Advanced IntelliEngineering

Vol. 1, No. 1, p. 21-44, 2026

Invited review

Large language models for intelligent healthcare and medicine: A systematic survey

Chaolei Wu[1], Yanhui Zhu[2]*

[1]*College of Information Engineering, Hubei University of Chinese Medicine, Wuhan 430065, P. R. China*
[2]*School of Computing, Montclair State University, Montclair NJ 07043, United States*

Keywords:
Large language models
medical artificial intelligence
intelligent healthcare
medical engineering
intelligent medical systems

Cited as:
Wu, C., Zhu, Y. Large language models for intelligent healthcare and medicine: A systematic survey. Advanced IntelliEngineering, 2026, 1(1): 21-44.
https://doi.org/10.46690/aie.2026.01.03

Abstract:
Against the backdrop of healthcare digitalization, intelligent medical systems, and the rapid expansion of multi-source medical data, large language models are reshaping medical artificial intelligence from a traditional task-driven paradigm into a generative, interactive, and multimodal intelligence paradigm. This paper systematically reviews recent progress in large language models and multimodal large language models for intelligent healthcare and medicine, with emphasis on technological evolution, model architectures, training data, evaluation systems, medical applications, and trustworthy deployment. The review first traces the development of medical artificial intelligence from supervised learning and pre-trained language models to large language models and multimodal large language models. It then summarizes the backbone architectures and domain adaptation methods of medical large language models, as well as modality encoding, alignment mechanisms, and task interfaces in multimodal systems. In addition, we discuss medical data resources, fine-tuning strategies, evaluation frameworks, and their practical value in diagnosis augmentation, clinical document generation, medical education, mental health services, and specialized clinical scenarios. Although existing studies have demonstrated substantial potential, medical large language models still face critical challenges, including hallucinations, insufficient factual reliability, privacy and data security risks, fairness and bias concerns, limited interpretability, and barriers to clinical deployment. Finally, we outline promising future directions, including retrieval-augmented generation, multimodal fusion, privacy-preserving training, human-machine collaboration, and general medical artificial intelligence. This survey aims to provide a comprehensive reference for trustworthy research, engineering implementation, and intelligent healthcare applications of medical large language models.

1. Introduction

Population aging, the rising burden of chronic diseases, and persistent disparities in the distribution of medical resources are placing growing pressure on contemporary healthcare systems[1, 2]. In routine clinical practice, physicians must synthesize heterogeneous patient information, including symptoms, laboratory tests, medical imaging, longitudinal medical histories, and treatment responses. They must also integrate rapidly expanding biomedical literature, clinical guidelines, and real-world evidence in a timely manner[1, 3–5]. At the same time, healthcare systems face an urgent need to improve efficiency, accessibility, and personalized care[6, 7]. Intelligent technologies that support medical knowledge understanding, clinical information processing, and medical decision-making have therefore become an important research direction in healthcare and medicine.

The development of medical digitalization has created a

Yandy Scientific Press

*Corresponding author.
E-mail address: charley5@stmail.hbucm.edu.cn (C. Wu); zhuya@montclair.edu (Y. Zhu).

Received April 10, 2026; revised May 17, 2026; accepted June 1, 2026; available online June 5, 2026.

strong foundation for the use of artificial intelligence (AI) in medicine. Electronic medical records, medical images, pathology images, genomic data, physiological signals, wearable-device data, medical literature, and clinician-patient communication records together form a large reservoir of multi-source and multimodal medical data[8, 9]. Traditional machine learning and deep learning methods have achieved success in tasks such as medical image recognition, disease risk prediction, biomedical information extraction, clinical text classification, and computer-aided diagnosis. However, these methods usually rely on task-specific annotations and predefined model designs, and they often address only narrowly defined problems[9, 10]. When applied to complex clinical scenarios, cross-modal data fusion, open-ended medical question answering, and dynamic clinician-patient interactions, traditional approaches still face limitations in generalization, scalability, and human-computer interaction.

Large language models (LLMs) have developed rapidly in recent years and have created a new research paradigm for medical AI[11]. Compared with earlier pre-trained language models that mainly focused on discriminative tasks, LLMs offer stronger capabilities in natural language processing (NLP), knowledge representation, semantic reasoning, text generation, and instruction following[12]. Through large-scale pre-training, supervised fine-tuning, instruction optimization, reinforcement learning with human feedback, retrieval-augmented generation (RAG), and multimodal alignment, LLMs can process medical knowledge and clinical information in a more flexible way. This shift has enabled medical AI to move progressively from conventional task-specific models toward generative, interactive, and knowledge-driven intelligent systems.

LLMs have now been widely investigated for medical applications such as biomedical information extraction, medical text classification, clinical question answering, electronic medical record analysis, and multimodal medical understanding[13]. In medical consultation and conversational scenarios, LLMs can interact with users through natural language and provide informational assistance to physicians, patients, and medical researchers[14]. In clinical text and electronic medical record analysis, they can extract key information, including diseases, medications, symptoms, test results, and treatment protocols. In medical imaging and report generation, multimodal LLMs strengthen the ability to jointly interpret images, text, and clinical context[8]. These applications show that LLMs are evolving from text-processing tools into a foundational paradigm for organizing medical knowledge, processing clinical information, and enabling intelligent medical services.

The value of LLMs in medicine goes beyond completing individual medical tasks. Their broader significance lies in their role as components of intelligent medical systems and medical engineering applications[15, 16]. In a practical medical environment, a usable intelligent system must not only answer medical questions but also integrate hospital information systems, fuse multi-source patient data, retrieve external medical evidence, generate explainable auxiliary suggestions, and interact effectively with clinicians and patients. From this perspective, LLMs can be viewed as an interface among medical data, clinical knowledge, and healthcare workflows. They are expected to become core components of intelligent clinical decision support systems, digital health platforms, remote medical services, and personalized health management systems. Studying LLMs from the perspectives of medical engineering and system transformation is therefore important for understanding both their technical potential and their deployment value in real clinical settings.

Despite their substantial potential, LLMs face several barriers to practical implementation in medicine. First, medical data are highly sensitive, and privacy, ethics, and regulatory requirements restrict data sharing, model training, and inter-institutional collaboration. Second, LLMs may generate factual errors, hallucinations, biases, or outdated knowledge, which can have serious consequences in high-stakes medical situations[16, 17]. Third, the interpretability, traceability, and accountability of model outputs remain insufficient, limiting clinicians' trust in model-assisted suggestions. Fourth, performance may vary across hospitals, regions, languages, populations, and disease contexts, so the generalization and clinical adaptability of these models require further validation. In addition, the integration of LLMs with existing hospital information systems, clinical workflows, and regulatory frameworks remains underdeveloped, making the transition from experimental research to clinical application highly challenging.

Compared with previous surveys on general-purpose or medical AI, this review focuses on the technical evolution, applications, and engineering implementation of LLMs and multimodal large language models (MLLMs) in medicine. Existing reviews have discussed progress in medical LLMs, including medical question answering, clinical dialogue, biomedical NLP, model evaluation, and ethical concerns. However, a more integrated and systematic analysis is still needed to connect technological evolution, model architecture, training data, medical applications, and trustworthy deployment. The main contributions of this study are summarized as follows:

1) A structured overview of the evolution, architectures, and multimodal development of medical LLMs;
2) A comprehensive summary of data resources, training methodologies, and evaluation frameworks to assist researchers in understanding model adaptation and performance assessment;
3) A review of representative achievements in medical LLM applications, alongside an analysis of challenges regarding engineering implementation, including data privacy, model reliability, interpretability, and ethical and regulatory issues;
4) A summary of future trends and research directions to provide a reference for future research.

To organize these topics, the overall research framework of this review is illustrated in Fig. 1. Section 2 introduces the evolution and paradigm shifts of medical AI technologies. Section 3 focuses on the architectures of LLMs and MLLMs. Section 4 reviews strategies for data construction, model training, and performance evaluation. Sections 5 and 6 discuss medical applications and challenges in engineering implementation. Section 7 outlines future development directions.

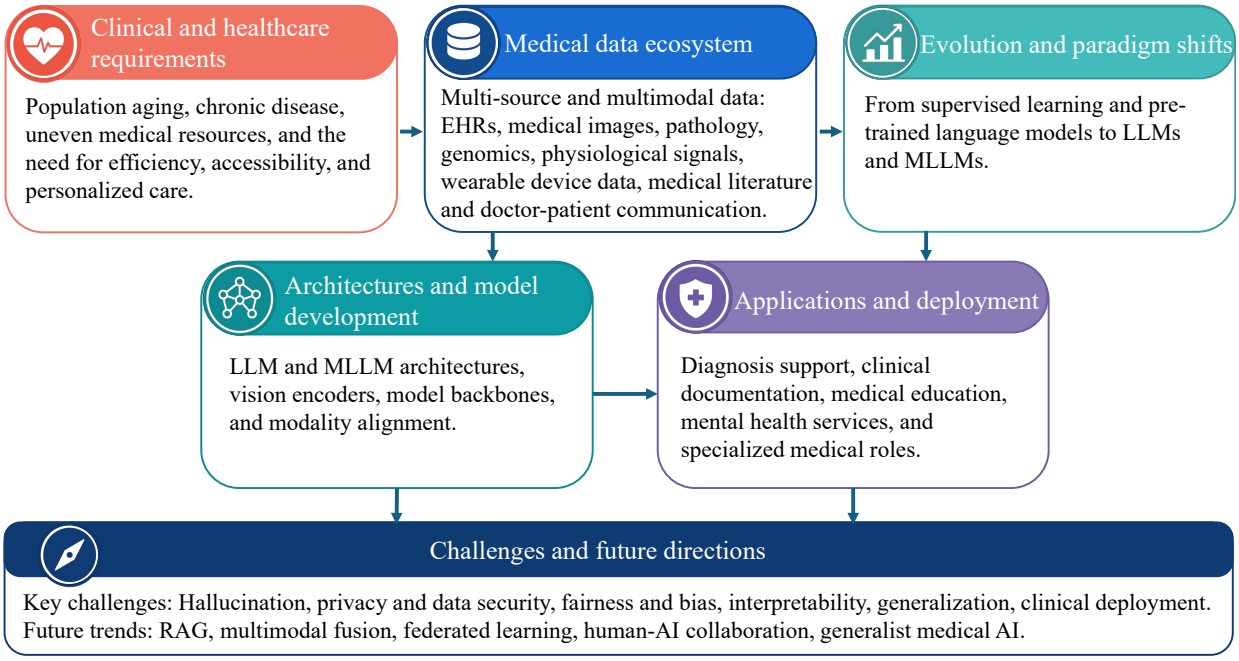

Fig. 1. Overall research framework of this review on LLMs and MLLMs in healthcare and medicine.

2. Evolution and paradigm shifts in medical AI

Medical AI evolution reflects the transition from task-specific prediction to generative, multi-modal and trustworthy intelligent systems. The early medical AI systems were built on hand-crafted features and supervised learning for narrowly defined tasks, while the pre-trained language models (PLMs) enhanced the contextual representation in biomedical and clinical NLP. With the emergence of promptable and instruction-aligned LLMs, medical AI has increasingly supported open-ended question answering, clinical interpretation, and interactive assistance. More recently, MLLMs, long-context modeling, retrieval-enhanced reasoning, and agent-based systems have further expanded the scope of medical AI toward cross-modal evidence integration and generalist medical intelligence. This developmental trajectory and its corresponding engineering focus are summarized in Fig. 2.

2.1 From supervised learning to PLMs

Before the emergence of LLMs, medical AI primarily developed within supervised learning frameworks[18]. During this stage, models were designed for specific medical tasks, including disease classification[19], biomedical entity recognition[20], relation extraction[21], and medical image analysis[22]. Their performance depended heavily on task-specific labeled data, medical domain knowledge, and manually designed modeling strategies. From the perspective of technological evolution, this stage can be understood as a transition from feature engineering to structure engineering, namely, from early machine learning methods to deep learning methods. Early machine learning relied on manually constructed features, whereas deep learning emphasized neural network architecture design and enabled models to learn features automatically from data. With the advent of PLMs, medical NLP began to shift from task-specific modeling toward contextual semantic representation, laying the foundation for medical LLMs[23].

2.1.1 Feature engineering and representation learning

In the early stage of medical AI, supervised learning was typically used for specialized tasks that depended on manually specified features. In medical text processing, bag-of-words models, term frequency-inverse document frequency, rule-based word segmentation, and statistical language models were used to classify diseases, symptoms, drugs, examination items, and clinical events[24–26]. For medical images and physiological signals, researchers first extracted texture features, shape features, gray-level distributions, frequency-domain features, or signal-shape features. These features were then used as inputs to classifiers or regression models such as support vector machines, logistic regression, and random forests. The advantage of this approach was that it could incorporate medical expert knowledge and task-specific rules, while also providing a degree of interpretability when the data scale was limited[27, 28].

