Intelligent Oncology 1 (2025) 308–325

Contents lists available at ScienceDirect

Intelligent Oncology

journal homepage: www.keaipublishing.com/en/intonc/

Review

Medical multimodal large language models: A systematic review

Yuan Hu[a,#], Chenhan Xu[b,#], Bo Lin[b,*], Weibin Yang[b,*], Yuan Yan Tang[c,d,*]

[a] School of Medicine, Chongqing University, Chongqing 400030, China
[b] Chongqing Key Laboratory for Intelligent Oncology in Breast Cancer (iCQBC), Chongqing University Cancer Hospital, Chongqing 400030, China
[c] Faculty of Science and Technology, University of Macau, Macao SAR 999078, China
[d] Faculty of Science and Technology, UOW College Hong Kong, Hong Kong SAR 999077, China

ARTICLE INFO

Keywords:
Multimodal large language model
Hallucination
Medical multimodal dataset
Clinical evaluation

ABSTRACT

The rapid advancement of artificial intelligence (AI) has ushered in a new era of medical multimodal large language models (MLLMs), which integrate diverse data modalities such as text, imaging, physiological signals, and genomics to enhance clinical decision-making. This systematic review explores the core methodologies and applied research frontiers of medical MLLMs, focusing on their architecture, training methods, evaluation techniques, and applications. We highlight the transformative potential of MLLMs in achieving cross-modal semantic alignment, medical knowledge integration, and robust clinical reasoning. Despite their promise, challenges such as data heterogeneity, hallucination, and computational efficiency persist. By reviewing state-of-the-art solutions and future directions, this paper provides a comprehensive technical guide for developing reliable and interpretable medical MLLMs, ultimately aiming to bridge the gap between AI and clinical practice.

1. Introduction

In recent decades, the exponential growth of medical data has driven significant advancements in analytical methodologies for intelligent healthcare systems. The multimodal paradigm has become fundamental to modern healthcare analytics, enabling a more comprehensive and integrated analysis of complex medical information.[1] Medical data is characterized by substantial heterogeneity, spanning a wide range of modalities such as text (e.g., medical literature), imaging, physiological signals, and genomics data. Conventional medical artificial intelligence (AI) models are typically designed to operate on a single modality. Although these models demonstrate proficiency in specific tasks, they struggle to replicate real-world clinical decision-making, which inherently relies on synthesizing multiple sources of information.[2,3] This limitation underscores the need for more integrated multimodal approaches.

The emergence of Transformer-based architectures,[4] notably due to their self-attention (SA) mechanism and parallel processing capabilities, has profoundly influenced natural language processing (NLP) and computer vision, paving the way for large language models (LLMs) and multimodal LLMs (MLLMs). These models exhibit transformative potential in medical contexts, owing to their advanced semantic understanding, contextual reasoning, and generative capabilities. As shown in Figure 1, the development of medical MLLMs lags behind general-purpose MLLMs by about two years, while also highlighting the urgent need to deploy medical MLLMs in clinical settings. The integration of LLMs and MLLMs with diverse medical data modalities promises to enable efficient, reliable, and interpretable multimodal intelligent systems for healthcare. For instance, Google's Med-PaLM 2[5] achieved an expert-level score of 86.5% on the United States Medical Licensing Examination,[6] highlighting the considerable potential of LLMs in medicine.[7] Additionally, medical-specific systems such as ChatDoctor,[8] ChatCAD,[9] and LLaVA-Med[10] signify growing adoption of AI-driven medical intelligence. However, critical challenges remain, including achieving cross-modal semantic alignment and knowledge integration, handling domain-specific and specialized medical data, ensuring model robustness and interpretability to avoid "black box" decision-making,[11–13] and addressing issues such as data privacy,[14,15] limited generalization in low-data scenarios,[16] and mitigating multimodal hallucinations.[17–19]

This review focuses on three main areas: the design of multimodal encoders for cross-modal representation alignment, the injection and

* Corresponding authors.
E-mail addresses: linbo@cqu.edu.cn (B. Lin), ywb@cqu.edu.cn (W. Yang), yytang@um.edu.mo (Y.Y. Tang).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.intonc.2025.09.005
Received 21 August 2025; Received in revised form 25 September 2025; Accepted 30 September 2025
Available online 9 October 2025
2950-2616/© 2025 The Author(s). Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

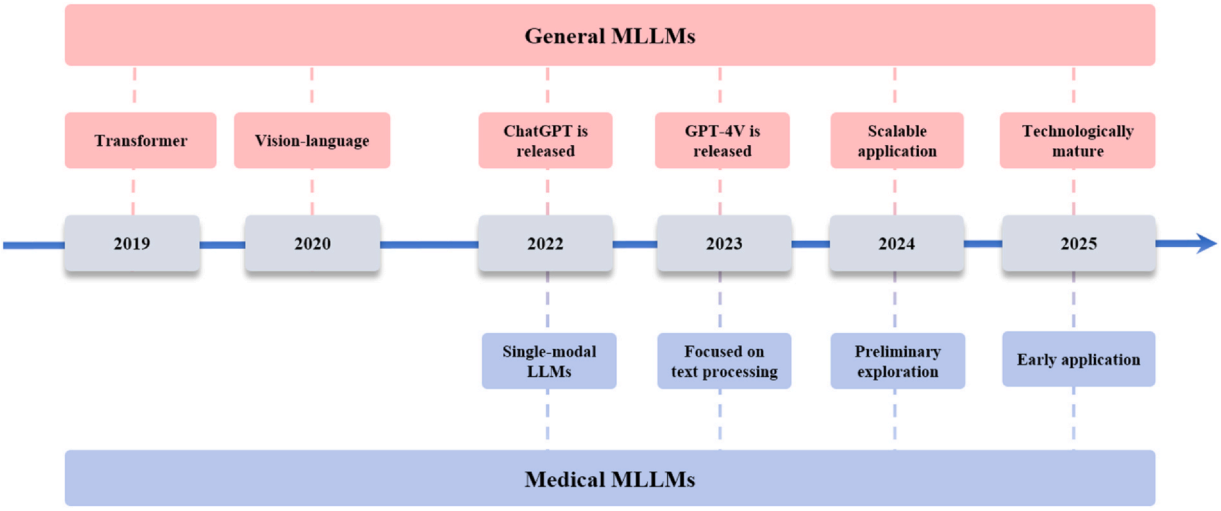


Figure 1. Timeline of general vs. medical MLLMs. Abbreviations: 1. LLMs: large language models 2. MLLMs: multimodel LLMs.

adaptation of medical knowledge through ontologies or knowledge graphs, and human-computer collaboration paradigms augmented with retrieval-augmented generation[20] and prompt engineering[21] to improve practical utility in tasks such as diagnostic assistance and report generation. The remainder of this survey is structured as follows. Section 2 introduces related work, including deep learning, LLMs, and MLLMs. Section 3 provides a brief summary of medical multimodal datasets. Methods and applications are presented in Sections 4 and 5, respectively. Potential challenges are outlined in Section 6, and future directions are provided in Section 7. Finally, the conclusions are presented in Section 8.

2. Related work

2.1. Deep learning

Deep learning represents a novel research direction within the domain of machine learning, which interprets data such as text, images, and sounds by mimicking the manner in which information is transmitted and processed between neurons in the brain. The establishment of a functional relationship from input to output is achieved by the establishment of neuron computing nodes and multi-level computing structures, the selection of the appropriate input layer and output layer, and the learning and tuning of the neural network to establish a functional relationship from input to output. The purpose of this process is to achieve a high correlation between the output result and reality. It should be noted that deep learning technologies are data-driven and require supervision to learn from data. The degree of training of deep learning models varies from fully supervised learning to fully self-supervised learning (SSL). SSL pertains to the process of data labeling that is autonomously performed by machines, thus obviating the necessity for human annotation. In contrast, weak supervised learning occupies a position intermediate between supervised learning and SSL, requiring a limited number of manual data labels. The overarching objective of deep learning is to construct a fully automated, unsupervised system that facilitates precise predictions.[22] Table 1 presents the primary deep learning models and their medical examples.

2.2. LLMs

In recent years, there has been a significant shift in the field of AI from a focus on discriminative processes to a greater emphasis on generative capabilities. Discriminative AI primarily involves the execution of tasks such as recognition, classification, analysis, and prediction through the process of learning the relationship between inputs and outputs. In contrast, generative AI has the capacity to generate new content, including text, images, and code, by leveraging the joint probability distribution of data to emulate the creative process observed in human beings.[32] This shift has resulted in a considerable expansion of the scope of AI applications and innovation potential. In this context, LLMs, as representatives of generative AI, have demonstrated strong language understanding and generative capabilities. These models are based on the Transformer architecture, which learns complex language patterns through SSL based on extensive text data.[33] This mechanism enables the model to dynamically compute the correlation weight between any two elements in a sequence, thereby efficiently capturing long-range dependencies and constructing a global contextual understanding. Its typical structure comprises an encoder for encoding input information and a decoder for generating the target sequence. Building upon this architecture, the autoregressive training paradigm endows the model with powerful generative capabilities. During training, the model learns to predict the next word in a sequence

Table 1
Classification of deep learning models and their applications in medicine.

Model	Medical example
CNN	U-Net and its enhanced variants for lung nodule segmentation.[23] A two-stage framework integrating detection and segmentation models.[24] A dual-path network based on U-Net for the segmentation of infected regions.[25]
ViT	The Transformer branch providing global context information to enhance the CNN backbone network's ability to localize lesions in CT images.[25] OcuViT system for automated classification and severity grading of diabetic retinopathy and age-related macular degeneration.[26] Hybrid model combining ViT, CNN, and angular radial partitioning for skin lesion segmentation and classification achieving 98.2% accuracy on HAM10000 dataset.[27] Low-rank adaptive multi-task learning network using ViT as image encoder for precise breast tumor segmentation and medical report generation.[28]
GAN	GANs employed to generate synthetic medical images for data augmentation and cross-modal transfer learning.[29] StyleGAN3-adaptive-augmentation yielding 512 × 512 malignant melanoma body-lesion images, offering high-fidelity synthetic data to reduce big-dataset dependence and downstream overdiagnosis.[30] StyleGAN2 for generating synthetic melanoma images with systematic validation combining quantitative metrics and qualitative expert evaluation.[31]

Abbreviations: 1. CNN: convolutional neural network 2. ViT: Vision Transformer 3. GAN: generative adversarial network 4. CT: computed tomography.

given the preceding context, optimizing parameters by maximizing the likelihood estimate of the entire sequence. This training approach allows the model to internalize the statistical patterns of language, enabling it to iteratively generate coherent and semantically coherent text, starting from an initial prompt during inference.

Presently, LLMs are extensively employed in image recognition, knowledge question answering, intelligent writing, and other domains, propelling economic and social transformations through multimodal and efficient automatic generation. In the field of medical and health LLMs, substantial attention is directed toward two primary aspects: data architecture and AI agents. With respect to data architecture, large healthcare models increase the accuracy of the underlying model in medical text and multimodal tasks through pretraining and tuning on extensive medical datasets,[34] or by connecting to external knowledge bases.[35–37] Models such as Med-PaLM[38] and Med-PaLM 2[5] can further strengthen their performance and reliability in medical text processing by adopting more advanced training architectures and promote the transparency of model design and operation processes, which is particularly important in the medical field.[5,39] The role of AI agents in the medical field is important, with particular regard to diagnosis and decision support, drug discovery, and medical management. The potential of AI agents extends beyond the realm of diagnosis and treatment, encompassing the simulation of the social impact of medicine. This aspect is of particular relevance to public health and policy-making. For instance, an AI agent developed by Stanford University in the United States simulates group behavior in a community,[40] and Tsinghua University in China has built a more comprehensive simulated hospital environment, Agent Hospital.[41]

Notably, reasoning models significantly enhance the capabilities of LLMs in tackling complex problems. These models adopt sophisticated reasoning paradigms to strengthen logical reasoning, causal inference, and multistep problem decomposition abilities. Key approaches include the following: (1) chain of thought reasoning enables models to break down complex problems into intermediate steps, mimicking human reasoning processes.[42] This technique has been highly effective in medical domains where diagnostic reasoning requires sequential logical steps. Medical adaptations such as Med-CoT[43] demonstrate improved diagnostic accuracy by explicitly modeling clinical decision pathways. (2) tree-of-thoughts extends chain of thought by exploring multiple reasoning paths simultaneously, allowing models to backtrack and evaluate alternative diagnostic hypotheses.[44] This approach shows promise in differential diagnosis scenarios where multiple potential conditions must be considered. (3) graph-of-thoughts further advances reasoning capabilities by representing knowledge as interconnected nodes, enabling more sophisticated relational reasoning about medical concepts and their interactions.[45] Self-consistency methods consolidate reasoning reliability by generating multiple reasoning paths and selecting the most consistent answer, reducing stochastic errors in medical decision-making.

2.3. MLLMs

In recent years, there have been considerable advancements in the field of MLLMs, exemplified by the GPT-4 model.[46] In comparison to their predecessors, these models have demonstrated remarkable proficiency in the management of intricate data structures and the integration of multiple data sources. They are capable of uncovering profound associations and patterns obscured within disparate modal data. This capacity encompasses the amalgamation of diverse information sources, including text, images, video, and audio. The ability to integrate these diverse sources of information enables a more comprehensive description of an event and the extraction of more valuable knowledge. The crux of multimodal tasks lies in the effective integration of these varied sources and the subsequent extraction of valuable knowledge.

