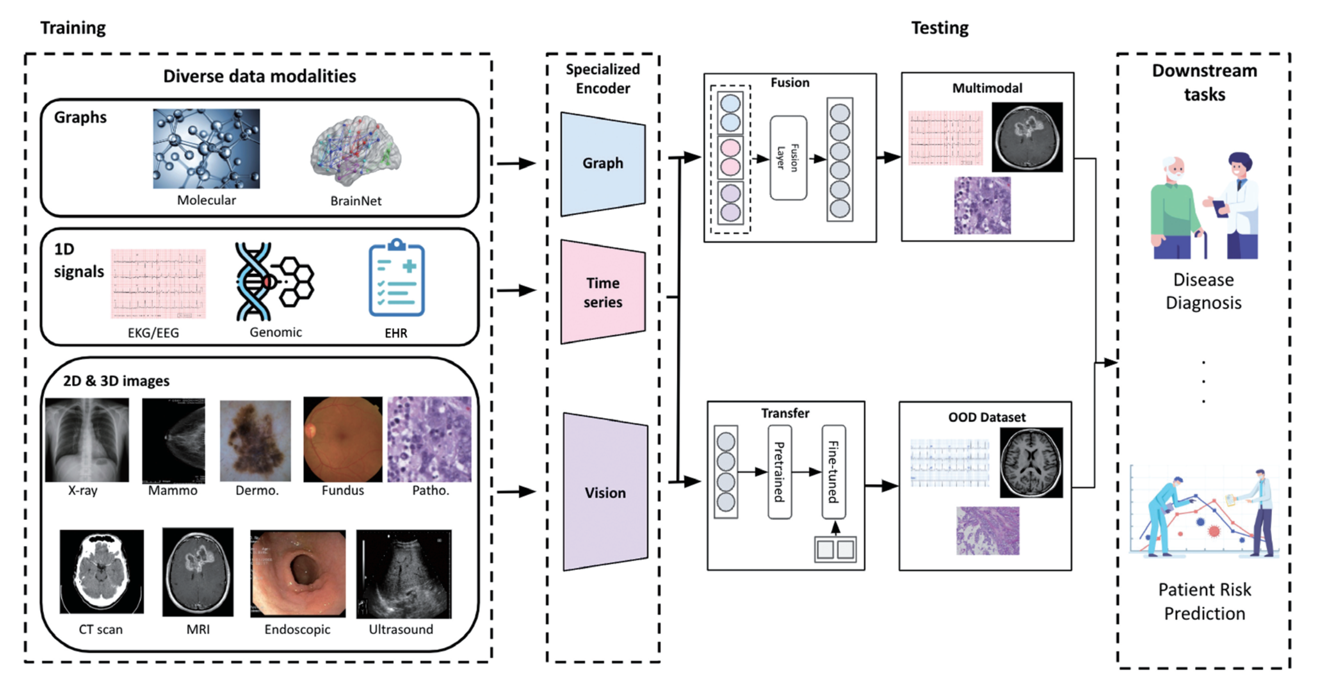


Fig. 8. Multimodal foundation model for integrated clinical reasoning. This figure illustrates the CLIMB framework, a multimodal foundation model that processes diverse medical data types (graphs, time-series signals, and 2D/3D images) via specialized encoders and unified fusion or transfer modules. Trained across multiple modalities, the model supports downstream tasks such as diagnosis and risk prediction with improved generalization, including for out-of-distribution data. Adapted from Dai et al., a preprint published under a CC BY SA license [103]. CLIMB = Continual Learning in Multimodality Benchmark, EKG = electrocardiogram, EEG = electroencephalography, EHR = electronic health record

Beyond MLLMs

As multimodal AI continues to evolve, VLA models have emerged as a new paradigm that integrates medical image perception, language-based interpretation, and embodied action planning within a unified framework of embodied AI [77]. Unlike traditional modular pipelines, VLA models aim to combine these components end-to-end, enabling systems to not only interpret vision findings but also actively support or carry out procedural tasks [78]. The fully autonomous robotic ultrasound platform reported by Su et al. [108] integrates real-time image acquisition, anatomical recognition, and robot-assisted probe manipulation, enabling thyroid scans to be performed without human interventions. Although most existing systems concentrate on perception and low-level control, recent survey articles have underscored the potential of embedding language-guided reasoning, whereby procedural instructions are parsed and translated into executable actions [78].

Furthermore, recent advances in collaborative learning paradigms among MLLMs have introduced novel approaches to enhance reasoning and generalization capabilities in medical imaging applications [79,80]. Notably, the model context protocol (MCP) and agent-to-agent (A2A) interaction frameworks facilitate effective inter-model communication and cooperation, offering new avenues for collaborative decision-making and task execution [109,110].

MCP establishes a unified client-server architecture that standardizes how MLLMs interact with external tools and data sources [111]. By enabling autonomous tool discovery and orchestration, MCP transforms passive models into active agents capable of context-aware operations, including API invocations and complex reasoning chains [109,112,113]. This protocol facilitates improved reasoning by allowing models to dynamically access and integrate diverse knowledge sources and computational resources. AI agents are autonomous systems designed to execute tasks independently by orchestrating workflows and leveraging available tools [114]. In multi-agent systems (MAS), these agents engage in A2A interactions to communicate and coordinate their activities [115]. MAS comprises multiple autonomous agents that collaborate to solve complex problems through distributed processing [116]. In healthcare settings, where diverse heterogeneous systems must integrate seamlessly, multi-agent architectures prove particularly valuable [110]. These agents can leverage their specialized capabilities and share domain-specific knowledge.

Challenges and Open Questions

MLLMs show great potential for medicine by integrating imaging with clinical data to enhance diagnostic accuracy and streamline workflows [117]. However, realizing this potential requires overcoming substantial hurdles spanning data acquisition, model reliability, technical implementation, evaluation, and clinical integration [7,8].

A major barrier is the lack of large-scale, high-quality multimodal medical datasets, particularly for 3D/4D imaging [118]. Developing such datasets is labor-intensive and requires expert annotation to support models capable of precise spatial reasoning. Additionally, data heterogeneity across institutions hinders the development of generalizable models [5]. Privacy regulations further complicate data sharing, necessitating advanced de-identification strategies or alternative approaches such as federated learning and synthetic data generation—methods whose effectiveness requires further validation [119-121].

