Sha et al. Mil Med Res 2026
https://doi.org/10.1016/j.mmr.2026.100050

MMR MILITARY MEDICAL RESEARCH



REVIEW Open Access

# A roadmap for medical large language models: a review of foundations, applications, and challenges

Yu-Yang Sha[1†], Li Yu[2†], Ze-Hui Lin[1], Amandeep Kaur[3], Yan-Yan Lou[4], Shivanand S Gornale[5], Tian-Yu Zhang[6], Ling Shing Wong[7], Zhi-Wen Wang[8], Yan Yan[9], Xian-Bin Zhang[10], Rui Hong[11,12], Ka Li[11], Sio Kei Im[1], Paulo de Carvalho[13*], Tao Tan[1*] and Ke-Feng Li[1*]

**Abstract**

Medical large language models (Med-LLMs) have shown considerable promise across a broad range of clinical tasks, including decision support, medical documentation, patient communication, multimodal analysis, and telemedicine. Their rapid development has generated growing interest in how large language models (LLMs) may support healthcare practice, while also raising important questions about reliability, clinical validity, and safe deployment. This review provides a structured overview of recent progress in Med-LLMs by examining their major application areas, key challenges, and emerging future directions. Current evidence shows that the clinical usefulness of Med-LLMs cannot be judged by model performance alone. Their value in practice depends on whether they are supported by reliable evidence, remain consistent with current medical knowledge, and can be integrated into clinical workflows. Important challenges remain in evaluation, safety, knowledge updating, and real-world deployment. These issues reflect a gap between performance in controlled settings and clinical practice. Future progress will require stronger clinical validation, better alignment with medical practice, and more careful deployment across different settings. The clinical impact of Med-LLMs will depend on whether they can be used as reliable tools in clinical care.

**Key words** Medical large language models (Med-LLMs), Artificial intelligence (AI), Clinical applications, Model evaluation, Clinical deployment

## Background

The launch of ChatGPT by OpenAI on November 30, 2022, marked a pivotal milestone in artificial intelligence (AI), propelling large language models (LLMs) to global prominence [1]. Unlike earlier breakthroughs such as AlphaFold [2], which primarily garnered attention within specific scientific communities, ChatGPT demonstrated exceptional capabilities in natural language understanding (NLU) and generation (NLG), capturing widespread interest across academia, industry, and the general public [3]. LLMs now exhibit sophisticated abilities in contextual reasoning, coherent text generation, and multi-turn dialogue, eliciting intense discussions regarding the future direction and societal implications of artificial general intelligence (AGI) [4]. The introduction of LLMs has not only transformed human-computer interaction but also catalyzed a surge of interdisciplinary research aimed at enhancing model capabilities, ensuring safety, and aligning AI behavior with human ideals [5]. Consequently, LLMs are increasingly recognized as functional infrastructure for building more adaptive, generalizable, and cognitively intelligent systems.

In recent years, research on LLMs has expanded rapidly, with leading research institutions and technology companies releasing models such as LLaMA, Mistral, Qwen, DeepSeek, Gemini, and Grok [6-9]. These models demonstrate strong performance across diverse natural language processing (NLP) tasks such as generation, translation, summarization, and information extraction. Figure 1 illustrates the chronological evolution of major LLMs, highlighting continuous architectural refinement and capability expansion. A central driver of this progress is the scaling law, which suggests that increasing the model size and training data often leads to substantial enhancements in model performance [10]. Under this paradigm, numerous organizations have pushed the boundaries of model size. For instance, the LLaMA family currently covers model sizes ranging from approximately 7B

†Yu-Yang Sha and Li Yu contributed equally to this work
*Correspondence: Paulo de Carvalho, carvalho@dei.uc.pt; Tao Tan, taotan@mpu.edu.mo; Ke-Feng Li, kefengl@mpu.edu.mo
[1]Faculty of Applied Sciences, Macao Polytechnic University, Macau 999078, Macao SAR, China
[13]Informatics Engineering Department, Faculty of Science and Technology, University of Coimbra, 3030-290 Coimbra, Portugal
Full list of author information is available at the end of the article

Received: 22 October 2025 Accepted: 23 June 2026

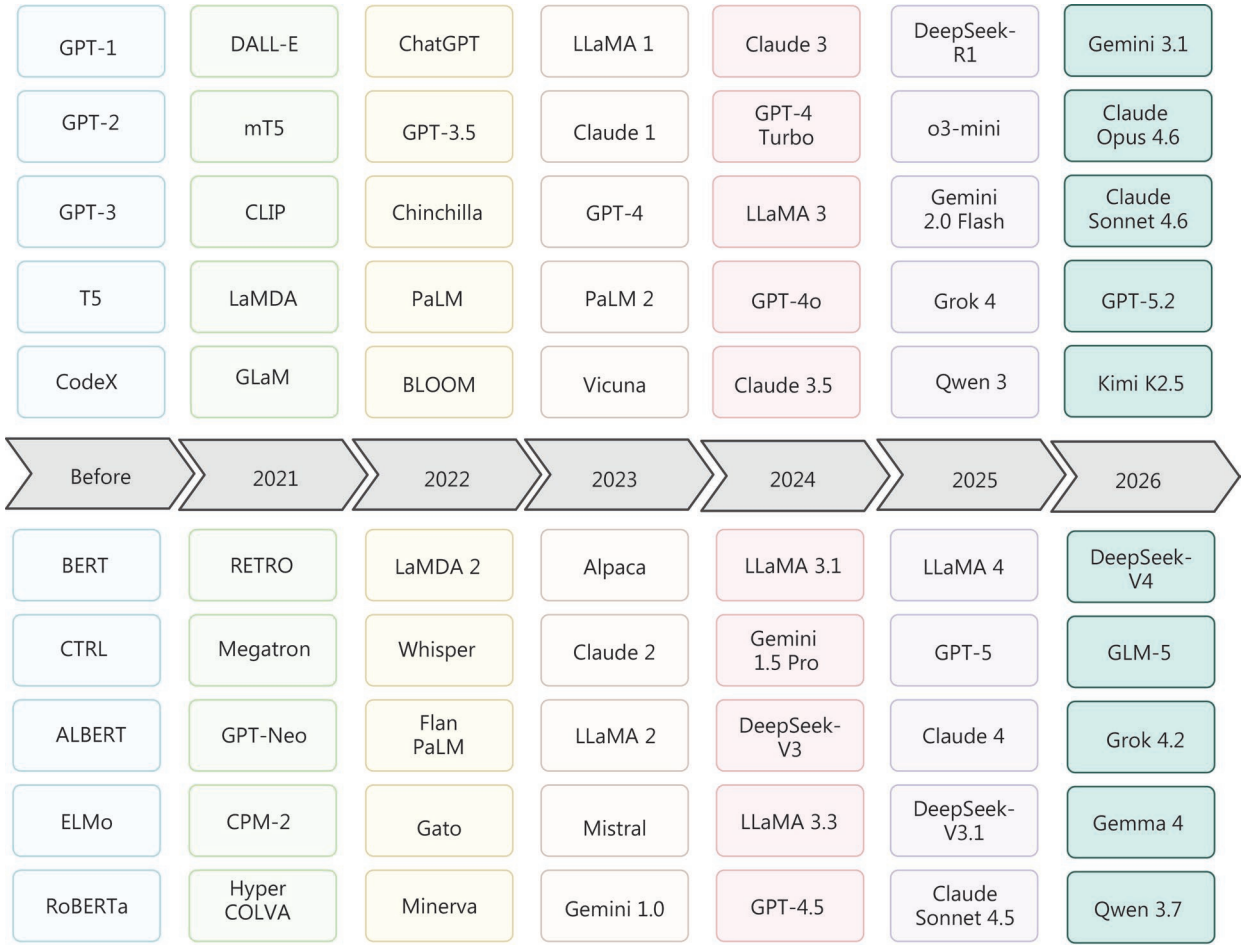


**Fig. 1 Chronological overview of major large language models (LLMs) developed in recent years.**
ALBERT. A lite BERT; BERT. Bidirectional encoder representations from transformers; CLIP. Contrastive language-image pre-training; CTRL. Conditional transformer language model; ELMo. Embeddings from language model; GPT. Generative pre-trained transformer; LLaMA. Large language model meta AI; PaLM. Pathways language model; T5. Text-to-text transfer transformer

to 405B parameters [8,11,12], while DeepSeek-V3 has been reported as a mixture of experts (MoE) model with 671B total parameters [13]. Beyond parameter scaling alone, emerging paradigms such as MoE architectures and retrieval-augmented generation (RAG) frameworks are increasingly explored to improve efficiency, adaptability, and domain specialization, signaling important directions for the next generation of LLMs [14,15].

Building upon the success of general-purpose LLMs, research has increasingly shifted toward domain-specific applications, including disease diagnosis, medical image understanding, and other clinically oriented tasks [16-18]. Prior studies suggest that LLM-based systems can achieve competitive or improved performance over traditional AI methods in certain medical tasks, benefiting from their contextual reasoning and language comprehension capabilities [3,4,16]. Among these domains, the medical domain has garnered particular attention due to its crucial societal importance and the intrinsic complexity of clinical reasoning. Clinical decision-making often involves uncertainty, multi-step inference, and risk-sensitive judgments that directly affect patient outcomes. Consequently, AI systems deployed in such settings must satisfy stringent requirements for safety, privacy, reliability, and accountability [5]. These challenges have motivated the development of medical LLMs (Med-LLMs), which are specifically adapted to medical knowledge and clinical workflows and are increasingly explored as intelligent support systems for clinical practice [19,20].

Med-LLMs are typically derived from general-purpose foundation models and further adapted for medical use through domain-adaptive pretraining and fine-tuning strategies. These strategies include supervised fine-tuning (SFT), parameter-efficient fine-tuning (PEFT), and instruction tuning and domain alignment techniques [21,22]. Such adaptations improve the models' understanding of medical terminology, clinical narratives, and diagnostic reasoning patterns. In

addition, RAG and domain-specific knowledge integration are increasingly incorporated to enhance factual grounding and reduce hallucinations in high-stakes clinical settings [15,23]. Early applications of these adaptation strategies led to representative Med-LLMs such as ChatDoctor, MedAlpaca, and PMC-LLaMA, which demonstrated the feasibility of aligning general-purpose LLMs with medical knowledge and clinical dialogue tasks [21,24,25]. Building upon these initial efforts, recent advancements have moved Med-LLMs beyond early experimental prototypes toward more clinically oriented and robust systems. For example, the Me-LLaMA family builds on open-source LLaMA models with continued medical pretraining and demonstrates strong capability in medical text analysis tasks [26]. Hulu-Med represents a multimodal extension, integrating medical text with imaging and video to support more complex clinical reasoning [27]. Variants of general models, such as MedGemma derived from multimodal Gemma architectures, are also being explored for clinical assistance and health-oriented analysis [28]. Together, these developments indicate that Med-LLMs are evolving toward systems with practical relevance for medical practice and healthcare delivery. Nevertheless, ensuring their safe and reliable deployment in real-world medical environments remains a critical challenge, motivating the structured analysis presented in this review.

Despite rapid progress, translating Med-LLMs from promising research prototypes into routine medical practice remains non-trivial. Most current models are still trained primarily for next-token prediction, which supports fluent generation but may be brittle under the uncertainty, causal reasoning, and multi-step decision-making demands of clinical care [29,30]. Moreover, commonly used evaluation protocols and metrics often fail to reflect risk-sensitive clinical outcomes. In medical settings, false reassurance and omission errors may lead to asymmetric consequences for patient safety [15,18,31]. Real-world deployment further introduces governance constraints, including patient data privacy, auditability, and responsibility allocation between clinicians and AI systems [32,33]. These gaps highlight the need for a roadmap, which we organize around six closely interconnected challenges: hallucination and factual reliability; evaluation, benchmarks, and clinical validity; safety, ethics, privacy, and regulation; robustness, generalization, and distribution shift; knowledge updating, model maintenance, and lifecycle challenges; and human oversight, workflow integration, and behavior alignment.

Motivated by the rapid advances of Med-LLMs and the persistent gaps between benchmark progress and real-world clinical deployment, this review presents a roadmap for translating Med-LLMs into clinical practice. We synthesize evidence from recent studies using a structured and transparent review approach, organizing the discussion across the development lifecycle of Med-LLMs, including foundational architectures, adaptation strategies, representative applications, evaluation standards, and governance considerations. Given the high-risk nature of medical decision-making, particular attention is devoted to issues essential for safe and reliable clinical adoption, such as hallucination mitigation, risk-aware evaluation, patient privacy protection, auditability, continual knowledge updating, and behavior alignment. The remainder of this paper is organized as follows: 1) describes the search strategy and selection criteria used to identify relevant studies; 2) introduces the foundations of LLMs and summarizes key training and adaptation approaches for Med-LLMs; 3) examines representative applications across major medical domains and their relevance to clinical practice; 4) discusses the main challenges that affect their clinical use, including issues related to reliability and evaluation; 5) outlines future research directions for the development of Med-LLMs, with a focus on clinical validation and real-world deployment; and 6) concludes the review by summarizing the main findings and their implications for clinical practice.