Research on multimodal large models originates from Transformer models in the field of NLP. Based on this structure, with efficient computing power and scalability, researchers have gradually expanded the scale of model parameters. After that, researchers began to build a multimodal large model of image-text modal fusion based on massive text and image data to achieve cross-modal understanding and generation, such as Stable diffusion and GPT-4.[46,47]

The construction of multimodal large models can be divided into three distinct paradigms, as outlined below. The first involves the understanding model paradigm, which is based on understanding models. The second is the generative model paradigm, which is based on generative models. The third is the codec construction method, which is based on codecs (Figure 2). The multimodal understanding model based on Transformer encoder borrows the bidirectional encoder representations from the Transformer method and adopts the encoder part of the Transformer as the model architecture to learn and understand the semantics of multimodal data and its association. The research focuses on pretraining data collection, basic model construction, SSL and model optimization training, downstream task fine-tuning and transfer learning, and large model parallel computing and inference acceleration.[48]

Li et al.[49] proposed VisualBERT, which aligns the text with the content of the image through the attention layer and improves the performance of visual question answering and visual common sense reasoning tasks on the basis of the learning objective function of the language model based on visual information. In their study, Ge et al.[50] investigated the potential of employing NLP technology in the domain of neurological data processing. Their findings indicate that this technology can substantially enhance the efficiency and efficacy of clinical

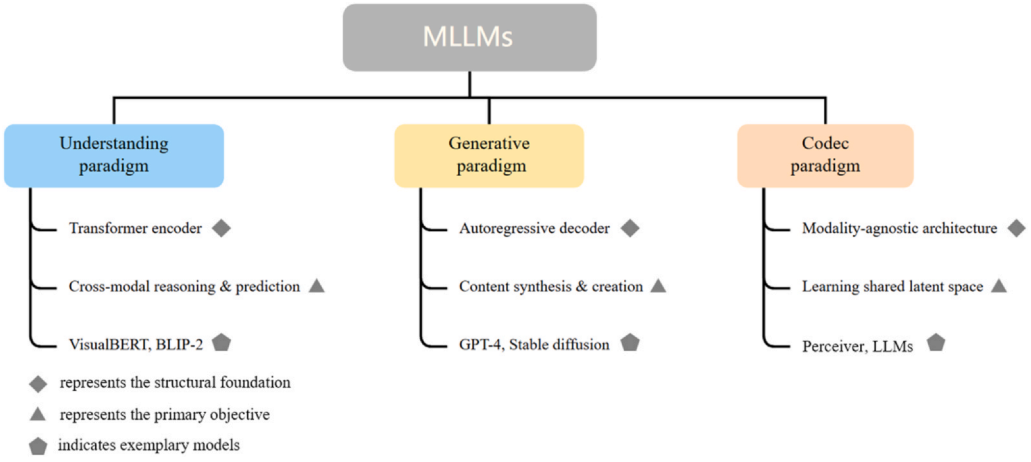


Figure 2. Classification of MLLMs. Abbreviations: 1. LLMs: large language models 2. MLLMs: multimodel LLMs.

care, provided that data security is guaranteed. Following the refinement of GPT-4 on a private dataset, Waisberg et al.[51] determined that the model could compose discharge summaries for postoperative patients and extract and analyze retinal vascular images to detect lesions, demonstrating specific capabilities in medical image data analysis. Giuffre et al.[52] found significant differences in the accuracy of GPT when processing text data in different formats, including both structured and unstructured data. Specifically, the more complex and structured the data format in the prompt message, the better the model is able to understand and process the information, which significantly improves the accuracy of its diagnosis and prediction. Brin et al.[53] found substantial discrepancies in the recognition accuracy of GPT-4V[54] for different imaging modalities. The diagnostic accuracy of computed tomography and X-ray images ranged from 87.6% to 97.0%, while the diagnostic accuracy of ultrasound image recognition, particularly in cases of blurring, was only 9.1%.

3. Datasets

Multimodal data fusion in medicine involves leveraging computational methods to synthesize diverse data types, integrating information from various modalities for improved target prediction.[55] Medical data encompasses a broad spectrum, including (1) text: detailed patient information such as medical history, diagnostic records, and treatment plans found in electronic health record (EHR) systems, as well as research findings from medical literature and clinical case studies; (2) images: examples including X-ray imaging and magnetic resonance imaging (MRI) scans; (3) time series: categories including electrocardiogram (ECG) and electroencephalogram; (4) audio: the changes in the frequency and rhythm of heart sounds, as well as recordings of consultations capturing patient complaints, symptoms, and other vital information.[56] These rich and diverse data forms contain a vast amount of medical information, providing a robust foundation for developing advanced medical multimodal large models. However, they also present significant challenges in processing and analysis. The ability to accurately and efficiently extract and fuse the hidden information within this multimodal medical data is critical for tackling current data analysis tasks and profoundly impacts disease diagnosis and treatment.[57] Therefore, understanding how to effectively acquire multimodal medical data is crucial.

The medical multimodal datasets serve as crucial resources for training large medical models, offering a vast amount of data and enabling the model to learn the correlations and features between different modal data, which can be categorized in several ways. These datasets can be distinguished by their intended use, such as training-specific datasets such as GMAI-VL-5.5M,[58] which enhances a model's grasp of medical knowledge, or evaluation-specific datasets such as GMAI-MMBench,[59] designed to comprehensively assess the reliability and stability of different models. Furthermore, classification can occur by the type of modality contained, such as image and text, or by the medical domains and tasks covered, including medical imaging and clinical department datasets.[60] Tables 2 and 3 list examples of valuable multimodal medical datasets for constructing MLLMs. Table 2 presents datasets with fewer than 10K samples, while Table 3 features image-text datasets containing 10K or more samples.

Current medical multimodal datasets exhibit several significant limitations that hinder the development of comprehensive MLLMs. The most prominent gaps include the following: (1) 4D dynamic imaging deficiency. Lack of temporal resolution data such as real-time surgical videos with instrument tracking, dynamic functional MRI sequences, and longitudinal treatment monitoring datasets; (2) Rare disease scarcity. Severe underrepresentation of orphan diseases and rare medical conditions across all modalities, creating bias in model training and evaluation; (3) Cross-modal alignment quality. Inconsistent and often poor synchronization between different modalities, particularly for time series data such as electroencephalogram-video pairing and intraoperative multimodal recordings; (4) Ethical access constraints. Restrictive data sharing policies that limit the availability of high-quality clinical multimodal data, especially for sensitive patient information and protected health data; (5) Multimodal imbalance. Overrepresentation of radiology and pathology image-text pairs compared to other critical modalities such as genomic-imaging fusion, proteomic-visual data, and multi-omics integration. These gaps substantially constrain the development of robust medical MLLMs that can handle the complexity and diversity of real-world clinical scenarios.

4. Methods

4.1. Architecture

The architecture of MLLMs is structured around three interconnected components: modality encoders, LLM backbones, and modality projectors, which collectively enable the integration and interpretation of heterogeneous medical data. The overview of the architecture is shown in Figure 3.

4.1.1. Modality encoder

In MLLMs, the modality encoder serves as a critical component for processing and integrating heterogeneous data types, acting as a foundational bridge between raw data and the symbolic-reasoning capabilities of LLMs. Its design directly affects the model's ability to achieve human-level multimodal understanding, making it a pivotal area of study. In medical contexts, modality encoders transform raw, domain-specific data (e.g., medical images, biological signals, or genomic sequences) into structured embeddings. The encoder is typically pretrained on large-scale datasets and then fine-tuned for specific tasks. Outputs from modality-specific encoders are mapped to a unified dimensionality and structure compatible with the LLM's Transformer architecture.

Table 2
Small-sample medical multimodal dataset (Sample size less than 10K).

Dataset	Modality	Description	License/URL
Montgomery CXR[61]	Image + Text	2D X-Ray datasets and diagnostic reports	https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/Montgomery-County-CXRSet/MontgomerySet/index.html
VQA-RAD[62]	Image	Radiation-related Med-VQA dataset	https://osf.io/89kps/
Linguistic Meaning[63]	Image + Text	fMRI and semantic vector representations	https://osf.io/crwz7/
VQA-Med[64]	Image	Medical VQA dataset	https://zenodo.org/records/10499039
ZuCo 2.0[65]	EEG + Eye tracking	High-density EEG and eye-tracking data	https://osf.io/2urht/
MMMU Health[66]	Image + Text	Medical image QA data	https://huggingface.co/datasets/MMMU/MMMU
MedVidQA[67]	Video	Medical instructional video QA dataset	https://github.com/deepaknlp/MedVidQACL/tree/master/MedVidQA
FACE[68]	EEG + Video	EEG dataset for emotional computing	https://www.synapse.org/Synapse:syn50614194/files/
WSI-VQA[69]	Image + Text	Pathology whole slice image VQA dataset	https://github.com/cpystan/WSI-VQA/tree/master
ChineseEEG[70]	EEG + Eye tracking	EEG and eye-tracking for Chinese reading	https://github.com/ncclabsustech/Chinese_reading_task_eeg_processing
EAV[71]	EEG + Audio + Video	Multimodal emotion recognition dataset	https://zenodo.org/records/10205702

Abbreviations: 1. EEG: electroencephalogram 2. VQA: visual question answering 3. fMRI: functional magnetic resonance imaging 4. QA: question answering.

Table 3
Multimodal medical image and text dataset (Sample size exceeds 10K).

Year	Dataset	Description	License/URL
2018	ROCOV2[72]	Integration of radiologic images with medical concepts	https://zenodo.org/records/8333645
2019	MIMIC-CXR[73]	Chest radiology report dataset	https://physionet.org/content/mimic-cxr/2.0.0/
2020	MedICaT[74]	Medical images with captions and annotations	https://docs.google.com/forms/d/e/1FAIpQLSdB6w2HHNtD-v6SJr3wFMQl8WxR-wigrfVJPvqI-RR
2020	PathVQA[75]	Pathology visual-quiz task dataset	https://drive.google.com/drive/folders/1G2C2_FUCyYQKCkSeCRRiTTsLDvOAjFj5
2021	SLAKE[76]	English-Chinese bilingual Med-VQA dataset	https://drive.google.com/file/d/1EZ0WpO5Z6BJUqC3iPBQJJS1INWSMsh7U/view?pli=1
2023	RP3D-Caption[77]	3D medical imaging graphic dataset	https://github.com/chaoyi-wu/RadFM
2023	PMC-OA[78]	Biomedical field multimodal dataset	https://huggingface.co/datasets/axiong/pmc-oa
2023	PMC-VQA[79]	Large-scale medical VQA dataset	https://huggingface.co/datasets/xmcmic/PMC-VQA
2023	Quilt-1M[80]	Visual-verbal histopathology dataset	https://zenodo.org/records/8239942
2023	MIMIC-Diff-VQA[81]	Chest X-ray differential VQA dataset	https://physionet.org/content/medical-diff-vqa/1.0.0/#files
2023	CAD-Chest[82]	Detailed labeled chest X-ray dataset	https://physionet.org/content/cad-chest/1.0/
2023	PMC-Inline[77]	Literature with embedded images from PubMed Central	https://huggingface.co/datasets/chaoyi-wu/PMC-Inline/tree/main
2024	CT-RATE[83]	Chest CT images and diagnostic reports	https://huggingface.co/datasets/ibrahimhamamci/CT-RATE
2024	OmniMedVQA[84]	Comprehensive medical VQA dataset	https://openxlab.org.cn/datasets/GMAI/OmniMedVQA
2024	PathMMU[85]	Pathology image understanding dataset	https://huggingface.co/datasets/jamessyx/PathMMU
2024	EIT-1M[86]	Multimodal EEG, images and text dataset	https://huggingface.co/datasets/eit−1m/EIT−1M
2025	BIOMEDICA[87]	Annotated data from PubMed Central Open Access database	https://huggingface.co/BIOMEDICA

Abbreviations: 1. VQA: visual question answering 2. CT: computed tomography 3. EEG: electroencephalogram.

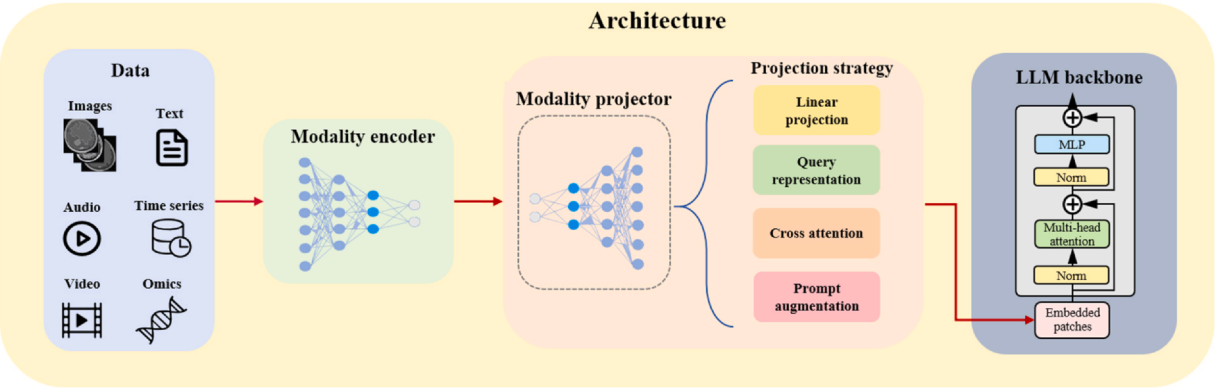


Figure 3. The architecture of medical MLLMs. Abbreviations: 1. MLP: Multilayer Perceptron 2. LLM: large language model.

Image encoders, also known as vision encoders, leverage pretrained general models or domain-specific variants to extract clinically relevant features from medical images. Specifically, the image encoder E encodes the visual input I_x into visual features V_x, illustrated as $V_x = E(I_x)$. Widely adopted pretrained image encoders for general-purpose tasks encompass architectures such as convolutional neural network,[88,89] Vision Transformer,[90] CLIP-based ViT,[91] NFNet,[92] EVA-CLIP,[93] LLaVa-Med,[10] MAIRA,[94] and PathChat.[95] To optimize the integration of medical knowledge into latent representations, domain-specific image encoders are integrated into medical MLLMs. These encoders are often fine-tuned on medical datasets to enhance their ability to capture diagnostic patterns (e.g., tumor boundaries or anatomical anomalies). For example, PathAsst,[96] LLaVA-Med,[10] and RaDialog[97] leverage vision modules from image-text pretrained frameworks, including CLIP,[91] BioMedCLIP,[98] and BioViL-T,[99] respectively. Furthermore, RAD-DINO[100] implements chest X-ray images pretraining by adopting the self-supervised strategy of DINO,[101] demonstrating enhanced adaptability to medical imaging analysis.

For nonvisual modalities, similarly, various specialized encoders are employed to obtain latent representations. Text encoders, also known as language encoders, aim to comprehend complex medical terminology and patient histories, which will be introduced in the next section since LLM backbones play this role in MLLMs.

Audio encoders[102,103] are specifically designed to capture and analyze nuanced acoustic patterns imperceptible to humans. These models leverage hierarchical feature extraction to discern clinically significant auditory signatures, such as identifying pathological deviations in cardiopulmonary auscultation (e.g., abnormal heart murmurs or crackling lung sounds) that differ from baseline physiological recordings. Its optimization framework emphasizes discriminative capabilities in the time-frequency domain, enabling automated medical diagnosis through the robust classification of health-critical audio biomarkers.

Temporal encoders (such as Transformer-based models)[104–107] are optimized to model time-dependent physiological trajectories, enabling real-time detection of deteriorating vital signs or pathological progression. They hierarchically extract temporal dependencies through SA mechanisms or recurrent gating, identifying precursors to acute clinical events such as septic shock or cardiac arrest. By correlating longitudinal patterns with emergent risks, they provide probabilistic early warnings for clinical teams while simultaneously generating interpretable biomarkers for automated triage prioritization in critical care settings. Their capacity for sequential anomaly detection bridges predictive analytics with actionable intervention protocols in time-sensitive medical workflows.[108,109]

Omics encoders (e.g., Basset,[110] DeepCpG[111] and DeepVariant[112]) are engineered to decode high-dimensional biological signals, including

Table 4
Comparison of different modality projectors in medical MLLMs.