Model trustworthiness is critical for clinical deployment. Multimodal LLMs are prone to hallucination, producing fluent but factually incorrect statements that may misguide diagnostic decisions and compromise patient safety [122]. While retrieval-augmented generation (RAG) can reduce this risk, it does not eliminate it entirely [123]. These models also inherit biases from their training data, potentially exacerbating performance disparities across demographic groups and deepening existing health inequities [124]. Effective risk mitigation hinges on balanced, expertly curated datasets, continuous post-deployment auditing, and cross-institutional benchmarking [125,126]. Furthermore, LLMs must transparently communicate uncertainty to prevent clinician over-reliance and to preserve critical human oversight.

The "black box" nature of complex MLLMs poses a significant barrier to clinical trust and adoption [127-129]. Interpretability—the ability to understand how a model reaches its conclusions—is essential for verification and responsible integration into clinical workflows [121,130,131]. Although explainable AI techniques such as attention maps offer some insight, achieving meaningful interpretability in multimodal radiology tasks remains challenging [132,133]. Effective explanations must bridge visual evidence and linguistic reasoning, ideally through fine-grained visual grounding that links specific image regions to generated text, mirroring established radiological practice.

Evaluating MLLMs in medicine requires moving beyond conventional natural language processing and computer vision metrics (e.g., BLEU, ROUGE, accuracy), which fail to capture clinically relevant outcomes (Fig. 9) [134-136]. Specialized frameworks that assess key dimensions such as factual consistency, clinical utility, safety, fairness, and spatial grounding are urgently needed [137,138]. In radiology, emerging domain-specific metrics, such as RadGraph [139], GREEN [140], SFVE [141], and LAVE [142], exemplify this shift toward more meaningful, task-aligned assessments. Developing robust, standardized benchmarks covering diverse tasks, modalities, and clinical scenarios is also vital for progress. Despite advancements in automated evaluation, expert radiologist review remains essential for determining whether MLLM outputs are clinically acceptable and trustworthy [137].

Finally, successful clinical integration of MLLMs requires more than technical performance alone [143,144]. These models must be seamlessly integrated into existing clinical workflows, such as PACS/RIS and EHR systems, without causing disruption. This requires careful attention to human-computer interaction and interface design (Fig. 10) [145,146]. In parallel, clear regulatory frameworks must be established to guide the validation, deployment, and ongoing monitoring of these adaptive, generative models, addressing issues like model drift, accountability, and legal liability [147]. Numerous ethical considerations, including patient privacy, bias mitigation, accountability, transparency, the impact on the profession, and preventing misuse, must be proactively addressed through interdisciplinary collaboration [148,149].

While MLLMs hold significant promise for transforming medical practice, their safe and effective clinical deployment depends on solving foundational challenges in data quality, reliability, explainability, technical design, evaluation, and ethical implementation [150]. Collaborative innovation across these areas is essential for responsibly implementing these AI systems to enhance medical practice and improve patient outcomes [143,144].