With the development of deep learning, medical AI gradually shifted from feature engineering to representation learning. Unlike traditional methods that require manually defined features, deep neural networks can learn multi-level representations automatically from raw or lightly preprocessed medical data. In medical imaging, convolutional neural networks extract local visual patterns and high-level semantic features from radiological scans, pathological slides, and other medical images[29]. In biomedical NLP and clinical text processing, recurrent neural networks, long short-term memory networks, attention mechanisms, and Transformer architectures are used to model contextual dependencies and semantic correlations[30]. By reducing dependence on manually engineered features, representation learning improves the modeling of complex

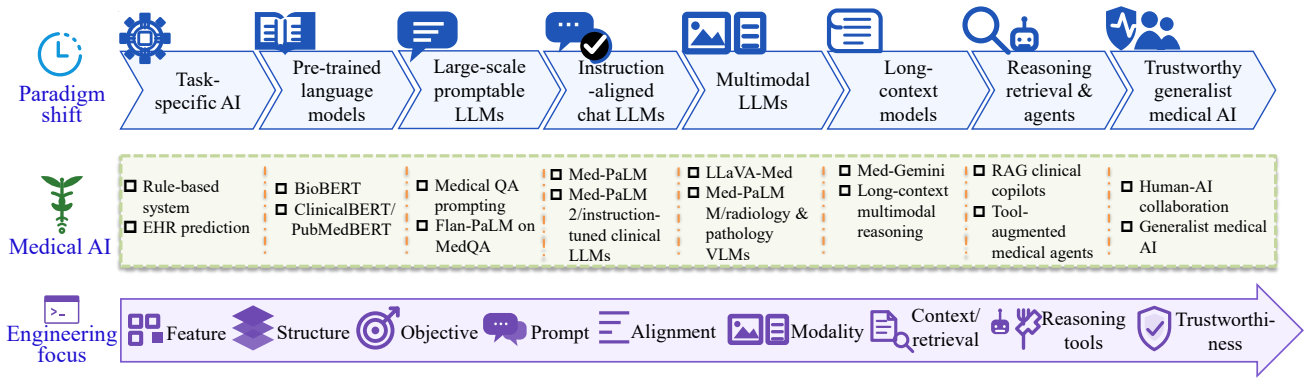

Fig. 2. The evolution of LLMs and MLLMs in medicine.

nonlinear patterns and further advances medical AI in image recognition, disease prediction, clinical text classification, and information extraction[31].

During this period, representation learning systems mainly served specific tasks and relied heavily on supervised training[18, 27, 28]. Models typically required separate training or fine-tuning for each task. As a result, their generalization was often limited when they were applied to new hospitals, disease types, patient populations, or clinical settings. These models also lacked the ability to handle open-ended medical inquiries, generate clinical explanations, or support natural-language interaction[9, 10, 32]. These limitations promoted the evolution of medical NLP toward PLMs, which can learn more generalizable contextual semantic representations through pre-training on large-scale biomedical literature and clinical corpora. This development laid the foundation for LLMs in the medical domain.

2.1.2 The era of BERT and PLMs in fundamental medical NLP tasks

The advent of BERT and the subsequent wave of PLMs[33, 34] advanced medical NLP into the "pre-training–fine-tuning" phase. Compared with traditional machine learning and deep learning methods, BERT-like models can learn context-dependent semantic representations from large-scale corpora during pre-training and then adapt to specific medical scenarios through downstream fine-tuning. The medical field has developed domain-specific pre-trained models such as BioBERT, ClinicalBERT, BlueBERT, and PubMedBERT[33, 34]. These models are generally trained on biomedical literature, clinical records, or electronic medical records, which improves their understanding of medical terminology, clinical abbreviations, and professional contexts.

PLMs are widely used in fundamental medical NLP tasks, including named entity recognition (NER), relation extraction, medical text classification, semantic similarity assessment, information retrieval, and preliminary medical question answering[21, 34, 35]. They reduce dependence on manual feature design and improve performance in medical text understanding. However, these models mainly serve discriminative tasks and usually require separate fine-tuning for each downstream application. Their generative ability, open-ended question-answering capability, multi-turn interaction, and multimodal information integration remain limited. Consequently, BERT-like PLMs cannot fully meet the demand for generative, interactive, and cross-modal intelligence in complex medical contexts, which has further pushed medical AI toward LLMs.

2.2 LLMs and multimodal intelligence in medical AI

With the development of PLM-driven medical NLP, LLMs have further shifted the research paradigm of medical AI. Through larger-scale corpus pre-training and parameter expansion, LLMs demonstrate stronger capabilities in language understanding, knowledge representation, semantic reasoning, and text generation[12, 32, 36]. This enables medical AI to support open-ended tasks such as medical question answering, clinical interpretation, medical record summarization, report generation, and clinician-patient interaction. In this sense, LLMs have moved medical AI from task-specific predictive models toward generative, interactive, and knowledge-based intelligent systems.

Traditional medical AI usually requires labeled datasets and separate model training or fine-tuning for each downstream task. In the LLM era, however, many medical tasks can be reformulated as generative tasks using natural language prompts, instructions, or a small number of examples. Prompt-based methods[37, 38] blur task boundaries and allow models to adapt more flexibly to diverse medical scenarios. This interaction pattern is closer to the natural-language communication among clinicians, patients, and medical researchers in real-world settings.

The development of LLMs has also shifted medical AI from model-centered methods toward data-centered methods. In deep learning, performance improvements were often achieved through advances in network architecture, attention mechanisms, loss functions, or classifier design. In the LLM era, the basic model architecture has become more stable, and model capabilities are increasingly shaped by training-data scale, data quality, domain coverage, instruction data, expert feedback, and evaluation systems. This pattern is especially important in medicine, where specialized knowledge, timeliness, and safety risks are central concerns. High-quality medical literature, clinical practice guidelines, electronic health

records (EHRs), clinician-patient dialogues, expert-annotated data, and preference feedback directly influence the knowledge reliability, practical applicability, and safety of medical LLMs. Therefore, medical LLM development cannot rely only on parameter scale or general model capabilities. It also requires systematic methods for medical data governance, domain adaptation, expert alignment, and reliable evaluation.

Real-world medical scenarios do not depend exclusively on text. Text-based LLMs can process clinical narratives, medical literature, and clinician-patient conversations, but they cannot directly interpret essential non-textual evidence, such as computed tomography (CT) scans, magnetic resonance imagings (MRIs), X-rays, ultrasounds, fundus images, pathological sections, and electrocardiograms[8, 9]. This limitation prevents purely text-based medical intelligence from fully supporting clinical workflows. Because medical data are inherently multimodal, the development of medical AI toward multimodal intelligence is an unavoidable trend.

MLLMs provide a new framework for medical AI by integrating text, images, and other medical modalities[8, 10, 39]. By combining visual encoders, language-model backbones, and modality-alignment modules, MLLMs can map medical images, pathological images, clinical texts, and other modality-specific information into a shared semantic space. This capability supports tasks such as medical visual question answering, image interpretation, report generation, multimodal diagnostic assistance, and patient interaction. Compared with traditional single-modal models, MLLMs more closely reflect the way clinicians integrate multiple sources of evidence in practice. They also provide a new technical foundation for intelligent clinical decision support, digital health platforms, and personalized medical services. The development of LLMs and multimodal intelligence shows that medical AI is moving from task-specific modeling toward generative reasoning, natural-language interaction, and cross-modal medical understanding. This paradigm shift motivates a closer examination of model architectures, modality-alignment mechanisms, and medical adaptation methods for medical LLMs and MLLMs.

2.3 Long-context modeling and reasoning

Long-context modeling has become an important direction in the evolution of LLMs and MLLMs, especially for medical applications in which relevant information is distributed across longitudinal electronic health records, laboratory results, imaging reports, clinical notes, treatment histories, guidelines, and biomedical literature. Earlier LLMs were constrained by relatively short context windows, which required clinicians or downstream systems to compress patient information into limited prompts. Recent long-context models, such as Gemini 1.5, substantially increase the amount of text, image, audio, video, and document-level information that can be processed within a single interaction. This enables models to retrieve and reason over fine-grained information from long documents and multimodal clinical records[40]. In medicine, this capability is valuable because clinical decisions often depend not on an isolated symptom or test result, but on temporal patterns, comorbidities, prior interventions, contraindications, and evolving patient responses. Med-Gemini further demonstrates the promise of combining long-context processing, multimodal inputs, medical specialization, and search-based grounding for medical question answering, clinical summarization, radiology interpretation, and reasoning over long de-identified health records[41].

Reasoning ability is another major axis of recent progress. In medical settings, models are expected not only to retrieve relevant information but also to compare differential diagnoses, reconcile conflicting evidence, follow clinical guidelines, express uncertainty, and generate safe recommendations. Prompting strategies such as chain-of-thought reasoning, self-consistency, ensemble refinement, and RAG have been used to improve performance on complex question-answering and knowledge-intensive tasks[37, 38, 42, 43]. More recently, reasoning-oriented models trained with reinforcement learning, such as OpenAI o1 and DeepSeek-R1, suggest that test-time reasoning and self-verification may become increasingly important for difficult medical tasks that require multi-step inference[44, 45]. Nevertheless, longer context and stronger reasoning do not automatically guarantee clinical reliability. Long-context models may still overlook relevant evidence, overemphasize irrelevant context, or generate plausible but unsupported explanations. Medical long-context reasoning systems should therefore be coupled with explicit evidence retrieval, calibrated uncertainty estimation, source attribution, abstention mechanisms, and clinician-in-the-loop verification.

2.4 Trustworthy generalist medical AI

The long-term goal of medical LLMs and MLLMs is to move from narrow, task-specific systems toward trustworthy generalist medical AI. Such models should flexibly process multiple types of biomedical data and support a broad range of clinical, educational, administrative, and research tasks. Unlike specialized models designed for a single modality or disease category, generalist medical AI aims to integrate clinical language, medical images, laboratory data, waveforms, genomics, and other patient-specific information into a unified or coordinated model system. Med-PaLM, Med-PaLM 2, Med-PaLM M, and Med-Gemini are representative milestones in this direction, showing progress from medical question answering to multimodal biomedical reasoning and more general clinical assistance[12, 38, 39, 41]. Such systems could support diagnosis augmentation, clinical documentation, patient education, triage, literature search, and specialist consultation. In practice, however, generalist medical AI should be viewed as a collaborative assistant rather than an autonomous clinician, because real-world care requires contextual judgment, ethical responsibility, patient preference elicitation, and accountability.

Trustworthiness is the central requirement for deploying generalist medical AI in safety-critical environments. Medical LLMs and MLLMs must address hallucinations, factual inconsistencies, outdated knowledge, privacy leakage, demographic bias, poor calibration, limited interpretability, distribution shift, and user overreliance. Evaluation should therefore move beyond benchmark accuracy and include clinically grounded

Table 1. Comparison of language-model backbones for medical tasks.

Backbone type	Mechanism	Typical model	Strength	Limitation
Encoder-only	Bidirectional contextual encoding	BERT-like models[33, 34]	Text understanding	Weak generation
Decoder-only	Autoregressive generation	GPT/LLaMA like models[36, 38, 46]	Strong generation and interaction	Hallucination and safety risks
Encoder-Decoder	Sequence-to-sequence modeling	T5/BART-like models[47–50]	Good for controlled generation	Less dominant in current instruction-following LLMs

assessment of factuality, reasoning quality, uncertainty, potential harm, subgroup fairness, robustness to adversarial inputs, and consistency with medical consensus[12, 38]. Governance frameworks also emphasize transparency, accountability, privacy protection, equity, human oversight, and post-deployment monitoring for large multimodal models in healthcare[51]. Recent discussions of responsible AI in healthcare further stress that deployment should be supported by clinical evidence generation, workflow integration, bias mitigation, prospective validation, and continuous auditing[52]. Thus, trustworthy generalist medical AI requires not only more capable models but also rigorous evaluation pipelines, domain-specific safety alignment, privacy-preserving training, interpretable evidence grounding, and institutional governance mechanisms.

3. Architectures of medical LLMs and MLLMs

3.1 Backbone architectures of language models in medical AI

The architectures of medical LLMs and MLLMs are largely derived from general language-model backbones. Encoder-only, encoder-decoder, and decoder-only architectures represent three major backbone types. They differ in contextual modeling mechanisms, input-output patterns, and generation capabilities, and these differences shape their roles in medical AI. Rather than revisiting the general Transformer architecture, this section summarizes the key characteristics of these backbones and their relevance to medical language modeling. Table 1 provides a concise comparison of their mechanisms, representative models, strengths, and limitations.

Encoder-only models learn semantic representations of input text through bidirectional context encoding. They are well suited to text understanding tasks and are useful for biomedical and clinical NLP problems that require accurate semantic representation, such as entity recognition, relation extraction, text classification, and semantic matching[33–35]. In medical scenarios, bidirectional encoding helps capture terminology, abbreviations, and context-dependent expressions in clinical notes and biomedical literature[33, 34]. However, encoder-only models are mainly designed for discriminative outputs, such as labels, entity spans, or matching scores. They are therefore less suitable for open-ended generation, long-form clinical explanation, and interactive medical dialogue.

Encoder-decoder models follow a sequence-to-sequence framework[47, 48], in which the encoder represents the input and the decoder generates the target sequence. This structure is suitable for controlled text-generation tasks, including summarization, report generation, text rewriting, and question generation. In medical AI, encoder-decoder models provide an important architectural basis for transforming clinical records, medical reports, or biomedical documents into more structured and readable outputs[49, 50, 53]. Nevertheless, as decoder-only LLMs have become dominant in instruction following, open-ended generation, and conversational interaction, encoder-decoder models have become less central to recent medical LLM development.

Decoder-only models have become the predominant backbone for contemporary LLMs[36, 38, 46]. They generate text autoregressively from preceding context and are well suited to instruction following, medical question answering, dialogue, explanation, summarization, and reasoning-oriented tasks. Compared with encoder-only and encoder-decoder architectures, decoder-only LLMs provide greater flexibility for unifying multiple medical tasks within a generative framework. Their use in medicine also raises critical concerns, including hallucinations, outdated knowledge, factual inconsistencies, limited evidence traceability, and unsafe responses[53–55]. These concerns are especially important when model outputs involve diagnosis, treatment suggestions, medication information, or patient education.

Overall, these backbone architectures provide the structural foundation for medical language models, but they do not by themselves guarantee clinical reliability or domain suitability. To support medical applications, general language-model backbones usually require further adaptation with medical corpora, clinical knowledge, instructional data, output constraints, and safety-alignment mechanisms.