Characteristics	Linear projection	Query representation	Cross-attention	Prompt augmentation
Mechanism	Linear transformation	Learnable queries interact with visual features	Visual features as keys/values, text as Queries	Offloads vision to expert models, uses text prompts
Parameter count	Moderate	Low to Moderate	High	Very Low (relies on external models)
Inference speed	Fast	Very Fast	Slow	Fast (but depends on expert model)
Representation fidelity	High (preserves all tokens)	Low (compressive)	Very high (fine-grained interaction)	Medium (limited by expert model)
Example models	LLaVA-Med,[10] XrayGPT[137]	MedBLIP,[138] BLIP − 2[139]	Flamingo,[140] CoCa[141]	OphGLM,[142] ChatCAD[9]

detecting pathogenic mutations, methylation patterns, or non-coding regulatory elements linked to disease etiology. These frameworks employ advanced feature abstraction to map genotype-phenotype correlations, enabling systematic identification of molecular signatures for precision medicine applications. By prioritizing causal variants and epigenetic drivers, they facilitate biomarker discovery for stratified patient cohorts and accelerate the development of tailored therapeutic strategies, including gene-editing targets or pharmacogenomic optimization in oncology and rare genetic disorders.

Furthermore, emerging parameter-efficient paradigms in multimodal learning explore alternative architectures that eliminate dedicated encoder modules. A notable example is Fuyu-8B,[113] which implements visual tokenization through direct linear projection of image patches into LLM-compatible embeddings, bypassing conventional convolutional or Transformer-based feature extractors. This streamlined approach inherently enables dynamic resolution adaptation mechanisms, as the model processes raw pixel grids without rigid resizing or padding constraints. Such architectures preserve spatial fidelity across arbitrary input dimensions while maintaining computational tractability, proving advantageous for medical imaging workflows requiring native multiscale analysis of high-resolution scans or irregularly cropped regions of interest.

4.1.2. LLM backbone

By leveraging advanced LLMs from NLP, MLLMs extend these capabilities to analyze textual inputs and produce contextually appropriate outputs. The architectural foundation for text processing in both general-domain systems (e.g., OPT,[114] Vicuna,[115] LLaMA,[116] LLaMA2,[117] Chinese-LLaMA,[118] Flan-T5,[119]) and medical-specialized implementations predominantly relies on Transformer-based encoders. Notably, the medical field has seen dedicated adaptations, including BioGPT,[120] BioMedLM,[121] and PMC-LLaMA,[122] which demonstrate specialized proficiency in processing biomedical language while maintaining broad adoption across research and clinical applications.

In addition, the integration of a mixture of expert (MoE) architectures into LLMs has recently gained significant traction in computational linguistics research.[123–130] This sparse modeling paradigm offers distinct advantages over conventional dense architectures, particularly in achieving parameter scalability while preserving computational efficiency through conditional activation mechanisms. In the field of medicine, the application of MoE in large models has also been verified. Experimental validations across multiple studies (e.g., MoE-LLaVA,[131] Med-MoE,[132] and MicarVLMoE[133]) demonstrate enhanced performance metrics compared to their dense counterparts on various benchmarks.

4.1.3. Modality projector

A modality projector, also known as a cross-modal adapter or connector, plays a pivotal role in establishing semantic correspondence among heterogeneous data streams (e.g., linguistic, visual, and acoustic) in multimodal learning frameworks. These cross-modal adapters typically employ parameterized transformations—ranging from learned linear mappings to deep neural architectures—designed to encode high-dimensional sensory inputs into a shared latent space with consistent topological properties. Utilizing techniques such as attention,[4] cross-modal attention,[134] tensor fusion,[135] and concatenation,[136] connectors can achieve semantic coherence between heterogeneous embedding spaces generated by modality-specific encoders, synthesizing a unified knowledge representation that preserves complementary information across multimodal data. Unlike the self-supervised paradigm governing modality-specific encoding, this synchronization process necessitates cross-modally annotated training pairs for effective parameter learning. Domain-specific knowledge infusion enables the derivation of diagnostically coherent representations that maintain pathophysiological consistency, ultimately yielding multimodal embeddings with enhanced clinical interpretability and decision-making utility.

Generally, modality projectors can be categorized into four types: linear projection, query representation, cross-attention, and prompt augmentation. A comparison of four different modality projectors in MLLMs is presented in Table 4.

4.1.3.1. Linear projection. This projection-based approach serves as a multimodal conversion mechanism by aligning visual representations with textual embedding spaces through simple linear transformation layers.[143] Distinguished from query-based methods, this framework employs computationally efficient architectures, typically single or dual-layer multilayer perceptron, to bridge vision-language modalities. Pioneered by LLaVA[12] and LLaVA-1.5[144] in general MLLMs, this paradigm has been widely adopted in medical AI systems, including PathAsst,[96] XrayGPT,[137] and MAIRA-1,[145] Qilin-Med-VL,[146] LLaVA-Med,[10] and PMC-VQA.[79] They utilize multilayer perceptron to preserve full-dimensional visual tokens, thereby minimizing information loss during cross-modal alignment. While this design advantage maintains richer visual semantics compared to token-reducing query-based approaches, it concomitantly escalates training complexity due to higher-dimensional feature processing. These empirical validations across medical domains confirm the effectiveness of this streamlined projection mechanism in balancing architectural simplicity with multimodal representational capacity.

4.1.3.2. Query representation. Query-based methods have emerged as an efficient mechanism for extracting and aligning visual representations with language models through a set of learnable query vectors.[143] Notably, the Q-Former introduced in BLIP-2[139] exemplifies this approach by employing a frozen vision encoder to derive a fixed-length set of visual features that can be seamlessly integrated with LLMs via cross-attention. This method has been extended beyond 2D image understanding to 3D medical imaging, as demonstrated by MedBLIP,[138] highlighting the versatility of the querying mechanism. The core advantage of these approaches lies in their ability to reduce computational overhead by representing high-dimensional visual inputs with a limited number of queries. For instance, architectures such as the perceiver resample in Flamingo[140] and the Q-Former in BLIP-2[139] both maintain a constant number of visual outputs, typically 32, regardless of the input size, significantly improving inference efficiency. However, this compression often comes at the cost of losing fine-grained visual details, with empirical evidence

from DeCo[147] indicating that the Q-Former may function primarily as an inefficient token compressor. Alternative techniques, such as adaptive average pooling, have been shown to yield more effective token reduction while preserving essential visual information.

4.1.3.3. Cross-attention. Building upon the SA principles of Transformers, cross-modal attention mechanisms dynamically assign distinct functional roles (query, key, and value) to visual and linguistic representations during multimodal computation. This paradigm enables learnable intermodal interactions by strategically positioning image features as key-value pairs and textual features as queries. A representative implementation, the GATED XATTN-DENSE layers in Flamingo,[140] injects visual information into frozen pretrained LLMs through dense cross-attention layers. To mitigate computational overhead from high-dimensional visual data, these systems typically employ feature compression modules such as the perceiver resample[148] before cross-modal alignment. Subsequent medical adaptations, such as Med-Flamingo,[149] demonstrate the framework's efficacy in clinical multimodal reasoning by refining these attention-based fusion strategies. Parallel implementations, including CoCa[141] and CONCH,[150] leverage cross-attention in text decoders to achieve progressive visual-semantic integration via prompt augmentation techniques. While such dynamic interaction mechanisms enable finer-grained multimodal reasoning compared to static projection methods, they incur substantial computational costs, particularly when processing long visual token sequences that pose scalability challenges for resource-constrained applications.

4.1.3.4. Prompt augmentation. Prompt augmentation establishes visual-textual integration through a hybrid pipeline combining expert models with structured linguistic templates. This approach processes raw images using domain-specific analytical models (e.g., classification or segmentation networks) and converts their outputs into formatted text via predefined prompts, bypassing direct modality alignment. For instance, OphGLM[142] synthesizes diagnostic reports by embedding retinal lesion classifications and segmentation masks into standardized textual frameworks, which are then processed by LLMs. Similarly, ChatCAD[9] and ChatCAD + [151] employ computer-aided diagnosis models to analyze radiographic features, translating detection results into natural language narratives through template-based prompting for subsequent LLM interpretation. Distinct from projection-based or attention-driven methods, this paradigm eliminates cross-modal training requirements by outsourcing visual understanding to specialized models. While offering computational efficiency and domain adaptability, its performance is intrinsically constrained by two factors: the accuracy of upstream expert models in feature extraction and the semantic completeness of prompt templates in bridging visual analytics with linguistic contexts. Such dependencies pose a risk of cascading errors—inaccuracies in initial image processing or suboptimal template designs, which may irreversibly propagate to subsequent LLM inference stages.

4.2. Training methods

As two primary parts of the training step, pretraining and fine-tuning constitute essential procedures in the development and refinement of MLLMs (Figure 4). Rather than being isolated steps, they are integral and continuous processes that influence each component of an MLLM architecture[152].

4.2.1. Pretraining

The pretraining process begins with modality encoders, whose primary objective is to learn generalizable features and representations for each individual modality. Typically, these modality encoders undergo pretraining on large-scale datasets to achieve sufficient understanding of their respective modalities. Within the modality projector, pretraining enables the model to acquire preliminary methods for aligning and integrating embeddings derived from disparate modalities, thereby establishing the foundational groundwork for learning intricate cross-modal relationships. Although pretraining does not directly yield final decision outcomes, it constructs essential capabilities and knowledge infrastructure that facilitate more effective execution of downstream tasks in subsequent stages. This preparatory phase equips the model with crucial cross-modal comprehension prerequisites while maintaining task-agnostic characteristics prior to task-specific fine-tuning.

In this paper, we define the transition from pretraining to fine-tuning as model customization. This process begins with obtaining a foundational model and pretraining it on massive datasets, endowing it with universal cross-modal representation and alignment capabilities. Subsequently, the model undergoes calibration and adaptation using small-scale, high-quality, task-specific data, enabling it to perform effectively in specific domains, tasks, or under specific instructions.

4.2.2. Fine-tuning

The adaptation phase refines pretrained models for specialized downstream applications by calibrating their parametric configurations and architectural elements using limited, domain-focused training data. During this process, the model undergoes targeted optimization to promote its competence in comprehending and executing task-specific directives. In MLLMs, such a refinement procedure entails recalibrating modality-specific encoding mechanisms, strengthening cross-modal integration pathways, and optimizing context-aware decision-making protocols derived from enhanced multimodal representations. This stage transitions the model from generalized pretrained capabilities to domain-adapted proficiency while preserving the foundational knowledge acquired during the initial training phases. The resultant architecture exhibits an improved capacity to synthesize heterogeneous inputs and generate contextually grounded outputs through systematic parameter-space exploration and structural adaptation. In terms of the implementation method, fine-tuning methods can be grouped into four categories: addition-based, selection-based, reparameterization-based, and hybrid methods.[153–155]

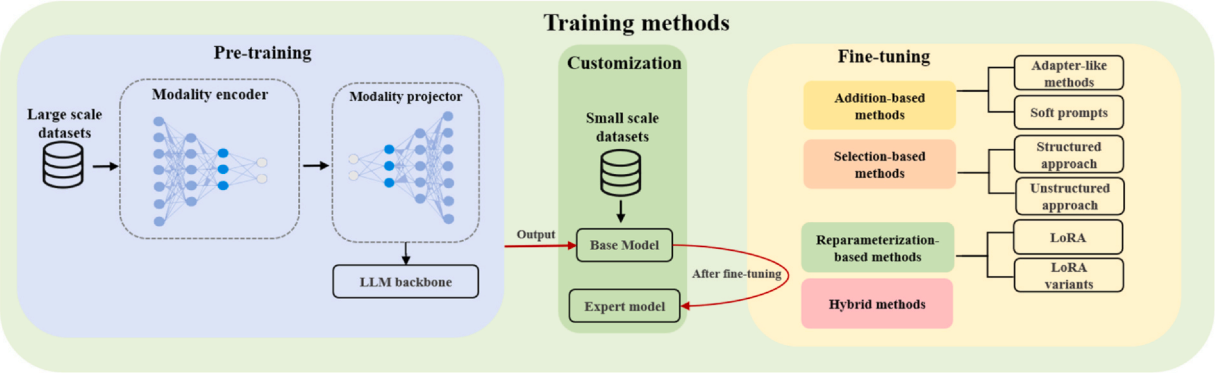


Figure 4. Overview of the training methods. Abbreviations: 1. LLM: large language model 2. LoRA: low-rank adaptation.

4.2.2.1. Addition-based methods. Addition-based methods enhance pretrained models by introducing supplementary trainable parameters or architectural layers while keeping the original weights frozen, which represents the most extensively investigated fine-tuning strategies. Notably, they can be categorized into two types: adapter-like methods and soft prompts. On the one hand, adapters[156] lie in the integration of fully connected layers between Transformer submodules, which maintains the original competencies of the LLMs while facilitating task-specific adaptation for specialized domains. Adapters have demonstrated that it is possible to achieve full fine-tuning performance by tuning less than 4% of the total model parameters. Adapter-like methods improve adapters from different aspects, such as relocating,[157,158] pruning,[159] reparameterization[160] and selective activation.[161] However, soft prompt was initially proposed to solve the problem of finite input sequence length in language model prompting,[162] which optimizes a subset of input embeddings through gradient-based methods. This reformulates prompt design from discrete token selection to continuous parameter tuning. While some methods restrict soft prompt training to the input layer,[163,164] others extend it across all Transformer layers.[165] The adaptive prefix tuning method[166] utilizes both fine-grained gated weight assignment and coarse-grained scaled weight specification for the prefix at each layer, which can be illustrated as follows:

$$\hat{P}_i = \lambda_i \odot \alpha_i \cdot [P_{ik}, P_{iv}], \tag{1}$$

where $\hat{P}_i$ denotes the pseudo prefix token in the ith layer; $[P_{ik}, P_{iv}]$, the key-value pair of the original pseudo prefix tokens; λ_i, a learnable scaled weight; $\odot$, element-wise multiplication; and α_i, the gated weights, which can be computed as:

$$\alpha_i = \text{sigmoid}(h_{i-1} W_i), \tag{2}$$

where h_{i-1} is the hidden states from the previous layer, and W_i are the parameters to be learned. Recent advancements show that soft prompts have obviously improved fine-tuning in medical MLLMs.[167–169]

4.2.2.2. Selection-based methods. Selection-based methods make only specific parameters in the original model trainable while keeping the rest frozen. They trace their origins to early neural-network adaptation strategies, notably the practice of fine-tuning only the uppermost layers of pretrained models. Contemporary methodologies have evolved to use more sophisticated selection paradigms, typically organized along two dimensions: structured and unstructured. Structured methods focus on particular parameter subsets such as bias terms[170] or strategically chosen weight-matrix rows,[171] whereas unstructured approaches target specific layer types, such as attention mechanisms or feed-forward components.[172] For example, the second-order approximation method[173] employs a second-order Taylor expansion to approximate the original optimization problem, transforming it into an analytically tractable form whose parameters can be obtained directly by solving the resulting approximation, which can be illustrated as:

$$\min_{\Delta\theta} \left[L(\theta_0) + \nabla L(\theta_0)^T M \Delta\theta + \frac{1}{2} (M\Delta\theta)^T H M \Delta\theta \right], \tag{3}$$

subject to $||M||_0 = \lfloor mp \rfloor; M_{ij} = 0,7D1 \forall i \neq j; M_{ii} \in \{0,1\}$, where θ_0 are the pretrained parameters, $\Delta\theta$ is the difference vector, M is the parameter mask matrix, L is the loss function, $\nabla L(\theta_0)^T$ is the gradient of the loss function at θ_0, and H is an approximated diagonal Hessian matrix. An alternative paradigm emerges in sparse training techniques, which constitutes a distinct form of selection-based methods. These methods typically disregard inherent network architecture, instead employing algorithmic criteria to identify optimal parameter subsets for adaptation.[174] However, the implementation of sparse parameter updates introduces significant computational challenges, particularly regarding hardware compatibility and training efficiency. Recent advancements in parameter reconfiguration[171] and structured $M \times N$ sparsity patterns[175] have partially mitigated these issues. Despite such progress, fully unstructured sparsity remains computationally intractable in modern hardware architectures due to memory access inefficiencies and parallel processing limitations.