## Search strategy and selection criteria

A structured literature search was conducted using bibliographic databases, including PubMed, Web of Science, Scopus, and the IEEE Xplore Digital Library, supplemented by Google Scholar to cross-check relevant studies and identify additional records. The search focused primarily on peer-reviewed journal articles and major conference proceedings related to Med-LLMs, their clinical applications, multimodal medical AI, evaluation, benchmarking, reliability, and deployment in healthcare settings. Given the rapid development of LLM research, selected influential preprints from arXiv were also considered when they reported important methodological advances, newly released models, or emerging directions not yet fully represented in the peer-reviewed literature. Search terms were selected to capture both methodological and clinical perspectives. A representative search strategy combined LLM-related terms with medical-domain and application-specific terms using Boolean operators. The core search query was: ("large language model" OR "LLM" OR "foundation model" OR "generative artificial intelligence") AND ("medicine" OR "medical" OR "clinical" OR "healthcare" OR "biomedical"). Additional terms were combined as appropriate to cover major application and challenge areas,

including "multimodal medical AI", "clinical decision support", "clinical documentation", "medical question answering", "mental health", "telemedicine", "military medicine", "evaluation", "benchmark", "hallucination", "reliability", "safety", "privacy", "regulation", "EU AI Act", "governance", "robustness", "knowledge updating", and "workflow integration". The search strategy was adapted according to the syntax and indexing functions of each database. Emphasis was placed on studies published between January 2021 and May 2026, with selected foundational works published before 2021 included where necessary to provide essential background for understanding the development of Med-LLMs.

From the retrieved literature, studies were manually reviewed and selected according to their relevance to the scope of this review. Priority was given to studies with clear clinical relevance, substantial influence in the field, and sufficient methodological or conceptual contribution. Peer-reviewed journal articles and conference papers were prioritized, while influential preprints were included when appropriate to capture recent advances not yet fully represented in the formal literature. Studies that were less relevant to medical contexts, provided limited methodological detail, or substantially overlapped with better-established work were not emphasized. Given the rapid evolution and heterogeneous nature of the literature in this field, rather than conducting a PRISMA-based systematic review or meta-analysis, we used a structured narrative approach to synthesize the selected studies thematically. Focusing on clinical applications, key challenges, and emerging future directions, this strategy aims to provide broad and structured coverage of a rapidly evolving field while maintaining conceptual transparency in how the reviewed evidence was identified and incorporated into the present review.

## Overview of medical large language models

Recent advances in Med-LLMs are fundamentally rooted in architectural and training paradigms developed for general-purpose LLMs [10,28,34,35]. However, applying Med-LLMs in healthcare settings introduces additional requirements related to reliability, safety, data governance, and continual knowledge updating [16,34,36,37]. Accordingly, rather than providing a comprehensive overview of general-purpose LLM development, this section focuses on technical foundations that are most relevant to medical adaptation and trustworthiness [38]. We first summarize core modeling architectures and emerging paradigms, followed by training and adaptation strategies that enable domain specialization, and finally outline technical principles that support the safe and reliable development of Med-LLMs.

### Foundations of large language models

LLMs are built upon a combination of architectural innovations and large-scale training paradigms that enable contextual representation and generative reasoning [35,39]. Understanding these technical foundations is essential for explaining how such models acquire capabilities later adapted to medical applications. Therefore, the key modeling principles underlying contemporary LLMs are examined, including architectural design, scaling behavior, and emerging paradigms, followed by their implications for medical AI and the development of Med-LLMs [40,41].

#### *Transformer-based architecture*

Transformer architectures form the technical foundation of LLMs [35,42]. Unlike earlier sequence modeling approaches that rely on strictly sequential processing, Transformers employ self-attention mechanisms that enable tokens to interact with other relevant tokens within a given contextual window. In autoregressive decoder-only models, which constitute the dominant architecture for LLMs, attention is typically constrained by causal masking so that predictions depend only on preceding context [9,12,43]. This context-aware interaction allows models to capture long-range semantic relationships across extended text, facilitating the integration of dispersed information within complex narratives. Such capabilities are particularly important for medical language processing, where clinically meaningful evidence may appear across multiple sentences, heterogeneous terminologies, or temporally separated clinical descriptions.

In addition to improved contextual representation, Transformer architectures exhibit strong scalability properties that have enabled the development of large-scale foundation models. The attention-based design supports parallel computation during training, allowing efficient optimization on massive datasets compared with recurrent architectures that require sequential updates. As model parameters and training corpora increase, these architectures demonstrate improved generalization across diverse linguistic and knowledge-intensive tasks [10,43]. This scalability has established Transformers as the dominant backbone for LLMs and has enabled their transfer to specialized domains, including biomedical text analysis, clinical documentation understanding, and medical decision-making scenarios [36].

At the same time, the architectural principles underlying Transformer-based models also influence their behavioral characteristics [44]. Most widely used generative LLMs are trained using autoregressive next-token prediction objectives, producing outputs through probabilistic estimation

conditioned on prior context rather than explicit causal reasoning processes. While this mechanism enables flexible language generation and knowledge synthesis, it does not inherently guarantee factual consistency or logical validity [36,45]. Consequently, both the strengths and limitations of LLMs originate from the same architectural foundation, an observation that becomes particularly relevant when adapting these systems to safety-critical medical applications. Understanding these properties provides an essential basis for subsequent discussions on the development and application of Med-LLMs [40,46].

### *Modeling paradigms of large language models*

With Transformer architectures serving as the common backbone, LLMs can be broadly categorized into three principal modeling paradigms according to how contextual information is encoded and generated: encoder-only, decoder-only, and encoder-decoder models [7,47,48]. These paradigms primarily differ in information flow patterns and learning objectives, which subsequently influence their applicability across downstream tasks [49,50]. Figure 2 illustrates the structural characteristics of these modeling paradigms, while Table 1 [8,9,11-13,43,51-61] summarizes representative LLMs associated with each configuration. Rather than representing competing designs, these paradigms provide complementary strategies for language understanding and generation that support a wide range of downstream applications [47,50].

Encoder-only models. Encoder-only models consist of stacked Transformer encoder layers designed to learn bidirectional contextual representations from input sequences [51,52]. By simultaneously attending to preceding and succeeding tokens, these models emphasize semantic re-

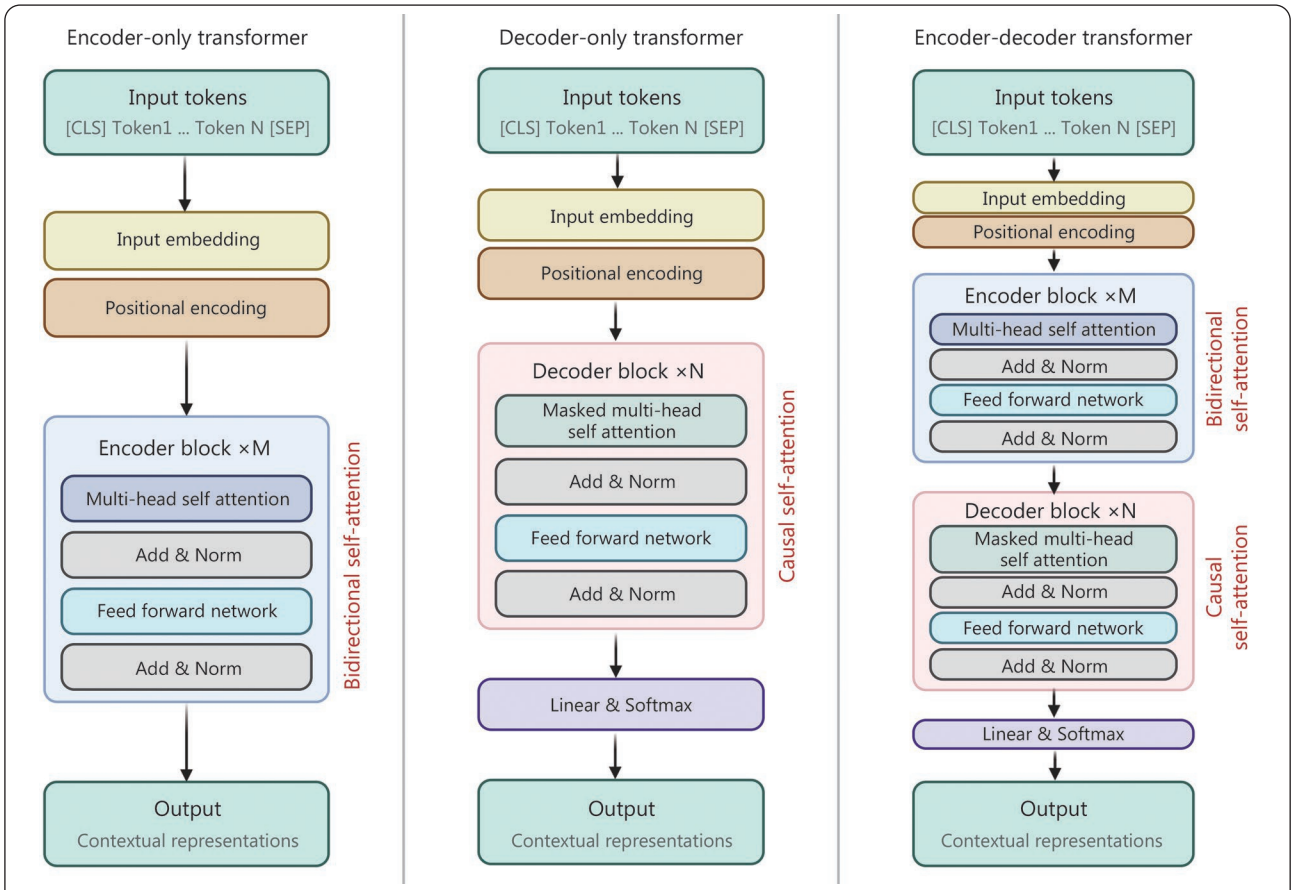


**Fig. 2 The three basic Transformer architectures: encoder-only Transformer, decoder-only Transformer, and encoder-decoder Transformer.**

In an encoder-only transformer, input tokens are processed simultaneously through stacked encoder blocks using bidirectional self-attention, allowing each token to attend to all other tokens and produce contextualized representations. In a decoder-only transformer, representations are computed autoregressively from left to right under causal masking, where each token attends only to previously generated tokens for next-token prediction. In an encoder-decoder transformer, source tokens are first encoded into contextual representations by the encoder, and the decoder then generates target representations conditioned on both previous target tokens and the encoder outputs via cross-attention. Arrows indicate the direction of information flow throughout the architectures. CLS. Classification token; SEP. Separator token

**Table 1 Representative Transformer-based language models across encoder-only, decoder-only, and encoder-decoder paradigms**

| Architecture | Representative model | Year | Parameter scale | Training data scale | Reference |
|---|---|---|---|---|---|
| Encoder-only | BERT | 2018 | 110 M/340 M | 3.3 B tokens | [51] |
| | RoBERTa | 2019 | 355 M | 160 GB text | [52] |
| | DeBERTa | 2020 | 1.5 B | 160 GB text | [53] |
| | ELECTRA | 2020 | 14 M/110 M/335 M | 3.3 B tokens | [54] |
| Decoder-only | GPT-3 | 2020 | 175 B | 300 B tokens | [55] |
| | GPT-4 | 2023 | Not reported | Not reported | [43] |
| | LLaMA | 2023 | 7 B/13 B/33 B/65 B | 1.4 T tokens | [8] |
| | LLaMA 2 | 2023 | 7 B/13 B/70 B | 2 T tokens | [11] |
| | LLaMA 3 | 2024 | 8 B/70 B | 15 T tokens | [12] |
| | Qwen3 | 2025 | 0.6 B–32 B 235 B (22 B active, MoE) | 36 T tokens | [9] |
| | Gemma3 | 2025 | 270 M/1 B/4 B/12 B/27 B | Not reported | [56] |
| | DeepSeek-V3 | 2024 | 671 B (37 B active, MoE) | 14.8 T tokens | [13] |
| | GLM-5 | 2026 | 744 B (40 B active, MoE) | 28.5 T tokens | [57] |
| Encoder-decoder | BART | 2019 | 140 M/400 M | 160 GB text | [58] |
| | T5 | 2020 | 11 B | 750 GB text | [59] |
| | FLAN-T5 | 2022 | 3 B/11 B | 780 B tokens | [60] |
| | UL2 | 2022 | 20 B | 1 T tokens | [61] |

Not reported indicates information not specified in the original technical documentation. Parameter values without additional annotation denote dense Transformer models. For MoE architectures, the value outside parentheses represents the total number of model parameters, whereas the value inside parentheses indicates the subset of activated expert parameters involved in computation per token. M. Million; B. Billion; T. Trillion; GB. Gigabyte; MoE. Mixture of experts

presentation learning rather than open-ended text generation. Such properties are particularly beneficial for medical language understanding tasks requiring precise contextual interpretation, including clinical entity recognition, biomedical information extraction, and electronic health record (EHR) analysis. Early models, including BERT [51] and RoBERTa [52], established the foundation of bidirectional language representation learning, while recent developments such as ModernBERT [62] and domain-adapted variants [63] improve efficiency and long-context modeling for large-scale biomedical text understanding. Despite the rise of generative LLMs, encoder-only models continue to play an important role in medical language understanding pipelines [62-64].