3.2 Domain adaptation of medical LLMs

Although current medical LLMs are usually built on general LLM backbones, these models cannot directly satisfy the requirements of medical scenarios. General LLMs are trained on broad corpora and have strong language understanding and generation capabilities, but their mastery of medical terminology, clinical abbreviations, medical record writing styles, diagnostic and treatment logic, and guideline knowledge remains limited[12, 23, 36]. Domain adaptation is therefore a key step in transforming general LLMs into medical LLMs. Its goal is not to redesign the backbone architecture, but to improve the model's ability to understand medical language, clinical knowledge, medical task instructions, and safety requirements.

This enables the model to better support tasks such as medical question answering, medical record summarization, clinical interpretation, patient education, and medical report generation.

Domain adaptation of medical LLMs is first reflected in medical knowledge and terminology[33, 34, 56]. Medical content contains many professional terms, abbreviations, examination indicators, drug and disease names, diagnostic and treatment procedures, and forms of implicit clinical logic. For example, EHRs, discharge summaries, imaging reports, clinical guidelines, biomedical literature, and clinician-patient dialogues all have distinct domain characteristics in language style, information density, and expression norms. High-quality medical corpora enable models to learn semantic relationships among medical concepts, understand clinical context, and distinguish information about symptoms, diseases, examinations, treatments, and prognosis. This adaptation to medical terminology lays the foundation for subsequent tasks and clinical applications.

Second, medical LLMs require adaptation to medical tasks and instructions[36, 57, 58]. Conventional medical AI often relies on classification or recognition tasks with fixed label spaces, whereas LLMs usually perform tasks through natural language instructions. Medical LLMs must therefore understand medical information, infer user intent, and organize their outputs according to the needs of specific clinical contexts. For example, clinical summarization requires models to extract key information from medical records, including chief complaints, past medical history, examination results, diagnoses, and treatment plans. Medical question answering requires models to distinguish evidence-based factual explanations, diagnostic suggestions, and risk advisories according to question type. Patient education requires the model to express information clearly, understandably, and prudently. By adapting to medical instruction data and task examples, models can better meet medical task requirements and improve usability across scenarios.

Third, medical LLM adaptation must address output control and safety alignment[17, 38, 59]. Model outputs in medical scenarios may involve disease diagnosis, medication explanations, treatment recommendations, and health management. Linguistic fluency is therefore not sufficient; factual consistency, evidence traceability, uncertainty expression, and clinical safety are also essential. Medical LLMs should avoid unsupported conclusions, clearly express uncertainty when evidence is insufficient, and prompt users to seek professional assistance in high-risk situations. They also need to generate structured or semi-structured content for tasks such as differential diagnosis lists, examination summaries, discharge summaries, and medication instructions. These output constraints distinguish medical LLMs from general language models and directly affect their reliability in healthcare and medicine.

Although domain adaptation can improve the medical knowledge representation, clinical language understanding, and task execution capabilities of text-based medical LLMs, these capabilities remain primarily limited to text processing. Real medical decisions often require comprehensive multi-source information, including medical images, pathological images, physiological signals, laboratory tests, structured electronic medical records, and longitudinal follow-up data[8, 9]. Text-only medical LLMs can explain medical knowledge and generate clinical text, but they cannot directly interpret non-textual evidence such as X-rays, CT, MRI, fundus images, pathological sections, and electrocardiograms[8, 60, 61]. Therefore, although domain adaptation of medical LLMs is an important foundation for medical intelligence, it is insufficient for the multimodal requirements of real-world clinical scenarios. This limitation has driven the development of medical MLLMs, which introduce modality encoders and modality-alignment modules on top of the language-model backbone to fuse information across text, images, and other medical modalities.

3.3 Architectures of MLLMs

Building on text-based medical LLMs, medical MLLMs extend model capabilities to non-textual medical information. Unlike LLMs that primarily rely on clinical text, medical literature, or EHRs, medical MLLMs must understand multi-source information, including medical images, physiological signals, pathological images, structured data, and natural-language questions[8, 9, 63]. Their architecture is therefore not a simple expansion of language-model scale. Instead, it combines a medically adapted LLM backbone with modality-specific encoders, modality-alignment modules, and task interfaces to achieve cross-modal information fusion and medical semantic generation[60, 61, 63]. The architecture of medical MLLMs is shown in Fig. 3.

The modality-specific encoder[60–62] is one of the key components of medical MLLMs. Because different modalities have different data structures and semantic features, modality-specific encoders are usually needed to obtain effective representations. For example, medical images such as X-rays, CT, MRI, ultrasound, and fundus images can be represented with CNNs, Vision Transformers, or CLIP-like visual encoders. Pathological images often require stronger local-region modeling because of their high resolution and fine-grained tissue structure. Physiological signals such as electrocardiograms, electroencephalograms, and wearable-device data rely more on time-series encoders, one-dimensional CNNs, or Transformer-based signal encoders. Compared with general image or text data, medical modalities often have high annotation costs, small lesion regions, device heterogeneity, substantial noise, and complex clinical semantics. These characteristics make encoder selection and medical adaptation central issues in MLLM architecture design.

The domain-adapted medical LLM backbone[36, 38, 55] is used to understand inputs, organize medical knowledge, perform reasoning, and generate final responses. In the MLLM architecture, different encoders provide non-textual evidence, while the LLM backbone performs cross-modal semantic integration and natural language generation. For example, in medical visual question answering, the model must identify abnormal image regions and generate medical explanations based on the semantics of the question. In image report generation, the model must convert visual findings into report text that conforms to clinical expression norms. The medi-

Table 2. Key architectural components of medical MLLMs.

Component	Function	Main challenges
Modality-specific encoders	Extract non-textual modality features	High resolution, fine-grained lesions, device heterogeneity, noisy signals[60–62]
Domain-adapted medical LLM backbone	Support medical semantic understanding, reasoning, and generation	Medical knowledge reliability, long-context modeling, safe generation[38, 55]
Modality alignment module	Map non-textual features into the semantic space	Fine-grained semantic alignment, visual grounding, evidence traceability[60, 61, 63]
Task interface and output module	Organize task input and generate clinical outputs	Output standardization, factual consistency, interpretability, safety[53, 64]

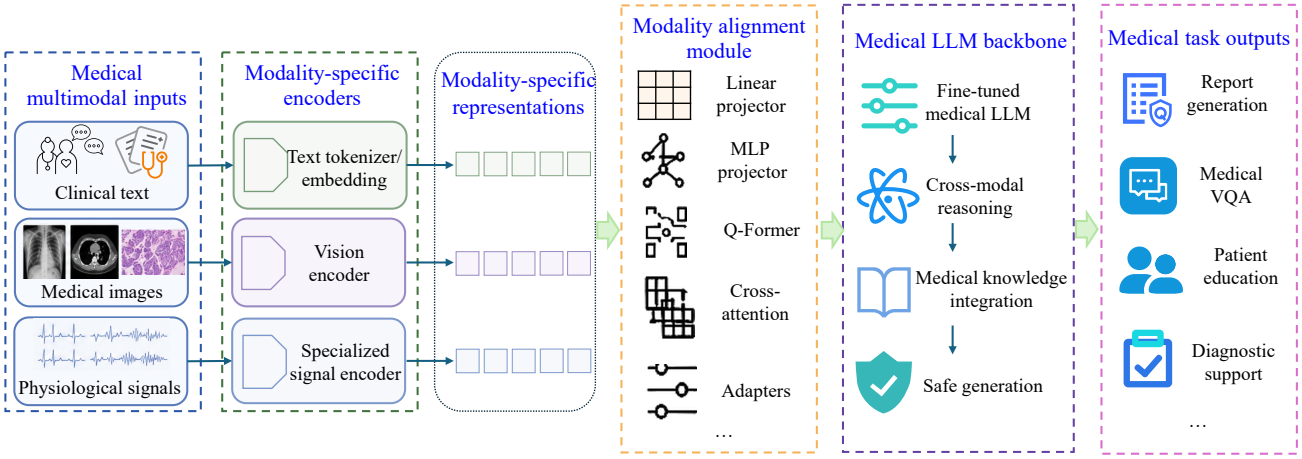

Fig. 3. Architecture of medical MLLM.

cal knowledge coverage, instruction-following ability, long-context modeling, and safe generation capability of the LLM backbone therefore directly affect the clinical usability of medical MLLMs.

Modality alignment[60, 61, 63] maps features extracted by different modality encoders into a shared semantic space and enables the language model to reason and generate from medical images, signals, or structured data. Common alignment methods include linear projection, MLP projectors, cross-attention, and Q-Former. For medical MLLMs, alignment requires fine-grained relationships among medical evidence, local lesions, clinical questions, and textual descriptions. For example, localized opacity in chest X-rays, abnormal tissue regions in pathology images, and irregular waveforms in electrocardiograms must be linked to relevant medical terminology, diagnostic indicators, or report descriptions. Fine-grained visual grounding, cross-modal semantic consistency, and traceability of medical information are therefore essential for distinguishing medical MLLMs from general MLLMs.

Medical MLLMs also require suitable task interfaces and output formats for different medical tasks[53, 55, 64]. Depending on task requirements, the model may generate answers to medical visual questions, image descriptions, structured reports, diagnostic suggestions, explanations of examination results, or patient-oriented health education content. Medical visual question answering emphasizes cross-modal question understanding and evidence identification. Report generation prioritizes factual consistency and standardized medical terminology. Clinical decision-support tasks require multi-source evidence integration, uncertainty expression, and safety constraints. Therefore, the architecture of medical MLLMs must consider input modalities, alignment methods, language generation capability, and clinical output specifications at the same time. Table 2 summarizes the functions and key challenges of the main architectural components in medical MLLMs.

Overall, medical MLLMs extend model capabilities from text comprehension and generation to cross-modal integration of medical evidence by adding modality encoders and alignment modules to the medical LLM backbone. Their primary advantage is that they more closely resemble clinical decision-making, in which medical images, pathological findings, physiological signals, medical record text, and natural-language questions are integrated for interpretation and generation. The applicability of MLLMs depends on the input modalities, output forms, reasoning depth, and clinical risk levels of specific medical tasks. It is therefore necessary to evaluate the advantages and limitations of different architectures for medical task adaptation, as discussed in Section 3-D.

Table 3. Comparison of model architectures in medical task adaptation.

Task	Applicable architecture	Input	Output	Challenges
Medical NER and relation extraction	Encoder-only models	Clinical text, biomedical literature	Entities, labels, relations	Weak generation capability, limited cross-task generalization[35, 65]
Medical text classification and semantic matching	Encoder-only models	Clinical records, question-answering texts, literature abstracts	Categories, matching scores, relevance judgments	Difficulty in handling open-ended explanations and complex reasoning[32, 55]
Clinical summarization and medical record paraphrasing	Encoder-decoder models/Decoder-only LLMs	EHRs, discharge summaries, clinical records	Summaries, structured text	Fact inconsistency and risk of key information omission[53, 64]
Medical question answering and clinical dialogue	Decoder-only LLMs	Medical questions, medical record texts, medical literature	Natural language answers, explanatory texts	Hallucination, outdated knowledge, and safety risks[38, 55]
Medical report generation	Decoder-only LLMs/Medical MLLMs	Medical images, examination results, clinical text	Imaging reports, examination reports	Evidence grounding and report fact consistency[61, 64]
Medical VQA and image interpretation	Medical MLLMs	X-ray, CT, MRI, pathological images, question texts	Image interpretations, question and answer (QA) results	Fine-grained lesion recognition and cross-modal alignment[61, 63]
Multimodal clinical decision support	Medical MLLMs/Hybrid systems	EHRs, medical images, laboratory tests, physiological signals	Capable of integrating multi-source medical evidence	High requirements for interpretability, reliability, and clinical safety[32, 55]

3.4 Comparative analysis of architectures for medical task adaptation

The applicability of different model architectures in medical AI depends mainly on input modalities, output formats, reasoning requirements, and clinical risk levels. Encoder-only models are better suited to discriminative text-understanding tasks such as medical NER, relation extraction, text classification, and semantic matching[34, 35, 65]. These tasks require the model to accurately identify medical terms, clinical entities, and contextual relationships rather than generate open-ended responses. Encoder-decoder models are better suited to controlled text-to-text generation tasks, such as clinical text summarization, medical report generation, and medical record restructuring[48, 50, 53]. Decoder-only LLMs are suitable for open-ended tasks such as medical question answering, clinical conversations, medical explanations, literature understanding, and complex reasoning, because their autoregressive generation and instruction-following capabilities support natural-language interaction[12, 36, 38]. In contrast, medical MLLMs are suitable for tasks involving medical images, pathological images, physiological signals, and clinical texts, including medical visual question answering, image interpretation, multimodal report generation, and multi-source evidence-assisted decision-making[61, 63, 64]. The compatibility between architectures and medical tasks is shown in Table 3.

From the perspective of medical task requirements, architecture selection is usually determined by four factors. The first is input modality. If a task involves only clinical text or biomedical literature, text-based language models can often meet basic needs. If it involves non-textual information such as X-rays, CT, MRI, pathological images, fundus images, or electrocardiograms, modality-specific encoders and alignment modules are needed. The second factor is output form. Medical entity recognition and text classification typically produce labels, entities, or relations and are better suited to encoder-only models. Medical summarization, report generation, and patient education require continuous text generation and are more suitable for encoder-decoder models or decoder-only LLMs. Medical visual question answering and multimodal explanation usually require MLLMs that integrate visual evidence with language reasoning. The third factor is reasoning complexity. Simple classification tasks emphasize local semantic recognition, whereas clinical question answering, differential diagnosis, and multimodal decision support require cross-sentence, cross-document, and cross-modal evidence integration. The fourth factor is clinical risk. Tasks related to diagnosis, treatment recommendations, and medication decisions require higher factual consistency, evidence traceability, uncertainty expression, and safety constraints.

More complex architectures are not necessarily better for all medical tasks[35, 65]. For basic tasks such as NER, relation extraction, and medical text classification, encoder-only models still provide high efficiency and stable performance. For medical summarization and report generation, encoder-decoder models or decoder-only LLMs can be selected according to data size, output length, and interaction requirements. For open-ended medical question answering,

Table 4. Summary of medical datasets for LLM pre-training and fine-tuning.