4.2.2.3. Reparametrization-based methods. Reparameterization-based methods transform existing parameters into a more parameter-efficient form, e.g., a low-rank parameter matrix, to adapt to specific downstream tasks. Low-rank adaptation (LoRA)[176] is a typical well-known method that introduces low-rank trainable matrices to update the pretrained weight matrix via a low-rank decomposition algorithm. KronA[177] is a Kronecker product-based adapter module. Subsequently, more studies have proposed improvements and optimizations to the LoRA algorithm, primarily by introducing novel constraints or optimization strategies to enhance its performance. Further research has focused on task-specific LoRA optimization, designing more refined low-rank structures or incorporating domain-specific constraints to enhance model performance in specialized downstream tasks (Table 5). These advancements not only broaden the applicability of LoRA but also drive further progress in the field of parameter-efficient fine-tuning.

4.2.2.4. Hybrid methods. Given the substantial task-dependent variability in fine-tuning methods, recent research has increasingly focused on hybrid approaches that integrate complementary strengths from multiple adaptation strategies. This trend stems from the observation that no single fine-tuning technique universally dominates across diverse applications, prompting efforts to develop more robust solutions through strategic combination. UniPELT[184] employs a gating mechanism to dynamically select optimal submodules for each task, while MAM Adapter[157] reformulates adaptation as targeted modifications to hidden states in pretrained models. S_4[185] organizes layers in a spindle pattern with uniform parameter allocation and group-specific strategies, demonstrating consistent performance gains across NLP tasks. The LLM-Adapter framework[186] systematically integrates various adaptation mechanisms (Series/Parallel Adapters, LoRA) at strategically selected architectural locations. For vision models, NOAH[187] automates prompt module design through neural architecture search, enhancing transfer learning capabilities. Most recently, AUTOPEFT[188] has advanced the field through multi-objective Bayesian optimization, automatically discovering Pareto-optimal configurations that maximize both performance and parameter efficiency with minimal computational overhead.

Table 5
A cheat-sheet of representative LoRA variants.

Method	Key idea	Application context
DyLoRA[178]	Dynamic rank selection during training.	Task-adaptive fine-tuning
QdyLoRA[179]	4-bit quantization + dynamic rank adjustment.	Memory-efficient tuning with rank flexibility
DoRA[180]	Splits weights and uses LoRA for direction.	High-performance adaptation
AdaLoRA[181]	Adaptive rank allocation via singular value pruning.	High-performance adaptation
Sparse LoRA[182]	Incorporates sparsity constraints on LoRA parameters.	Efficient and scalable tuning
LoRA+[183]	Uses differential learning rates for the projection layers.	Improved convergence

Abbreviation: LoRA: low-rank adaptation.

4.3. Evaluation methods

Medical MLLMs are fundamentally different from their application in other domains. Healthcare is a high-stakes arena in which the consequences of errors can be severe, directly impacting patient safety and well-being. This inherent criticality elevates the importance of evaluation from a purely technical exercise to a paramount safeguard. Unlike general-purpose AI, medical MLLMs require exceptionally rigorous assessments across a wide spectrum of clinical scenarios and patient populations to ensure both safety and efficacy. The intricate nature of medical data, the inherent variability across clinical contexts, and the profound ethical implications of integrating AI into patient care necessitate a multifaceted evaluation approach that extends considerably beyond conventional accuracy metrics.

Current benchmarking studies consistently reveal a considerable variability in performance levels across different models.[189] For instance, in the MMMU benchmark,[66] GPT-4V[54] demonstrates a leading performance in medical tasks, particularly in its adept handling of complex multimodal content. It is followed by models such as Sense-Chat-Vision-0423-Preview and Qwen-VL-MAX,[190] which also perform commendably but exhibit certain limitations. Similarly, the GMAI-MMBench[59] positions GPT-4o[191] at the forefront of medical tasks, with Gemini 1.5 and GPT-4V closely trailing. Intriguingly, specialized medical models such as MedDr,[192] despite their focused development, generally lag behind these top-performing general-purpose models. This observed variability underscores the intrinsic complexity of medical tasks and highlights the continuous need for advanced development in this specialized domain. Furthermore, dedicated benchmarks, such as M3D,[193] illuminate specific capabilities within the multimodal landscape. M3D-LaMed, for instance, exhibits superior performance in 3D medical imaging tasks, including the generation of detailed reports and visual question answering, indicating its strong potential for direct clinical assistance. While general-purpose models often demonstrate superior benchmark scores, their broad applicability raises questions regarding their inherent safety and reliability within specialized medical contexts, particularly when compared to domain-specific models that, despite lower current performance, are explicitly designed for medical applications. This highlights that raw quantitative metrics, while indicative of capability, do not fully address the nuanced requirements of clinical safety and ethical appropriateness, underscoring the necessity for evaluation to extend beyond mere accuracy to encompass clinical appropriateness and patient safety.

Generally, the evaluation of medical MLLMs could be grouped into three types: metrics evaluation, AI evaluation, and human evaluation.[194] As depicted in Figure 5, in the medical field, model evaluation may go beyond traditional metrics and incorporate clinical reliability and patient safety as core considerations. To enable meaningful comparisons, the figure assumes performance scores for three models (Models A, B, and C) across key indicators.

4.3.1. Metrics evaluation

Metrics evaluation offers a foundational approach to assessing model performance, providing standardized, objective, and quantitative measurements. These metrics are indispensable for systematic comparisons across diverse tasks and offer an efficient, scalable method for initial performance assessments. The landscape of automatic evaluation metrics is diverse, reflecting the varied capabilities of MLLMs in medical applications. For tasks involving the identification or localization of specific features within images, such as recognizing anatomical structures or pathological findings, metrics including accuracy and average precision are commonly employed to gauge a model's recognition capabilities. When evaluating a model's perception capabilities, especially in tasks such as image segmentation or object detection, more specialized metrics are adopted, including mean intersection over union, mean average precision, and the Dice coefficient.[195] These metrics provide objective measures of how well a model

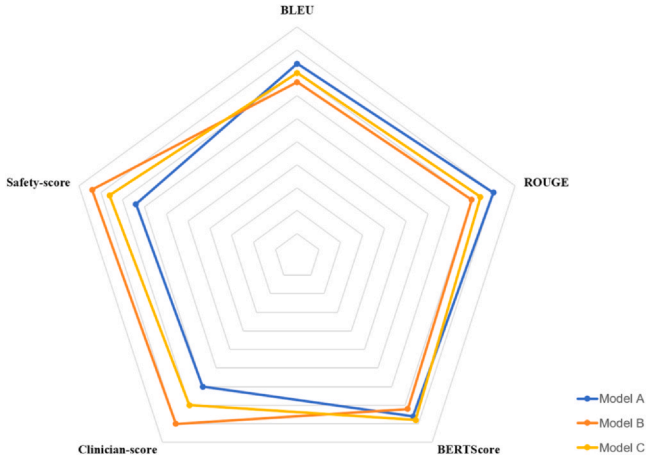


Figure 5. Multi-dimensional model performance evaluation. Models A, B, and C are hypothetical models used for evaluation. Abbreviations: 1. BLEU: bilingual evaluation understudy 2. ROUGE: recall-oriented understudy for summarization evaluation.

can delineate and localize relevant regions within medical imagery. To evaluate text generation capabilities, which are central to many medical MLLM applications, a range of metrics are utilized. Accuracy serves as a primary metric for multiple-choice question benchmarks, such as MedQA[6] and MedMCQA,[196] where the goal is to evaluate the correctness of discrete answers. N-gram overlap metrics are widely used to assess the linguistic quality and similarity of generated text to reference outputs.

The bilingual evaluation understudy (BLEU)[197] metric quantifies the quality of generated text by measuring the similarity of n-grams (consecutive word sequences) between the model's output and one or more reference texts. BLEU is computed as follows:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4}$$

where N is the maximum n-gram order and w_n are the weights assigned to each n-gram precision (usually uniform weights). BP is defined as

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \le r \end{cases} \tag{5}$$

where c is the length of the candidate translation and r is the effective reference corpus length. p_n indicates the modified n-gram precision, which is computed as follows:

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n-gram \in C} \text{Count}_{\text{clip}}(n-gram)}{\sum_{C' \in \{\text{Candidates}\}} \sum_{n-gram' \in C'} \text{Count}(n-gram')} \tag{6}$$

where Candidates is the set of candidate sentences. Count is to obtain all n-grams in the candidate sentence. $Count_{clip}$ counts clipped n-grams in the candidate that do not exceed the counts in any reference. Variants such as BLEU-1 reflect word-level accuracy, while BLEU-4 places greater emphasis on textual continuity. Recall-oriented understudy for generalized evaluation (ROUGE)[198] encompasses several metrics, including ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. Similar to BLEU, ROUGE-N evaluates n-gram similarity but prioritizes recall, whereas BLEU focuses on precision. ROUGE-N is computed as follows:

$$ROUGE\text{-}N = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{(gram_n)}} \tag{7}$$

where n stands for the length of the n-gram $gram_n$. $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. ROUGE-L assesses textual coherence by calculating the longest common subsequence between the generated

and reference texts. ROUGE-W extends ROUGE-L by applying weighted scoring to emphasize continuous and accurate matches, and ROUGE-S accommodates non-contiguous words in n-grams. Diversity metrics are also crucial, for generative models in particular.

Distinct-n[199] quantifies the diversity of generated text by calculating the ratio of unique n-grams to the total number of n-grams, providing an indication of the variety and non-repetitiveness of the generated language. For multimodal tasks involving image captioning, consensus-based image description evaluation[200] is specifically designed. It combines n-gram recall and precision, assigning higher weights to rare n-grams to better assess the capture of key information within image descriptions. More advanced approaches, such as BERTScore,[201] leverage pretrained bidirectional encoder representations from Transformer embeddings to compute token-level similarity scores between generated and reference sentences. Medical models such as HuaTuoGPT,[202] ClinicalGPT,[203] SoulChat,[204] and BianQue[205] commonly employ these metrics to assess their generative performance. Table 6 lists specialized evaluation metrics for the medical domain tailored for medical MLLMs.

4.3.2. AI evaluation

In response to the resource-intensive nature of human evaluation, a burgeoning trend involves the utilization of advanced AI models, particularly LLMs such as GPT-4,[46] as evaluators.[10,211] This methodology capitalizes on the sophisticated instruction-following capabilities of these models to assess the quality of outputs generated by other MLLMs. Advanced AI models, including ChatGPT[212] and GPT-4,[46] are increasingly being employed as evaluators, often demonstrating an alignment with human values in their assessments. These AI evaluators are tasked with assessing the performance of other models across several critical dimensions. These typically include helpfulness, relevance, factual accuracy, and the level of detail provided in the response, with scores often assigned on a numerical scale, such as 1–10. A notable advantage of AI evaluators is their capacity to offer detailed explanations for their assessments, moving beyond a simple numerical score to provide fine-grained insights into a model's specific strengths and areas requiring improvement. Empirical studies have indicated that AI evaluation can, in certain contexts, surpass the performance of traditional automatic metrics and, in some cases, achieve parity with human evaluation, especially in general natural language generation tasks. Within the medical domain, AI models have been specifically instructed to evaluate model responses based on criteria such as helpfulness, relevance, accuracy, and the depth of information provided. Furthermore, its capabilities extend to assessing the overall acceptability and congruence of an MLLM's conversational tone with that of human counselors.

4.4. Human evaluation

Human evaluation remains an indispensable and foundational approach for assessing medical MLLMs.[194] It possesses a unique capacity to capture critical nuances and subjective aspects that automatic metrics and even advanced AI evaluators may not fully grasp. This method is vital for ensuring clinical quality, patient safety, and alignment with human ethical values in the sensitive domain of healthcare. Human evaluation enables a thorough analysis of MLLMs by examining various key aspects: (1) Relevance: Experts determine whether the model's response directly aligns with the intended instruction or clinical query. (2) Coherence: The assessment includes evaluating the logical structure and internal consistency of the generated response. (3) Fluency: This dimension focuses on the naturalness and grammatical soundness of the generated output, ensuring that it is comprehensible and well formed. (4) Accuracy: This is a paramount criterion that assesses the factual correctness of the medical information provided by the model. (5) Appropriateness: Experts evaluate the suitability of the response within a specific clinical context, considering patient safety and standard medical practice. (6) Comprehensiveness: This involves judging the completeness of the information provided, ensuring that all necessary details are included without being overly verbose. (7) Safety: This is a critical aspect that ensures that the response does not pose risks to patient health or provide potentially harmful or misleading advice. (8) Ethical Implications: Assessment extends to adherence of the model's output to established ethical guidelines and principles pertinent to healthcare. (9) Helpfulness: Experts gauge the practical utility of the response in a real-world clinical scenario, determining its value to a clinician or patient.

5. Applications

The advent of LLMs has marked a pivotal moment in AI, with systems such as ChatGPT demonstrating near-human capabilities in processing and generating text.[213,214] Initially developed for textual tasks, the technology frontier has rapidly advanced toward multimodality, enabling newer models such as GPT-4 and Med-Gemini to interpret and synthesize information from a diverse array of inputs, including text, images, and other data formats.[215–217] This evolution is consequential for medicine, a domain inherently reliant on the integration of disparate data streams, from narrative clinical notes and structured laboratory values to complex medical imaging.[218] While text-only LLMs have shown proficiency in knowledge-based assessments such as the United States Medical Licensing Examination,[38,219] their inability to directly process visual data has remained a significant limitation.[9] The development of MLLMs addresses this gap, creating new paradigms for clinical support, research, and education by bridging the divide between visual and textual medical information.[216] These sophisticated systems are being engineered to not only interpret complex data but also to engage in diagnostic reasoning, thereby augmenting the capabilities of healthcare professionals.[216,220] As illustrated in Figure 6, we introduced the related applications of medical MLLMs from these aspects: radiology report generation, diagnostic reasoning support, treatment decision support, education and training, pathology and

Table 6
Medical-specific evaluation metrics for medical MLLMs.