Decoder-only models. Decoder-only models are built upon autoregressive Transformer decoders that generate tokens sequentially under causal attention constraints [65]. This modeling paradigm has become the dominant configuration of LLMs due to its scalability and adaptability across heterogeneous tasks [66]. Through large-scale pretraining on diverse corpora, decoder-only models acquire strong generative capabilities and flexible contextual adaptation, enabling interactive language use, knowledge synthesis, and task generalization without explicit task-specific optimization. Representative foundation model families include GPT [43,55], Gemini [6,42], LLaMA [8,11,12], Qwen [9], and DeepSeek [7,13], which increasingly serve as backbones for domain-adapted Med-LLMs. Owing to their capacity for instruction following and in-context adaptation, decoder-only models provide a practical basis for conversational assistance, medical question answering, and clinical documentation support [19,66,67].

Encoder-decoder models. Encoder-decoder models combine bidirectional input encoding with autoregressive output generation via cross-attention mechanisms, forming a sequence-to-sequence learning framework [59,61]. This paradigm is particularly suitable for structured transformation tasks requiring faithful mapping between inputs and outputs. In medical contexts, encoder-decoder models remain widely used for clinical report generation, medical summarization, and medical language translation [68]. Beyond the widely used T5 [59] and instruction-tuned FLAN-T5 [60], encoder-decoder variants such as UL2 [61] and FLAN-UL2 further strengthen generalization through unified pretraining and instruction tuning, while LongT5 [69] extends the framework to longer clinical documents. This advantage is also consistent with evidence from general sequence-to-sequence evaluations, where representative models such as T5 and UL2 have shown strong performance on summarization, translation,

and other text-to-text benchmarks [59-61]. Such properties make them particularly suitable for medical tasks that require faithful transformation from clinical inputs to structured or narrative outputs. In the biomedical domain, encoder-decoder adaptations including BioBART [70] and SciFive [71] are frequently used backbones for generation-oriented biomedical NLP tasks.

The coexistence of multiple modeling paradigms reflects the heterogeneous requirements of medical AI applications [72]. Encoder-only models emphasize reliable semantic representation, decoder-only models enable flexible knowledge synthesis and interaction, and encoder-decoder models facilitate structured information transformation. These distinctions influence task suitability as well as controllability and robustness considerations when adapting LLMs to safety-critical domains [38,40]. Understanding paradigm-level differences, therefore, provides an essential foundation for analyzing the development and application of Med-LLMs discussed in subsequent sections [19].

#### *Scaling and emergent capabilities*

A central empirical observation in the development of LLMs is that model performance often improves predictably with scale [39,73]. Prior work has described power-law relationships between training loss and key scaling factors, including model parameters, dataset size, and training compute, suggesting that better performance can be achieved by increasing resources under a fixed optimization recipe [10,74]. Subsequent studies further emphasized that compute-optimal training depends on jointly scaling model size and the amount of training data, rather than expanding parameters alone, highlighting the practical role of data and compute allocation in shaping achievable capability at a given budget [75,76].

As scaling progresses, LLMs can exhibit qualitative capability shifts on certain tasks, sometimes described as emergent behaviors [77]. These behaviors may include improved multi-step problem solving and reasoning-like patterns that are not evident at smaller scales, although their onset and measurability can depend on task design and evaluation granularity [77,78]. In parallel, instruction following is largely enabled by post-training alignment processes, such as supervised instruction tuning and preference-based optimization with human feedback, which substantially improve helpfulness and intent-following relative to base next-token predictors [79-81]. Scaling and instruction alignment do not necessarily guarantee reliability in safety-critical domains [38,40]. LLMs remain probabilistic systems and may produce unstable or incorrect outputs when facing distribution shifts, ambiguous inputs, or incomplete clinical information. Improvements observed on benchmark evaluations, therefore, do not directly translate into reliable real-world medical performance [31]. These limitations highlight the need for evaluation frameworks and governance mechanisms specifically designed for risk-sensitive healthcare applications [37,46,82].

#### **Emerging directions in large language models design**

As LLMs are scaled to support broader capabilities and longer contexts, attention-based dense Transformers face practical constraints in compute, memory, and knowledge updating. In response, several design directions have gained momentum that complement standard dense Transformer scaling, with particular relevance to safety-critical domains such as medicine [83,84].

Recent research has explored alternative sequence modeling approaches that reduce reliance on full self-attention mechanisms [85]. Among these, state-space models (SSMs) have emerged as a promising direction for efficient long-sequence modeling [86,87]. Instead of explicitly computing pairwise attention across tokens, SSMs model sequences through recurrent state transitions, enabling computational costs that scale more favorably with input length. Selective state-space architectures such as Mamba demonstrate competitive language modeling performance while improving efficiency in long-context processing. These properties are particularly relevant for domains involving extended textual or temporal information, where maintaining long-range dependencies under practical computational constraints remains challenging. Although SSM-based models are still under active investigation and have not replaced attention-based architectures, they represent an important complementary direction for improving scalability and efficiency in next-generation LLM design [88].

Another major direction is to decouple part of factual knowledge from model parameters by integrating non-parametric memory, typically via retrieval [15]. RAG combines a retriever with a generator to condition outputs on retrieved evidence for knowledge-intensive tasks [15,89]. Retrieval-Enhanced Transformer (RETRO) further illustrates retrieval integrated into the modeling pipeline, reporting competitive performance with substantially fewer parameters by conditioning generation on retrieved chunks from a large corpus [90]. For Med-LLMs, retrieval-integrated designs provide a principled mechanism to ground responses in updatable sources such as guidelines and institutional knowledge bases, which can improve traceability and support

governance requirements when combined with appropriate evaluation and auditing [91,92].

MoE architectures introduce sparsity into LLMs by activating only a portion of model parameters during each computation step [14]. This design allows model capacity to scale substantially while maintaining manageable computational cost, and it has been adopted by several recent high-performance LLMs [14,93]. Sparse activation improves computational efficiency under practical resource constraints, which is relevant for the adaptation of Med-LLMs in real-world settings. However, increased system complexity may introduce challenges in optimization stability and deployment management [94]. Consequently, while MoE improves scalability, it does not inherently address reliability or safety requirements in medical applications.

These developments suggest that advances in LLM capability are no longer driven solely by increases in model scale [87,92,93]. Emerging design strategies emphasize computational efficiency, improved handling of long-context information, and closer integration with external knowledge sources. Such directions provide important technical foundations for improving adaptability and controllability in Med-LLMs, particularly in environments where computational resources and knowledge updating remain practical concerns [95,96]. Nevertheless, architectural innovation alone cannot ensure reliable behavior in medical settings. Further progress is needed in evaluation, alignment, and governance frameworks, as discussed in subsequent sections of this review.

**Domain requirements for medical large language models**

Although general-purpose LLMs are trained on web-scale corpora and show strong language competence, their direct transfer to medicine is limited by a pronounced domain shift [83,97]. Clinical text differs from general-domain language in vocabulary and style, with dense jargon, abbreviations, idiosyncratic documentation practices, and mixed structured and free-text formats [98,99]. These characteristics make clinical narratives harder to model reliably than general public-domain text and can degrade extraction and understanding when models are not adapted to clinical distributions. Beyond language mismatch, medical knowledge and clinical reasoning impose distinct constraints [40]. Clinical decisions often require multi-step inference under uncertainty, reconciliation of incomplete or noisy evidence, and risk-sensitive judgment where different error types have asymmetric consequences for patient safety and resource allocation. As a result, improvements on general benchmarks or in fluent generation do not necessarily translate into safe clinical behavior, especially when models encounter ambiguous presentations or out-of-distribution cases [38,100].

Medical applications operate in safety-critical contexts where model errors may directly affect clinical decisions and patient outcomes [84]. Strict privacy protection and data governance frameworks limit the collection, sharing, and secondary use of medical data, thereby constraining large-scale model training and evaluation [101]. At the same time, medical knowledge evolves continuously as new evidence and clinical guidelines emerge, requiring mechanisms for controlled updating and traceable knowledge integration. These characteristics distinguish medical settings from general language modeling environments and place additional demands on reliability, transparency, and accountable model behavior in the development of Med-LLMs [37,102].

**Training and adaptation strategies for medical large language models**

Med-LLMs are typically developed by adapting general-purpose LLMs to address the linguistic, knowledge, and safety constraints of medical domains [26,83]. This adaptation does not rely on a single modification step but instead involves a coordinated set of training and alignment strategies. Collectively, these approaches aim to reduce domain mismatch, guide clinically appropriate model behavior, enable resource-aware customization, and support knowledge grounding through external evidence [91,103,104]. Representative technical strategies for Med-LLM training and adaptation are summarized in Table 2 [16,21,22,24-26,105-123].

A common starting point is domain-adaptive or continued pretraining on biomedical and clinical corpora [26,104,124]. This stage mitigates distributional differences between general web-based text and medical language by exposing the model to domain-specific terminology, documentation patterns, and contextual semantics [108]. Although the core objective typically remains language modeling, the shift in training data improves representation quality for medical concepts and discourse. Continued pretraining has therefore become a practical and widely adopted approach for medical adaptation, particularly when training LLMs from scratch is constrained by data access and computational cost [26,83,108].

Following domain adaptation, SFT and instruction-based alignment are applied to shape how models respond in medical contexts [125]. SFT leverages curated medical tasks or instruction-formatted data to improve output relevance, structure, and consistency. Instruction tuning enhances the model's ability to follow clinically meaningful prompts, including appropriate expression of uncertainty and

**Table 2 Representative technical strategies for medical large language models (Med-LLMs) training and adaptation**

| Category | Model | Year | Backbone | Modality | Adaptation method | Medical data type | Reference |
|---|---|---|---|---|---|---|---|
| Continued pretraining | BioGPT | 2022 | GPT-style LM | Text | Domain-adaptive pretraining | Biomedical literature | [105] |
| | GatorTron | 2022 | Transformer LM | Text | Domain-adaptive pretraining | Clinical text and biomedical literature | [106] |
| | GatorTronGPT | 2023 | GPT-style LM | Text | General domain pretraining | Clinical corpora | [107] |
| | PMC-LLaMA | 2023 | LLaMA | Text | Continued pretraining | Biomedical literature | [21] |
| | MEDITRON | 2023 | LLaMA-2 | Text | Continued pretraining | Biomedical literature and guidelines | [108] |
| | BioMedLM | 2024 | GPT-style LM | Text | Domain-adaptive pretraining | Biomedical literature | [109] |
| Supervised fine-tuning | ChatDoctor | 2023 | LLaMA | Text | Dialogue instruction tuning | Clinical consultations | [24] |
| | MedAlpaca | 2023 | LLaMA | Text | Instruction tuning | Medical QA data | [25] |
| | HuatuoGPT | 2023 | General LLM | Text | Instruction tuning | Clinical consultations | [110] |
| | ClinicalGPT | 2023 | General LLM | Text | Task-specific fine-tuning | Clinical text and consultations | [111] |
| | Med-PaLM 2 | 2023 | PaLM-2 | Text | Instruction tuning | Medical QA data | [16] |
| | LLaVA-Med | 2023 | Vision-language model | Multimodal | Multimodal instruction tuning | Medical image-text pairs | [112] |
| | MAIRA-2 | 2024 | Vision-language model | Multimodal | Supervised report alignment | Chest X-ray reports | [113] |
| Parameter-efficient adaptation | Clinical Camel | 2023 | LLaMA-2 | Text | QLoRA | Clinical dialogue | [22] |
| | MedAdapter | 2023 | BERT | Text | Adapter tuning | Biomedical literature | [114] |
| | Med42 | 2024 | LLaMA-2 | Text | PEFT | Medical benchmark | [115] |
| | MedLoRA | 2024 | LLaMA | Text | LoRA | Medical instruction | [116] |
| | Me-LLaMA | 2025 | LLaMA-2 | Text | LoRA | Medical QA data | [26] |
| | PeFoMed | 2024 | Vision-language model | Multimodal | Multimodal PEFT | Medical image-text pairs | [117] |
| | AnyRef | 2024 | Vision-language model | Multimodal | Multimodal LoRA | Medical image-text pairs | [118] |
| Retrieval-augmented generation | Almanac | 2024 | GPT-4 | Text | Guideline-grounded retrieval | Clinical guidelines | [119] |
| | MedRAG | 2025 | Multiple LLMs | Text | Evidence-grounded retrieval | Biomedical literature | [120] |
| | ClinicalRAG | 2024 | Multiple LLMs | Text | EHR-grounded retrieval | Clinical records | [121] |
| | i-MedRAG | 2024 | Multiple LLMs | Text | Iterative retrieval | Biomedical literature | [122] |
| | Omni-RAG | 2025 | Multiple LLMs | Multimodal | Multimodal RAG | Medical image-text pairs | [123] |