Dataset	Type	Size	AI synthesis
MIMIC-III[66]	EHR	Approximately 2 million de-identified notes.	No
MIMIC-IV[67]	EHR	About 300 thousand patients and 430 thousand admissions.	No
CPRD[68]	EHR	Anonymized medical records for over 11.3 million patients.	No
PubMed	Literature	Over 34 million citations and abstracts of biomedical literature, about 4.5 billion words.	No
PMC	Literature	Provides free full-text access to PubMed, about 13.5 billion words.	No
PubMedQA[69]	QA	1 thousand labeled, 612 thousand unlabeled, and 211.3 thousand manually generated QA pairs.	No
MedQA[70]	QA	61,097 multiple-choice QA pairs.	No
MedMCQA[71]	QA	194 thousand multiple-choice QA pairs.	No
MultiMedQA[12]	QA	Includes six existing datasets and one new dataset.	No
CMtMedQA[72]	Dialogue	70 thousand multi-round conversation datasets from real doctor-patient conversations.	Yes
MedDialog[73]	Dialogue	3.4 million Chinese conversations and 0.6 million English conversations.	No
GenMedGPT-5k[74]	Dialogue	5 thousand generated conversations between patients and physicians from ChatGPT.	Yes
MedC-I[75]	Instruction-following data	202 million tokens.	Yes
MedInstruct-52k[76]	Instruction-following data	52 thousand instruction-response pairs generated by GPT-4.	Yes
MedS-Ins[58]	Instruction-following data	5 million instances with 19 thousand instructions across 122 clinical tasks, collected from 58 medically oriented corpora and five text domains.	Partly
UMLS[77]	Knowledge base	2 million entities for 900 thousand concepts.	No
CMeKG[78]	Knowledge base	Chinese medical knowledge graph.	No
COMETA[79]	Web data	Consisting of 20 thousand English biomedical entity mentions.	No

clinical conversations, and medical reasoning, decoder-only LLMs usually have an advantage because of their generation and instruction-following capabilities. However, when a task requires direct understanding of medical images, pathological images, or physiological signals, text-only LLMs are insufficient and medical MLLMs are more appropriate[8, 60, 63]. Thus, medical model architectures should be selected according to specific clinical tasks rather than model size or architectural complexity alone.

The architectural evolution of medical LLMs and MLLMs reflects the broader movement of medical AI from text understanding to text generation and then to cross-modal evidence integration. Encoder-only models, encoder-decoder models, decoder-only LLMs, and medical MLLMs correspond to different levels of medical task requirements, ranging from basic medical NLP to clinical text generation, open-ended interaction, and multimodal medical intelligence. Future medical AI systems may not rely on a single architecture. Instead, they may combine modules according to clinical workflow, for example, using encoder-only models for information extraction, decoder-only LLMs for explanation and interaction, and MLLMs for images or other non-textual evidence. However, architecture alone cannot determine the performance or reliability of a medical intelligent system. Medical data quality, training strategies, domain alignment methods, and evaluation systems are equally important and are discussed in Section 4.

4. Data, training and evaluation strategies

In medical scenarios, whether a model can understand medical semantics, acquire clinical knowledge, follow medical task instructions, and generate reliable outputs depends largely on the quality of training data, domain fine-tuning strategies, and the design of the evaluation system[55, 58, 80]. Data, training, and evaluation therefore form the core technical chain for developing medical LLMs and MLLMs.

Compared with general LLMs, medical LLMs and MLLMs impose stricter requirements on data and evaluation. Medical data are highly specialized, heterogeneous, and privacy-sensitive. They include medical literature, clinical guidelines, EHRs, medical question-answering data, medical images, pathological images, physiological signals, and corresponding

Table 5. Summary of medical datasets for MLLM pre-training and fine-tuning.

Dataset	Type	Size	AI synthesis
PMC-VQA[81]	QA	Contains 149 thousand images and 227 thousand VQA pairs.	Yes
PathVQA[82]	QA	4,998 pathology images with 32,799 QA pairs.	No
Medical-CXR-VQA[83]	QA	780,014 VQA pairs derived from MIMIC-CXR, covering abnormality, location, type, level, view, and presence questions.	Yes
MIMIC-CXR[84]	Image-report	227,835 imaging studies for 65,379 patients.	No
CheXpert[85]	Image-report	224,316 chest X-rays with reports.	No
MEETI[86]	Signal-image-text	Raw electrocardiogram (ECG) waveforms, plotted ECG images, beat-level quantitative parameters, and detailed textual interpretations for more than 800 thousand ECG recordings.	Yes
PathCap[87]	Image-caption	142 thousand high-quality pathology image-caption pairs.	Yes
PMC-OA[88]	Image-caption	1.6 million image-caption pairs.	No
MedTrinity-25 million[89]	Image-caption/ multigranular annotation	Over 25 million medical image-region of interest-description triplets across 10 modalities and more than 65 diseases.	Partly
Derm1M[90]	Image-text	1,029,761 dermatology image-text pairs, more than 390 skin conditions, and 130 clinical concepts aligned with expert ontology knowledge.	No
PathInstruct[87]	Instruction-following data	180 thousand instruction-following data.	Yes
CheXinstruct[91]	Instruction-following data	An instruction-tuning dataset curated from 28 publicly available datasets.	Yes

textual descriptions[8, 80]. At the same time, model outputs may influence disease understanding, patient education, clinical decision-making, and health management[16, 17, 55]. Evaluation therefore cannot rely only on general language-generation metrics. It must also consider factual consistency, medical accuracy, safety, explainability, and clinical usability.

4.1 Medical training datasets

Medical training datasets form the basis for developing medical LLMs and MLLMs. Compared with general text corpora, medical data contain specialized terms, abbreviations, clinical reasoning patterns, structured examination information, and heterogeneous multimodal sources. Different data types play different roles in model training. Medical literature and clinical guidelines mainly strengthen the model's ability to express medical knowledge. EHRs and clinical texts help the model understand real-world clinical scenarios. Medical question answering, clinician-patient dialogues, and instruction data are used mainly for task fine-tuning and instruction following. Medical image-text pairs, pathology images with diagnostic descriptions, and visual question-answering data are essential for training medical MLLMs. Tables 4 and 5 summarize typical datasets used to train medical LLMs and MLLMs and their main purposes, respectively.

Literature resources such as PubMed and PubMed Central[33, 34] contain large collections of biomedical abstracts, full-text papers, and medical research evidence. They provide models with information about disease mechanisms, therapeutic relationships, pharmacological knowledge, and evidence-based medicine. Such data have high medical knowledge density and standardized terminology, so they are often used for domain adaptation of biomedical PLMs and medical LLMs. Medical knowledge bases such as UMLS[77] are also frequently used for medical concept normalization, entity linking, and knowledge enhancement. However, biomedical literature mainly reflects research-oriented and normative medical language, which differs from real-world clinical records. Models trained only on literature corpora may therefore struggle to adapt fully to clinical workflows and clinician-patient communication scenarios[33, 56, 67].

Compared with literature and guidelines, clinical texts[66, 67] and EHRs[66, 67] more directly reflect real clinical practice and are essential for clinical adaptation of medical LLMs. MIMIC-III[66] and MIMIC-IV[67] are widely used datasets in critical care and EHR research. MIMIC-IV[67] includes patient measurements, orders, diagnoses, procedures, treatments, and de-identified clinical text from the real-world clinical system at Beth Israel Deaconess Medical Center. MIMIC-IV[67] covers emergency and intensive care unit patients from 2008 to 2019 and is often used for clinical prediction, case analysis, clinical summarization, and medical NLP. EHR data capture dynamic patient conditions, physicians' diagnostic reasoning, and treatment trajectories, making them more consistent with real-world clinical practice. However, their access and use are restricted by privacy protection, ethical review, de-identification requirements, and cross-institutional sharing limitations.

Medical question answering, clinician-patient dialogues,

and medical instruction data further enhance the task execution and interaction capabilities of LLMs[57, 69–71, 73]. PubMedQA[69] is developed from PubMed abstracts and is used to answer biomedical research questions. It includes 1,000 expert-labeled samples, 61,200 unlabeled samples, and 211,300 algorithm-generated samples. MedQA[70] is derived from medical licensing examination questions and includes English, simplified Chinese, and traditional Chinese questions, making it useful for assessing medical knowledge and examination-style reasoning. MedMCQA[71] contains more than 194,000 medical multiple-choice questions covering 2,400 medical topics and 21 medical disciplines.

Medical dialogue and instruction data are important for improving the interaction capability and task compliance of LLMs. Dialogue data can help models adapt to patient inquiries, conversational expression, multi-turn communication, and health consultation scenarios. Instruction data can organize tasks such as question answering, summarization, explanation, report generation, and patient education into natural-language instruction formats. Compared with simple medical question-answering data, instruction data better reflects practical LLM use because it helps the model understand user intent and generate responses that satisfy task requirements[36, 57, 58]. However, the quality of medical dialogue and instruction data varies substantially[74–76]. Some data may come from internet question-answering platforms or synthetic samples, leading to problems such as insufficient medical accuracy, non-standard terminology, and unclear safety boundaries. Such data therefore require expert review, quality assessment, and safety verification before use.

Medical MLLMs rely on image-text data, multimodal question-answering data, and specialized image datasets for training[63, 83, 89]. MIMIC-CXR[84] is a representative chest X-ray image-report dataset containing 377,110 images and 227,835 imaging studies with radiology reports. CheXpert[85] contains 224,316 chest X-ray images with radiology reports and is commonly used for chest X-ray interpretation and multi-label classification. Beyond radiology, biomedical image-text datasets such as PMC-OA[88] extract images and descriptions from open-access literature and can train medical image-text alignment in vision-language models. These multimodal data are crucial for medical image interpretation, report generation, and medical visual question answering. However, they often suffer from weak image-text alignment, insufficient abnormal-region annotation, report noise, and device-related variation[61, 63, 64].

Governance and quality management of medical training data are essential for model trustworthiness. Data types such as EHRs, medical literature, question-answering datasets, and image-text data should undergo de-identification, noise reduction, duplicate removal, terminology standardization, label verification, modality alignment, and bias evaluation before training[92–94]. Low-quality data may cause models to learn erroneous knowledge. At the same time, medical knowledge and clinical guidelines are continually updated, which requires version control and ongoing data maintenance. The construction of medical training datasets should therefore not be regarded as simple data collection. It is a systematic engineering process involving medical knowledge organization, privacy protection, quality control, and risk management.

4.2 Fine-tuning methodologies for the medical domain

After medical training data are available, the key question is how to adapt general LLMs or MLLMs to medical environments. In medicine, fine-tuning means enabling a model to understand clinical context, follow medical task instructions, generate standardized outputs, and operate safely in high-risk situations[36, 58, 59]. Medical fine-tuning is more demanding than general-domain fine-tuning. Training data are often limited and privacy-sensitive, medical knowledge is specialized and time-sensitive, and model outputs may affect patient understanding, physicians' judgments, and health management[16, 17, 92]. Common medical adaptation methods include domain-continued pre-training, supervised fine-tuning, instruction fine-tuning, parameter-efficient fine-tuning, preference alignment, and multimodal alignment. Table 6 summarizes common fine-tuning and adaptation methods for medical LLMs and MLLMs.

- Domain-specific continue pre-training[33, 34, 36, 56]: In medical model construction, this method is often used as a preparatory stage for domain adaptation. Starting from a general model, continued pre-training uses medical literature, clinical guidelines, electronic health records, imaging reports, and medical textbooks to help the model learn medical terminology, disease knowledge, diagnostic and therapeutic expressions, and clinical writing styles.
- Supervised fine-tuning[38, 50, 64]: By constructing samples for tasks such as medical question answering, medical record summarization, clinical interpretation, patient education, report generation, and literature understanding, supervised fine-tuning teaches the model to generate outputs that match task objectives. Unlike traditional classification models, it often unifies multiple medical tasks into a text-input and text-output format, thereby improving task generalization.
- Instruction fine-tuning[36, 57, 58]: In medical scenarios, user requests often appear not as fixed task labels but as natural-language questions, case descriptions, patient inquiries, or clinicians' instructions. Instruction fine-tuning constructs diverse medical instruction-response data so that the model can recognize task intent, organize response structure, and adjust expression style for different users.
- Parameter efficient fine-tuning[96, 97]: Because LLMs have many parameters, full-parameter fine-tuning is computationally expensive and may cause catastrophic forgetting or deployment difficulties. Parameter-efficient methods such as LoRA, adapter tuning, prefix tuning, and prompt tuning adapt models to medical tasks while preserving base-model capabilities by updating only a small number of parameters or low-rank matrices. These methods are particularly suitable for medical institutions or research teams that need specialized models but have limited data and computing resources.

Table 6. Fine-tuning and adaptation methods for medical LLMs and MLLMs.

Method	Data	Applicable scenario	Strength	Limitation
Domain-specific continue pre-training	Medical literature, clinical guidelines, EHRs, report texts	Domain adaptation of medical LLMs[33, 34, 36, 56]	Improves understanding of medical terminology and clinical context	High training cost; risk of learning outdated or biased knowledge
Supervised fine-tuning	Medical QA datasets, medical record summaries, report generation samples	Question answering, summarization, interpretation, report generation[38, 50, 61, 64]	Clear training objective and stable training process	Relies on high-quality annotations; limited task coverage
Instruction fine-tuning	Medical instruction-response pairs	Multi-task medical LLMs, patient education, clinical interpretation[12, 36, 57, 95]	Enhances interactivity and cross-task generalization	Unstable instruction quality; issues with safety boundaries
Parameter efficient fine-tuning	Small-scale specialty-specific data, task-specific datasets	Local adaptation in medical institutions, specialty-specific models[96, 97]	Low computational cost and flexible deployment	Adaptation performance is limited by base model and data quality
Direct preference optimization and safety alignment	Expert preference data, safety feedback, refusal samples	High-risk medical QA, clinical decision support systems[38, 59, 98, 99]	Reinforces prudence, reliability and responsible boundaries	High cost of expert feedback; difficulty in unifying evaluation criteria

- Direct preference optimization and safety alignment[59, 98, 99]: Medical AI models must not only answer questions but also recognize when questions cannot be answered directly, when ambiguity should be communicated, and when advice must be accompanied by a recommendation to seek professional medical help. Safety alignment should include hallucination control, evidence citation, risk disclaimers, refusal mechanisms, and protection of sensitive health information.