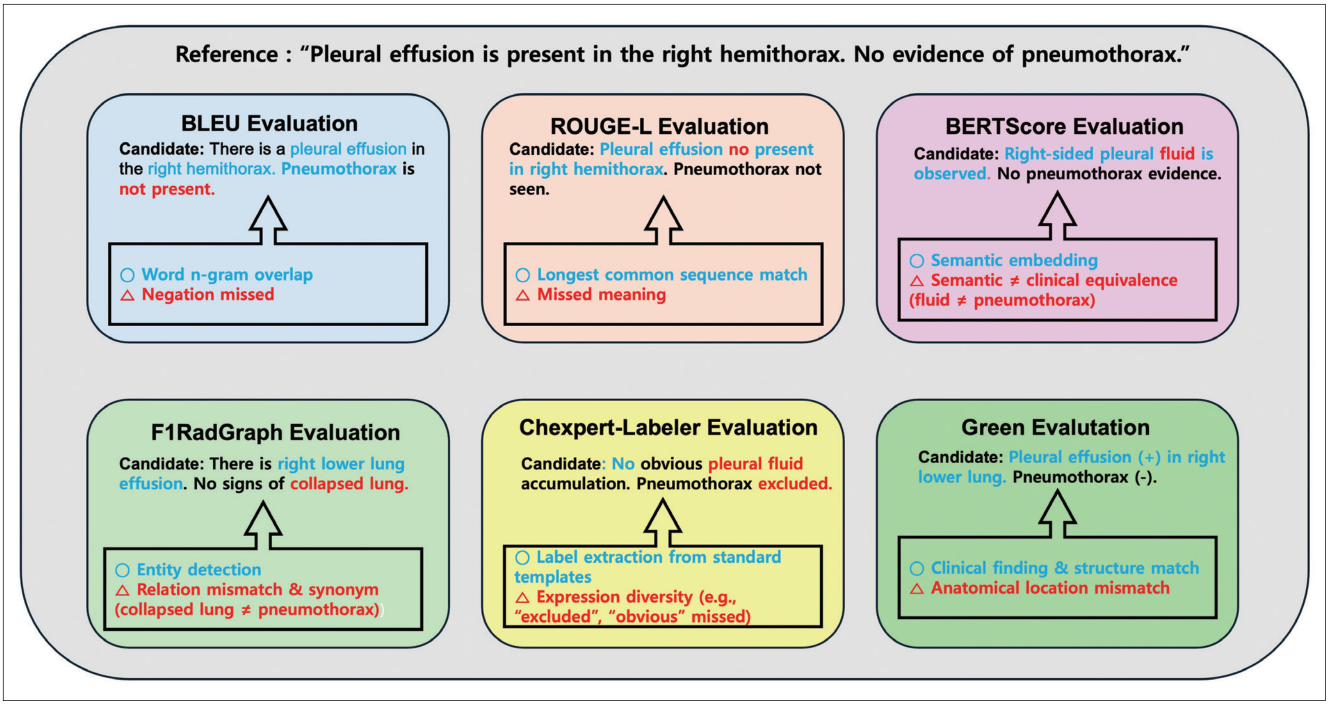

Reference : "Pleural effusion is present in the right hemithorax. No evidence of pneumothorax."
BLEU Evaluation
Candidate: There is a pleural effusion in the right hemithorax. Pneumothorax is not present.
○ Word n-gram overlap
△ Negation missed
ROUGE-L Evaluation
Candidate: Pleural effusion no present in right hemithorax. Pneumothorax not seen.
○ Longest common sequence match
△ Missed meaning
BERTScore Evaluation
Candidate: Right-sided pleural fluid is observed. No pneumothorax evidence.
○ Semantic embedding
△ Semantic ≠ clinical equivalence (fluid ≠ pneumothorax)
F1RadGraph Evaluation
Candidate: There is right lower lung effusion. No signs of collapsed lung.
○ Entity detection
△ Relation mismatch & synonym (collapsed lung ≠ pneumothorax)
Chexpert-Labeler Evaluation
Candidate: No obvious pleural fluid accumulation. Pneumothorax excluded.
○ Label extraction from standard templates
△ Expression diversity (e.g., "excluded", "obvious" missed)
Green Evalutation
Candidate: Pleural effusion (+) in right lower lung. Pneumothorax (-).
○ Clinical finding & structure match
△ Anatomical location mismatch

Fig. 9. Comparison of evaluation metrics for radiology report generation. Six evaluation methods are applied to a candidate report, highlighting their respective strengths (blue) and blind spots (red) relative to the reference: "Pleural effusion is present in the right hemithorax. No evidence of pneumothorax." Token-based (BLEU, ROUGE), semantic (BERTScore), and rule-based (CheXpert) metrics often fail to detect clinically significant errors, such as incorrect negation or anatomical misclassification. In contrast, entity- and structure-aware metrics (e.g., F1RadGraph, Green Score) better capture clinical correctness but still lack full interpretability. Blue = correctly recognized content, Red = clinically important errors not penalized, ○ = metric strengths (e.g., token match, clinical entity), Δ = metric limitations (e.g., missed negation or relation)

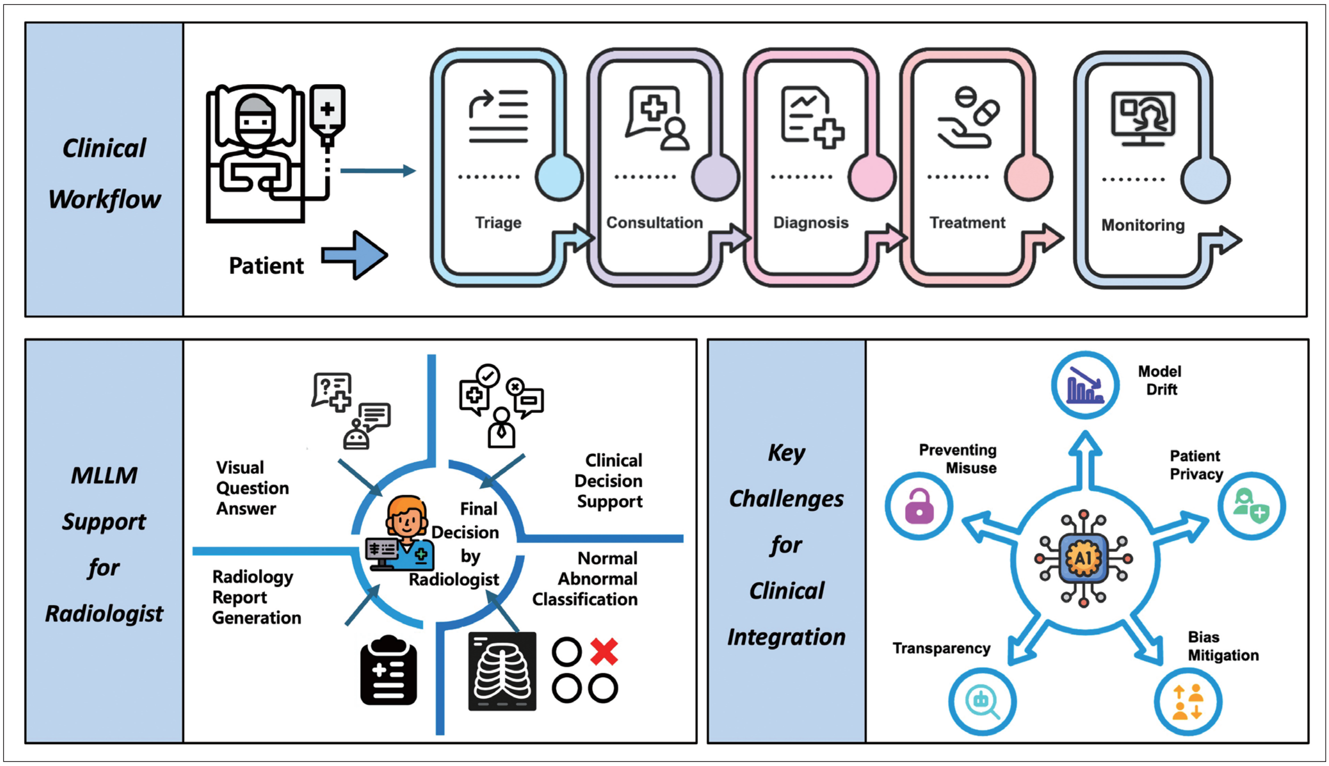

Clinical Workflow
Patient
Triage
Consultation
Diagnosis
Treatment
Monitoring
MLLM Support for Radiologist
Visual Question Answer
Clinical Decision Support
Final Decision by Radiologist
Radiology Report Generation
Normal Abnormal Classification
Key Challenges for Clinical Integration
Model Drift
Preventing Misuse
Patient Privacy
AI
Transparency
Bias Mitigation

Fig. 10. Clinical integration of MLLMs in radiology: workflow, support, and challenges. This figure illustrates the potential roles of MLLMs in supporting radiologists across the clinical workflow—from triage and diagnosis to treatment planning and monitoring. Key applications include report generation, visual question answering, abnormality detection, and clinical decision support. Furthermore, key challenges include model drift, patient privacy, bias mitigation, preventing misuse, and transparency, which must be addressed for safe integration into clinical systems. MLLM = multimodal large language model

CONCLUSION

MLLMs hold significant potential to transform medical practice by integrating imaging and clinical data within a unified inference framework. For example, FMs pre-trained on extensive visual-text corpora provide broad prior knowledge that enables MLLMs to connect subtle radiographic cues with patient context, propose coherent differential diagnoses, and draft preliminary reports. Spatially grounded reasoning—where textual outputs are anchored to specific image regions—further enhances interpretability and fosters clinician trust. Collectively, these advances position MLLMs as cognitive co-pilots capable of improving diagnostic accuracy, streamlining routine documentation, and delivering interactive decision support in daily clinical workflows.