Metric	Formula	Task applicability	Recommended threshold
Diagnostic accuracy[206]	$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$	Medical VQA, Diagnostic Assistance	> 0.90 (Expert-level)
Hallucination rate[19]	$\text{HR} = \frac{\textit{Hallucinated sentences}}{\textit{Verifiable factual sentences}}$	Report Generation, Dialogue	< 0.02 (Very low)
Lesion-wise Dice similarity coefficient[207]	$\text{DSC} = \frac{2TP}{2TP+FP+FN}$	Medical Image Segmentation	> 0.8 (Excellent)
Sensitivity (Recall)[208]	$\text{Sensitivity} = \frac{TP}{TP+FN}$	Pathology Screening, Detection	> 0.95 (High sensitivity)
Specificity[208]	$\text{Specificity} = \frac{TN}{TN+FP}$	Pathology Screening, Triage	> 0.90 (High specificity)
RadGraph F1 score[209]	$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$	Medical Report Generation	> 0.40 (Competitive)
CheXbert F1 score[210]	$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$	Chest X-ray Report Generation	> 0.40 (Competitive)

Abbreviations: 1. TP: true positive 2. TN: true negative 3. FP: false positive 4. FN: false negative 5. HR: hallucination rate 6. DSC: dice similarity coefficient 7. VQA: visual question answering.

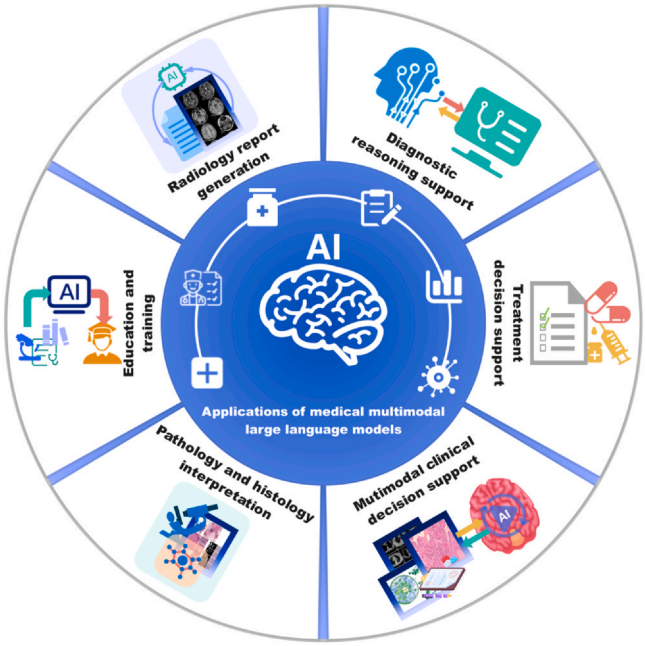


Figure 6. Overview of applications of medical multimodal large language models.

histology interpretation, and multimodal clinical decision support. Although pathological histology interpretation and multimodal clinical decision support essentially belong to the diagnostic and therapeutic phases, respectively, they have become the core focus of current MLLM research. Therefore, this paper addresses them independently. The application of various modality data in MLLMs is shown in Table 7.

5.1. Radiology report generation

This application primarily maps to generating structured reports from medical images, a task that demands both nuanced visual interpretation and precise linguistic articulation. The process of manually drafting these reports is time consuming and can be a significant contributor to clinician workload.[221] MLLMs are being developed to streamline this workflow by directly translating radiological images into structured, coherent, and clinically relevant narrative reports. Foundational models such as Med-Flamingo, LLaVA-Med, MAIRA-1, and RadFM are being engineered through advanced techniques such as instruction fine-tuning on extensive datasets of paired medical images and textual descriptions.[216,217] This training enables them to identify abnormalities, describe pathological findings, and generate diagnostic impressions that are contextually aware.[216] The architectural design often involves a synergistic process in which a vision-based model first analyzes the input image to generate an annotation or tensor output, which is then translated into natural language and provided as a supplementary input to the LLMs. For instance, the ChatCAD model leverages this approach to summarize findings from a computer-aided diagnosis model and formulate a final diagnosis, achieving a recall score substantially higher than prior state-of-the-art models.[9] While some MLLMs such as MAIRA-1 and RadFM are specialized for radiology, others such as Med-Gemini and LLaVA-Med are being developed as versatile foundation models applicable to a broader range of medical domains, including pathology.[216] The capacity of these systems to convert unstructured visual data into structured textual reports not only promises to strengthen efficiency but also to improve the standardization and quality of clinical documentation.[222]

Table 7
Application matrix of medical multimodal large language models.

Modality	Report	Diagnosis	Treatment	Education
Image	✓	✓	✓	✓
Text	✓	✓	✓	✓
Signal	—	✓	✓	—
Genomics	—	✓	✓	—

5.2. Diagnostic reasoning support

This application supports diagnosis by integrating imaging, clinical text, physiological signals, and genetic information. Beyond automated documentation, MLLMs are poised to function as powerful adjuncts in the complex cognitive task of diagnostic reasoning. Clinical diagnosis is fundamentally a multimodal process, requiring the synthesis of patient history, physical examination findings, laboratory results, and, crucially, medical imaging.[218] MLLMs are uniquely positioned to assist in this process by integrating and reasoning over these heterogeneous data types. Models such as Med-Gemini are explicitly trained with curated datasets of medical instructions and multimodal data, enabling them to comprehend not only visual features but also the intricate medical concepts, procedures, and lines of reasoning that underpin a diagnosis.[216] This capacity for cross-modal synthesis allows MLLMs to support clinicians by suggesting differential diagnoses based on a holistic view of the patient's data. For example, by combining its interpretation of a chest X-ray with information from the EHR, such as symptoms and comorbidities, an MLLM can help formulate a more accurate and comprehensive diagnostic assessment.[218] This capability is not limited to radiology; researchers are exploring the use of MLLMs in dentistry, where models could integrate patients' oral images with their spoken symptom descriptions to achieve fully automated diagnostic support.[220] The reasoning is not a "black box"; advanced prompting techniques can guide models like GPT-4 to simulate the cognitive steps of a physician, providing an interpretable chain of thought that enhances trust and transparency.[223] By acting as a logical reasoning tool, MLLMs can improve the accuracy of diagnoses derived from both visual and textual evidence.[9]

5.3. Treatment decision support

MLLMs empower data-driven, personalized treatment planning by synthesizing complex patient information. They analyze genomics, transcriptomics, and proteomics data to identify disease-specific biomarkers, predict individual patient responses to therapies (e.g., targeted cancer drugs), and forecast potential side effects, directly informing precision medicine strategies. For complex cases, they integrate clinical notes, diagnostic images, lab results, and temporal progression data to suggest evidence-based treatment options, weighing risks and benefits specific to the patient's profile. In procedural settings, they interpret real-time video feeds (e.g., endoscopy, surgery) to provide intraoperative insights and decision support. Furthermore, MLLMs analyze longitudinal data from EHR and wearable sensors to optimize chronic disease management plans (e.g., diabetes, heart failure), dynamically adjusting recommendations based on patient responses and trends. They also aid in drug discovery by predicting compound efficacy and interactions based on multimodal biological data.

5.4. Education and training

This application maps to the generation of educational content from images and text. The educational paradigm in medicine is also poised for transformation through the integration of MLLMs. By generating detailed reports from medical images and providing clear, context-aware explanations, models including Med-Gemini can create a virtually limitless supply of dynamic and realistic case studies for medical trainees.[217,224] This functionality allows students to engage with a wider array of clinical scenarios than might be available in traditional

textbooks or clinical rotations, thereby enriching their learning experience.[225] Furthermore, MLLMs can be leveraged to create interactive and personalized learning environments. They can simulate conversations with virtual patients, allowing students to practice and refine their skills in history-taking and clinical reasoning in a controlled, low-stakes setting.[226] The ability of these models to adapt their language and complexity renders them ideal as personalized tutors; they can break down complex topics including protein biosynthesis into simple analogies for a lay audience or provide in-depth, expert-level explanations for advanced learners.[227] This flexibility enables the generation of customized educational materials, from practice questions to entire learning modules, that are tailored to an individual student's progress and areas of weakness.[225] As these tools become more sophisticated, they will not only augment existing educational methods but also foster new approaches to medical training that emphasize critical thinking and interactive problem-solving.[227]

5.5. Pathology and histology interpretation

The application of MLLMs extends into the microscopic realm of pathology and histology, where the interpretation of tissue samples is central to cancer diagnosis and other disease assessments. Similar to radiology, this field relies heavily on the expert analysis of visual data—in this case, histopathological slides. Foundational MLLMs, including Med-Gemini and LLaVA-Med, are being adapted for this domain, and they are trained on vast archives of digitized pathology slides paired with their corresponding diagnostic reports.[216] These systems are being trained to perform tasks such as identifying malignant cells, classifying tumor types, and quantifying biomarkers directly from images. By correlating these visual findings with textual information from a patient's clinical history, MLLMs can provide valuable second opinions and decision support to pathologists. The ability of MLLMs to analyze images at a granular level may also enable the identification of subtle morphological patterns or quantitative metrics that are not easily discernible by the human eye but hold prognostic or predictive value.[77] In this capacity, MLLMs are not envisioned as a replacement for pathologists but as a powerful augmentation tool designed to enhance diagnostic accuracy, reduce inter-observer variability, and streamline the analytical workflow in pathology laboratories.

5.6. Multimodal clinical decision support

The ultimate trajectory of MLLM development in medicine is toward the creation of comprehensive, multimodal clinical decision support systems. Modern healthcare is characterized by an explosion of data, and clinicians are increasingly challenged with integrating information from diverse sources to make optimal treatment decisions.[218,221] MLLMs are being architected to serve as a central hub for this data integration, creating a unified and interactive view of the patient. The successful application of Med-PaLM, which can flexibly encode and interpret multiple biomedical data types, showcases the potential of this approach for tasks ranging from diagnosis to drug development.[220] Future MLLMs are expected to incorporate an even wider range of modalities beyond static images and text. There is a recognized need to process time series data, such as ECG and continuous glucose monitoring streams, as well as other data types such as genomic and proteomic profiles.[216] Integrating these inputs would enable MLLMs to support real-time monitoring, predict disease progression, and recommend personalized treatment plans based on a dynamic and continuously updated understanding of a patient's health status. However, the realization of such systems faces substantial technical hurdles, including the computational challenges of aligning and processing multiple modalities simultaneously and the scarcity of large, curated, and accessible multimodal clinical datasets. Addressing these challenges will be critical to unlocking the full potential of MLLMs to serve as indispensable assistants to the next generation of clinicians.

6. Challenges

Despite the immense potential of medical MLLMs to transform healthcare, their journey from development to clinical deployment faces numerous interconnected challenges. These bottlenecks hinder model performance, reliability, and scalability, necessitating systematic solutions to ensure their safe and effective application in real-world settings. This section categorizes primary obstacles into four key dimensions: data challenges, output trustworthiness, modal alignment and information complementarity, and computational efficiency. The bottlenecks of medical MLLMs are shown in Figure 7.

6.1. Data challenges

There are significant differences between different modal data in terms of format, resolution, and time scale, leading to serious heterogeneity problems faced by the data during fusion processing. Specifically, text data are usually stored in text file format, while image data are in specialized medical imaging formats such as DICOM,[228] and time series data are presented in specific signal data formats. These format differences neccessitate the data require complex conversion and alignment operations during preprocessing and fusion. In addition, there are significant differences in the pixel resolution of the image data and the temporal resolution of the time series, which further increases the difficulty of data fusion. Differences in time scales are also significant. Text data are usually static records, while time series data have dynamic changes at different time intervals, which makes it necessary

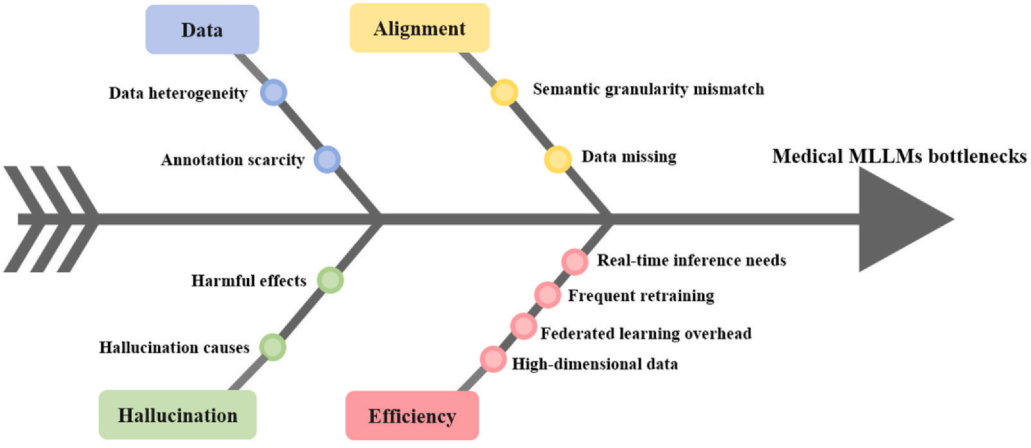


Figure 7. Challenges faced by medical MLLMs. Abbreviation: MLLMs: multimodal large language models.

to consider the processing of time alignment and dynamic changes during fusion. These heterogeneity issues not only increase the complexity of data preprocessing and fusion but may also lead to the loss or misuse of information.[229]

The annotation of medical multimodal data is highly dependent on the knowledge and experience of professional medical experts.[230] For instance, the accurate pairwise annotation of pathological images and diagnostic reports needs to be carried out by pathologists based on their professional knowledge. Due to the limited resources of experts and the time-consuming and laborious labeling work, the scarcity of medical multimodal labeled data seriously restricts the training and performance improvement of the model, which has become one of the key problems to be solved in the process of medical MLLMs.

To tackle these issues, researchers have offered diverse solutions.[231] SSL leverages data's inherent structure and features to produce supervisory signals, lessening the need for manual labeling in multimodal data processing. Many crucial studies have underscored SSL's progress and promise in healthcare,[231–234] positioning it as a tool for comprehensive medical data analysis. Liu et al.[231] examined SSL's role in cardiac, abdominal, and brain MRI segmentation. Several studies[235–237] have delved into SSL's medical applications. These works stress SSL methods in medical imaging and their progress, boosting medical intelligence without vast labeled data. Shurrab et al.[237] reviewed recent SSL methods in medical imaging, dividing them into predictive, generative, and comparative approaches and detailed their workings and imaging applications. With these SSL methods, models are better able to grasp and use data's intrinsic traits in multimodal processing, enhancing performance and generalization.

In addition, cross-modal representation alignment techniques enable matching and correspondence between different modal data (e.g., image, text, and audio.) at the feature, semantic, or representation level. For example, cross-modal prototype memories are built under the Unified Medical Language System,[238] and feature distributions are aligned through optimal transmission theory. By mapping imaging signs and laboratory indicators to the same semantic space, the pathological correlation between them can be quantified. This mapping not only helps improve the interpretability of the model but also enhances its reliability in clinical applications.[239] Chan et al.[240] proposed a generalized LLM framework that can efficiently combine time series data with rich textual contextual information to analyze physiological signals. MTS-LOF[241] utilizes contrast learning and mask auto encoder methods to provide a unique approach for representation learning of medical time series data. Cross-modal representation alignment plays an important role in medical multimodal data processing by providing more accurate and comprehensive feature representations for models, thus improving their performance and generalization capabilities.