Multimodal fine-tuning and modality alignment are crucial for medical MLLMs. These models must learn to align and fuse medical images, pathology images, physiological time series, and clinical texts to produce semantically consistent medical answers. Typical practices include aligning vision and language with image-report pairs, training cross-modal question-answering ability with medical VQA datasets, and reducing training costs by freezing parts of the visual encoder and LLM backbone while training only projection layers, Q-Former modules, or other alignment components. A major difficulty is that images and text often do not have strict one-to-one correspondence. Reports may mention only important abnormalities and omit normal structures, while lesion regions may lack direct annotations[61, 63, 64]. Fine-tuning medical MLLMs therefore requires not only large-scale multimodal data but also fine-grained grounding, report factuality, and cross-modal evidence traceability.

Overall, the choice of fine-tuning method in medicine should be aligned with model type, data conditions, task objectives, and clinical risk level. For MLLMs, multimodal fine-tuning and modality alignment determine whether the model can truly integrate medical images, clinical texts, and other non-textual evidence. Importantly, fine-tuning alone cannot guarantee clinical usability. Model performance must be validated through systematic evaluation, including automatic metrics, expert assessment, clinical consistency evaluation, and safety testing, as discussed in Section 4-C.

4.3 Evaluation metrics and frameworks

The evaluation of medical LLMs and MLLMs cannot rely only on general natural language processing metrics. Because model outputs may involve disease interpretation, medication information, test result interpretation, patient education, and clinical decision support, evaluation must address task performance, medical correctness, factual consistency, safety, interpretability, and clinical usability[16, 55, 100]. For multiple-choice, classification, and information extraction tasks, metrics such as accuracy, precision, recall, F1-score, and AUC remain useful. For medical summarization, report generation, and open-ended responses, metrics such as BLEU, ROUGE, METEOR, and BERTScore can assess textual similarity. However, automatic metrics cannot fully verify the accuracy, safety, or clinical appropriateness of medical content and should be treated only as preliminary evaluation tools[53, 55, 100].

Medical question-answering benchmarks are important for evaluating the medical knowledge and reasoning capabilities of LLMs. Benchmarks such as MedQA[70], MedMCQA[71], PubMedQA[69], and MultiMedQA[12, 38] are commonly used to test performance on medical licensing examinations, literature-based question answering, and health consultation queries. These benchmarks are reproducible and convenient for cross-model comparison. However, their well-structured questions do not fully capture real clinical complexity, including incomplete information, multimorbidity, and individualized decision-making. Medical question-answering benchmarks are therefore useful for preliminary capability screening but cannot alone validate clinical usability.

Expert evaluation and clinical safety assessment are more critical for evaluating medical models[16, 55, 100]. Open-ended medical question answering, clinical explanation, patient ed-

ucation, and case summarization often require medical experts to judge whether responses are correct, complete, cautious, and potentially harmful. Evaluation frameworks such as QUEST[101] assess medical LLMs across dimensions such as information quality, understanding and reasoning, expression style, safety, and trustworthiness. For high-risk clinical scenarios, evaluation should also consider guideline consistency, emergency identification, medication safety, hallucination control, and uncertainty expression[53, 55, 100]. For medical MLLMs, evaluation should further cover image-text consistency, visual evidence grounding, lesion identification, and multimodal reasoning[61, 63, 64]. Benchmarks such as OmniMedVQA[63] can be used to assess cross-modal understanding in medical vision-language models.

In summary, evaluation of medical LLMs and MLLMs should follow a multi-level framework. Automatic metrics can support basic performance screening, medical question-answering benchmarks can compare knowledge and reasoning capabilities, expert evaluation can assess medical rationality and expression quality, safety assessment can identify clinical risks, and multimodal evaluation can verify cross-modal evidence integration. Recent comprehensive frameworks such as MedHELM further expand evaluation to clinical decision support, medical documentation generation, patient communication, and medical research, offering a more systematic approach to real-world model assessment. Table 6 summarizes common evaluation metrics and representative frameworks for medical LLMs and MLLMs.

5. Advanced applications of LLMs and MLLMs in medicine

With the development of domain adaptation, instruction following, multimodal understanding, and evaluation systems, research on medical LLMs and MLLMs has expanded from basic medical text processing to more complex clinical and healthcare scenarios[6, 13, 102]. Existing work focuses mainly on diagnostic improvement[38, 95, 103], clinical documentation and report generation[53, 61, 64], medical education[14, 16], mental health services[104, 105], and specialized medical tasks[6, 13, 102]. These studies examine how models can support medical knowledge understanding[12, 32, 36], clinical information integration[13, 102], natural-language interaction[14, 74, 95], and multimodal evidence interpretation[8, 63, 64].

5.1 Enhancing medical diagnosis

Diagnostic enhancement[38, 95, 100] is a major entry point for LLMs and MLLMs in clinical practice. Its goal is not to let models diagnose diseases independently, but to use their strengths in information integration, natural-language interaction, and generative reasoning to support information collection, knowledge organization, and candidate-hypothesis generation. Clinical diagnosis[3–5] typically requires clinicians to integrate chief complaints, medical history, physical examination findings, laboratory results, imaging data, and prior treatment responses within limited time. It also requires dynamic judgment under incomplete information. A key advantage of LLMs is their ability to organize dispersed textual information into coherent clinical summaries, problem lists, or differential diagnosis frameworks, thereby providing useful clues for subsequent clinical decision-making.

In diagnostic research, medical question answering and clinical reasoning tasks are widely used to evaluate LLMs' medical knowledge and diagnostic logic[12, 38, 70]. Rather than merely answering isolated factual questions, models are assessed on symptom interpretation, disease-mechanism explanation, test-result interpretation, and summaries of diagnostic and therapeutic principles. These capabilities support diagnostic enhancement because clinicians often need to retrieve and interpret medical knowledge quickly during the diagnostic workflow. However, question-answering proficiency does not imply real-world diagnostic capability[55, 95, 100]. Standardized tasks have clear boundaries, whereas real patient information is often ambiguous, incomplete, and affected by age, comorbidities, medication history, and healthcare context. The value of LLMs in diagnosis should therefore be framed as knowledge assistance and reasoning support rather than replacement of clinical judgment[5, 16, 51].

Studies of diagnostic dialogue also show the potential of LLMs for information collection and collaborative reasoning[95, 103]. Unlike one-time input of complete case data, real clinical consultation requires iterative follow-up questions based on patient responses, so that possible diagnoses can be narrowed or excluded. Existing studies[95, 103] have examined LLMs in simulated consultations, history taking, and diagnosis-related dialogues, focusing on whether models can ask relevant follow-up questions, maintain coherent patient information, and generate reasonable diagnostic hypotheses. These applications are closer to clinical workflows than traditional medical question answering, but they also expose limitations in risk identification, context retention, and responsibility boundaries. In acute conditions, rare diseases, complex comorbidities, and special populations, failure to detect danger signals in time may create serious clinical risks.

MLLMs extend diagnostic enhancement from textual information processing to multimodal evidence interpretation[8, 9, 60]. Many diagnoses depend not only on clinical text but also on medical imaging, pathology slides, fundus photographs, skin images, physiological time series, and laboratory findings. The potential value of medical MLLMs lies in translating non-textual medical evidence into interpretable language outputs and linking such evidence with clinical questions, medical record context, and diagnostic hypotheses. In radiology or pathology, for example, multimodal models can generate explanatory descriptions of imaging findings or answer questions about abnormal regions, lesion characteristics, and clinical implications. However, multimodal diagnostic enhancement requires strong visual grounding and evidential consistency[61, 63, 64]. Model-generated explanations must align with actual medical evidence; otherwise, they may produce plausible but unsupported diagnostic narratives.

From a translational perspective, diagnostic-enhancement systems should be designed as clinician-auditable, traceable, and intervenable auxiliary modules[5, 102, 106]. A practical direction is to use LLMs and MLLMs for case-information condensation, diagnostic hypothesis generation, evidence re-

trieval, guideline interpretation, and multimodal result explanation, while calibrating outputs through clinician feedback. Future research should pay greater attention to real workflow challenges, including how models handle missing information, express uncertainty, avoid overdiagnosis, maintain consistent performance across hospitals and patient populations, and align outputs with guidelines, evidence, and raw data. Only after these issues are systematically validated can diagnostic enhancement move from model-capability demonstrations to reliable clinical support.

5.2 Clinical documentation and report generation

Clinical documentation and report generation are relatively accessible entry points for integrating LLMs into healthcare workflows. These tasks mainly involve structuring, compressing, transcribing, and standardizing information[13, 53, 102]. In current clinical practice, physicians spend substantial time on documentation outside direct patient care, including outpatient records, progress notes, discharge summaries, consultation opinions, referral letters, and diagnostic examination reports[13, 102]. Documentation is essential for continuity of care and quality control, but it also creates a heavy workload for clinicians. LLMs extend clinical text processing beyond information extraction and template filling toward workflow-driven summarization, draft writing, and multi-version text transformation[53, 102, 107].

EHR summarization illustrates the foundational value of LLMs in clinical documentation[53, 107]. Clinical notes are often lengthy, redundant, temporally complex, and filled with abbreviations and non-standard expressions[53, 66, 67]. LLMs can generate summaries at different levels of granularity for different use cases, such as clinician handoffs, patient referrals, plain-language explanations, research curation, and quality control. Unlike the training resources discussed in the data section, the focus here is on how models support the review and reuse of clinical information. To provide practical value, an EHR summarization system must preserve critical clinical facts while condensing text and must clearly distinguish facts, inferences, and care plans in the original records.

Generating clinical notes from clinician-patient dialogues is a central direction in automated clinical documentation[13, 55, 102]. Such systems transform natural speech from outpatient consultations or ward rounds into SOAP notes, visit summaries, or medical record drafts, thereby reducing post-consultation paperwork. This task requires more than dialogue summarization. Models must extract clinically relevant information from colloquial conversations and organize the patient's subjective description, the clinician's assessment, objective examination results, and follow-up plan into text that meets clinical standards. The main difficulty comes from ellipses, repetitions, negations, and implicit meanings in conversational speech[53, 102]. Models may misinterpret a patient's subjective experience as an objective diagnosis or miss important negative symptoms and safety information.

In medical report generation, increasing attention has been placed on the consistency between generated text and underlying evidence[61, 64]. Medical reports are not merely written outputs; they are also important bases for future diagnostic and therapeutic decisions in radiology, pathology, endoscopy, laboratory testing, and other settings. Text-based LLMs can summarize reports, rewrite them, explain them for patients, and translate them across languages[53, 107], whereas MLLMs can generate report drafts or explanatory descriptions directly from images and other non-textual evidence[61, 64]. Unlike general text generation, medical report quality cannot be evaluated by fluency alone. Review should assess whether the report accurately reflects key findings, avoids omitting important abnormalities, uses standardized terminology, and remains consistent with the original imaging or examination evidence.

The boundaries of clinical documentation generation should also be clearly defined. Readable model-generated text does not necessarily imply trustworthy medical content[16, 17, 53]. Common but subtle errors include misaligned timelines, incorrect drug doses, inaccurate summaries of test results, excessive extrapolation from imaging findings, and overly decisive diagnostic language. Because clinical records have legal, quality-of-care, and continuity-of-care implications, LLMs and MLLMs are better positioned as documentation assistants and report-draft generators than as independent producers of finalized records. Future systems should include links to original evidence, structured templates, clinician review, revision tracking, and quality assessment so that generated text supports clinical workflows rather than creating an additional review burden.

5.3 Medical education and mental health services

Compared with traditional textbooks, question banks, and static learning platforms, LLMs can support medical concept explanation, case discussion, exam-question interpretation, clinical reasoning training, and real-time feedback through natural-language interaction. For medical students, resident physicians, and continuing medical education, models can adjust the depth of explanation according to learner questions and generate differential diagnosis ideas, examination rationales, or comparisons of treatment principles based on case facts. The main value of these applications is not to replace teachers or clinical instructors, but to provide a flexible way to organize information and create interactive learning environments.

LLMs used for medical education must still operate within clear limits. Medical learning involves not only knowledge memorization but also clinical reasoning, evidence evaluation, communication skills, and professional responsibility[14, 16, 17]. If model explanations contain factual errors, logical gaps, or oversimplifications, they may mislead learners about disease mechanisms or treatment principles. LLMs are therefore best used as supplementary tools for concept clarification, case discussion, and preliminary feedback, rather than as definitive instructional sources. Such systems should be combined with reviewed textbooks, guidelines, question banks, and teacher feedback, and they should clearly identify knowledge sources and uncertain content.

LLMs also show promise in mental healthcare[104, 105]. Dialogue models have been studied for low-barrier scenarios such as emotional support, mental health education, stress management, prescreening, and follow-up communication. These systems are accessible and can provide initial support when professional resources are limited or when users are reluctant to seek care. However, mental health applications raise serious concerns related to self-harm, crisis intervention, privacy protection, and emotional dependency. If a model fails to detect high-risk signals or provides inappropriate guidance, the consequences can be severe. LLMs for mental health should therefore be treated as complementary support tools and paired with risk identification, crisis referral, human oversight, and professional service linkage.