Despite this promise, several critical challenges must be addressed before MLLMs can be safely and effectively adopted in clinical settings. Progress is hampered by limited, well-annotated multimodal datasets; the propensity to hallucinate or perpetuate biases learned from non-representative training data; and opaque decision pathways that undermine clinical explainability. Safe deployment, therefore, will require rigorous, multi-institutional validation to uncover performance gaps, accompanied by techniques that surface model rationale and uncertainty. Practical barriers, such as substantial computational demands and inference latency, must also be reduced to enable seamless integration into fast-paced imaging environments. Furthermore, clear regulatory and ethical guidance is essential to govern evaluation, monitoring, and accountability as these systems transition from laboratory to routine patient care.

Looking forward, interdisciplinary efforts should focus on refining model design and instituting safeguards to bridge the gap between experimental performance and clinical reliability. Promising directions include incorporating domain-specific constraints and modular expert components to balance generalizability with specialty-level accuracy, as well as implementing safety mechanisms such as output verification and clinician-in-the-loop review. With sustained innovation and careful governance, MLLMs could become trusted partners in clinical practice rather than opaque oracles, supporting practitioners in diagnosis and decision-making while maintaining appropriate human oversight. In time, these technologies have the potential to substantially augment clinicians' capabilities, improve diagnostic accuracy and efficiency, and ultimately enhance patient outcomes—but realizing this potential will require a measured, transparent approach to their integration into clinical practice.

Conflicts of Interest

Namkug Kim, who is on the editorial board of the *Korean Journal of Radiology*, was not involved in the editorial evaluation or decision to publish this article. The remaining author has declared no conflicts of interest.

Author Contributions

Conceptualization: Namkug Kim. Investigation: Yoojin Nam, Dong Yeong Kim, Sunggu Kyung, Jinyoung Seo, Jeong Min Song, Jimin Kwon, Jihyun Kim, Wooyoung Jo, Hyungbin Park, Jimin Sung, Taehee Kwon, Kanghyun Kim. Project administration: Yoojin Nam, Dong Yeong Kim. Supervision: Namkug Kim. Visualization: Yoojin Nam, Dong Yeong Kim, Jeong Min Song, Jimin Kwon, Jihyun Kim, Wooyoung Jo, Hyungbin Park, Jimin Sung, Sangah Park, Heeyeon Kwon. Writing—original draft: Yoojin Nam, Dong Yeong Kim. Writing—review & editing: Namkug Kim, Sunggu Kyung, Jinyoung Seo, Jeong Min Song, Jimin Kwon, Jihyun Kim, Wooyoung Jo, Hyungbin Park, Jimin Sung, Sangah Park, Heeyeon Kwon.

ORCID IDs

Yoojin Nam
https://orcid.org/0000-0001-8565-1360
Dong Yeong Kim
https://orcid.org/0000-0002-8548-7377
Sunggu Kyung
https://orcid.org/0000-0002-7582-9484
Jinyoung Seo
https://orcid.org/0009-0000-6828-2508
Jeong Min Song
https://orcid.org/0009-0009-0335-6903
Jimin Kwon
https://orcid.org/0009-0001-5923-9900
Jihyun Kim
https://orcid.org/0009-0003-0378-4170
Wooyoung Jo
https://orcid.org/0009-0000-7413-1351
Hyungbin Park
https://orcid.org/0009-0007-7762-3432
Jimin Sung
https://orcid.org/0009-0001-4053-4803
Sangah Park
https://orcid.org/0009-0004-1243-2625
Heeyeon Kwon
https://orcid.org/0009-0005-1283-6600
Taehee Kwon
https://orcid.org/0009-0000-2707-2176
Kanghyun Kim
https://orcid.org/0009-0009-6720-8045
Namkug Kim
https://orcid.org/0000-0002-3438-2217

Funding Statement

None

REFERENCES

1. Cheng PM, Montagnon E, Yamashita R, Pan I, Cadrin-Chênevert A, Perdigón Romero F, et al. Deep learning: an update for radiologists. *Radiographics* 2021;41:1427-1445
2. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: a radiologist's guide. *Radiology* 2019;290:590-606
3. Yildirim N, Richardson H, Wetscherek MT, Bajwa J, Jacob J, Pinnock MA, et al. Multimodal healthcare AI: identifying and designing clinically relevant vision-language applications for radiology [accessed on May 1, 2025]. Available at: https://doi.org/10.1145/3613904.3642013
4. Nakaura T, Ito R, Ueda D, Nozaki T, Fushimi Y, Matsui Y, et al. The impact of large language models on radiology: a guide for radiologists on the latest innovations in AI. *Jpn J Radiol* 2024;42:685-696
5. Tariq A, Banerjee I, Trivedi H, Gichoya J. Multimodal artificial intelligence models for radiology. *BJR Artif Intell* 2025;2:ubae017
6. Soni N, Ora M, Agarwal A, Yang T, Bathla G. A review of the opportunities and challenges with large language models in radiology: the road ahead. *AJNR Am J Neuroradiol* 2025;46:1292-1299
7. Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, et al. A survey on multimodal large language models. *Natl Sci Rev* 2024;11:nwae403
8. AlSaad R, Abd-Alrazaq A, Boughorbel S, Ahmed A, Renault MA, Damseh R, et al. Multimodal large language models in health care: applications, challenges, and future outlook. *J Med Internet Res* 2024;26:e59505