Multiple LLMs indicates that the framework can operate with different underlying LLMs (e.g., GPT series, LLaMA series, or other general-purpose LLMs), depending on deployment configuration. Vision-language model refers to a multimodal model jointly trained to process visual and textual inputs and to learn aligned representations across image and text modalities. EHR. Electronic health record; KG. Knowledge graph; LM. Language model; LLM. Large language model; MoE. Mixture of experts; PEFT. Parameter-efficient fine-tuning; QA. Question answering; QLoRA. Quantized low-rank adaptation; RAG. Retrieval-augmented generation

adherence to requested output formats [104,126]. In addition, preference-based alignment methods are increasingly used to constrain undesirable outputs and promote more consistent behavior under complex or ambiguous prompts [80]. In medical settings, these alignment stages serve to bound model behavior within clinically acceptable expectations rather than merely improving task performance [103,125]. To address practical constraints in computation and governance, PEFT methods are often employed [127]. Instead of updating all parameters of a large model, PEFT approaches modify only a small subset while keeping the majority of the base model fixed. This strategy substantially reduces memory and training costs, facilitating iterative adaptation across institutions, specialties, or evolving clinical subdomains [128,129]. This efficiency is especially important in medical contexts where data governance constraints and resource limitations make repeated full-parameter fine-tuning impractical.

Because medical knowledge evolves and clinical recommendations are periodically updated, many Med-LLMs incorporate RAG and related knowledge integration mechanisms [91,95]. Instead of relying only on knowledge stored in model parameters, RAG-based approaches enable Med-LLMs to retrieve relevant information from external sources, such as clinical guidelines, biomedical literature, or institutional knowledge bases, during response generation. This design can improve traceability and make outputs easier to update when medical evidence changes [15]. Retrieval-based grounding does not eliminate all potential failure modes. However, it can improve factual consistency and help align responses with current clinical references when supported by appropriate source control and validation procedures [92,95]. Continued pretraining, behavioral alignment, parameter-efficient adaptation, and retrieval-based grounding together define the primary technical pathways through which general LLMs are transformed into Med-LLMs. Each pathway addresses distinct domain-specific constraints, including distributional mismatch, behavioral control, resource limitations, and knowledge updating. However, these strategies function as complementary components rather than standalone solutions, and their effectiveness ultimately depends on careful integration and context-sensitive evaluation [83,125,130].

### Principles for medical large language models

The previous sections outlined how Med-LLMs develop capability through model architecture and domain adaptation [31,84]. However, stronger model performance does not necessarily translate into reliable behavior in medical settings [131]. The requirement for trustworthiness arises from intrinsic characteristics of current LLM training paradigms and evaluation practices. Most contemporary LLMs are optimized for next-token prediction, a learning objective that favors statistical coherence rather than calibrated factual correctness or evidence-based reasoning [132]. In medical settings, where inputs are often incomplete or uncertain, this optimization can produce plausible yet unsupported outputs [133]. At the same time, commonly used evaluation benchmarks rely on static datasets and aggregate accuracy metrics, which may inadequately capture asymmetric clinical risks or context-dependent decision pathways [31,134].

These challenges are further shaped by structural properties of medical data [135]. Clinical information is sensitive, institution-specific, and heterogeneous across populations and documentation standards [136,137]. These constraints limit large-scale data sharing and introduce distributional variability between training and deployment environments. As a result, Med-LLMs may encounter domain shifts or incomplete coverage of rare conditions, complicating reliability assessment and cross-institutional generalization [138]. Behavioral alignment introduces additional complexity. Clinical recommendations are context-dependent and embedded within responsibility structures that vary across specialties and operational settings [131]. Alignment signals obtained during training cannot fully represent this variability, and acceptable risk thresholds may differ across use cases. Consequently, the need for human oversight and boundary-setting emerges from the institutional nature of medical decision-making rather than from model capability alone [82].

Finally, the dynamic evolution of medical knowledge creates a structural tension between static model parameters and continuously updated clinical evidence [134]. Guidelines, therapeutic standards, and regulatory frameworks change over time, whereas model representations remain fixed unless explicitly updated. This temporal mismatch makes continual knowledge updating a foundational concern. Taken together, these factors demonstrate that the performance and deployment of Med-LLMs are shaped by training objectives, evaluation assumptions, governance constraints, alignment processes, and knowledge dynamics [139]. Building on this foundation, the following section examines how these structural characteristics manifest across diverse medical applications.

## Applications of medical large language models

Recent progress in Med-LLMs has stimulated growing interest in their potential roles across a range of medical and clinical

tasks. Building on the literature identified through the review methodology described in Section 2 (Search strategy and selection criteria), this section examines several key application domains of Med-LLMs, including clinical decision support, medical documentation, question answering, multimodal medical data analysis, medical translation, mental health care, and traditional Chinese medicine (TCM). To provide a coherent overview, the discussion is organized around major categories of medical tasks that broadly reflect common clinical and healthcare workflows. Within each domain, we examine how Med-LLMs are used in practice and summarize the data sources and evaluation approaches reported in existing studies. We also highlight practical factors that may affect their adoption in clinical or healthcare environments. Rather than cataloguing individual datasets or systems, the goal is to identify recurring patterns in how these models are integrated into different medical contexts. Figure 3 presents a comprehensive overview of major application scenarios of Med-LLMs, which will be discussed in detail in the following sections. Table 3 [16,140-151] and Table 4 [26,28,66,68,106,112,120,152-176] provide an overview of representative datasets, benchmark tasks, and evaluation metrics for Med-LLM development and assessment, as well as their applications across major clinical domains.

**Clinical decision support**

Clinical decision support refers to systems designed to assist clinicians in synthesizing patient information and supporting diagnostic reasoning, rather than replacing medical judgment [40,84]. In clinical practice, physicians must integrate heterogeneous evidence, including symptoms, laboratory results, medical history, and imaging findings, often under considerable time pressure and uncertainty [20,177]. Med-LLMs have recently been explored as tools capable of interpreting clinical narratives, summarizing longitudinal patient records, and proposing differential diagnostic hypotheses based on available information [154,178]. Unlike earlier rule-based systems, these models can process unstructured medical language and capture contextual relationships within patient records. This capability makes them potentially useful for tasks such as triage support, diagnostic reasoning assistance, and risk stratification. At the same time, these applications highlight the importance of reliability and interpretability, because even seemingly plausible suggestions may influence clinical judgment if presented without appropriate context [40,179].

Med-LLMs are commonly incorporated into clinical decision support through two complementary patterns. In the conversational setting, a model supports case discussion by producing structured summaries, identifying missing or conflicting information, and outlining plausible diagnostic explanations in response to clinician queries [180]. In parallel, many systems couple Med-LLMs with external medical knowledge sources such as clinical guidelines, curated databases, or local institutional resources, so that outputs can be grounded in retrievable evidence rather than relying solely on parametric memory [181]. RAG is frequently used in this evidence-grounded setting to facilitate traceable synthesis and reduce unsupported statements, especially when recommendations depend on up-to-date or institution-specific guidance [91,182]. Beyond free-form dialogue, recent work has also explored more structured workflows in which intermediate decisions are made explicit and can be audited [152,181]. TrialGPT illustrates this direction by parsing eligibility criteria, assessing criterion-level patient eligibility, and ranking candidate trials to support clinical trial matching [152]. Together, these developments show how Med-LLMs can function not only as conversational assistants, but also as evidence-aware components within workflow-oriented decision support.

Evidence for Med-LLMs' decision support is currently derived from two main forms of evaluation. One approach relies on benchmark-style datasets that assess medical knowledge and reasoning ability through structured questions, including multiple-choice examinations and curated question-answering tasks [66,183]. Evaluations such as MultiMedQA [16] and MedQA [140] have been widely used to measure whether models can retrieve and apply biomedical knowledge in controlled settings. The second approach focuses on tasks derived from clinical documentation, where the goal is to assess whether generated outputs are clinically meaningful and safe in practical workflows [184]. Studies in emergency medicine and other clinical contexts have examined tasks such as summarizing EHRs or assisting with clinical handoffs [173,185,186]. In these settings, evaluation often emphasizes usefulness, clarity, and safety rather than accuracy alone [187]. The choice of evaluation metrics also depends on the clinical decision support task. In structured prediction tasks such as triage support, risk stratification, or diagnostic classification, model performance is commonly evaluated using metrics such as precision, recall, and F1 score, which help characterize different types of diagnostic errors and their potential clinical implications [143,147,148]. For example, high recall may be particularly important when the goal is to avoid missing high-risk patients, whereas low precision may increase unnecessary alerts or downstream review burden. In documentation-

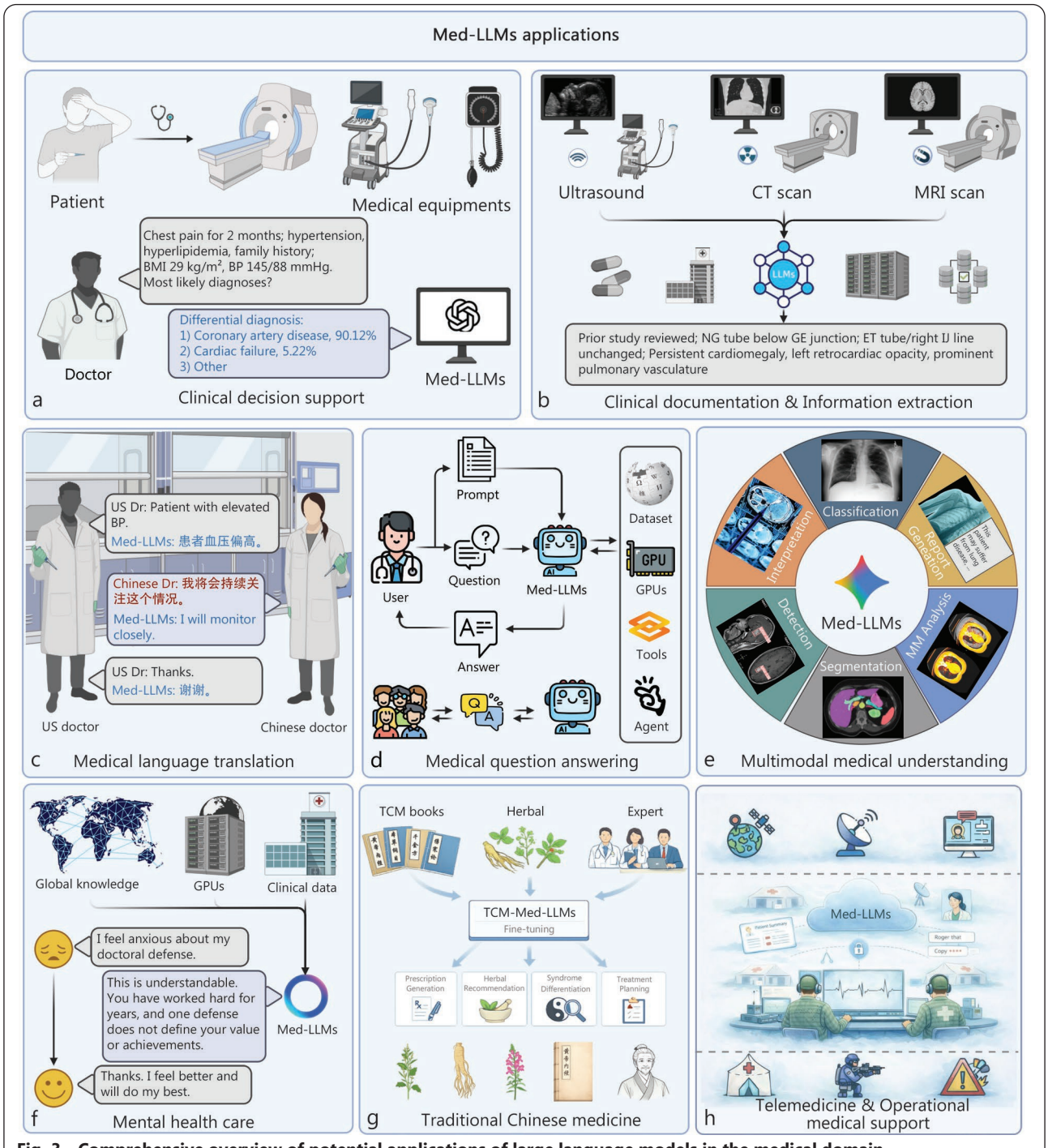


**Fig. 3 Comprehensive overview of potential applications of large language models in the medical domain.**

**a** Clinical decision support. **b** Clinical documentation and information extraction. **c** Medical language translation. **d** Medical question answering. **e** Multimodal medical understanding. **f** Mental health care. **g** Traditional Chinese Medicine. **h** Telemedicine and operational medical support. BMI. Body mass index; BP. Blood pressure; CT. Computed tomography; ET tube. Endotracheal tube; GE junction. Gastroesophageal junction; IJ. Internal jugular; LLM. Large language model; Med-LLM. Medical large language model; MRI. Magnetic resonance imaging; NG tube. Nasogastric tube; TCM. Traditional Chinese medicine

oriented tasks such as clinical summarization or handoff support, fluent output alone is insufficient, because omission of critical findings or medications may reduce clinical safety even when the generated text appears coherent. In such settings, factual consistency, completeness, and clinician judgment are often needed alongside conventional performance metrics to assess whether outputs are truly useful in practice. Together, these two evaluation paradigms provide complementary