To achieve cross-institutional collaborative training while protecting data privacy, model training can be performed using the federated learning enhancement paradigm, which improves the efficiency and effectiveness of multimodal data processing through heterogeneous architecture collaboration and knowledge distillation enhancement.[242,243] Federated learning, as a distributed machine learning paradigm that achieves privacy preservation, allows multiple nodes (hospitals) to collaborate on the resulting different training models by a central server after training the models locally and separately to improve model generalization.[244]

6.2. Output trustworthiness

The trustworthiness of model outputs is paramount in medical applications, where errors can have severe real-world consequences. A primary threat to output reliability is the phenomenon of hallucination. Hallucination of LLMs is a significant phenomenon in which model outputs deviate from factual accuracy, exhibit logical inconsistencies, or contradict established knowledge. This behavior primarily stems from data biases, inadequate training, or the progressive accumulation of reasoning errors, especially in multistep inference contexts.[243,245] The presence of outdated, incorrect, or biased information within training datasets can lead the model to internalize and reproduce flawed correlations. Furthermore, if the model has not been sufficiently trained on domain-specific knowledge—such as in medicine—it becomes prone to generating implausible or erroneous content. Additionally, during extended dialogues or complex reasoning tasks, minor initial inaccuracies may compound, leading to increasingly unrealistic or invented outputs.[246]

In medical applications, the implications of LLM hallucinations are severe. Model-generated inaccuracies can result in incorrect diagnostic suggestions, potentially delaying correct treatment and harming patient outcomes. Similarly, flawed treatment recommendations may introduce unnecessary risks or even endanger lives. Beyond individual patient care, hallucination can contribute to the spread of misinformation in public health contexts, distorting health literacy and undermining trust in medical AI systems.[243]

To mitigate these challenges, researchers have proposed several remedial approaches. Developing specialized benchmark datasets aids in detecting and evaluating LLM hallucinations, especially in medical settings.[17,19,243] Data cleansing and curation of pretraining corpora aim to enhance data quality by removing biased, outdated, or erroneous information.[247,248] Model fine-tuning using high-quality expert-annotated medical data can enhance factual reliability and domain adaptability.[249] Self-verification methods (e.g., verification chains) enable models to autonomously detect and correct errors,[250] while retrieval-augmented generation techniques integrate real-time information retrieval from authoritative external knowledge bases, grounding model responses in verifiable evidence.[251] These hybrid approaches offer promising research directions for reducing hallucinations and enhancing the reliability of LLMs.

6.3. Modal alignment and information complementarity

The effectiveness of LLMs in medicine is further challenged by the need to process and integrate diverse data modalities, such as medical images, EHRs, genomic sequences, and sensor signals. These modalities exist in inconsistent feature spaces and possess varying semantic granularity, making alignment—the process of establishing meaningful connections between them—a complex task.

Key challenges include the frequent absence of paired data (e.g., an image without its corresponding report), high annotation costs leading to scarcely aligned samples, and difficulties in spatiotemporal alignment (e.g., matching continuous ECG signals to discrete lab results). Furthermore, inconsistencies within modalities, such as terminology variations in text or artifacts in images, introduce noise that complicates cross-modal learning.

To address these issues, advanced methods focus on creating unified representation spaces. Medical knowledge-enhanced contrastive learning approaches, such as MedCLIP,[252] use medical ontologies to construct positive and negative sample pairs, effectively aligning image and text embeddings. Hierarchical alignment strategies first align coarse-grained semantics (e.g., disease category) before refining to local features (e.g., linking lesion regions to specific descriptive phrases).

For dynamic information fusion, graph neural networks are employed to build intermodal relationship graphs. Here, features from different modalities (e.g., image regions and text concepts) interact through attention mechanisms to achieve context-aware fusion.[253] Uncertainty-aware models quantify the reliability of each modality and dynamically adjust fusion weights, for instance, prioritizing pathology findings over other data sources. To mitigate reliance on manual labels, SSL techniques leverage inherent temporal or structural correlations in EHR data as signals for weak supervision, facilitating effective alignment, even with limited annotated examples.

6.4. Computational efficiency

In medical data processing, the problem of high-dimensional data loading is particularly prominent. The huge data volume generated by

Table 8
Three-phase future direction for medical MLLMs.

Horizon	Key directions
Short-term	1. Enhanced multimodal fusion: deeper integration of imaging, genomics, and clinical text for comprehensive diagnostics. 2. Federated learning systems: privacy-preserving model training across healthcare institutions using blockchain for data security. 3. Interpretability techniques: development of expert modeling and cue augmentation to improve model transparency, especially for sensor data.
Mid-term	1. Edge intelligence with 6G: deployment of MLLMs on efficient edge architectures leveraging 6G networks for low-latency, real-time multi-agent coordination in clinical environments. 2. IoT and wearable integration: continuous health monitoring in non-clinical settings, enabling telemedicine and smart diagnosis. 3. Drug discovery and clinical trials: application of MLLMs for molecular optimization, patient recruitment, and trial data management.
Long-term	1. AGI in medicine: pursuit of generalist AI models capable of holistic patient care planning based on genetic, historical, and lifestyle data. 2. Intelligent medical education: AI-driven platforms with virtual patients and simulated scenarios for personalized training. 3. Ethical and regulatory frameworks: establishment of comprehensive norms and regulatory sandboxes for responsible medical AI.

Abbreviations: 1. MLLMs: medical multimodal large language models 2. IoT: Internet of things 3. AGI: artificial general intelligence 4. AI: artificial intelligence.

3D/4D medical images (e.g., computed tomography, positron emission tomography, and dynamic scans) leads to a sharp increase in the model input dimensions, and the graphics memory occupancy is too high, which puts huge pressure on computational resources. Meanwhile, the demand for real-time reasoning in emergency scenarios is extremely urgent. For example, the joint image analysis of patients with chest pain requires a second response, but the latency of multimodal joint reasoning often makes it difficult to meet this requirement. In addition, in the context of data privacy isolation between hospitals, the communication overhead of multimodal models in federated learning increases significantly and becomes a bottleneck for distributed training. Rare disease models require frequent retraining due to sparse data, leading to inefficient allocation of computational resources and further exacerbating the long-tail computational load problem.

To cope with these challenges, lightweight model design has become a key strategy. Specifically, modality-specific compression techniques use dynamic sparse convolution for modalities with high computational loads,[254] while distilled compact language models are used for text modalities, thus effectively reducing computational complexity. In addition, the staged inference pipeline first screens key features by lightweight modalities and then triggers high-precision multimodal joint inference to improve inference efficiency. In terms of hardware and algorithm co-optimization, model arithmetic-level acceleration is achieved based on a medical-specific AI chip to support lossless accuracy under INT8 quantization, while the asynchronous update mechanism in federated learning is utilized to allow local training of each hospital modality and sharing only the encrypted fusion-layer gradient to protect data privacy and reduce communication overhead. Finally, edge-cloud collaborative computing deploys a lightweight feature extractor at the edge device in the hospital, and the cloud performs only the multimodal fusion decision, thus reducing the data transmission delay and improving overall system responsiveness and efficiency.

7. Future directions

The evolution of medical MLLMs is poised to progress through distinct phases, shaped by continuous advancements across technical capabilities, operational integration, and ethical governance. These developments will collectively enhance the utility, reliability, and accessibility of MLLMs in real-world medical practice. To systematically outline the anticipated trajectory and provide a clear development roadmap, we categorize future directions into three sequential horizons—short-term, mid-term, and long-term—each building on the previous phase's achievements. These horizons, along with their key research and application priorities, are summarized in Table 8.

The short-term horizon concentrates on strengthening the foundational capabilities of MLLMs, emphasizing enhanced multimodal fusion for comprehensive diagnostics, privacy-preserving federated learning systems, and improved interpretability techniques. In the mid-term, the scope expands toward scalable deployment and broader application, including edge intelligence leveraging 6G networks, integration with IoT and wearable devices for continuous monitoring, and support for drug discovery and clinical trials. The long-term vision looks toward transformative impacts, aiming for artificial general intelligence in medicine, intelligent medical education platforms, and the establishment of robust ethical and regulatory frameworks to guide responsible and widespread adoption.

8. Conclusions

This review has systematically outlined the methodologies, applications, and challenges of medical MLLMs. Key bottlenecks—such as data heterogeneity, LLM hallucination, and computational efficiency—must be addressed to enable robust clinical adoption. Future work should prioritize the development of federated benchmarks and interpretable frameworks to ensure safe and scalable deployment in real-world healthcare settings.

CRediT authorship contribution statement

Chenhan Xu: Writing – original draft, Validation, Investigation. **Bo Lin:** Writing – original draft, Validation, Methodology, Formal analysis. **Weibin Yang:** Writing – original draft, Resources, Methodology, Conceptualization. **Yuan Yan Tang:** Writing – review & editing, Supervision, Resources, Funding acquisition, Formal analysis. **Yuan Hu:** Writing – original draft, Resources, Methodology, Investigation.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No.: 62172458).

Declaration of Competing Interest

Yuan Yan Tang is an editorial board member for Intelligent Oncology and was not involved in the editorial review or the decision to publishthis article. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Fahrner LJ, Chen E, Topol E, Rajpurkar P. The generative era of medical AI. *Cell.* 2025;188(14):3648–3660.
2. Lin Q, Zhu Y, Mei X, et al. Has multimodal learning delivered universal intelligence in healthcare? A comprehensive survey. *Inf Fusion.* 2025;116:102795.
3. Krones F, Marikkar U, Parsons G, Szmul A, Mahdi A. Review of multimodal machine learning approaches in healthcare. *Inf Fusion.* 2025;114:102690.
4. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17: Curran Associates Inc; 2017:6000–6010.
5. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. *Nat Med.* 2025;31(3):943–950.
6. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl Sci.* 2021;11(14).