Although interactive medical communication is useful for both medical education and mental health services, the two areas require different safety boundaries[17, 104, 105]. Medical education emphasizes knowledge accuracy, instructional adaptation, and teacher supervision, whereas mental health services emphasize risk identification, ethical responsibility, and crisis-management capacity. Future studies should further examine model effectiveness across user groups, cultural backgrounds, and risk levels, and should clearly define the working relationships among medical educators, clinical physicians, and mental health professionals.

5.4 Specialized roles

In addition to diagnostic support, document generation, medical education, and mental health, LLMs and MLLMs are increasingly being explored for specialized medical roles[6, 13, 102]. These applications are usually embedded in specific clinical departments, research processes, or medical management procedures, where they support information retrieval, workflow assistance, evidence summarization, risk alerts, or cross-modal interpretation. In surgery, anesthesiology, emergency medicine, and critical care, for example, models can assist with preoperative case summaries, perioperative risk organization, guideline retrieval, critical patient-course summaries, and clinical handovers. The value of these applications lies mainly in compressing information and retrieving knowledge in high-intensity workflows, not in replacing professionals in real-time decision-making.

In pharmacy and medication management, LLMs can support drug-information interpretation, drug-interaction queries, medication-precaution organization, and patient medication education[108, 109]. These tasks may appear to be simple information answering, but they require strict attention to dosage, contraindications, special populations, drug interactions, and guideline updates. A model that cannot access current pharmacological information or detect patient-specific risks may produce dangerous medication suggestions. Medication-related applications should therefore be linked to drug databases, clinical pharmacist review, and hospital information systems rather than relying solely on the model's internal knowledge.

Medical research and literature analysis also provide important specialized applications[23, 32, 36]. LLMs can assist researchers with literature summarization, research topic organization, entity-relation extraction, clinical trial curation, and hypothesis generation. In biomedical research, they may improve information-processing efficiency and support tasks such as drug repositioning, disease-mechanism discovery, and knowledge graph development[32, 36, 46]. However, LLMs can still generate incorrect citations, overgeneralizations, or misinterpretations of research findings. Their outputs are therefore more useful as supporting resources for organizing literature and ideas than as scientific evidence or research conclusions.

A common feature of specialized applications is that task boundaries are clearer, but requirements for professionalism, real-time performance, and reliability are higher. Different departments and workflows require different model capabilities, and general LLMs cannot directly cover all specialized scenarios. A more feasible direction is to combine LLMs, MLLMs, medical knowledge bases, retrieval systems, structured medical records, and expert feedback to build auxiliary systems tailored to specific scenarios. The clinical value of such systems depends not only on model capability, but also on whether they can be integrated into existing workflows, reviewed by professionals, and operated within clear responsibility boundaries.

6. Challenges, ethics, and future directions

Despite their potential in medical question answering, diagnostic improvement, clinical document generation, multimodal interpretation, and patient communication, LLMs and MLLMs face substantial technical, ethical, and clinical governance challenges[16, 17, 51]. Medical scenarios are high-risk and require strong accountability. Factual errors, privacy leaks, group biases, inadequate explanations, or deployment failures may affect patient safety, clinicians' judgments, and health equity. Discussion of medical LLMs and MLLMs should therefore not focus only on model capabilities and application outcomes. It must also systematically examine reliability, data security, fairness, transparency, and clinical implementation risks. This section discusses hallucinations and factual reliability, privacy and data security, fairness and health equity, explainability and clinical transparency, and practical risks in clinical deployment.

6.1 Hallucination and factual reliability

Hallucination and factual reliability are core risks in medical LLM and MLLM applications[53, 54, 110]. Medical hallucinations can appear as fabricated disease mechanisms, inaccurate medication recommendations, unsupported diagnostic conclusions, invented clinical guidelines, or misleading interpretation of images and reports. In general open-domain tasks, such errors may affect information quality. In medicine, however, they can directly influence patients' understanding, clinicians' judgments, and treatment safety. Therefore, hallucination control is not only a technical issue in text generation but also a core requirement for medical safety.

The formation of medical hallucinations is related to training data, generative mechanisms, and lack of clinical context[23, 54, 110]. If the training data contain outdated guidelines, inconsistent terminology, low-quality internet health

information, or incomplete clinical evidence, the model may learn unreliable associations. The autoregressive generation mechanism may also produce fluent but unsupported answers when evidence is insufficient. In addition, many medical questions require patient-specific information, including comorbidities, medication history, laboratory indicators, allergies, age, sex, and pregnancy status. Without these details, a model may still generate confident but inappropriate advice. For MLLMs, hallucinations are also related to cross-modal limitations[61, 63, 64]. Weak image-text alignment, insufficient lesion localization, or inaccurate interpretation of visual evidence may cause the model to generate explanations that conflict with the image, miss critical lesions, or misapply general medical knowledge to a specific case.

Mitigating medical hallucinations requires a combination of data governance, evidence augmentation, model constraints, and human evaluation. Model knowledge can be improved through high-quality medical corpora, expert-reviewed instruction data, clinical guidelines, and structured knowledge bases. Retrieval-augmented generation, evidence citation, and access to external databases can improve response traceability and reduce the risk of overly definitive answers without sufficient evidence. Xu et al. [110] introduced the MEGA-RAG framework to mitigate hallucinations in public health question answering by combining dense retrieval, BM25 keyword search, biomedical knowledge graphs, reranking, and discrepancy-aware refinement. Asgari et al. [53] proposed a framework for assessing clinical safety and hallucination rates in medical text summarization, including error classification, output comparison, clinical safety assessment, and the CREOLA graphical interface. These studies suggest that hallucination management should include not only model optimization but also error discovery, risk classification, and clinical expert review. Outputs related to diagnosis, therapy, and medication should therefore be treated as supplementary information rather than definitive medical conclusions.

6.2 Privacy and data security

Privacy and data security are critical to the application of medical LLMs and MLLMs in clinical settings. Unlike general-domain data, medical data contain explicit personal identifiers and highly sensitive information about diseases, medications, genetic characteristics, family history, mental health, and lifestyle. Even after de-identification, re-identification may remain possible through dates, rare diseases, specific test findings, or correlations across multiple data sources. Privacy concerns for medical LLMs and MLLMs should therefore be analyzed throughout the entire data lifecycle, including data collection, annotation, pre-training, fine-tuning, inference, log storage, and third-party platform access[93].

Privacy risks in LLMs are closely tied to both training and generation. Large-scale models may memorize sensitive content from training data and may reveal personal information when prompted or attacked in specific ways. Current research indicates that language models can be vulnerable to data-extraction attacks on the training set, increasing the risk associated with fine-tuning on real clinical information[111]. When medical institutions use external closed-source models, patient data may also pass through cloud interfaces, logging systems, or third-party service providers, which creates concerns about cross-domain data flows and ambiguous responsibility boundaries. For MLLMs, privacy protection must extend beyond text de-identification because medical images, pathology slides, and physiological signals may contain implicit identifiers.

Recent research has begun to examine technical and system-level approaches to privacy protection. Jonnagaddala and Wong [92] studied privacy protection for electronic health records in the LLM era and showed that local deployment, synthetic data generation, differential privacy, and de-identification can reduce exposure risks for sensitive health information. Wiest et al. [107] presented a locally deployed LLM-based method for extracting clinical free-text information, showing that structured information extraction can be performed without relying on external commercial interfaces. These studies indicate that privacy protection does not necessarily require sacrificing model capability, but the appropriate technical approach should be chosen according to task risk, data sensitivity, and deployment setting. However, de-identification, synthetic data, federated learning, and differential privacy are not free; they may introduce information loss, performance degradation, higher technical complexity, or increased costs for cross-institutional collaboration.

Data-security governance for medical LLMs and MLLMs should therefore go beyond ad hoc technical safeguards. It should establish an integrated framework that combines system design, policy constraints, institutional responsibility, and continuous monitoring. Medical data used in training, fine-tuning, retrieval, and post-deployment feedback should be traceable and auditable. The scope of model access to patient information should be clearly restricted. For cross-institutional collaboration, privacy-preserving computation and standardized governance processes are needed to balance model generalization with patient data protection[112]. Without this foundation, the deployment of medical LLMs and MLLMs may create new risks in addition to improving medical intelligence.

6.3 Fairness, bias and health equity

Fairness, bias, and health equity are key ethical issues for medical LLMs and MLLMs. Beyond performance differences across groups, bias in medical models can affect whether patients receive timely consultation, accurate health explanations, appropriate recommendations, and equitable access to medical resources. The language groups most visible in clinical records, medical guidelines, and online health information tend to be affluent, mainstream populations and common disease groups. Model training can therefore devalue or misrepresent low-resource regions, minority groups, older adults, children, patients with rare diseases, people with disabilities, and non-English speakers[94, 113]. Equity issues in medical LLMs and MLLMs should therefore be addressed not only as algorithmic performance gaps but also in relation to health disparities, social determinants, and unequal access to medical resources.

The formation of medical model bias has multiple sources:

1) At the data level: EHR, medical images, clinical trials, and medical literature themselves may have insufficient population representation, annotation differences, and institutional biases;
2) At the model level: LLMs may learn surface correlations between disease risks, treatment compliance, economic status, education level, or social identity, and generate explanations or recommendations with stereotypical assumptions when there is a lack of sufficient medical basis;
3) At the deployment level: Even if the model performs well overall, it may show performance degradation in specific subgroups, medical institutions, or language and cultural environments.

Existing work has begun to develop specialized evaluation tools for this problem. For example, Pfohl et al. [114] developed a toolkit for identifying health-equity harms and bias in medical LLM responses. The toolkit includes a multifactor human evaluation framework, the adversarial medical question-answering dataset EquityMedQA, and an evaluation process involving reviewers from multiple backgrounds. Ji et al. [115] proposed EquityGuard, which evaluates health-inequity risks in LLM medical applications from the perspective of both detection and mitigation. These efforts show that fairness evaluation should move beyond overall accuracy and focus on risks affecting specific groups, tasks, and usage scenarios.

To reduce bias in medical LLMs and MLLMs, mitigation should cover the entire pipeline, from data collection to model training, evaluation, validation, and clinical deployment. During training and evaluation, data sources, disease types, language and cultural backgrounds, and demographic characteristics should be more representative. Subgroup performance analysis, counterfactual testing, and fairness sensitivity assessment should also be performed. During model development, bias detection, expert feedback, fairness constraints, and human preference alignment can be combined to reduce unsupported inferences about social identities or resource conditions. During deployment, differential model performance should be continuously monitored in real populations to prevent systematic disadvantages in low-resource environments or vulnerable groups. Health equity does not mean giving identical answers to all patients. Rather, it requires models to respect individual differences and social context while avoiding the misrepresentation of structural inequalities as individual responsibility or medical inevitability. Fairness governance should therefore be treated as a fundamental condition for trustworthy medical LLM and MLLM deployment, not as an optional ethical supplement.

A promising future direction is to formulate fairness governance for medical LLMs and MLLMs as a constrained submodular optimization problem over training samples, instruction-following data, evaluation cases, retrieval evidence, and expert-feedback budgets. In this formulation, the objective can capture clinical utility, disease coverage, institutional diversity, linguistic diversity, and representation of rare or underserved conditions, while the constraints enforce minimum or balanced coverage for demographic groups, low-resource hospitals, language communities, and intersectional populations. Recent work on fair and streaming submodular maximization, fair data summarization, and fair k-submodular maximization suggests that fairness constraints can be incorporated into subset selection with explicit utility-fairness trade-offs, rather than being treated only as post-hoc subgroup reporting[116–118]. For medical LLMs and MLLMs, this perspective could support fairness-aware construction of pre-training and fine-tuning corpora, subgroup-balanced benchmark design, and deployment-time allocation of retrieval or clinician-review resources to populations for whom model errors may cause the greatest harm. Combined with health-equity evaluation tools such as EquityMedQA and EquityGuard, submodular fairness methods may help move fairness research from descriptive bias detection toward optimization-based, auditable, and continuously monitored equity control[94, 114, 115].

6.4 Explainability and clinical transparency

Interpretability and clinical transparency are essential for moving medical LLMs and MLLMs from experimental performance to reliable use. Unlike traditional structured prediction models, LLMs often generate natural-language outputs, which may lead users to assume that they are receiving sufficient explanations. However, fluency does not guarantee sound reasoning or traceable evidence. In medical applications, interpretability requires a model not only to present findings but also to explain its reasoning, conditions of applicability, and limits of uncertainty. For high-stakes tasks such as diagnostic interpretation, medication recommendation, image analysis, and patient education, clinicians must judge whether the output is consistent with patient history, examination findings, guideline evidence, and clinical reasoning[119].

For MLLMs, clinical transparency also includes traceability of cross-modal evidence. A model that generates explanations from medical images, pathology images, or physiological signals should identify the relevant region or signal segment, specify the source of evidence, and clarify the relationship between visual or signal information and the generated text. It should avoid producing plausible but unsupported conclusions. Current methods, including evidence citation, visual grounding, attention visualization, retrieval-enhanced generation, and human review, can partially improve the auditability of model outputs[119, 120]. However, interpretability alone does not guarantee accuracy, and overly simple explanations may create new misunderstandings. The focus of transparency for medical LLMs and MLLMs should therefore be evidence traceability, output auditability, and human reviewability rather than more persuasive explanatory prose.