**Table 3 Representative datasets, benchmark tasks, and evaluation metrics for medical large language models (Med-LLMs) development and assessment**

| Dataset | Year | Modality | Task category | Data scale | Evaluation metrics | Capability assessed | Reference |
|---|---|---|---|---|---|---|---|
| MedQA | 2020 | Text | Medical QA, exam reasoning | 61,097 questions | Accuracy | Medical reasoning | [140] |
| PubMedQA | 2019 | Text | Biomedical evidence inference | 1000 labeled questions; 61,200 unlabeled instances; 211,300 artificial instances | Accuracy, F1-score | Medical evidence interpretation | [141] |
| MedMCQA | 2022 | Text | Medical QA, exam reasoning | 194,000 questions | Accuracy | Medical reasoning | [142] |
| MultiMedQA | 2023 | Text | Medical QA, exam reasoning | 7 integrated QA datasets | Accuracy | Clinical reasoning | [16] |
| MIMIC-III | 2016 | Text, tabular | Clinical prediction, risk modeling | 58,000 hospital admissions | AUROC, AUPRC, F1-score, C-index | Clinical risk prediction | [143] |
| UK Biobank | 2015 | Image, tabular | Clinical prediction, risk modeling | 502,616 participants | AUROC, AUPRC, F1-score, C-index | Population risk prediction | [144] |
| MIMIC-CXR | 2019 | Multimodal | Report generation | 377,110 images; 227,835 studies | BLEU, ROUGE, METEOR, BERTScore, BARTScore, RadGraph F1, CheXbert F1 | Radiology report generation | [145] |
| IU X-Ray | 2016 | Multimodal | Report generation | 7470 images; 3955 reports | BLEU, ROUGE, METEOR, BERTScore, BARTScore | Radiology report generation | [146] |
| CheXpert | 2019 | Image | Image classification, label prediction | 224,316 images; 65,240 patients | AUROC, AUPRC, F1-score, sensitivity, specificity | Disease classification | [147] |
| CXPMRG | 2025 | Multimodal | Report generation | 223,462 image-report pairs; 187,711 studies | BLEU, BERTScore, RadGraph F1, CheXbert F1, RadCliQ | Radiology report generation | [148] |
| VQA-RAD | 2018 | Multimodal | Medical VQA | 315 images; 3515 QA pairs | Accuracy, F1-score | Radiology vision-language reasoning | [149] |
| PathVQA | 2020 | Multimodal | Medical VQA | 4998 images; 32,799 QA pairs | Accuracy, F1-score | Pathology vision-language reasoning | [150] |
| OmniMedVQA | 2024 | Multimodal | Medical VQA | 73 source datasets; 12 modalities; 20 anatomical regions | Accuracy, F1-score | General medical vision-language reasoning | [151] |

AUPRC. Area under the precision-recall curve; AUROC. Area under the receiver operating characteristic curve; BLEU. Bilingual evaluation understudy; C-index. Concordance index; METEOR. Metric for evaluation of translation with explicit ordering; QA. Question answering; ROUGE. Recall-oriented understudy for gisting evaluation; VQA. Visual question answering

**Table 4 Representative medical large language models (Med-LLMs) applications in major clinical domains, together with their development methods, medical tasks, and clinical roles**

| Application | Model | Year | Modality | Method | Medical task | Clinical role | Reference |
|---|---|---|---|---|---|---|---|
| Clinical decision support | GatorTron | 2022 | Text | Clinical-domain pretraining | Clinical text understanding | EHR-based decision support | [106] |
| | TrialGPT | 2024 | Text | RAG | Trial eligibility assessment | Clinical trial matching | [152] |
| | Med-PaLM 2 | 2025 | Text | Instruction tuning | Medical question answering | Diagnostic decision-making | [66] |
| | ClinicalGPT-R1 | 2025 | Text | RL | Diagnostic reasoning | General clinical decision support | [153] |
| Clinical documentation & extraction | ACS-LLM | 2024 | Text | PEFT | Clinical text summarization | Clinical note drafting | [68] |
| | EHR-Tools | 2025 | Text/EHR | Prompt ensemble | Hospital course generation | Workflow-integrated documentation support | [154] |
| | Flamingo-CXR | 2025 | Multimodal | Multimodal fine-tuning | Radiology report generation | Radiology documentation drafting | [155] |
| | CLEAR | 2025 | Text | RAG | Clinical information extraction | Structured clinical entity extraction | [156] |
| Medical language translation | MMed-Llama3 | 2024 | Text | Continued pretraining | Multilingual medical translation | Multilingual medical communication | [157] |
| | LLMs-in-MT | 2024 | Text | PEFT | Biomedical text translation | Domain-specific translation | [158] |
| | MedCOD | 2025 | Text | LoRA | English-to-Spanish medical translation | Terminology-preserving translation | [159] |
| | MultiMed-ST | 2025 | Speech/Text | Instruction tuning | Medical speech translation | Cross-lingual clinical communication | [160] |
| Medical question answering | LLaVA-Med | 2023 | Multimodal | Multimodal instruction tuning | Biomedical visual question answering | Biomedical image interpretation | [112] |
| | Me-LLaMA | 2024 | Text | Continued pretraining | Biomedical question answering | Biomedical knowledge retrieval assistance | [26] |
| | MedRAG | 2024 | Text | RAG | Evidence-grounded question answering | Medical evidence retrieval and answering | [120] |
| | LINS | 2025 | Text | Multi-agent RAG | Evidence-grounded question answering | Credible evidence-based answering | [161] |
| Multimodal medical understanding | scGPT | 2024 | Multimodal | Multimodal pretraining | Multi-omics representation learning | Molecular phenotype characterization | [162] |
| | MedGemma | 2025 | Multimodal | Multimodal pretraining | Medical image-text understanding | Open medical model development | [28] |
| | HONeYBEE | 2025 | Multimodal | Foundation-model embeddings | Multimodal oncology representation learning | Patient-level oncology profiling | [163] |
| | MOFS | 2025 | Multimodal | Intermediate & late multimodal fusion | Multi-omics subtype discovery | Precision glioma stratification | [164] |
| Mental health care | MentaLLaMA | 2024 | Text | Instruction tuning | Depression risk detection | Social-media mental health screening | [165] |
| | CBT-LLM | 2024 | Text | Instruction tuning | Mental health question answering | CBT-aligned psychoeducational dialogue | [166] |
| | MLlm-DR | 2026 | Multimodal | PEFT | Depression severity assessment | Multimodal mental health evaluation | [167] |
| | MDD-Thinker | 2026 | Text | PEFT & RL | Clinical depression diagnosis reasoning | Diagnostic decision support | [168] |
| Traditional Chinese medicine | Qibo | 2024 | Text | Instruction tuning | TCM question answering | TCM knowledge consultation | [169] |
| | JingFang | 2025 | Text | Multi-agent reasoning | Syndrome differentiation and treatment recommendation | TCM knowledge consultation | [170] |
| | ShizhenGPT | 2025 | Multimodal | Multimodal pretraining | Image-text TCM understanding | Multimodal TCM interpretation | [171] |
| | TCM-KLLaMA | 2025 | Text | Knowledge graph integration | Herbal prescription generation | Prescription planning | [172] |

(Continued)

| Application | Model | Year | Modality | Method | Medical task | Clinical role | Reference |
|---|---|---|---|---|---|---|---|
| Telemedicine and operational medical support | ED Handoff LLM | 2024 | Text | Prompt engineering | Handoff note generation | Transition-of-care communication | [173] |
| | GPT-4o TCCC | 2025 | Text | Guideline-grounded prompting | Ventilator management decision support | Battlefield casualty management | [174] |
| | MedAgentBench | 2025 | Text/EHR | LLM-agent orchestration | Longitudinal EHR task execution | Operational workflow automation | [175] |
| | SELSM-Agent | 2026 | Text/EHR | Retrieval-guided agent reasoning | FHIR-based clinical decision tasks | Edge-deployable medical agent support | [176] |

CBT. Cognitive behavioral therapy; ED. Emergency department; EHR. Electronic health record; FHIR. Fast Healthcare Interoperability Resources; LLM. Large language model; LoRA. Low-rank adaptation; MDD. Major depressive disorder; MT. Machine translation; PEFT. Parameter-efficient fine-tuning; QA. Question answering; RAG. Retrieval-augmented generation; RL. Reinforcement learning; ST. Speech translation; TCCC. Tactical combat casualty care; TCM. Traditional Chinese medicine

perspectives, although neither fully captures the complexity of real clinical decision-making.

Despite these promising developments, the deployment of Med-LLMs in clinical decision support faces several practical constraints. Clinical data vary substantially across institutions, specialties, and documentation practices, which can introduce distribution shifts that affect model reliability [136,185]. The probabilistic nature of language generation also means that confident but unsupported statements may occasionally appear in outputs, creating potential risks if results are interpreted without verification [20]. In addition, healthcare environments impose strict requirements related to privacy protection, data governance, and accountability. Models must therefore operate within regulatory and institutional frameworks that limit data sharing and require traceability of system behavior [152]. Clinical knowledge also evolves continuously as new evidence emerges and treatment guidelines change. Maintaining the relevance of Med-LLM systems in such environments requires mechanisms for updating knowledge and monitoring model behavior over time. These considerations illustrate why evaluating Med-LLMs' decision support requires not only performance metrics but also careful attention to integration within clinical workflows [20,84,188].

### Clinical documentation and information extraction

Clinical documentation is a fundamental part of clinical practice, but it is also widely recognized as a major source of administrative workload for clinicians [136,185]. Recent research suggests that Med-LLMs can assist with drafting and revising clinical text while also helping transform narrative records into structured information that can be used for care coordination, billing, and research activities [189,190]. In this section, clinical documentation is considered in a broad sense. It includes the generation of narrative reports such as discharge summaries and procedure notes, as well as the extraction and normalization of information from free-text clinical records [156,186]. These processes are closely connected in routine practice. Clinical notes are often created from multiple sources of information and later converted into structured elements such as problem lists, medication histories, allergy records, and diagnostic codes [128,191]. Considering documentation from this workflow perspective helps explain why reliability and safety are particularly important. Errors in documentation may not only affect the medical record itself but may also influence clinical decisions, quality reporting, and administrative processes [192].

Across the literature, Med-LLMs are incorporated into clinical documentation workflows through both generation-

oriented and structuring-oriented designs, and these approaches are often combined in practice. In generation-oriented settings, models assist in drafting clinical narratives from encounter information, including discharge summaries, progress notes, procedure notes, and radiology reports [185,186,193]. These systems aim to reduce the time clinicians spend composing routine documentation while helping organize key findings and summarize patient encounters in a clearer narrative form. In most proposed deployments, the generated text is reviewed and edited by clinicians rather than used directly, and the models function primarily as drafting assistants instead of fully autonomous documentation systems [194]. Generating clinical documentation also presents distinct challenges, because notes must remain faithful to the underlying patient record, reflect diagnostic uncertainty when appropriate, and avoid introducing unsupported information [186,195]. Alongside narrative drafting, Med-LLMs are increasingly used to extract clinically relevant information from free-text records and convert it into structured elements such as problem lists, medication histories, and allergy records. These structured outputs can support downstream tasks, including coding, cohort identification, and clinical research data preparation [191,196]. In practice, documentation generation and information extraction are often integrated within the same workflow, allowing narrative notes to be produced while key information is simultaneously organized into structured representations that can be reused across clinical systems [41].