7. Zhou H, Liu F, Gu B, et al. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:231105112*. Published online 2023.
8. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a Large Language Model Meta-AI (LLaMA) using medical domain knowledge. *Cureus.* 2023;15(6):e40895.
9. Wang S, Zhao Z, Ouyang X, Liu T, Wang Q, Shen D. Interactive computer-aided diagnosis on medical image using large language models. *Commun Eng.* 2024;3(1):133.
10. Li C, Wong C, Zhang S, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:230600890*. Published online 2023.
11. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:220302155*. Published online 2022.
12. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. *arXiv preprint arXiv:230408485*. Published online 2023.
13. Zhang S, Dong L, Li X, et al. Instruction tuning for large language models: a survey. *arXiv preprint arXiv:230810792*. Published online 2023.
14. Liu Z, Dou G, Jia M, et al. Protecting privacy in multimodal large language models with mllmu-bench. *arXiv preprint arXiv:241022108*. Published online 2024.
15. Huang X, Wang X, Zhang H, et al. Medical MLLM is vulnerable: cross-modality jailbreak and mismatched attacks on medical multimodal large language models. *arXiv preprint arXiv:240520775*. Published online 2024.
16. Lau GKR, Dao H, Low BKH. Uncertainty quantification for MLLMs. In: *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI.;* 2025.
17. Pal A, Umapathi LK, Sankarasubbu M. Med-halt: medical domain hallucination test for large language models. *arXiv preprint arXiv:230715343*. Published online 2023.
18. Rawte V, Sheth A, Das A. A survey of hallucination in large foundation models. *arXiv preprint arXiv:230905922*. Published online 2023.
19. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Comput Surv.* 2023;55(12):1–38.
20. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP Tasks. *arXiv preprint arXiv:*200511401. Published online 2021.
21. White J, Fu Q, Hays S, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:230211382*. Published online 2023.
22. Sultan AS, Elgharib MA, Tavares T, Jessri M, Basile JR. The use of artificial intelligence, machine learning and deep learning in oncologic histopathology. *J Oral Pathol Med.* 2020;49(9):849–856.
23. Liu X, Song L, Liu S, Zhang Y. A review of deep-learning-based medical image segmentation methods. *Sustainability.* 2021;13(3):1224.
24. Banu SF, Sarker MdMK, Abdel-Nasser M, Puig D, Raswan HA. AWEU-Net: an attention-aware weight excitation U-Net for lung nodule segmentation. *arXiv preprint arXiv:211005144*. Published online 2021.
25. Peng Y, Zhang T, Guo Y. Cov-TransNet: dual branch fusion network with transformer for COVID-19 infection segmentation. *Biomed Signal Process Control.* 2023;80:104366.
26. Ahmed F, Uddin MDJ. OcuViT: a vision transformer-based approach for automated diabetic retinopathy and AMD classification. *J Imaging Inform Med.* Published online September 19, 2025.
27. ThangaPurni JS, Braveen M. Unified ARP-ViT-CNN system: hybrid deep learning approach for segmenting and classifying multiple skin cancer lesions. *Array.* 2025;28:100515.
28. Tagnamas J, Ramadan H, Yahyaouy A, Tairi H. Multi-task approach based on combined CNN-transformer for efficient segmentation and classification of breast tumors in ultrasound images. *Vis Comput Ind Biomed Art.* 2024;7(1):2.
29. Pu W. Shuffle GAN with autoencoder: a deep learning approach to separate moving and stationary targets in SAR imagery. *IEEE Trans Neural Netw Learn Syst.* 2022;33(9):4770–4784.
30. Luschi A, Tognetti L, Cevenini G, Rubegni P, Iadanza E. *A region-specific GAN-based solution for data augmentation in dermatology. International Conference on Artificial Intelligence in Medicine.* Springer; 2025:240–244.
31. Luschi A, Tognetti L, Cartocci A, Cevenini G, Rubegni P, Iadanza E. Advancing synthetic data for dermatology: GAN comparison with multi-metric and expert validation approach. *Health Technol.* 2025;15(3):553–562.
32. Wu J, Gan W, Chen Z, Wan S, Lin H. AI-generated content (aigc): a survey. *arXiv preprint arXiv:230406632.* Published online 2023.
33. Shanahan M. Talking about large language models. *arXiv preprint arXiv:221203551*. Published online 2023.
34. He K, Mao R, Lin Q, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Inf Fusion.* 2025;118:102963.
35. Pais C, Liu J, Voigt R, Gupta V, Wade E, Bayati M. Large language models for preventing medication direction errors in online pharmacies. *Nat Med.* 2024;30(6):1574–1582.
36. Gao Y, Li R, Croxford E, et al. Large language models and medical knowledge grounding for diagnosis prediction. *medRxiv preprint medRxiv: 23298641*. Published online 2023.
37. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. *Informatics.* 2024;11(3):57.
38. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature.* 2023;620(7972):172–180.
39. Kim C, Gadgil SU, DeGrave AJ, et al. Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nat Med.* 2024;30(4):1154–1165.
40. Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative agents: interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442.* Published online 2023.
41. Li J, Lai Y, Li W, et al. Agent hospital: a simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:240502957*. Published online 2024.
42. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:220111903*. Published online 2023.
43. Medical large language model for diagnostic reasoning across specialties. *Nat Med.* 2025;31(3):743-744.
44. Yao S, Yu D, Zhao J, et al. Tree of thoughts: deliberate problem solving with large language models. *arXiv preprint arXiv:230510601*. Published online 2023.
45. Besta M, Blach N, Kubicek A, et al. Graph of thoughts: solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence.* 2024;38(16):17682–17690.
46. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report. *arXiv preprint arXiv:230308774*. Published online 2023.
47. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:211210752*. Published online 2022.
48. Liu Y, Zhu G, Zhu B, et al. Taisu: a 166m large-scale high-quality dataset for chinese vision-language pre-training. *Adv Neural Inf Process Syst.* 2022;35:16705–16717.
49. Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. Visualbert: a simple and performant baseline for vision and language. *arXiv preprint arXiv:190803557*. Published online 2019.
50. Ge W, Rice HJ, Sheikh IS, et al. Improving neurology clinical care with natural language processing tools. *Neurology.* 2023;101(22):1010–1018.
51. Waisberg E, Ong J, Masalkhi M, et al. GPT-4: a new era of artificial intelligence in medicine. *Irish J Med Sci (1971).* 2023;192(6):3197–3200.
52. Giuffrè M, Masutti F, Gulotta M, et al. Optimizing AI performance for autoimmune hepatitis guidelines: a case study on enhanced GPT model accuracy. *Dig Liver Dis.* 2024;56:S28.
53. Brin D, Sorin V, Barash Y, et al. Assessing GPT-4 multimodal performance in radiological image analysis. *Eur Radiol.* 2025;35(4):1959–1965.
54. Wu Y, Wang S, Yang H, et al. An early evaluation of gpt-4v (ision). *arXiv preprint arXiv:231016534*. Published online 2023.
55. Moujahid H, Cherradi B, Al-Sarem M, et al. Combining CNN and Grad-Cam for COVID-19 disease prediction and visual explanation. *Intell Autom Soft Comput.* 2022;32(2):723–745.
56. Hemker K, Simidjievski N, Jamnik M. HEALNet: Multimodal fusion for heterogeneous biomedical data. *arXiv preprint arXiv:231109115*. Published online 2024.
57. Fan FL, Xiong J, Li M, Wang G. On interpretability of artificial neural networks: a survey. *IEEE Trans Radiat Plasma Med Sci.* 2021;5(6):741–760.
58. Li T, Su Y, Li W, et al. Gmai-vl & gmai-vl-5.5 m: a large vision-language model and a comprehensive multimodal dataset towards general medical ai. *arXiv preprint arXiv:241114522*. Published online 2024.
59. Chen P, Ye J, Wang G, et al. GMAI-MMBench: a comprehensive multimodal evaluation benchmark towards general medical AI. *arXiv preprint arXiv:240803361*. Published online 2024.
60. Sun K, Xue S, Sun F, et al. Medical multimodal foundation models in clinical diagnosis and treatment: applications, challenges, and future directions. *arXiv preprint arXiv:241202621*. Published online 2024.
61. Jaeger S, Candemir S, Antani S, Wáng YXJ, Lu PX, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant Imaging Med Surg.* 2014;4(6):475.
62. Lau JJ, Gayen S, ben Abacha A, Demner-Fushman D. A dataset of clinically generated visual questions and answers about radiology images. *Sci Data.* 2018;5(1):180251.
63. Pereira F, Lou B, Pritchett B, et al. Toward a universal decoder of linguistic meaning from brain activation. *Nat Commun.* 2018;9(1):963.
64. Abacha AB, Hasan SA, Datla VV, Liu J, Demner-Fushman D, Müller H. VQA-MED: overview of the medical visual question answering task at ImageClef 2019. *CEUR Workshop Proceedings.* 2019;2380.
65. Hollenstein N, Troendle M, Zhang C, Langer N. ZuCo 2.0: a dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:191200903*. Published online 2019.
66. Yue X, Ni Y, Zhang K, et al. MMMU: a massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *arXiv preprint arXiv:231116502*. Published online 2023.
67. Gupta D, Attal K, Demner-Fushman D. A dataset for medical instructional video classification and question answering. *arXiv preprint arXiv:220112888*. Published online 2022.
68. Chen J, Wang X, Huang C, Hu X, Shen X, Zhang D. A large finer-grained affective computing EEG dataset. *Sci Data.* 2023;10(1):740.
69. Chen P, Zhu C, Zheng S, Li H, Yang L. WSI-VQA: interpreting whole slide images by generative visual question answering. *arXiv preprint arXiv:240705603*. Published online 2024.
70. Mou X, He C, Tan L, et al. ChineseEEG: a Chinese linguistic corpora EEG dataset for semantic alignment and neural decoding. *Sci Data.* 2024;11(1):550.
71. Lee MH, Shomanov A, Begim B, et al. EAV: EEG-audio-video dataset for emotion recognition in conversational contexts. *Sci Data.* 2024;11(1):1026.
72. Pelka O, Koitka S, Rückert J, Nensa F, Friedrich CM. Radiology Objects in COntext (ROCO): a multimodal image dataset. In: Stoyanov D, Taylor Z, Balocco S, et al., eds. *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis.* Springer International Publishing; 2018:180–189.

73. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data.* 2019;6(1):317.
74. Subramanian S, Wang LL, Mehta S, et al. Medicat: a dataset of medical images, captions, and textual references. *arXiv preprint arXiv:201006000*. Published online 2020.
75. He X, Zhang Y, Mou L, Xing E, Xie P. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:200310286*. Published online 2020.
76. Liu B, Zhan LM, Xu L, Ma L, Yang Y, Wu XM. SLAKE: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. *arXiv preprint arXiv:* 210209542. Published online 2021.
77. Wu C, Zhang X, Zhang Y, Hui H, Wang Y, Xie W. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nat Commun.* 2025;16(1):7866.
78. Lin W, Zhao Z, Zhang X, et al. Pmc-clip: contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:230307240*. Published online 2023.
79. Zhang X, Wu C, Zhao Z, et al. Pmc-vqa: visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:230510415*. Published online 2023.
80. Ikezogwo WO, Seyfioglu MS, Ghezloo F, et al. Quilt-1M: one million image-text pairs for histopathology. *arXiv preprint arXiv:230611207*. Published online 2023.
81. Hu X, Gu L, An Q, et al. Expert Knowledge-Aware Image Difference Graph Representation Learning for Difference-Aware Medical Visual Question Answering. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '23. ACM; 2023:4156–4165.
82. Zhang M, Hu X, Gu L, et al. Expert uncertainty and severity aware chest X-ray classification by multi-relationship graph learning. *arXiv preprint arXiv:230903331*. Published online 2023.
83. Hamamci IE, Er S, Almas F, others. A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv240317834*. Published online 2024.
84. Hu Y, Li T, Lu Q, et al. OmniMedVQA: a new large-scale comprehensive evaluation benchmark for medical LVLM. *arXiv preprint arXiv:240209181*. Published online 2024.
85. Sun Y, Wu H, Zhu C, et al. PathMMU: a massive multimodal expert-level benchmark for understanding and reasoning in pathology. *arXiv preprint arXiv:240116355*. Published online 2024.
86. Zheng X, Wang L, Chen K, Lyu Y, Zhou J, Wang L. EIT-1M: one million EEG-image-text pairs for human visual-textual recognition and more. *arXiv preprint arXiv:240701884*. Published online 2024.
87. Lozano A, Sun MW, Burgess J, et al. BIOMEDICA: an open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. *arXiv preprint arXiv:250107171*. Published online 2025.
88. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:14091556*. Published online 2014.
89. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.; 2016:770-778.
90. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:201011929*. Published online 2020.
91. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:210300020*. Published online 2021.
92. Brock A, De S, Smith SL, Simonyan K. High-performance large-scale image recognition without normalization. *arXiv preprint arXiv:210206171*. Published online 2021.
93. Fang Y, Wang W, Xie B, et al. EVA: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:221107636*. Published online 2022.
94. Srivastav S, Ranjit M, Pérez-García F, et al. MAIRA at RRG24: A specialised large multimodal model for radiology report generation. In: Demner-Fushman D, Ananiadou S, Miwa M, Roberts K, Tsujii J, eds. *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics; 2024:597-602.
95. Lu MY, Chen B, Williamson DFK, et al. A multimodal generative AI copilot for human pathology. *Nature.* 2024;634(8033):466–473.
96. Sun Y, Zhu C, Zheng S, et al. PathAsst: a generative foundation AI assistant towards artificial general intelligence of pathology. *arXiv preprint arXiv:221107636*. Published online 2024.
97. Pellegrini C, Özsoy E, Busam B, Navab N, Keicher M. Radialog: a large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:231118681*. Published online 2023.
98. Zhang S, Xu Y, Usuyama N, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:230300915*. Published online 2023.
99. Bannur S, Hyland S, Liu Q, et al. Learning to exploit temporal structure for biomedical vision-language processing. *arXiv preprint arXiv:230104558*. Published online 2023.
100. Pérez-García F, Sharma H, Bond-Taylor S, et al. Exploring scalable medical image encoders beyond text supervision. *Nat Mach Intell.* 2025;7(1):119–130.
101. Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers. Published online 2021. *arXiv preprint arXiv:210414294*.
102. Hannun A, Case C, Casper J, et al. Deep speech: scaling up end-to-end speech recognition. *arXiv preprint arXiv:14125567*. Published online 2014.
103. Albaqami H, Hassan GM, Datta A. Automatic detection of abnormal eeg signals using wavenet and lstm. *Sensors.* 2023;23(13):5960.
104. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805*. Published online 2019.
105. Baidya R, Jeong H. Anomaly detection in time series data using reversible instance normalized anomaly transformer. *Sensors.* 2023;23(22):9272.
106. Bai N, Wang X, Han R, Wang Q, Liu Z. PAFormer: anomaly detection of time series with parallel-attention transformer. *IEEE Trans Neural Netw Learn Syst.* 2025;36(2):3315–3328.
107. Zhu S, Zheng J, Ma Q. MR-Transformer: multiresolution transformer for multivariate time series prediction. *IEEE Trans Neural Netw Learn Syst.* 2025;36(1):1171–1183.
108. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med.* 2021;4(1):86.
109. Zhang X, Chowdhury RR, Gupta RK, Shang J. Large language models for time series: a survey. *arXiv preprint arXiv:240201801*. Published online 2024.
110. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 2016;26(7):990–999.
111. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* 2017;18(1):67.
112. Poplin R, Chang PC, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol.* 2018;36(10):983–987.
113. Bavishi R, Elsen E, Hawthorne C, et al. Introducing our multimodal models. *Adept Blog.* Published online 2023.
114. Zhang S, Roller S, Goyal N, et al. OPT: open pre-trained transformer language models. *arXiv preprint arXiv:220501068*. Published online 2022.
115. Chiang WL, Li Z, Lin Z, et al. Vicuna: an open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. Published online March 2023.
116. Touvron H, Lavril T, Izacard G, et al. Llama: open and efficient foundation language models. *arXiv preprint arXiv:230213971*. Published online 2023.
117. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. *arXiv preprint arXiv:230709288*. Published online 2023.
118. Cui Y, Yang Z, Yao X. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:230408177*. Published online 2023.
119. Chung HW, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:221011416*. Published online 2022.
120. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinforma.* 2022;23:6.
121. Bolton E, Venigalla A, Yasunaga M, et al. BioMedLM: a 2.7B parameter language model trained on biomedical text. *arXiv preprint arXiv:240318421*. Published online 2024.
122. Wu C, Lin W, Zhang X, Zhang Y, Wang Y, Xie W. PMC-LLaMA: towards building open-source language models for medicine. *arXiv preprint arXiv:230414454*. Published online 2023.
123. Fedus W, Zoph B, Shazeer N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:210103961*. Published online 2022.
124. Shen S, Hou L, Zhou Y, et al. Mixture-of-experts meets instruction tuning: a winning combination for large language models. *arXiv preprint arXiv:230514705*. Published online 2023.
125. Jiang AQ, Sablayrolles A, Roux A, et al. Mixtral of experts. *arXiv preprint arXiv:240104088*. Published online 2024.
126. McKinzie B, Gan Z, Fauconnier JP, et al. *MM1: methods, analysis and insights from multimodal LLM pre-training. European Conference on Computer Vision.* 2024; 2024:304–323.
127. Yang L, Shen D, Cai C, et al. Solving token gradient conflict in mixture-of-experts for large vision-language model. *arXiv preprint arXiv:240619905*. Published online 2024.
128. Zhao H, Wang Z, Sun Q, et al. LLaVA-CMoE: towards continual mixture of experts for large vision-language models. *arXiv preprint arXiv:250321227*. Published online 2025.
129. Mu S, Lin S. A Comprehensive survey of mixture-of-experts: algorithms, theory, and applications. *arXiv preprint arXiv:250307137*. Published online 2025.
130. Cai W, Jiang J, Wang F, Tang J, Kim S, Huang J. A survey on mixture of experts in large language models. *IEEE Trans Knowl Data Eng.* 2025;37(7):3896–3915.
131. Lin B, Tang Z, Ye Y, et al. Moe-llava: mixture of experts for large vision-language models. *arXiv preprint arXiv:240115947*. Published online 2024.
132. Jiang S, Zheng T, Zhang Y, Jin Y, Yuan L, Liu Z. Med-moe: mixture of domain-specific experts for lightweight medical vision-language models. *arXiv preprint arXiv:240410237*. Published online 2024.
133. Izhar Ja, Japar N, Idris N, Dang T. MicarVLMoE: A modern gated cross-aligned vision-language mixture of experts model for medical image captioning and report generation. *arXiv preprint arXiv:250420343*. Published online 2024.
134. Kiela D, Bhooshan S, Firooz H, Perez E, Testuggine D. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:190902950*. Published online 2019.
135. Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:170707250*. Published online 2017.
136. Wang X, Jiang Y, Bach N, et al. Automated concatenation of embeddings for structured prediction. *arXiv preprint arXiv:201005006*. Published online 2020.
137. Thawakar O, Shaker A, Mullappilly SS, et al. XrayGPT: chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:230607971*. Published online 2025.