9. Kim S, Lee CK, Kim SS. Large language models: a guide for radiologists. *Korean J Radiol* 2024;25:126-133
10. Wang J, Jiang H, Liu Y, Ma C, Zhang X, Pan Y, et al. A comprehensive review of multimodal large language models: performance and challenges across different tasks. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2408.01319
11. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology* 2024;310:e232756
12. Matarazzo A, Torlone R. A survey on large language models with some insights on their capabilities and limitations. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2501.04040
13. David M, Tom E, Paul B. Major large language models (LLMs): ranked by capabilities, sized by billion parameters used for training [accessed on May 1, 2025]. Available at: https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt
14. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv [Preprint]. 2020 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2010.11929
15. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. *Proc Mach Learn Res* 2021;139:8748-8763
16. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2304.08485
17. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol* 2024;30:80-90
18. Paschali M, Chen Z, Blankemeier L, Varma M, Youssef A, Bluethgen C, et al. Foundation models in radiology: what, how, why, and why not. *Radiology* 2025;314:e240597
19. Laurençon H, Tronchon L, Cord M, Sanh V. What matters when building vision-language models? arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2405.02246
20. Gholami A, Kim S, Dong Z, Yao Z, Mahoney MW, Keutzer K. *A survey of quantization methods for efficient neural network inference*. In: Thiruvathukal GK, Lu YH, Kim J, Chen Y, Chen B, eds. *Low-power computer vision*. New York: Chapman and Hall/CRC, 2022:291-326
21. Han S, Mao H, Dally WJ. Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv [Preprint]. 2015 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.1510.00149
22. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv [Preprint]. 2015 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.1503.02531
23. Fang Y, Wang W, Xie B, Sun Q, Wu L, Wang X, et al. EVA: exploring the limits of masked visual representation learning at scale [accessed on May 1, 2025]. Available at: http://doi.org/10.1109/CVPR52729.2023.01855
24. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. *End-to-end object detection with transformers*. In: Vedaldi A, Bischof H, Brox T, Frahm JM, eds. *Computer vision – ECCV 2020*. Cham: Springer, 2020:213-229
25. Li J, Li D, Savarese S, Hoi S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *Proc Mach Learn Res* 2023;202:19730-19742
26. Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: a visual language model for few-shot learning. arXiv [Preprint]. 2022 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2204.14198
27. Wang W, Lv Q, Yu W, Hong W, Qi J, Wang Y, et al. CogVLM: visual expert for pretrained language models. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2311.03079
28. Li K, He Y, Wang Y, Li Y, Wang W, Luo P, et al. VideoChat: chat-centric video understanding. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2305.06355
29. Yin S, Fu C, Zhao S, Xu T, Wang H, Sui D, et al. Woodpecker: hallucination correction for multimodal large language models. *Sci China Inf Sci* 2024;67:220105
30. Li Z, Wu X, Du H, Liu F, Nghiem H, Shi G. A survey of state of the art large vision language models: alignment, benchmark, evaluations and challenges. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2501.02189
31. Liu E, Neubig G, Andreas J. An incomplete loop: instruction inference, instruction following, and in-context learning in language models. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2404.03028
32. Caffagni D, Cocchi F, Barsellotti L, Moratelli N, Sarto S, Baraldi L, et al. The revolution of multimodal large language models: a survey. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2402.12451
33. Chen L, Li J, Dong X, Zhang P, He C, Wang J, et al. *ShareGPT4V: improving large multi-modal models with better captions*. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, eds. *Computer Vision – ECCV 2024*. Cham: Springer, 2024:370-387
34. Hong GZ, Cui Y, Fuxman A, Chan S, Luo E. Why fine-grained labels in pretraining benefit generalization? arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2410.23129
35. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv [Preprint]. 2021 [accessed on May 1, 2025]. Available at:

https://doi.org/10.48550/arXiv.2106.09685
36. Zhu D, Chen J, Shen X, Li X, Elhoseiny M. MiniGPT-4: enhancing vision-language understanding with advanced large language models. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2304.10592
37. Xu Z, Shen Y, Huang L. MultiInstruct: improving multi-modal zero-shot learning via instruction tuning. arXiv [Preprint]. 2022 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2212.10773
38. Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, et al. Fine-tuning language models from human preferences. arXiv [Preprint]. 2019 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.1909.08593
39. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2303.08774
40. Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, Schalkwyk J, et al. Gemini: a family of highly capable multimodal models. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2312.11805
41. Anthropic A. The Claude 3 model family: Opus, Sonnet, Haiku [accessed on May 1, 2025]. Available at: https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf
42. Ebrahimi S, Arik SÖ, Nama T, Pfister T. CROME: cross-modal adapters for efficient multimodal LLM. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2408.06610
43. Marafioti A, Zohar O, Farré M, Noyan M, Bakouch E, Cuenca P, et al. SmolVLM: redefining small and efficient multimodal models. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2504.05299
44. Li CY, Chang KJ, Yang CF, Wu HY, Chen W, Bansal H, et al. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. *Nat Commun* 2025;16:2258
45. Bai F, Du Y, Huang T, Meng MQH, Zhao B. M3D: advancing 3D medical image analysis with multi-modal large language models. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2404.00578
46. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* 2019;6:317
47. Wu JT, Agu NN, Lourentzou I, Sharma A, Paguio JA, Yao JS, et al. Chest ImaGenome dataset for clinical reasoning. arXiv [Preprint]. 2021 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2108.00316
48. Lau JJ, Gayen S, Ben Abacha A, Demner-Fushman D. A dataset of clinically generated visual questions and answers about radiology images. *Sci Data* 2018;5:180251
49. Liu B, Zhan LM, Xu L, Ma L, Yang Y, Wu XM. Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering [accessed on May 1, 2025]. Available at: http://doi.org/10.1109/ISBI48211.2021.9434010
50. Zhang X, Wu C, Zhao Z, Lin W, Zhang Y, Wang Y, et al. PMC-VQA: visual instruction tuning for medical visual question answering. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2305.10415
51. Li Q, Li L, Li Y. Developing ChatGPT for biology and medicine: a complete review of biomedical question answering. *Biophys Rep* 2024;10:152-171
52. Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. *Proc Mach Learn Res* 2022;182:2-25
53. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: contrastive learning from unpaired medical images and text. arXiv [Preprint]. 2022 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2210.10163
54. Bannur S, Hyland S, Liu Q, Perez-Garcia F, Ilse M, Castro DC, et al. Learning to exploit temporal structure for biomedical vision-language processing. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2301.04558
55. Boecking B, Usuyama N, Bannur S, Castro DC, Schwaighofer A, Hyland S, et al. *Making the most of text semantics to improve biomedical vision–language processing*. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, eds. *Computer vision – ECCV 2022*. Cham: Springer, 2022:1-21
56. Zhang S, Xu Y, Usuyama N, Xu H, Bagga J, Tinn R, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2303.00915
57. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2306.00890
58. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, et al. Towards generalist biomedical AI. *NEJM AI* 2024;1:AIoa2300138
59. Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhery A, Ichter B, et al. PaLM-E: an embodied multimodal language model [accessed on May 1, 2025]. Available at: https://proceedings.mlr.press/v202/driess23a.html
60. Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, et al. Med-Flamingo: a multimodal medical few-shot learner. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2307.15189
61. Thawakar OC, Shaker AM, Mullappilly SS, Cholakkal H, Anwer RM, Khan S, et al. XrayGPT: chest radiographs summarization using large medical vision-language models [accessed on May 1, 2025]. Available at: http://doi.org/10.18653/v1/2024.bionlp-1.35
62. Agrawal A, Lu J, Antol S, Mitchell M, Zitnick CL, Batra D, et al. VQA: visual question answering. arXiv [Preprint]. 2015 [accessed on May 1, 2025]. Available at: https://doi.