Evaluation of documentation-oriented Med-LLMs commonly blends automatic metrics with expert judgment, reflecting the gap between text similarity and clinical usefulness [197]. For report generation, overlap metrics such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [198] and Bilingual Evaluation Understudy (BLEU) [199] remain common, but many studies emphasize that high overlap does not guarantee factual correctness, appropriate omissions, or safe phrasing in clinical contexts [193,195,200]. As a result, human evaluation is frequently used to assess correctness, completeness, and utility, often with clinician raters and workflow-relevant rubrics. For information extraction, standard measures such as Precision, Recall, and F1 score are widely used because outputs can be compared against annotated spans or structured labels [190,200]. Coding support and administrative applications often use accuracy and error analysis across large label spaces, where performance can be sensitive to note noise and local coding conventions [191]. Evidence increasingly points to the need for task-specific evaluation that distinguishes harmless stylistic variation from clinically meaningful errors.

Several practical constraints influence whether these systems can be used safely in real clinical environments [84,201]. First, generated documentation may contain unsupported additions or subtle distortions. This issue becomes particularly problematic when clinical notes are treated as authoritative records. As a result, many studies emphasize cautious deployment with clear clinician oversight and traceability [186,194,197,200]. Second, privacy and governance requirements limit data sharing and logging. These constraints also make model improvement more difficult because training signals are embedded in protected health information [136]. Third, documentation practices vary across institutions, specialties, and EHR systems. Such variation can create distribution shifts that reduce reliability when systems are transferred without local validation. Finally, administrative tasks such as medical coding illustrate that fluent text generation does not necessarily translate into reliable structured outputs. Empirical studies have shown that general LLMs may perform poorly in medical coding scenarios [189,191,196]. This observation highlights the importance of domain constraints and careful evaluation before these systems are used in operational settings. These issues motivate the trustworthiness challenges discussed later in this review, particularly those related to hallucination, evaluation, and governance [201].

**Medical language translation**

Medical language translation plays an important role in clinical communication across linguistic and cultural boundaries [202,203]. In healthcare settings, it involves not only transferring medical information between languages but also reformulating professional medical expressions into forms that can be more easily understood by different audiences. Such tasks arise in multilingual clinical encounters, informed consent, discharge communication, and the international exchange of medical knowledge [204]. In these contexts, the primary requirement is not only linguistic fluency but also the preservation of clinical meaning [203,205]. A grammatically correct translation may still distort terminology, omit uncertainty, or alter the practical implications of the original medical statement. For this reason, recent studies increasingly emphasize communication quality, clarity, and safety when evaluating language conversion in medical contexts [205,206].

Current Med-LLM applications in this area broadly follow several patterns. The first involves clinician-facing translation, where the goal is to preserve technical precision in multilingual clinical records, guideline interpretation, or cross-institutional

evidence exchange [157,207]. A second scenario involves translation and reformulation aimed at patient communication, where complex medical language is converted into clearer and more accessible explanations for non-expert audiences [202]. A third direction focuses on translation grounded in external medical knowledge sources, where terminology systems, curated references, or biomedical literature are incorporated to improve semantic fidelity [206]. Recent multilingual medical foundation models illustrate this trend. For example, MMed-Llama3 [157] was developed together with a large multilingual medical corpus and benchmark, reflecting a shift toward multilingual medical language modeling rather than English-centric adaptation alone. Other recent work, such as MedCOD [159], suggests that incorporating structured medical knowledge can further improve translation accuracy in specialized biomedical contexts.

The data and evaluation landscape for medical language translation remains narrower than the communication demands encountered in real clinical environments [203,207]. Classic biomedical translation resources, such as the European Medicines Agency (EMEA) corpus [208], continue to provide widely used multilingual medical text for training and evaluation [209]. In addition, the Conference on Machine Translation (WMT) Biomedical Translation Task offers standardized benchmarks for assessing translation performance across biomedical language pairs. More recent multilingual medical LLM research has also introduced broader evaluation resources, such as MMedBench, which assesses multilingual medical understanding across multiple tasks rather than focusing solely on sentence-level translation [157]. Evaluation typically relies on automatic metrics such as BLEU together with neural evaluation metrics like Crosslingual Optimized Metric for Evaluation of Translation (COMET) [210]. However, these metrics primarily measure textual similarity and may not fully capture terminology accuracy, preservation of clinical meaning, or the appropriateness of translated content for clinical use. For this reason, expert review remains particularly important in medical contexts, where even small translation errors can alter clinical interpretation [203].

Several practical limitations continue to constrain the use of Med-LLMs for medical language translation [204]. Medical terminology is highly context-dependent, and errors involving drug names, procedures, abbreviations, or risk statements may alter clinical meaning even when the translated text appears fluent [159,205]. Translation and reformulation for patient-facing communication introduce an additional tension between readability and fidelity, because simplification can improve comprehension while also removing nuance or uncertainty. Multilingual imbalance remains another concern, as English-dominant training data and limited resources for many languages can reduce robustness outside high-resource settings [157,210]. Real-world deployment also raises governance challenges related to privacy protection, informed consent, record consistency, and local terminology standards [37,41,137]. Taken together, these challenges suggest that medical language translation involves more than multilingual language processing. It is also a form of clinical communication where preserving medical meaning is essential.

**Medical question answering**

Medical question answering is one of the most widely studied applications of Med-LLMs [66]. It plays an important role in supporting knowledge access, clinical communication, and decision support [211]. In this setting, models are expected to respond to natural-language questions about symptoms, diagnoses, treatments, drug use, or biomedical evidence. The task spans two main use contexts. In clinician-facing settings, the goal is to support information retrieval, evidence synthesis, or guideline interpretation during care [212,213]. In patient-facing settings, models may be used to explain medical information or answer consumer health questions in an accessible language. This distinction matters because the same answer format may not be appropriate across both contexts, and the tolerance for uncertainty, incompleteness, or overstatement differs substantially between professional and consumer use [187,214]. As a result, medical question answering is not merely a knowledge test for LLMs, but a practical communication task with direct implications for trustworthiness and safety [84,215].

Current Med-LLMs for medical question answering generally fall into two broad patterns. The first relies on instruction-following models that answer questions directly from internal model knowledge, often performing strongly on benchmark-style exams and short factual questions [212]. Med-PaLM 2 [66] is a representative example and achieved strong performance across multiple tasks in the MultiMedQA [16] benchmark, including multiple-choice medical questions, long-form consumer health questions, and clinical consultation questions [66]. The second pattern emphasizes evidence grounding, often by combining LLMs with retrieval over clinical references, biomedical literature, or curated knowledge sources before answer generation [213]. This design is increasingly important because medical answers are expected to be not only fluent but also aligned with current evidence and able to communicate uncertainty appropriately. Recent work has also moved toward doctor-centered question answering

and workflow-aligned conversational settings, where models are evaluated as assistants that support physicians rather than as direct substitutes for clinical judgment [16,41,97].

Widely used benchmarks for medical question answering cover several complementary types of clinical knowledge and reasoning tasks [183]. Representative datasets include MedQA [140] and MedMCQA [142], both derived from medical licensing examinations. These benchmarks are widely used to evaluate clinical knowledge and diagnostic reasoning. Other benchmarks focus on understanding biomedical literature and generating evidence-based responses [141]. The MultiMedQA [16] benchmark suite further integrates multiple medical question answering datasets and has been widely adopted to assess the capabilities of LLMs across both professional medical questions and consumer health queries. Evaluation commonly relies on metrics such as accuracy, exact match, and F1 score for structured questions, while text similarity measures, including BLEU [199] or ROUGE [198], are usually used when responses involve longer explanations. However, automatic metrics alone cannot fully capture clinical correctness or safety. Therefore, several studies incorporate expert assessment to evaluate whether generated answers are medically accurate, consistent with available evidence, and appropriate for clinical or patient-facing settings [16,197,214].

The main limitations of medical question answering are closely tied to the risks of applying LLMs in high-stakes communication. Models may generate answers that are fluent but incomplete, outdated, or insufficiently grounded in evidence [216]. They may also respond inconsistently when the same question is phrased differently, which is especially problematic in patient-facing settings where users may interpret the response as authoritative advice [217]. In clinician-facing settings, the primary concerns extend beyond readability to include traceability of evidence, appropriate confidence calibration, and consistency with established clinical practice standards [47,212]. Recent safety studies show that evaluating these systems only for factual accuracy is not enough, because medically unsafe answers can still appear plausible [82,84,187]. This makes medical question answering a particularly important application area for Med-LLMs. It highlights the need for better safety-sensitive evaluation, clearer boundaries between assistance and advice, and stronger grounding in clinical evidence and workflow context.

**Multimodal medical understanding**

Many clinically important questions require information from more than one data modality rather than from a single source alone [67,218]. In practice, clinicians often interpret medical images together with narrative reports, structured laboratory results, physiological signals, and genomic or other molecular data [219]. This is particularly relevant in radiology, pathology, and precision medicine, where different modalities often contribute complementary information for diagnosis, prognosis, and treatment planning [220,221]. For Med-LLMs, the value of multimodal modeling lies in its ability to connect these heterogeneous signals and support clinically meaningful interpretation rather than treating each modality as an isolated input [218].

A prominent line of work in this area focuses on integrating medical images with natural language information [112,222]. Vision-language models link images with associated textual context, such as radiology reports, pathology descriptions, or clinician queries. This design enables systems to support tasks including report generation, image-grounded question answering, and structured interpretation of imaging findings [223]. Early medical adaptations of general vision-language architectures, such as LLaVA-Med [112] and Med-Flamingo [222], demonstrated that domain-specific instruction tuning and medical data curation can substantially improve performance on image-text reasoning tasks. More recent multimodal medical foundation models have extended this paradigm toward broader clinical settings. For example, Med-Gemini [224] integrates diverse medical modalities and has been explored for tasks that combine imaging evidence with clinical narratives and biomedical knowledge. Some emerging studies further suggest that such multimodal frameworks may incorporate molecular or genomic information alongside clinical and imaging data to support applications related to precision medicine [225-230]. These developments indicate a gradual shift from single-modality image analysis toward more integrated multimodal reasoning in medical AI [36,231].

The data and evaluation landscape for multimodal medical models remains diverse. Many studies rely on paired image-text datasets, where medical images are linked with radiology reports or clinical descriptions. Widely used resources include datasets derived from large clinical archives, such as Medical Information Mart for Intensive Care Chest X-ray (MIMIC-CXR) [145] or Indiana University Chest X-ray Collection (IU X-Ray) [146], which provide image-report pairs for tasks like report generation and image-grounded question answering. Evaluation typically combines automatic text similarity metrics, such as BLEU [199] and ROUGE [198], with expert assessment, because linguistic overlap alone does not guarantee clinical correctness or appropriate interpretation of imaging findings. Recent benchmarks have also begun to examine whether model outputs are properly grounded in

visual evidence, reflecting growing attention to reliability in multimodal medical AI [180,232,233]. Even so, no single evaluation framework yet captures all clinically relevant dimensions of multimodal performance.

Genomics represents an important extension of multimodal Med-LLMs beyond conventional image-text settings [227]. Early studies suggest that genomic information can be integrated with clinical narratives, imaging features, and other biomedical data to support tasks such as variant interpretation, disease risk assessment, and patient stratification [162,225,227,228]. This line of research is closely related to radiogenomics, which seeks to connect imaging phenotypes with underlying genomic or molecular patterns [234]. However, incorporating genomic data also introduces additional challenges. Genomic information is highly sensitive, difficult to standardize across institutions, and often probabilistic in interpretation [235]. More broadly, multimodal Med-LLMs must contend with cross-modal misalignment, uneven data quality, and the risk of producing outputs that appear plausible but are not adequately supported by the underlying evidence [180,232,233,235]. These constraints highlight both the promise and the complexity of multimodal medical understanding in clinical applications of Med-LLMs.

**Mental health care**

Mental health care relies heavily on language-based assessment and ongoing communication, which makes it a natural but high-risk setting for Med-LLM applications [236-238]. In many situations, clinicians rely on narrative descriptions, conversational exchanges, and patient-reported experiences to understand symptoms and psychological states [239]. For this reason, recent literature generally frames the role of LLMs in mental health as supportive rather than diagnostic. Current systems are most appropriately used to assist with tasks such as summarizing patient-reported concerns or helping communicate mental health information in an accessible language [236]. They may also support low-intensity follow-up communication in certain contexts. These functions can improve access to support and facilitate communication between patients and clinicians [237,238]. However, they do not replace professional judgment, particularly when situations involve elevated clinical risk or complex therapeutic decisions [168,236].