A future research direction is to shift clinical transparency from generating fluent explanations to selecting compact, verifiable, and clinically meaningful evidence sets. Instead of providing only a free-form rationale, a medical LLM or MLLM could return a small but diverse collection of supporting evidence, such as guideline passages, retrieved biomedical articles, similar patient cases, laboratory trends, image regions, signal segments, and counterfactual alternatives. This idea is closely related to Submodular Pick and submodular summa-

rization, where representative yet non-redundant examples are selected to explain model behavior or summarize complex information[121–123]. Extending this principle to medical LLMs and MLLMs would support both local and global audit trails. Each selected evidence item would have an interpretable marginal contribution, while cost-constrained, evolutionary, or weakly submodular algorithms could balance factual support, diversity, clinician review time, privacy constraints, and computational cost[124, 125]. For multimodal medicine, such evidence portfolios should be grounded at multiple levels, including textual citations, image regions, temporal signal segments, and patient-history elements, so that clinicians can verify not only the generated answer but also the evidence pathway supporting it. This would make explainability less dependent on persuasive language and more dependent on inspectable, minimal, and clinically relevant evidence aligned with clinical explainable AI criteria such as understandability, relevance, truthfulness, plausibility, and efficiency[126].

6.5 Clinical deployment and practical risks

The clinical utility of medical LLMs and MLLMs is a complex issue. Although models may perform well in benchmark evaluations or retrospective studies, real-world application is constrained by workflow integration, system stability, response time, cost management, liability attribution, and regulatory compliance. Medical models are not standalone question-answering systems. They must be integrated into electronic health records, image archiving systems, clinical decision support systems, or clinician-patient communication workflows. If model outputs cannot be embedded into existing clinical workflows, or if they increase clinicians' review burden, their practical value will be limited[13, 102].

The continuous evolution of medical knowledge also creates maintenance obligations for clinical applications. Disease guidelines, drug indications, adverse-reaction data, and diagnostic and therapeutic standards are constantly updated. Models based only on static training data are vulnerable to knowledge obsolescence. After deployment, version management, performance monitoring, error feedback, risk classification, and real-world validation are necessary to ensure that experimental performance gains are not lost in clinical practice[102, 106]. The deployment of medical LLMs and MLLMs should therefore be viewed not as simple model integration into hospital systems, but as a continuous governance process. The aim is not to replace physicians, but to provide manageable support for organizing clinical information, developing interpretations, and assisting decision-making while clearly defining responsibility boundaries and establishing human review procedures.

7. Conclusions

The future development of medical LLMs and MLLMs first requires improving knowledge reliability and factual traceability. Because medical knowledge is continuously updated, static parameterized knowledge cannot fully meet clinical requirements for accuracy and timeliness. Retrieval-augmented generation, access to medical knowledge bases, citation of clinical guidelines, and evidence-traceability mechanisms are expected to reduce hallucination and outdated-knowledge risks. These mechanisms can further shift model responses from generating plausible text toward generating evidence-based medical explanations.

Multimodal fusion remains a key direction for moving medical models toward real clinical scenarios. Clinical decision-making often relies on multi-source information, including medical images, pathology images, physiological signals, laboratory test results, electronic health records, and patients' natural-language descriptions. Future medical MLLMs should improve fine-grained alignment across modalities, especially the correspondence among image regions, lesion features, clinical texts, and medical terminology. Evaluation systems should also expand beyond single-question-answering accuracy to include image-text consistency, visual evidence localization, report factuality, and clinical interpretability.

Privacy-preserving training and cross-institutional collaboration are important foundations for sustainable medical large-model development. Medical data are highly sensitive and constrained by ethical review, data security, and regulatory compliance, so traditional centralized training is difficult to apply directly in real medical environments. Future research should further develop federated learning, differential privacy, localized deployment, secure multi-party computation, and synthetic data to improve model generalization and cross-institutional adaptability while protecting patient privacy.

Human-machine collaboration and general medical AI are two important directions for deepening the application of medical LLMs. At the current stage, medical LLMs and MLLMs are better suited to information organization, evidence retrieval, text generation, interpretation assistance, and decision support than to independent clinical decision-making. Future systems should strengthen physicians' roles in review, feedback, and intervention during model use, and should gradually explore general medical intelligent systems that adapt to multi-disease, multimodal, multi-scenario, and multi-role tasks. However, general medical artificial intelligence should not be understood as unrestricted model autonomy. It should be built on factual reliability, interpretability, safety, controllability, and clearly defined responsibility boundaries.

In conclusion, LLMs and MLLMs are driving the evolution of medical artificial intelligence from traditional task-specific models toward generative, interactive, and multimodal intelligent systems. This paper provides a systematic review of technological evolution, model architectures, data training, evaluation methods, medical applications, and ethical challenges. It shows that medical large models have substantial potential in diagnosis augmentation, clinical document generation, medical education, patient communication, and specialized tasks. However, their practical value depends not only on model scale or generation capability, but also on factual reliability, data security, fairness, interpretability, and adaptation to real clinical workflows. Future research should move beyond the pursuit of benchmark performance and focus on building clinically trustworthy systems. High-quality medical data, domain knowledge enhancement, multimodal alignment, privacy protection, expert feedback, and continuous

evaluation should be integrated into a unified framework. The goal of medical LLMs and MLLMs should not be to replace physicians, but to provide an important technical foundation for intelligent medical systems while maintaining clear safety constraints and responsibility boundaries.

Acknowledgements

We acknowledge the funding support from the Hubei Provincial International Science and Technology Cooperation Project (No. 2025EHA011).

Conflicts of interest

The authors declare no competing interest.

References

[1] Vollset, S. E., Ababneh, H. S., Abate, Y. H., et al. Burden of disease scenarios for 204 countries and territories, 2022-2050: A forecasting analysis for the global burden of disease study 2021. The Lancet, 2024, 403(10440): 2204-2256.

[2] Boniol, M., Kunjumen, T., Nair, T. S., et al. The global health workforce stock and distribution in 2020 and 2030: A threat to equity and 'universal' health coverage? BMJ Global Health, 2022, 7(6): e009316.

[3] Huesmann, L., Sudacka, M., Durning, S. J., et al. Clinical reasoning: What do nurses, physicians, and students reason about. Journal of Interprofessional Care, 2023, 37(6): 990-998.

[4] Schuler, K., Jung, I. C., Zerlik, M., et al. Context factors in clinical decision-making: A scoping review. BMC Medical Informatics and Decision Making, 2025, 25(1): 133.

[5] Sokol, K., Fackler, J., Vogt, J. E. Artificial intelligence should genuinely support clinical reasoning and decision making to bridge the translational gap. npj Digital Medicine, 2025, 8(1): 345.

[6] Fahim, Y. A., Hasani, I. W., Kabba, S., et al. Artificial intelligence in healthcare and medicine: Clinical applications, therapeutic advances, and future perspectives. European Journal of Medical Research, 2025, 30(1): 848.

[7] Narayan, S. M., Chung, M. K., Adedinsewo, D., et al. Access to digital health technologies: Personalized framework and global perspectives. Nature Reviews Cardiology, 2026, 23(1): 9-22.

[8] AlSaad, R., Abd-Alrazaq, A., Boughorbel, S., et al. Multimodal large language models in health care: Applications, challenges, and future outlook. Journal of Medical Internet Research, 2024, 26: e59505.

[9] Jandoubi, B., Akhloufi, M. A. Multimodal artificial intelligence in medical diagnostics. Information, 2025, 16(7): 591.

[10] Buess, L., Keicher, M., Navab, N., et al. From large language models to multimodal AI: A scoping review on the potential of generative AI in medicine. Biomedical Engineering Letters, 2025, 15(5): 845-863.

[11] Chen, S. F., Alyakin, A., Seas, A., et al. LLM-assisted systematic review of large language models in clinical medicine. Nature Medicine, 2026, 32(3): 1152-1159.

[12] Singhal, K., Azizi, S., Tu, T., et al. Large language models encode clinical knowledge. Nature, 2023, 620(7972): 172-180.

[13] Busch, F., Hoffmann, L., Rueger, C., et al. Current applications and challenges in large language models for patient care: A systematic review. Communications Medicine, 2025, 5(1): 26.

[14] Kung, T. H., Cheatham, M., Medenilla, A., et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2023, 2(2): e0000198.

[15] Kyu-Hwan, J. Large language models in medicine: Clinical applications, technical challenges, and ethical considerations. Healthcare Informatics Research, 2025, 31(2): 114-124.

[16] Bedi, S., Liu, Y., Orr-Ewing, L., et al. Testing and evaluation of health care applications of large language models: A systematic review. The Journal of the American Medical Association, 2025, 333(4): 319-328.

[17] Fareed, M., Fatima, M., Uddin, J., et al. A systematic review of ethical considerations of large language models in healthcare and medicine. Frontiers in Digital Health, 2025, 7: 1653631.

[18] Pantanowitz, L., Pearce, T., Abukhiran, I., et al. Non-generative artificial intelligence in medicine: Advancements and applications in supervised and unsupervised machine learning. Modern Pathology, 2025, 38(3): 100680.

[19] Shaikh, M. R., Jeyabose, A., Arjunan, R. V. Deep learning for Alzheimer's disease: Advances in classification, segmentation, subtyping, and explainability. BioMedical Engineering OnLine, 2025, 24(1): 150.

[20] Cao, L., Wu, C., Luo, G., et al. Online biomedical named entities recognition by data and knowledge-driven model. Artificial Intelligence in Medicine, 2024, 150: 102813.

[21] Huang, M., Han, J., Lin, P., et al. Surveying biomedical relation extraction: A critical examination of current datasets and the proposal of a new resource. Briefings in Bioinformatics, 2024, 25(3): bbae132.

[22] Prinzi, F., Currieri, T., Gaglio, S., et al. Shallow and deep learning classifiers in medical image analysis. European Radiology Experimental, 2024, 8(1): 26.

[23] Sahoo, S. S., Plasek, J. M., Xu, H., et al. Large language models for biomedicine: Foundations, opportunities, challenges, and best practices. Journal of the American Medical Informatics Association, 2024, 31(9): 2114-2124.

[24] De Santis, E., Martino, A., Ronci, F., et al. From bag-of-words to transformers: A comparative study for text classification in healthcare discussions in social media. IEEE Transactions on Emerging Topics in Computational Intelligence, 2025, 9(1): 1063-1077.

[25] Liang, Z., Zhao, Y., Xu, H., et al. A hybrid model integrating RoBERTa, TF-IDF, and attention mechanism for medical query intent classification. Scientific Reports, 2025, 15(1): 42847.

[26] Wang, C., Pan, D., Kumari, S., et al. Large language models driven health text information analysis in consumer electronics. IEEE Transactions on Consumer Electronics, 2025, 71(2): 3510-3521.

[27] Shortliffe, E. H., Davis, R., Axline, S. G., et al. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the mycin system. Computers and Biomedical Research, 1975, 8(4): 303-320.

[28] Kononenko, I. Machine learning for medical diagnosis: History, state of the art and perspective. Artificial Intelligence in Medicine, 2001, 23(1): 89-109.

[29] Lu, C., Zhang, J., Liu, R. Deep learning-based image classification for integrating pathology and radiology in AI-assisted medical imaging. Scientific Reports, 2025, 15(1): 27029.

[30] Kumar, M., Shubham, Saini, L. M., et al. Transformers for enhanced biomedical text classification. Paper Presented at the 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Tirupur, India, 3-5 September, 2025.

[31] Ahmed, S. F., Alam, M. S. B., Hassan, M., et al. Deep learning modelling techniques: Current progress, applications, advantages, and challenges. Artificial Intelligence Review, 2023, 56(11): 13521-13617.

[32] Zhou, J., Li, H., Chen, S., et al. Large language models in biomedicine and healthcare. npj Artificial Intelligence, 2025, 1(1): 44.

[33] Gu, Y., Tinn, R., Cheng, H., et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 2021, 3(1): 2.

[34] Lee, J., Yoon, W., Kim, S., et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 2020, 36(4): 1234-1240.

[35] Fraile Navarro, D., Ijaz, K., Rezazadegan, D., et al. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. International Journal of Medical Informatics, 2023, 177: 105122.

[36] Xie, Q., Chen, Q., Chen, A., et al. Medical foundation large language models for comprehensive text analysis and beyond. npj Digital Medicine, 2025, 8(1): 141.

[37] Wei, J., Wang, X., Schuurmans, D., et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[38] Singhal, K., Tu, T., Gottweis, J., et al. Toward expert-level medical question answering with large language models. Nature Medicine, 2025, 31(3): 943-950.

[39] Tu, T., Azizi, S., Driess, D., et al. Towards generalist biomedical AI. The New England Journal of Medicine Artificial Intelligence, 2024, 1(3): AIoa2300138.

[40] Team, G., Georgiev, P., Lei, V. I., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. ArXiv Preprint ArXiv: 2403.05530, 2024.

[41] Saab, K., Tu, T., Weng, W., et al. Capabilities of Gemini models in medicine. ArXiv Preprint ArXiv: 2404.18416, 2024.

[42] Wang, X., Wei, J., Schuurmans, D., et al. Self-consistency improves chain of thought reasoning in language models. ArXiv Preprint ArXiv: 2203.11171, 2022.

[43] Lewis, P., Perez, E., Piktus, A., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[44] Jaech, A., Kalai, A., Lerer, A., et al. OpenAI o1 system card. ArXiv Preprint ArXiv: 2412.16720, 2024.

[45] Guo, D., Yang, D., Zhang, H., et al. Deepseek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature, 2025, 645(8081): 633-638.

[46] Luo, R., Sun, L., Xia, Y., et al. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 2022, 23(6): bbac409.

[47] Raffel, C., Shazeer, N., Roberts, A., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 2020, 21(1): 140.

[48] Lewis, M., Liu, Y., Goyal, N., et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Paper Presented at the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 5-10 July, 2020.

[49] Phan, L., Anibal, J. T., Tran, H. T., et al. SciFive: A text-to-text transformer model for biomedical literature. ArXiv Preprint ArXiv: 2106.03598, 2021.

[50] Yuan, H., Yuan, Z., Gan, R., et al. BioBART: Pretraining and evaluation of a biomedical generative language model. Paper Presented at the 21st Workshop on Biomedical Language Processing, Dublin, Ireland, 26 May, 2022.

[51] WHO. Ethics and Governance of Artificial Intelligence for Health. World Health Organization, Geneval, Switzerland, 2021.