org/10.48550/arXiv.1505.00468
63. Hamamci IE, Er S, Menze B. *CT2Rep: automated radiology report generation for 3D medical imaging.* In: Linguraru MG, Dou Q, Feragen A, Giannarou S, Glocker B, Lekadir K, et al., eds. *Medical image computing and computer assisted intervention – MICCAI 2024.* Cham: Springer, 2024:476-486
64. Mello-Thoms C, Abbey CK, Krupinski EA. Introducing the special series on 2D and 3D imaging: perspectives in human and model observer performance. *J Med Imaging (Bellingham)* 2020;7:051201
65. Shi Y, Zhu X, Hu Y, Guo C, Li M, Wu J. Med-2E3: a 2D-enhanced 3D medical multimodal large language model. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2411.12783
66. Liu C, Wan Z, Wang Y, Shen H, Wang H, Zheng K, et al. Argus: benchmarking and enhancing vision-language models for 3D radiology report generation. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2406.07146
67. Hamamci IE, Er S, Wang C, Almas F, Simsek AG, Esirgun SN, et al. Developing generalist foundation models from a multimodal dataset for 3D computed tomography. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2403.17834
68. Zhang X, Wu C, Zhao Z, Lei J, Zhang Y, Wang Y, et al. RadGenome-chest CT: a grounded vision-language dataset for chest CT analysis. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2404.16754
69. Zhao Z, Zhang Y, Wu C, Zhang X, Zhou X, Zhang Y, et al. Large-vocabulary segmentation for medical images with text prompts. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2312.17183
70. Kyung S, Park H, Seo J, Sung J, Kim J, Kim D, et al. MedErr-CT: a visual question answering benchmark for identifying and correcting errors in CT reports. arXiv [Preprint]. 2025 [accessed on June 30, 2025]. Available at: https://doi.org/10.48550/arXiv.2506.19217
71. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2308.02463
72. Chen Q, Hu X, Wang Z, Hong Y. MedBLIP: bootstrapping language-image pre-training from 3D medical images and texts. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2305.10799
73. Guo Q, De Mello S, Yin H, Byeon W, Cheung KC, Yu Y, et al. RegionGPT: towards region understanding vision language model. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2403.02330
74. Ma C, Jiang Y, Wu J, Yuan Z, Qi X. *Groma: localized visual tokenization for grounding multimodal large language models.* In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, eds. *Computer vision – ECCV 2024.* Cham: Springer, 2024:417-435
75. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. *Nature* 2023;616:259-265
76. Shurrab S, Duwairi R. Self-supervised learning methods and applications in medical imaging analysis: a survey. *PeerJ Comput Sci* 2022;8:e1045
77. Yue Y, Wang Y, Jiang H, Liu P, Song S, Huang G. EchoWorld: learning motion-aware world models for echocardiography probe guidance. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2504.13065
78. Sapkota R, Cao Y, Roumeliotis KI, Karkee M. Vision-language-action models: concepts, progress, applications and challenges. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2505.04769
79. Chen W, Zhao Z, Yao J, Zhang Y, Bu J, Wang H. Multi-modal medical diagnosis via large-small model collaboration (ExHall D Poster #442). Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025; 2025 June 11-15; Nashville, TN, USA, Piscataway: IEEE, 2025
80. Tanno R, Barrett DGT, Sellergren A, Ghaisas S, Dathathri S, See A, et al. Collaboration between clinicians and vision-language models in radiology report generation. *Nat Med* 2025;31:599-608
81. Bannur S, Bouzid K, Castro DC, Schwaighofer A, Thieme A, Bond-Taylor S, et al. MAIRA-2: grounded radiology report generation. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2406.04449
82. Wang L, Wang H, Yang H, Mao J, Yang Z, Shen J, et al. Interpretable bilingual multimodal large language model for diverse biomedical tasks. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2410.18387
83. Huang X, Huang H, Shen L, Yang Y, Shang F, Liu J, et al. *A refer-and-ground multimodal large language model for biomedicine.* In: Linguraru MG, Dou Q, Feragen A, Giannarou S, Glocker B, Lekadir K, et al., eds. *Medical image computing and computer assisted intervention – MICCAI 2024.* Cham: Springer, 2024:399-409
84. Zhou HY, Acosta JN, Adithan S, Datta S, Topol EJ, Rajpurkar P. MedVersa: a generalist foundation model for medical image interpretation. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2405.07988
85. Tanida T, Müller P, Kaissis G, Rueckert D. Interactive and explainable region-guided radiology report generation [accessed on May 1, 2025]. Available at: https://doi.org/10.1109/CVPR52729.2023.00718
86. Sharma H, Salvatelli V, Srivastav S, Bouzid K, Bannur S, Castro DC, et al. MAIRA-Seg: enhancing radiology report generation with segmentation-aware multimodal large language models. *Proc Mach Learn Res* 2025;259:941-960