138. Chen Q, Hu X, Wang Z, Hong Y. MedBLIP: bootstrapping language-image pre-training from 3D medical images and texts. *arXiv preprint arXiv:230510799*. Published online 2023.
139. Li J, Li D, Savarese S, Hoi S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:230112597*. Published online 2023.
140. Alayrac JB, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:220414198*. Published online 2022.
141. Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, Wu Y. Coca: contrastive captioners are image-text foundation models. *arXiv preprint arXiv:220501917*. Published online 2022.
142. Gao W, Deng Z, Niu Z, et al. Ophglm: training an ophthalmology large language-and-vision assistant based on instructions and dialogue. *arXiv preprint arXiv:230612174*. Published online 2023.
143. Song S, Li X, Li S, et al. How to bridge the gap between modalities: a comprehensive survey on multimodal large language model. *arXiv preprint arXiv:231107594*. Published online 2023.
144. Liu H, Li C, Li Y, Lee YJ. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:231003744*. Published online 2024.
145. Hyland SL, Bannur S, Bouzid K, et al. Maira-1: a specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:231113668*. Published online 2023.
146. Liu J, Wang Z, Ye Q, Chong D, Zhou P, Hua Y. Qilin-med-vl: towards Chinese large vision-language model for general healthcare. *arXiv preprint arXiv:231017956*. Published online 2023.
147. Yao L, Li L, Ren S, et al. Deco: decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:240520985*. Published online 2024.
148. Jaegle A, Gimeno F, Brock A, Zisserman A, Vinyals O, Carreira J. Perceiver: general perception with iterative attention. *arXiv preprint arXiv:210303206*. Published online 2021.
149. Moor M, Huang Q, Wu S, et al. Med-Flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:230715189*. Published online 2023.
150. Lu MY, Chen B, Williamson DFK, et al. A visual-language foundation model for computational pathology. *Nat Med.* 2024;30(3):863–874.
151. Zhao Z, Wang S, Gu J, et al. ChatCAD+: toward a universal and reliable interactive CAD using LLMs. *IEEE Trans Med Imaging.* 2024;43(11):3755–3766.
152. AlSaad R, Abd-Alrazaq A, Boughorbel S, et al. Multimodal large language models in health care: applications, challenges, and future outlook. *J Med Internet Res.* 2024;26:e59505.
153. Lialin V, Deshpande V, Rumshisky A. Scaling down to scale up: a guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:230315647*. Published online 2023.
154. Ding N, Qin Y, Yang G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat Mach Intell.* 2023;5(3):220–235.
155. Wang L, Chen S, Jiang L, et al. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artif Intell Rev.* 2025;58(8):227.
156. Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP. *arXiv preprint arXiv:190200751*. Published online 2019.
157. He J, Zhou C, Ma X, Berg-Kirkpatrick T, Neubig G. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:211004366*. Published online 2021.
158. Zhu Y, Feng J, Zhao C, Wang M, Li L. Counter-interference adapter for multilingual machine translation. *arXiv preprint arXiv:210408154*. Published online 2021.
159. He S, Ding L, Dong D, Zhang M, Tao D. Sparseadapter: an easy approach for improving the parameter-efficiency of adapters. *arXiv preprint arXiv:221004284*. Published online 2022.
160. Mahabadi RK, Henderson J, Ruder S. Compacter: efficient low-rank hypercomplex adapter layers. *arXiv preprint arXiv:210604647*. Published online 2021.
161. Lei T, Bai J, Brahma S, et al. Conditional adapters: parameter-efficient transfer learning with fast inference. *arXiv preprint arXiv:2304.04947*. Published online 2023.
162. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI blog.* 2019;1(8):9.
163. Liu X, Zheng Y, Du Z, et al. GPT understands, too. *arXiv preprint arXiv:210310385*. Published online 2023.
164. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:210408691*. Published online 2021.
165. Li XL, Liang P. Prefix-tuning: optimizing continuous prompts for generation. *arXiv preprint arXiv:210100190*. Published online 2021.
166. Zhang ZR, Tan C, Xu H, Wang C, Huang J, Huang S. Towards adaptive prefix tuning for parameter-efficient language model fine-tuning. *arXiv preprint arXiv:230515212*. Published online 2023.
167. Chen Z, Diao S, Wang B, Li G, Wan X. Towards unifying medical vision-and-language pre-training via soft prompts. *arXiv preprint arXiv:230208958*. Published online 2023.
168. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*. Published online 2021.
169. Wang Y, Zhou L, Zhang W, Zhang F, Wang Y. A soft prompt learning method for medical text classification with simulated human cognitive capabilities. *Artif Intell Rev.* 2025;58(4):118.
170. Zaken E ben, Ravfogel S, Goldberg Y. Bitfit: simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:210610199*. Published online 2021.
171. Vucetic D, Tayaranian M, Ziaeefard M, Clark JJ, Meyer BH, Gross WJ. Efficient fine-tuning of BERT models on the edge. In: *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE; 2022:1838-1842.
172. Akbar-Tajari M, Rajaee S, Pilehvar MT. An empirical study on the transferability of transformer modules in parameter-efficient fine-tuning. *arXiv preprint arXiv:230200378*. Published online 2023.
173. Fu Z, Yang H, So AMC, Lam W, Bing L, Collier N. On the effectiveness of parameter-efficient fine-tuning. *arXiv preprint arXiv:221115583*. Published online 2022.
174. Ansell A, Ponti EM, Korhonen A, Vulić I. Composable sparse fine-tuning for cross-lingual transfer. *arXiv preprint arXiv:211007560*. Published online 2021.
175. Holmes C, Zhang M, He Y, Wu B. NxMTransformer: semi-structured sparsification for natural language understanding via ADMM. *arXiv preprint arXiv:211015766*. Published online 2021.
176. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. *arXiv preprint arXiv:210609685*. Published online 2021.
177. Edalati A, Tahaei M, Kobyzev I, Nia VP, Clark JJ, Rezagholizadeh M. KronA: parameter efficient tuning with Kronecker adapter. *arXiv preprint arXiv:221210650*. Published online 2022.
178. Valipour M, Rezagholizadeh M, Kobyzev I, Ghodsi A. Dylora: parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. *arXiv preprint arXiv:221007558*. Published online 2022.
179. Rajabzadeh H, Valipour M, Zhu T, et al. Qdylora: quantized dynamic low-rank adaptation for efficient large language model tuning. *arXiv preprint arXiv:240210462*. Published online 2024.
180. Liu SY, Wang CY, Yin H, et al. DoRA: weight-Decomposed Low-Rank Adaptation. *arXiv preprint arXiv:240209353*. Published online 2024.
181. Zhang Q, Chen M, Bukharin A, et al. Adalora: adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:230310512*. Published online 2023.
182. Ding N, Lv X, Wang Q, et al. Sparse low-rank adaptation of pre-trained language models. *arXiv preprint arXiv:231111696*. Published online 2023.
183. Hayou S, Ghosh N, Yu B. Lora+: efficient low rank adaptation of large models. *arXiv preprint arXiv:240212354*. Published online 2024.
184. Mao Y, Mathias L, Hou R, et al. Unipelt: a unified framework for parameter-efficient language model tuning. *arXiv preprint arXiv:211007577*. Published online 2021.
185. Chen J, Zhang A, Shi X, Li M, Smola A, Yang D. Parameter-efficient fine-tuning design spaces. *arXiv preprint arXiv:230101821*. Published online 2023.
186. Hu Z, Wang L, Lan Y, et al. Llm-adapters: an adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:230401933*. Published online 2023.
187. Zhang Y, Zhou K, Liu Z. Neural prompt search. *arXiv preprint arXiv:220604673*. Published online 2022.
188. Zhou H, Wan X, Vulić I, Korhonen A. AutoPEFT: automatic configuration search for parameter-efficient fine-tuning. *arXiv preprint arXiv:230112132*. Published online 2024.
189. Li J, Lu W, Fei H, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:240808632*. Published online 2024.
190. Bai J, Bai S, Yang S, et al. A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:230812966*. Published online 2023.
191. Islam R, Moushi OM. Gpt-4o: the cutting-edge advancement in multimodal llm. In: *Intelligent Computing-Proceedings of the Computing Conference*. Springer; 2025:47-60.
192. He S, Nie Y, Wang H, et al. GSCo: towards generalizable AI in medicine via generalist-specialist collaboration. *undefined*. *arXiv preprint arXiv:240415127*. Published online 2024.
193. Bai F, Du Y, Huang T, Meng MQH, Zhao B. M3d: advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:240400578*. Published online 2024.
194. Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:230703109*. Published online 2023.
195. Li B, Wang R, Wang G, Ge Y, Ge Y, Shan Y. Seed-bench: benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:230716125*. Published online 2023.
196. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. *arXiv preprint arXiv:220314371*. Published online 2022.
197. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Isabelle P, Charniak E, Lin D, eds. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2002:311-318.
198. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: *Text Summarization Branches Out.* Association for Computational Linguistics; 2004:74–81.
199. Li J, Galley M, Brockett C, Gao J, Dolan B. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:151003055*. Published online 2015.
200. Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.; 2015:4566-4575.
201. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. Bertscore: evaluating text generation with bert. *arXiv preprint arXiv:190409675*. Published online 2019.
202. Zhang H, Chen J, Jiang F, et al. HuatuoGPT, towards taming language model to be a doctor. *arXiv preprint arXiv:230515075*. Published online 2023.
203. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:230609968*. Published online 2023.
204. Chen Y, Xing X, Lin J, et al. SoulChat: improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. *arXiv preprint arXiv:231100273*. Published online 2023.

205. Chen Y, Wang Z, Xing X, et al. Bianque: balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:231015896*. Published online 2023.
206. Šimundić AM. Measures of diagnostic accuracy: basic definitions. *EJIFCC*. 2009;19(4):203–211.
207. Maier-Hein L, Eisenmann M, Reinke A, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat Commun*. 2018;9(1):5217.
208. Altman DG, Bland JM. Diagnostic tests. 1: sensitivity and specificity. *BMJ*. 1994;308(6943):1552.
209. Jain S, Agrawal A, Saporta A, et al. RadGraph: extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:210614463*. Published online 2021.
210. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *arXiv preprint arXiv:200409167*. Published online 2020.
211. Liu JM, Li D, Cao H, Ren T, Liao Z, Wu J. Chatcounselor: a large language models for mental health support. *arXiv preprint arXiv:230915461*. Published online 2023.
212. Welsby P, Cheung BMYChatGPT. *Postgrad Med J*. 2023;99(1176):1047–1048.
213. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med*. 2023;3(1):141.
214. Yu E, Chu X, Zhang W, et al. Large language models in medicine: applications, challenges, and future directions. *Int J Med Sci*. 2025;22(11):2792–2801.
215. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940.
216. Liu F, Zhou H, Gu B, et al. Application of large language models in medicine. *Nat Rev Bioeng*. 2025;3(6):445–464.
217. Tu T, Azizi S, Driess D, et al. Towards Generalist Biomedical AI. *arXiv preprint arXiv:230714334*. Published online 2023.
218. Meskó B. The impact of multimodal large language models on health care's future. *J Med Internet Res*. 2023;25:e52865.
219. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:230313375*. Published online 2023.
220. Huang H, Zheng O, Wang D, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. *Int J Oral Sci*. 2023;15(1):29.
221. Tripathi S, Sukumaran R, Cook TS. Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. *J Am Med Inform Assoc*. 2024;31(6):1436–1440.
222. Roberts K. Large language models for reducing clinicians' documentation burden. *Nat Med*. 2024;30(4):942–943.
223. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit Med*. 2024;7(1):20.
224. Saab K, Tu T, Weng WH, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:240418416*. Published online 2024.
225. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. *J Educ Eval Health Prof*. 2024;21:6.
226. Holderried F, Stegemann-Philipps C, Herschbach L, et al. A Generative Pretrained Transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. *JMIR Med Educ*. 2024;10:e53961.
227. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ*. 2024;17(5):926–931.
228. Jian X, Yuan L. Research and application of multimodal data fusion in medical images. In: *2023 IEEE 7th Information Technology and Mechatronics Engineering Conference (ITOEC)*. Vol 7.; 2023:128-132.
229. Hemker K, Simidjievski N, Jamnik M. HEALNet–hybrid multi-modal fusion for heterogeneous biomedical data. *arXiv preprint arXiv:231109115*. Published online 2023.
230. Shetty S, Ananthanarayana VS, Mahale A. Comprehensive review of multimodal medical data analysis: open issues and future research Directions. *Acta Informatica Pragensia*. 2022;11(3):423–457.
231. Liu Z, Kainth K, Zhou A, et al. A review of self-supervised, generative, and few-shot deep learning methods for data-limited magnetic resonance imaging segmentation. *NMR Biomed*. 2024;37(8):e5143.
232. Rani V, Kumar M, Gupta A, Sachdeva M, Mittal A, Kumar K. Self-supervised learning for medical image analysis: a comprehensive review. *Evol Syst*. 2024;15(4):1607–1633.
233. VanBerlo B, Hoey J, Wong A. A survey of the impact of self-supervised pretraining for diagnostic tasks in medical X-ray, CT, MRI, and ultrasound. *BMC Med Imaging*. 2024;24(1):79.
234. Pani K, Chawla I. Examining the quality of learned representations in self-supervised medical image analysis: a comprehensive review and empirical study. *Multimed Tools Appl*. 2025;84(9):6295–6325.
235. Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digit Med*. 2023;6(1):74.
236. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. *Nat Biomed Eng*. 2022;6(12):1346–1352.
237. Shurrab S, Duwairi R. Self-supervised learning methods and applications in medical imaging analysis: a survey. *Peer J Comput Sci*. 2022;8:e1045.
238. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004 32(Database issue):D267-70.
239. Afshar M, Gao Y, Gupta D, Croxford E, Demner-Fushman D. On the role of the UMLS in supporting diagnosis generation proposed by large language models. *J Biomed Inform*. 2024;157:104707.
240. Chan N, Parker F, Bennett W, et al. Medtsllm: leveraging llms for multimodal medical time series analysis. *arXiv preprint arXiv:240807773*. Published online 2024.
241. Li H, Carreon-Rascon AS, Chen X, Yuan G, Li A. MTS-LOF: medical time-series representation learning via occlusion-invariant features. *IEEE J Biomed Health Inform*. 2024:1–12 Published online.
242. Xie L, Lin M, Luan T, et al. Mh-pflid: model heterogeneous personalized federated learning via injection and distillation for medical data analysis. *arXiv preprint arXiv:240506822*. Published online 2024.
243. Sahoo P, Meharia P, Ghosh A, Saha S, Jain V, Chadha A. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *arXiv preprint arXiv:240509589*. Published online 2024.
244. Zhang F, Shuai Z, Kuang K, Wu F, Zhuang Y, Xiao J. Unified fair federated learning for digital healthcare. *Patterns*. 2024;5(1):100907.
245. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans Inf Syst*. 2025;43(2):1–55.
246. Zhang Y, Li Y, Cui L, et al. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:230901219*. Published online 2025.
247. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233–1239.
248. Cao Y, Kang Y, Wang C, Sun L. Instruction mining: instruction data selection for tuning large language models. *arXiv preprint arXiv:230706290*. Published online 2023.
249. Chen L, Li S, Yan J, et al. Alpagasus: training a better alpaca with fewer data. *arXiv preprint arXiv:230708701*. Published online 2023.
250. Dhuliawala S, Komeili M, Xu J, et al. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:230911495*. Published online 2023.
251. Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:210407567*. Published online 2021.
252. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:221010163*. Published online 2022.
253. Zhan F, Zhang J, Yu Y, Wu R, Lu S. Modulated contrast for versatile image synthesis. *arXiv preprint arXiv: 220309333*. Published online 2022.
254. Vooturi DT, Varma G, Kothapalli K. Dynamic block sparse reparameterization of convolutional neural networks. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*.; 2019:3046-3053.