[52] Welch, M. L., Grant, B., Deutschman, C., et al. A practical framework for operationalising responsible and equitable artificial intelligence in health care: Tackling bias, inequity, and implementation challenges. The Lancet Digital Health, 2026, 8(3): e100957.

[53] Asgari, E., Montaña-Brown, N., Dubois, M., et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digital Medicine, 2025, 8(1): 274.

[54] Farquhar, S., Kossen, J., Kuhn, L., et al. Detecting hallucinations in large language models using semantic entropy. Nature, 2024, 630(8017): 625-630.

[55] Bedi, S., Cui, H., Fuentes, M., et al. Holistic evaluation of large language models for medical tasks with MedHELM. Nature Medicine, 2026, 32(3): 943-951.

[56] Christophe, C., Raha, T., Maslenkova, S., et al. Beyond fine-tuning: Unleashing the potential of continuous pre-

training for clinical LLMs. Paper Presented at Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, USA, 12-16 November, 2024.

[57] Tran, H., Yang, Z., Yao, Z., et al. Bioinstruct: Instruction tuning of large language models for biomedical natural language processing. Journal of the American Medical Informatics Association, 2024, 31(9): 1821-1832.

[58] Wu, C., Qiu, P., Liu, J., et al. Towards evaluating and building versatile large language models for medicine. npj Digital Medicine, 2025, 8(1): 58.

[59] Giuffrè, M., You, K., Pang, Z., et al. Expert of experts verification and alignment (EVAL) framework for large language models safety in gastroenterology. npj Digital Medicine, 2025, 8(1): 242.

[60] Wu, C., Zhang, X., Zhang, Y., et al. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. Nature Communications, 2025, 16(1): 7866.

[61] Li, C., Chang, K., Yang, C., et al. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. Nature Communications, 2025, 16(1): 2258.

[62] Liu, C., Ouyang, C., Wan, Z., et al. Knowledge-enhanced multimodal ECG representation learning with arbitrary-lead inputs. Paper Presented at the Association for Computational Linguistics, Vienna, Austria, 27 July-1 August, 2025.

[63] Hu, Y., Li, T., Lu, Q., et al. OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. Paper Presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, United States, 16-22 June, 2024.

[64] Tanno, R., Barrett, D. G. T., Sellergren, A., et al. Collaboration between clinicians and vision-language models in radiology report generation. Nature Medicine, 2025, 31: 599-608.

[65] Hu, Y., Chen, Q., Du, J., et al. Improving large language models for clinical named entity recognition via prompt engineering. Journal of the American Medical Informatics Association, 2024, 31(9): 1812-1820.

[66] Johnson, A. E., Pollard, T. J., Shen, L., et al. MIMIC-III, a freely accessible critical care database. Scientific Data, 2016, 3(1): 160035.

[67] Johnson, A. E. W., Bulgarelli, L., Shen, L., et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 2023, 10(1): 1.

[68] Herrett, E., Gallagher, A. M., Bhaskaran, K., et al. Data resource profile: Clinical practice research datalink (CPRD). International Journal of Epidemiology, 2015, 44(3): 827-836.

[69] Jin, Q., Dhingra, B., Liu, Z., et al. PubMedQA: A dataset for biomedical research question answering. Paper Presented at the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3-7 November, 2019.

[70] Jin, D., Pan, E., Oufattole, N., et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences, 2021, 11(14): 6421.

[71] Pal, A., Umapathi, L. K., Sankarasubbu, M. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. Paper Presented at the Conference on Health, Inference, and Learning, Virtual, 7-8 April, 2022.

[72] Yang, S., Zhao, H., Zhu, S., et al. Zhongjing: Enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. Paper Presented at the 38th Annual AAAI Conference on Artificial Intelligence, Vancouver, Canada, 20-27 February, 2024.

[73] Zeng, G., Yang, W., Ju, Z., et al. Meddialog: Large-scale medical dialogue datasets. Paper Presented at the Conference on Empirical Methods in Natural Language Processing, Virtual, 16-20 November, 2020.

[74] Li, Y., Li, Z., Zhang, K., et al. ChatDoctor: A medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. Cureus, 2023, 15(6): e40895.

[75] Wu, C., Lin, W., Zhang, X., et al. PMC-LLaMA: Toward building open-source language models for medicine. Journal of the American Medical Informatics Association, 2024, 31(9): 1833-1843.

[76] Zhang, X., Tian, C., Yang, X., et al. Alpacare: Instruction-tuned large language models for medical application. ArXiv Preprint ArXiv: 2310.14558, 2023.

[77] Bodenreider, O. The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Research, 2004, 32(suppl_1): D267-D270.

[78] Ao, D., Yang, Y., Sui, Z., et al. Preliminary study on the construction of Chinese medical knowledge graph. Journal of Chinese Information Processing, 2019, 33(10): 9. (in Chinese)

[79] Basaldella, M., Liu, F., Shareghi, E., et al. COMETA: A corpus for medical entity linking in the social media. Paper Presented at the Conference on Empirical Methods in Natural Language Processing, Virtual, 16-20 November, 2020.

[80] Xiao, H., Zhou, F., Liu, X., et al. A comprehensive survey of large language models and multimodal large language models in medicine. Information Fusion, 2025, 117: 102888.

[81] Zhang, X., Wu, C., Zhao, Z., et al. PMC-VQA: Visual instruction tuning for medical visual question answering. ArXiv Preprint ArXiv: 2305.10415, 2023.

[82] He, X., Zhang, Y., Mou, L., et al. PathVQA: 30000+ questions for medical visual question answering. ArXiv Preprint ArXiv: 2003.10286, 2020.

[83] Hu, X., Gu, L., Kobayashi, K., et al. Medical-CXR-VQA dataset: A large-scale LLM-enhanced medical dataset for visual question answering on chest X-ray images, 2025.

[84] Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 2019, 6(1): 317.

[85] Irvin, J., Rajpurkar, P., Ko, M., et al. CheXpert: A

large chest radiograph dataset with uncertainty labels and expert comparison. Paper Presented at the AAAI Conference on Artificial Intelligence, Hawaii, United States, 27 January-1 February, 2019.

[86] Zhang, D., Lan, X., Geng, S., et al. MEETI: A multimodal ECG dataset from MIMIC-IV-ECG with signals, images, features and interpretations. Scientific Data, 2026, 13(1): 527.

[87] Sun, Y., Zhu, C., Zheng, S., et al. PathAsst: A generative foundation AI assistant towards artificial general intelligence of pathology. Paper Presented at the 38th Annual AAAI Conference on Artificial Intelligence, Vancouver, Canada, 20-27 February, 2024.

[88] Lin, W., Zhao, Z., Zhang, X., et al. PMC-CLIP: Contrastive language-image pre-training using biomedical documents. Paper Presented at International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, Canada, 8-12 October, 2023.

[89] Xie, Y., Zhou, C., Gao, L., et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. Paper Presented at the International Conference on Learning Representations (ICLR), Singapore, 24-28 April, 2025.

[90] Yan, S., Hu, M., Jiang, Y., et al. Derm1m: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. Paper Presented at the IEEE/CVF International Conference on Computer Vision, Hawaii, United States, 19-23 October, 2025.

[91] Chen, Z., Varma, M., Delbrouck, J.-B., et al. Chexagent: Towards a foundation model for chest X-Ray interpretation. Paper Presented at the Association for the Advancement of Artificial Intelligence's 2024 Spring Symposium Series, California, United States, 25-27 March, 2024.

[92] Jonnagaddala, J., Wong, Z. S.-Y. Privacy preserving strategies for electronic health records in the era of large language models. npj Digital Medicine, 2025, 8(1): 34.

[93] Zhong, X., Li, S., Chen, Z., et al. Considerations for patient privacy of large language models in health care: Scoping review. Journal of Medical Internet Research, 2025, 27: e76571.

[94] Chen, R. J., Wang, J. J., Williamson, D. F. K., et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature Biomedical Engineering, 2023, 7(6): 719-742.

[95] McDuff, D., Schaekermann, M., Tu, T., et al. Towards accurate differential diagnosis with large language models. Nature, 2025, 642(8067): 451-457.

[96] Gema, A., Minervini, P., Daines, L., et al. Parameter-efficient fine-tuning of LLaMA for the clinical domain. Paper Presented at the 6th Clinical Natural Language Processing Workshop, Mexico City, Mexico, 21 June, 2024.

[97] Shi, W., Xu, R., Zhuang, Y., et al. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. Paper Presented at Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, USA, 12-16 November, 2024.

[98] Rafailov, R., Sharma, A., Mitchell, E., et al. Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 2023, 36: 53728-53741.

[99] Kim, D., Lee, J., Yun, J., et al. Benchmarking direct preference optimization for medical large vision-language models. ArXiv Preprint Arxiv: 2601.17918, 2026.

[100] Zhang, M., Shen, Y., Li, Z., et al. LLMEval-Med: A real-world clinical benchmark for medical LLMs with physician validation. ArXiv Preprint ArXiv: 2506.04078, 2025.

[101] Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., et al. A framework for human evaluation of large language models in healthcare derived from literature review. npj Digital Medicine, 2024, 7(1): 258.

[102] Artsi, Y., Sorin, V., Glicksberg, B. S., et al. Large language models in real-world clinical workflows: A systematic review of applications and implementation. Frontiers in Digital Health, 2025, 7: 1659134.

[103] Gong, L., Fang, W., Yang, T., et al. Meddialogrubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models. ArXiv Preprint ArXiv: 2601.03023, 2026.

[104] Guo, Z., Lai, A., Thygesen, J. H., et al. Large language models for mental health applications: Systematic review. JMIR Mental Health, 2024, 11(1): e57400.

[105] Li, L., Huang, X., Ma, R., et al. LLM use for mental health: Crowdsourcing users' sentiment-based perspectives and values from social discussions. Paper Presented at the Web Conference, Dubai, The United Arab Emirates, 13-17 April, 2026.

[106] Vasey, B., Nagendran, M., Campbell, B., et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ-British Medical Journal, 2022, 377: e070904.

[107] Wiest, I. C., Ferber, D., Zhu, J., et al. Privacy-preserving large language models for structured medical information retrieval. npj Digital Medicine, 2024, 7(1): 257.

[108] Zhao, J., Xu, L., Tan, M., et al. Rxsafebench: Identifying medication safety issues of large language models in simulated consultation. ArXiv Preprint ArXiv: 2511.04328, 2025.

[109] Kazemzadeh, H., Dizaji, K. M., Tavakoli, S. R., et al. Drugrag: Enhancing pharmacy LLM performance through a novel retrieval-augmented generation pipeline. ArXiv Preprint ArXiv: 2512.14896, 2025.

[110] Xu, S., Yan, Z., Dai, C., et al. MEGA-RAG: A retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health. Frontiers in Public Health, 2025, 13: 1653381.

[111] Carlini, N., Tramèr, F., Wallace, E., et al. Extracting training data from large language models. Paper Presented at the 30th USENIX Security Symposium, Vancouver, Canada, 8-10 August, 2021.

[112] Yang, X., Wu, C., Yan, X., et al. Blockchain-based healthcare and medicine data sharing and service system. Paper Presented at the International Conference on

Blockchain and Trustworthy Systems, Chengdu, China, 4-5 August, 2022.

[113] Cross, J. L., Choma, M. A., Onofrey, J. A. Bias in medical AI: Implications for clinical decision-making. PLOS Digital Health, 2024, 3(11): e0000651.

[114] Pfohl, S. R., Cole-Lewis, H., Sayres, R., et al. A toolbox for surfacing health equity harms and biases in large language models. Nature Medicine, 2024, 30(12): 3590-3600.

[115] Ji, Y., Ma, W., Sivarajkumar, S., et al. Mitigating the risk of health inequity exacerbated by large language models. npj Digital Medicine, 2025, 8(1): 246.

[116] Celis, E., Keswani, V., Straszak, D., et al. Fair and diverse DPP-based data summarization. Paper Presented at the Thirty-Fifth International Conference on Machine Learning, Stockholm, Sweden, 10-15 July, 2018.

[117] El Halabi, M., Mitrović, S., Norouzi-Fard, A., et al. Fairness in streaming submodular maximization: Algorithms and hardness. Advances in Neural Information Processing Systems, 2020, 33: 13609-13622.

[118] Zhu, Y., Basu, S., Pavan, A. Fairness in monotone k-submodular maximization: Algorithms and applications. Paper Presented at the IEEE International Conference on Big Data, Washington, USA, 15-18 December, 2024.

[119] Mesinovic, M., Watkinson, P., Zhu, T. Explainability in the age of large language models for healthcare. Communications Engineering, 2025, 4(1): 128.

[120] Liu, X., Rivera, S. C., Moher, D., et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. The Lancet Digital Health, 2020, 2(10): e537-e548.

[121] Ribeiro, M. T., Singh, S., Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. Paper Presented at the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, USA, 13-17 August, 2016.

[122] Lin, H., Bilmes, J. A class of submodular functions for document summarization. Paper Presented at the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Oregon, USA, 19-24 June, 2011.

[123] Zhu, Y. Submodular optimization: Variants, theory and applications. Paper Presented at the 33rd ACM International Conference on Information and Knowledge Management, Idaho, USA, 21-25 October, 2024.

[124] Zhu, Y., Basu, S., Pavan, A. Improved evolutionary algorithms for submodular maximization with cost constraints. ArXiv Preprint ArXiv: 2405.05942, 2024a.

[125] Zhu, Y., Basu, S., Pavan, A. Regularized unconstrained weakly submodular maximization. Paper Presented at the 33rd ACM International Conference on Information and Knowledge Management, Idaho, USA, 21-25 October, 2024b.

[126] Jin, W., Li, X., Fatehi, M., et al. Guidelines and evaluation of clinical explainable AI in medical image analysis. Medical Image Analysis, 2023, 84: 102684.