87. Chen Z, Bie Y, Jin H, Chen H. Large language model with region-guided referring and grounding for CT report generation. *IEEE Trans Med Imaging* 2025;44:3139-3150
88. Kyung S, Seo J, Lim H, Kim D, Park H, Sung J, et al. MedRegion-CT: region-focused multimodal LLM for comprehensive 3D CT report generation. arXiv [Preprint]. 2025 [accessed on June 30, 2025]. Available at: https://doi.org/10.48550/arXiv.2506.23102
89. Jung KH. Uncover this tech term: foundation model. *Korean J Radiol* 2023;24:1038-1041
90. Huang SC, Jensen M, Yeung-Levy S, Lungren MP, Poon H, Chaudhari AS. Multimodal foundation models for medical imaging-a systematic review and implementation guidelines. medRxiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.1101/2024.10.23.24316003
91. Khan W, Leem S, See KB, Wong JK, Zhang S, Fang R. A comprehensive survey of foundation models in medicine. *IEEE Rev Biomed Eng* 2025 May 6 [Epub]. Available at: http://doi.org/10.1109/RBME.2025.3531360
92. Zhang S, Metaxas D. On the challenges and perspectives of foundation models for medical image analysis. *Med Image Anal* 2024;91:102996
93. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun* 2024;15:654
94. You K, Gu J, Ham J, Park B, Kim J, Hong EK, et al. *CXR-CLIP: toward large scale chest X-ray language-image pre-training*. In: Greenspan H, Madabhushi A, Mousavi P, Salcudean S, Duncan J, Syeda-Mahmood T, et al., eds. *Medical image computing and computer assisted intervention – MICCAI 2023*. Cham: Springer, 2023:101-111
95. Jang J, Kyung D, Kim SH, Lee H, Bae K, Choi E. Significantly improving zero-shot X-ray pathology classification via fine-tuning pre-trained image-text encoders. *Sci Rep* 2024;14:23199
96. Tak D, Garomsa BA, Chaunzwa TL, Zapaishchykova A, Pardo JCC, Ye Z, et al. A foundation model for generalized brain MRI analysis. medRxiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: http://doi.org/10.1101/2024.12.02.24317992
97. Huang W, Li C, Zhou HY, Yang H, Liu J, Liang Y, et al. Enhancing representation in radiography-reports foundation model: a granular alignment algorithm using masked contrastive learning. *Nat Commun* 2024;15:7620
98. Zhang X, Ou N, Basaran BD, Visentin M, Qiao M, Gu R, et al. *A foundation model for brain lesion segmentation with mixture of modality experts*. In: Linguraru MG, Dou Q, Feragen A, Giannarou S, Glocker B, Lekadir K, et al., eds. *Medical image computing and computer assisted intervention – MICCAI 2024*. Cham: Springer, 2024:379-389
99. Sobek J, Medina Inojosa JR, Medina Inojosa BJ, Rassoulinejad-Mousavi SM, Conte GM, Lopez-Jimenez F, et al. MedYOLO: a medical image object detection framework. *J Imaging Inform Med* 2024;37:3208-3216
100. Jiang Y, Shen Y. *M4oE: a foundation model for medical multimodal image segmentation with mixture of experts*. In: Linguraru MG, Dou Q, Feragen A, Giannarou S, Glocker B, Lekadir K, et al., eds. *Medical image computing and computer assisted intervention – MICCAI 2024*. Cham: Springer, 2024:621-631
101. Soenksen LR, Ma Y, Zeng C, Boussioux L, Villalobos Carballo K, Na L, et al. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digit Med* 2022;5:149
102. Taleb A, Kirchler M, Monti R, Lippert C. ContIG: self-supervised multimodal contrastive learning for medical imaging with genetics. arXiv [Preprint]. 2021 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2111.13424
103. Dai W, Chen P, Lu M, Li D, Wei H, Cui H, et al. CLIMB: data foundations for large scale multimodal clinical foundation models. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2503.07667
104. Cui H, Tejada-Lapuerta A, Brbić M, Saez-Rodriguez J, Cristea S, Goodarzi H, et al. Towards multimodal foundation models in molecular cell biology. *Nature* 2025;640:623-633
105. Wang D, Wang X, Wang L, Li M, Da Q, Liu X, et al. A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Sci Data* 2023;10:574
106. Schäfer R, Nicke T, Höfener H, Lange A, Merhof D, Feuerhake F, et al. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nat Comput Sci* 2024;4:495-509
107. Bian Y, Li J, Ye C, Jia X, Yang Q. Artificial intelligence in medical imaging: from task-specific models to large-scale foundation models. *Chin Med J (Engl)* 2025;138:651-663
108. Su K, Liu J, Ren X, Huo Y, Du G, Zhao W, et al. A fully autonomous robotic ultrasound system for thyroid scanning. *Nat Commun* 2024;15:4004
109. Hou X, Zhao Y, Wang S, Wang H. Model context protocol (MCP): landscape, security threats, and future research directions. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2503.23278
110. Moritz M, Topol E, Rajpurkar P. Coordinated AI agents for advancing healthcare. *Nat Biomed Eng* 2025;9:432-438
111. Narajala VS, Habler I. Enterprise-grade security for the model context protocol (MCP): frameworks and mitigation strategies. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2504.08623
112. Borgeaud S, Mensch A, Hoffmann J, Cai T, Rutherford E, Millican K, et al. Improving language models by retrieving from trillions of tokens. *Proc Mach Learn Res* 2022;162:2206-2240
113. Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y. HuggingGPT: solving AI tasks with ChatGPT and its friends in hugging face. arXiv [Preprint]. 2023 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2303.17580
114. Krishnan N. AI agents: evolution, architecture, and real-world applications. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2503.12687
115. Chen X, Yi H, You M, Liu W, Wang L, Li H, et al. Enhancing

diagnostic capability with multi-agents conversational large language models. *NPJ Digit Med* 2025;8:159