Recent studies have explored how Med-LLMs may assist mental health care in practice. One line of work focuses on screening and assessment support [168,240]. Models analyze interviews, self-reports, questionnaires, or clinical notes to identify linguistic patterns associated with depression, anxiety, or related psychological conditions [239]. An example is MentaLLaMA [165], a mental health-oriented LLM designed for interpretable analysis of mental health-related social media text. Another direction examines conversational support adapted for counseling-style interactions or cognitive behavioral therapy-inspired dialogue [166,241]. Cognitive Behavioral Therapy-based Large Language Model (CBT-LLM) [166] represents this approach and focuses on cognitive behavioral therapy-based mental health question answering in Chinese. A further area involves psychoeducational and communication support, where models help explain symptoms, summarize counseling content, or assist clinicians with documentation and follow-up communication [236,238]. Across these forms, the most defensible use case remains assistance to clinicians or service users rather than replacement of therapeutic judgment [237,238].

Data resources for mental health applications often rely on text-based data derived from interviews, question-answer interactions, and counseling-style conversations [241-243]. Representative resources include DAIC-WOZ, which contains interview-based data widely used in depression assessment [242]. Other examples include PsyQA [243], which contains question-answer pairs related to mental health knowledge, and dialogue datasets such as CounseLLMe [241] that simulate counseling-style interactions. Evaluation in this area typically combines task-specific metrics with qualitative assessment. Screening-oriented studies often report measures such as accuracy, F1 score, or AUROC, while conversational systems are more frequently examined using language-quality metrics together with expert review [168,236]. However, recent research emphasizes that automatic metrics alone are insufficient for mental health applications. They cannot fully capture whether a response is safe, supportive, or appropriate for vulnerable users [237,238]. As a result, many studies now incorporate expert clinical review to assess whether generated responses are clinically appropriate, context-sensitive, and safe for real-world use [236,237,244].

Mental health care imposes particularly strict practical constraints on the use of Med-LLMs [238]. Supportive or emotionally responsive language can easily create an impression of understanding that exceeds a model's actual capacity for clinical judgment. This gap becomes especially concerning in repeated or high-stakes interactions, where users may over-trust the system or rely on it in situations that require professional assessment [244]. Reviews in psychotherapy and psychiatry consistently caution that current LLM-based systems should not be treated as substitutes for licensed care,

and recent regulatory discussions reflect similar concerns. Additional challenges arise from the sensitive nature of mental health data and the diversity of language used to describe psychological experiences across cultures and populations [165,243,245]. In addition, maintaining consistent safety behavior during extended interactions remains difficult for current systems [82,187,246]. For these reasons, mental health care highlights both the potential value of Med-LLMs in supportive communication and the importance of clear boundaries, human oversight, and safety-oriented evaluation in responsible deployment [236-238].

**Traditional Chinese medicine**

TCM provides a distinctive application setting for Med-LLMs [169,247]. Its clinical practice relies on a specialized conceptual system and syndrome differentiation, which link symptoms, signs, constitution, and other contextual information to diagnostic and therapeutic decisions [248]. In this setting, LLM-based applications are not simply Chinese-language versions of general medical question answering [169,249]. They should also address key elements of TCM reasoning, including pattern differentiation, herbal formula selection, acupuncture-related decisions, and the interpretation of classical medical texts in modern clinical contexts. Recent studies have explored how Med-LLMs can support TCM practice, particularly in consultation settings, knowledge access, and clinical reasoning [110,169,187]. At the same time, TCM applications introduce challenges not typically encountered in mainstream biomedical settings. Model outputs must remain consistent with theory-specific terminology and accepted clinical practice [248,250].

Recent studies have begun to explore several forms of Med-LLMs in the context of TCM [169,249]. One important direction concerns consultation-oriented question answering, where models are adapted to answer questions from patients or practitioners using TCM knowledge and terminology. Systems such as MedChatZH [249] and HuatuoGPT [110] support multi-turn dialogue grounded in TCM literature and clinical consultation data. In these studies, models are used to support tasks such as syndrome differentiation, acupuncture point selection, and herbal prescription generation. More recent efforts increasingly incorporate structured medical knowledge into the modeling process. Knowledge-enhanced frameworks such as TCM-KLLaMA [172] integrate TCM knowledge graphs with LLMs to improve the consistency and interpretability of generated recommendations. Taken together, these studies suggest that Med-LLMs in TCM are gradually moving beyond general dialogue toward more structured and knowledge-grounded forms of clinical support [249]. An additional emerging direction involves hybrid Med-LLM frameworks that support alignment between TCM practice and Western medicine in integrative care settings [247,251]. In many real-world clinical environments, patients receive herbal prescriptions together with conventional medications and laboratory-guided treatment [251]. In such scenarios, Med-LLMs may assist clinicians by linking TCM syndrome differentiation with Western diagnoses and medication histories, and by supporting prescription alignment to identify potential inconsistencies or contraindications across treatment systems [172,248,252]. Such capabilities may help improve communication between practitioners trained in different medical traditions and facilitate more interpretable integrative decision-making [172,251].

Evaluation resources for TCM-oriented Med-LLMs are still emerging, with only a small number of dedicated benchmarks currently available [169,250,253]. One representative benchmark is TCMBench, which was constructed from TCM Licensing Examination questions and accompanied by a domain-specific scoring method known as TCMScore [250]. Other resources, such as Qibo-Benchmark, have also been proposed to evaluate general knowledge and reasoning in TCM contexts [169]. More recent work has introduced task-oriented benchmarks reflecting clinical scenarios. For example, TCMEval-PA [252] evaluates prescription auditing by assessing the normativity and clinical appropriateness of generated herbal formulas, while TCM-SED [253] focuses on stroke-related TCM tasks, including diagnosis, pattern differentiation, and formula selection. Evaluation commonly relies on metrics such as accuracy, as well as text similarity measures including ROUGE [198] and BERTScore [254] for generative tasks. However, many studies also rely on expert assessment, as automated metrics alone cannot determine whether outputs remain consistent with TCM theory and accepted clinical practice [250,252].

TCM-oriented applications present several practical constraints for the safe and reliable use of Med-LLMs [251,255]. Many tasks in this domain do not have a single easily verifiable answer. Syndrome differentiation, prescription selection, and interpretation of classical theory may vary across schools, regions, and clinical styles. In addition, responses that appear plausible may still misrepresent TCM concepts or disrupt the internal logic of herbal formulas [247,249]. In some cases, they may also produce recommendations that trained practitioners would not consider acceptable [256]. Standardization remains challenging because TCM language draws on diverse sources, ranging from classical

literature to modern textbooks and contemporary clinical records. Deployment in broader healthcare environments further introduces requirements related to interoperability, governance, and expert validation [247]. Taken together, these issues show that TCM applications extend Med-LLMs into highly specialized knowledge systems, but their real-world use still depends on domain-grounded evaluation, clinician oversight, and careful governance.

**Telemedicine and operational medical support**

Telemedicine and operational medical support represent an important application setting for Med-LLMs, particularly in military medicine [257,258]. In these contexts, medical care is often delivered across distance, under time pressure, and with limited access to specialist support [259,260]. Military telemedicine has long been used to extend specialist expertise to field hospitals, deployed units, and other austere settings [261]. In these environments, the value of Med-LLMs lies primarily in strengthening clinical communication rather than replacing medical decision-making. They may support remote consultation, structured case summarization, handoff preparation, and the organization of incomplete clinical information [260]. Such functions are closely connected to operational practice, where timely coordination and effective information transfer can help clinicians obtain specialist input and support more informed care decisions [262].

Recent studies have begun to examine how AI systems, including Med-LLMs, may support telemedicine and operational medical care [16,263]. One important direction involves tele-critical care and remote consultation, where digital platforms connect frontline providers with specialists located away from the point of care [261]. Such systems have been explored in military and other austere environments to extend clinical expertise to field hospitals, deployed units, and emergency response settings. In parallel, recent work has begun to explore AI-assisted decision support. In these settings, Med-LLMs are adapted to interpret clinical guidance and assist with triage decisions, treatment planning, and clinical documentation. Simulation studies suggest that systems aligned with operational medical guidelines, including Tactical Combat Casualty Care (TCCC), may improve decision consistency in selected scenarios [175,262]. Overall, these developments indicate a gradual shift from communication infrastructure alone toward AI-supported consultation and operational decision support [261,263].

In contrast to many other Med-LLM applications, evaluation in this area relies less on standardized public benchmarks [263,264]. One reason is that military and operational telemedicine often involves sensitive data, deployment-specific workflows, and rare but high-consequence scenarios that are difficult to capture in openly shared datasets [257]. As a result, many studies rely instead on simulated case scenarios, implementation studies, and expert-led workflow assessment [175,261]. Relevant evidence also comes from related resources, including emergency triage tasks, handoff-note studies, and teleconsultation workflows [265,266]. These alternatives allow researchers to examine whether systems improve communication, support documentation, and assist decision-making in realistic settings. In practice, evaluation therefore depends more on scenario-based testing, usability, and expert review than on benchmark accuracy alone [175,259].

Telemedicine and operational medical support often take place under demanding conditions [258]. Clinical information is often incomplete, and communication may be delayed or unreliable. In some situations, limited escalation pathways may lead clinicians to over-trust plausible but insufficiently supported recommendations [265]. Model performance may also vary in military and other austere environments because patient populations, available resources, and operational priorities differ from those seen in typical training data [16]. The deployment of Med-LLMs in remote military environments introduces additional constraints. Systems must operate under latency, edge-computing limits, cybersecurity requirements, and strict governance expectations [264,267]. Telemedicine and operational medical support highlight the potential of Med-LLMs to extend clinical expertise to remote and resource-limited environments, particularly in military settings. Their use in such contexts still requires careful oversight and secure deployment.

## Challenges

Although Med-LLMs have demonstrated broad potential across diverse clinical applications, their reliable deployment in real-world medical settings remains constrained by several persistent challenges. These challenges extend beyond model performance alone and involve issues such as factual reliability, limitations of current evaluation practices, safety and regulatory requirements, vulnerability to distribution shift, and the difficulty of maintaining up-to-date medical knowledge [84]. They also reflect the need for effective human oversight when these systems are integrated into clinical workflows [40]. Taken together, these factors define the central obstacles that must be addressed before Med-LLMs can be deployed at scale in high-stakes medical settings. The following sections, therefore, examine these challenges from six complementary

perspectives, spanning reliability, evaluation, governance, robustness, model maintenance, and workflow integration.

**Hallucination and factual reliability**

Hallucination in LLMs refers to the generation of content that appears plausible but is factually incorrect, unsupported by evidence, or inconsistent with established knowledge [268,269]. In Med-LLMs, this problem is closely linked to factual reliability, because a response may sound clinically appropriate while still lacking adequate evidentiary support [40]. This issue is particularly important in medical settings, where decisions often depend on accurate interpretation of clinical information [64]. Unlike many general NLP applications, even minor inaccuracies in diagnostic suggestions, treatment recommendations, or medical explanations can have significant consequences. As a result, the reliability of model-generated medical content has become one of the central concerns in the clinical use of Med-LLMs [270].

Hallucination and reduced factual reliability can affect multiple clinical applications [64,268]. In clinical decision support, models may generate diagnostic reasoning or therapeutic recommendations that are inconsistent with established clinical guidelines and available medical evidence [40,119]. In higher-risk procedural settings such as surgery or perioperative care, similar hallucinated recommendations may be especially concerning because incorrect suggestions may affect time-sensitive decisions, escalation pathways, or procedure-related planning. In medical question answering and knowledge-access scenarios, hallucinations may appear as incorrect references to biomedical studies or inaccurate descriptions of treatment protocols [271]. In communication-oriented tasks such as mental health support or telemedicine consultation, the problem can be more subtle [272]. Responses may appear reasonable while still providing misleading or overly confident suggestions that are not appropriate for the specific clinical situation. Similar concerns also arise in specialized domains such as TCM [253]. Errors in interpreting theoretical concepts or herbal relationships may produce recommendations that appear plausible but remain inconsistent with accepted clinical practice.

Various strategies have been proposed to mitigate hallucination in Med-LLMs [119,273]. Approaches such as RAG, integration with structured medical knowledge sources, and alignment-based methods can improve the consistency between generated outputs and available evidence [92,221,273]. A further promising direction is to combine RAG with domain-specific knowledge graphs in hybrid grounding frameworks [91,95]. In such systems, retrieved documents can provide up-to-date external evidence, while knowledge graphs can supply structured relations among diseases, biomarkers, drugs, and treatment pathways [98,182]. This may be particularly valuable in high-stakes settings such as oncology, where Med-LLMs may need to align generated outputs with tumor staging, biomarker profiles, and regimen-specific contraindications [163,164,182]. For example, a hybrid design could help reduce unsupported recommendations by jointly checking retrieved oncology guidelines against graph-structured knowledge of drug interactions and treatment eligibility. However, these strategies do not fully eliminate the problem. Models may still generate unsupported conclusions when evidence is incomplete or retrieved information is misinterpreted, especially in cases involving complex clinical reasoning [100,274]. These limitations mean that hallucination remains a fundamental challenge to the factual reliability of Med-LLMs. Addressing this issue requires improvements in model design as well as more rigorous evaluation of factual reliability in realistic clinical settings.