116. Ma C, Li A, Du Y, Dong H, Yang Y. Efficient and scalable reinforcement learning for large-scale network control. *Nat Mach Intell* 2024;6:1006-1020
117. Zha J, Fan Y, Yang X, Gao C, Chen X. How to enable LLM with 3D capacity? A survey of spatial reasoning in LLM. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2504.05786
118. Yi Z, Xiao T, Albert MV. A survey on multimodal large language models in radiology for report generation and visual question answering. *Information* 2025;16:136
119. Alkaeed M, Abioye S, Qayyum A, Mekki YM, Berrou I, Abdallah M, et al. Open foundation models in healthcare: challenges, paradoxes, and opportunities with genai driven personalized prescription. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2502.04356
120. Das A, Jha D, Sanjotra J, Susladkar O, Sarkar S, Rauniyar A, et al. Ethical framework for responsible foundational models in medical imaging. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2406.11868
121. Mastoi QU, Latif S, Brohi S, Ahmad J, Alqhatani A, Alshehri MS, et al. Explainable AI in medical imaging: an interpretable and collaborative federated learning model for brain tumor classification. *Front Oncol* 2025;15:1535478
122. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. *Interact J Med Res* 2025;14:e59823
123. Chu YW, Zhang K, Malon C, Min MR. Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2502.15040
124. Kim Y, Jeong H, Chen S, Li SS, Lu M, Alhamoud K, et al. Medical hallucinations in foundation models and their impact on healthcare. arXiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2503.05777
125. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *NPJ Digit Med* 2025;8:154
126. Karargyris A, Umeton R, Sheller MJ, Aristizabal A, George J, Wuest A, et al. Federated benchmarking of medical artificial intelligence with MedPerf. *Nat Mach Intell* 2023;5:799-810
127. Qin H, Tong Y. Opportunities and challenges for large language models in primary health care. *J Prim Care Community Health* 2025;16:21501319241312571
128. Omar M, Sorin V, Collins JD, Reich D, Freeman R, Gavin N, et al. Large language models are highly vulnerable to adversarial hallucination attacks in clinical decision support: a multi-model assurance analysis. medRxiv [Preprint]. 2025 [accessed on May 1, 2025]. Available at: https://doi.org/10.1101/2025.03.18.25324184
129. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol* 2024;25:865-868
130. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, et al. Explainable AI in medical imaging: an overview for clinical practitioners - Saliency-based XAI approaches. *Eur J Radiol* 2023;162:110787
131. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *NPJ Digit Med* 2022;5:156
132. Wang AQ, Karaman BK, Kim H, Rosenthal J, Saluja R, Young SI, et al. A framework for interpretability in machine learning for medical imaging. *IEEE Access* 2024;12:53277-53292
133. Champendal M, Müller H, Prior JO, Dos Reis CS. A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging. *Eur J Radiol* 2023;169:111159
134. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation [accessed on May 1, 2025]. Available at: https://aclanthology.org/P02-1040.Pdf
135. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments [accessed on May 1, 2025]. Available at: https://aclanthology.org/W05-0909.pdf
136. Lin CY. ROUGE: a package for automatic evaluation of summaries [accessed on May 1, 2025]. Available at: https://aclanthology.org/W04-1013.pdf
137. Park SH, Han K, Lee JG. Conceptual review of outcome metrics and measures used in clinical evaluation of artificial intelligence in radiology. *Radiol Med* 2024;129:1644-1655
138. Krishna S, Bhambra N, Bleakney R, Bhayana R. Evaluation of reliability, repeatability, robustness, and confidence of GPT-3.5 and GPT-4 on a radiology board-style examination. *Radiology* 2024;311:e232715
139. Jain S, Agrawal A, Saporta A, Truong SQ, Duong DN, Bui T, et al. RadGraph: extracting clinical entities and relations from radiology reports. arXiv [Preprint]. 2021 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2106.14463
140. Ostmeier S, Xu J, Chen Z, Varma M, Blankemeier L, Bluethgen C, et al. GREEN: generative radiology report evaluation and error notation. arXiv [Preprint]. 2024 [accessed on May 1, 2025]. Available at: https://doi.org/10.48550/arXiv.2405.03595
141. Ji H, Si Q, Lin Z, Wang W. Towards flexible evaluation for generative visual question answering [accessed on May 1, 2025]. Available at: https://doi.org/10.1145/3664647.3681400
142. Mañas O, Krojer B, Agrawal A. Improving automatic VQA evaluation using large language models [accessed on May 1, 2025]. Available at: https://doi.org/10.1609/aaai.

v38i5.28212

143. Jabbour S, Fouhey D, Shepard S, Valley TS, Kazerooni EA, Banovic N, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. *JAMA* 2023;330:2275-2284

144. Liu H, Ding N, Li X, Chen Y, Sun H, Huang Y, et al. Artificial intelligence and radiologist burnout. *JAMA Netw Open* 2024;7:e2448714

145. Juluru K, Shih HH, Keshava Murthy KN, Elnajjar P, El-Rowmeim A, Roth C, et al. Integrating Al algorithms into the clinical workflow. *Radiol Artif Intell* 2021;3:e210013

146. Park SH, Langlotz CP. Crucial role of understanding in human-artificial intelligence interaction for successful clinical adoption. *Korean J Radiol* 2025;26:287-290

147. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023;6:120

148. Menz BD, Kuderer NM, Bacchi S, Modi ND, Chin-Yee B, Hu T, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. *BMJ* 2024;384:e078538

149. Huang X, Wang X, Zhang H, Zhu Y, Xi J, An J, et al. Medical MLLM is vulnerable: cross-modality jailbreak and mismatched attacks on medical multimodal large language models. *Proc AAAI Conf Artif Intell* 2025;39:3797-3805

150. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol* 2021;31:3797-3804