**Evaluation, benchmarks, and clinical validity**

Evaluation remains a central challenge for Med-LLMs because strong benchmark performance does not necessarily translate into reliable clinical use [64,275,276]. Much of the current literature still relies on exam-style question sets, reference-based text metrics, or narrowly defined task benchmarks to summarize model capability [277]. These approaches can be useful for measuring specific forms of knowledge recall or output similarity, but they do not adequately capture whether a system is clinically valid, appropriate for a medical task, or safe in practice [131]. Recent commentaries and systematic reviews have argued that Med-LLM evaluation should move beyond leaderboard performance and place greater emphasis on construct validity, task realism, and the clinical meaning of model outputs [278-280]. This concern is particularly important in medicine, where an answer may appear technically plausible while still being incomplete or clinically inappropriate in specific patient contexts [278,281].

The limitations of current evaluation become more evident when Med-LLMs are examined across different clinical applications [16]. In the task of medical question answering, high accuracy on licensing-style benchmarks such as the United States Medical Licensing Examination (USMLE) [282] or MedQA [140] can reflect strong recall of medical knowledge, yet these scores may overestimate real-world clinical reasoning ability [275]. Representative benchmark comparisons further show that both frontier general-purpose LLMs and domain-adapted Med-LLMs can achieve strong

performance on licensing-style medical evaluations [16,66]. For example, on the MedQA benchmark of USMLE-style questions, GPT-3.5 achieved 60.2% accuracy, Flan-PaLM reached 67.6%, GPT-4-base reached 86.1%, and Med-PaLM 2 reached 86.5% [66]. However, such benchmark gains should still be interpreted cautiously because exam-style accuracy alone does not establish clinical validity or safe deployment in real-world medical settings [46,97]. This concern has led several recent studies to question the use of exam-style benchmarks as proxies for clinical competence, noting substantial gaps between benchmark performance and expert judgment [277,279]. Similar issues arise in other tasks. In multimodal and documentation settings, commonly used text-overlap metrics such as BLEU [199] or ROUGE [198] do not guarantee factual correctness or clinically grounded interpretation [283]. In mental health applications, responses that appear empathetic or well-written do not necessarily ensure that the advice is clinically safe [280,284]. In telemedicine and operational support settings, evaluation often focuses on how systems perform within realistic clinical workflows rather than on benchmark scores alone [278]. Taken together, these observations suggest that clinical validity is strongly task-dependent and requires evaluation strategies that reflect the specific clinical context.

Recent work is increasingly moving toward more clinically relevant evaluation frameworks [131]. Representative directions include expert review methodologies such as CLEVER, scenario-based assessment frameworks, and human-centered evaluation that focuses on safety, usefulness, and clinical workflow performance [279]. Some studies have also explored LLM-as-a-Judge approaches for clinical summarization, although these methods still depend on carefully designed human reference standards and evaluation rubrics [276,279,285]. Despite this progress, a unified framework for evaluating Med-LLMs remains lacking. Existing evidence remains fragmented across different tasks and clinical domains [97,278,279]. In addition, many studies still prioritize evaluation convenience over clinical realism [277,280,286]. More robust evaluation will require not only stronger benchmarks and metrics, but also broader use of expert-centered and prospectively validated assessment strategies.

**Safety, ethics, privacy, and regulation**

Beyond technical performance, the use of Med-LLMs also raises important questions about safety, ethics, privacy, and regulatory oversight [287,288]. A model may appear factually reliable and achieve strong benchmark performance, but these results do not guarantee safe use in high-stakes clinical settings [287,289]. In medicine, the risks extend beyond incorrect answers alone. They also involve over-reliance on persuasive outputs, unclear responsibility for harmful recommendations, and the handling of highly sensitive patient information [290]. In real clinical deployment, this uncertainty may also create legal liability concerns when harmful outputs influence decisions, particularly if the boundaries between clinician judgment, institutional governance, and system recommendations are not clearly defined. Med-LLMs are increasingly used in settings that involve vulnerable patients and high-stakes clinical decisions. Their use also raises strict requirements for privacy and professional accountability [290-292]. In patient-facing settings, these requirements extend to informed consent, because patients should understand when LLM-based systems are involved in communication, documentation, or decision support, as well as the limits of those systems in clinical care.

These concerns do not appear in the same way across clinical applications. In mental health settings, safety concerns are closely related to vulnerable users and emotionally sensitive interactions [287]. Model responses may seem supportive but remain clinically inappropriate. In telemedicine and operational medical support, privacy and security can be more difficult to manage, especially in remote or military settings [292,293]. These environments often involve sensitive health data, limited communication infrastructure, and operationally sensitive information. They also increase the risk of bias and uneven performance [294]. The populations, workflows, and resource conditions encountered in military or austere care settings may differ from those reflected in model training data [151]. Across these settings, accuracy alone is not sufficient for responsible clinical use. Med-LLMs must also protect patients, maintain confidentiality, and reduce the risk of unfair or unsafe outcomes [288,290,294].

Recent regulatory developments have made these governance requirements more explicit [295-297]. The European Commission notes that AI-based software intended for medical purposes may fall within the AI Act's high-risk framework [298]. In this context, Med-LLMs are expected to meet stricter requirements for safety, transparency, human oversight, and system security [288,296]. Recent guidance on the AI Act suggests that compliance for high-risk medical AI will require stronger attention to documentation, lifecycle management, and human oversight [295,296,298]. In practice, this also points to the need for clearer deployment documentation, such as model cards or related reporting standards, to specify intended use, known limitations, validation scope, and oversight requirements in clinical settings

[46,82,235]. For Med-LLMs, this points to a broader need for traceability and governance throughout design, validation, and real-world use [287,291]. At the same time, regulatory expectations remain uneven across jurisdictions and use cases [295,296,298].

**Robustness, generalization, and distribution shift**

Robustness and generalization are difficult to ensure in Med-LLMs, because strong benchmark performance does not necessarily translate into stable behavior in real clinical settings [299,300]. In practice, training, evaluation, and deployment often take place under different data distributions, creating the risk of distribution shift [301]. This issue is especially important in medicine, where variation in patients, documentation practices, or institutional workflows can change model behavior and affect clinical use. Recent work in medical AI has highlighted distribution shift as a persistent obstacle to robust clinical deployment [301,302].

This problem becomes clearer when viewed across several representative applications discussed earlier [84]. In medical language translation, model performance may vary across languages, patient expressions, and levels of resource availability [160]. In multimodal tasks, robustness may weaken with changes in image quality, acquisition settings, or accompanying documentation, even if benchmark performance appears strong [303,304]. Mental health applications make this problem especially visible. Model behavior can shift when the system is used in populations or interaction settings that differ from those seen during training [237,305]. A similar issue arises in telemedicine and operational medical support, where remote or austere care environments may differ from benchmark settings in patient populations, workflows, and resource conditions [302]. These examples suggest that distribution shift is not limited to any single task. It is a recurring challenge that can undermine both generalization and real-world reliability.

Several strategies may improve robustness, including external validation across institutions, scenario-based testing, and evaluation of model uncertainty [302]. These steps are useful, but they do not eliminate the problem. Clinical data, patient populations, and local workflows continue to change, and many deployment environments are still poorly represented in available training and evaluation resources [302]. Strong benchmark performance does not guarantee stable behavior in real clinical use [300]. The robustness of Med-LLMs must be assessed under changing conditions rather than inferred from results obtained in limited settings. This challenge is also closely related to model updating and maintenance, because changes in knowledge and deployment conditions can directly affect robustness over time.

**Knowledge updating, model maintenance, and lifecycle challenges**

Medical knowledge is not static, and this creates a distinct challenge for keeping Med-LLMs aligned with current evidence [306]. Unlike the cross-setting mismatch discussed above, the central issue here is change over time. Clinical recommendations may be updated, new safety information may emerge, and institutional practices may gradually shift [306]. As a result, a model that performs well at one point may become less reliable as external knowledge, professional standards, and local workflows evolve [307,308]. Recent work on real-world clinical AI deployment has emphasized that LLM-based systems cannot be treated as fixed tools once they enter practice [307].

The impact of this problem can be seen in several common medical tasks. In clinical decision support and medical question answering, models may continue to produce answers that no longer match current evidence or updated guidelines [306,309]. In documentation and summarization tasks, changes in note structure, coding practice, or institutional workflow can make existing model outputs less useful [310]. Similar difficulties can also appear in retrieval-supported systems [119,273]. Access to external sources may improve timeliness, but reliability still depends on whether the underlying knowledge base is complete, well-maintained, and regularly updated. These examples show that knowledge updating and model maintenance are not peripheral engineering concerns [273,309,311]. They directly affect whether Med-LLMs remain clinically reliable after deployment [312].

Several strategies may help reduce these risks, including post-deployment monitoring, version control, and retrieval-based access to updated information [313]. Keeping a medical model aligned with current knowledge requires substantial time, cost, and validation effort [312]. Updating a system may involve new data curation, repeated testing, and additional clinical review before the revised model can be used with confidence [307]. Model updates can also introduce new problems. They may improve performance on the target problem while weakening performance on other tasks [314,315]. For Med-LLMs, long-term reliability depends on continued access to current knowledge and on the ability to introduce changes in a controlled and clinically accountable manner [311,312]. Knowledge updating and model maintenance, therefore, remain closely linked to oversight,

validation, and workflow integration.

### Human oversight, workflow integration, and behavior alignment

Med-LLMs may show strong performance in evaluation settings, but safe clinical use still depends on human oversight and effective workflow design [316,317]. In medicine, Med-LLMs should be used in settings that allow clinicians to review model outputs, make corrections, and override suggestions when necessary [318]. Their value depends on output quality, compatibility with existing patterns of care, and continued clinician control over their use [319]. Med-LLMs should function as tools within clinical systems rather than as independent decision makers. They must also behave in ways that are appropriate for clinical use, including acknowledging uncertainty, avoiding unwarranted confidence, and supporting escalation when human intervention is needed [320,321].

These concerns become clearer in common clinical tasks. In clinical decision support, model suggestions may assist reasoning, but interpretation and final judgment must remain with clinicians [318]. In documentation and summarization tasks, the value of model outputs depends on whether they can be effectively reviewed, refined, and incorporated into established clinical documentation practices [322]. Mental health applications make the need for oversight especially clear. Responses may sound supportive while remaining clinically inappropriate, and safe use depends on timely supervision and clear referral mechanisms [272]. In mental health settings, human-in-the-loop oversight may be most appropriate when implemented through tiered review and escalation frameworks [236,238]. For example, Med-LLMs may be limited to low-risk supportive functions such as psychoeducation, documentation drafting, or summarization, while outputs involving diagnostic interpretation, medication-related suggestions, crisis-related language, or high-confidence therapeutic advice should be routed to qualified clinicians for review before use [238,244,272]. Such frameworks may also combine uncertainty signaling, audit trails, and referral-first design to reduce the risk that fluent but unsupported responses are treated as clinically appropriate guidance [235,279]. Several challenges arise in telemedicine and operational medical support, where remote settings can make supervision and workflow coordination more difficult [261,318]. In these settings, the same output may be helpful in one context but unsuitable in another [320]. Across these tasks, behavior alignment involves more than accurate output. It also depends on whether the model responds cautiously, defers when appropriate, and remains compatible with the clinical setting.

Several approaches may help address these challenges. Human-in-the-loop review, clear escalation rules, and workflow-aware deployment can improve oversight in practice [316,323]. However, supervision is not achieved simply by asking clinicians to check model outputs after they are generated. Human review may become superficial if model outputs enter the workflow at the wrong point, prove difficult to judge, or gain acceptance too quickly [322]. For Med-LLMs, behavior alignment needs to be judged in real clinical use rather than assumed from model design alone. A clinically aligned system should support care in a way that clinicians can understand, question, and incorporate into existing workflows [46,324]. In practical deployment, this reviewability may need to be supported by explanation mechanisms such as evidence traceability, source-linked outputs, or attribution-based analysis of influential inputs. In hybrid Med-LLMs that incorporate structured clinical variables, explainable AI techniques such as SHAP may also help reveal which factors most strongly influence model-supported risk estimates or recommendation components [325]. It should also express uncertainty in a usable way, avoid overconfident recommendations, and support timely referral or escalation in higher-risk situations [133]. In this sense, the effective clinical use of Med-LLMs depends on durable forms of supervision and workflow integration, rather than stronger models alone.

## Future work

Med-LLMs have shown broad promise across a range of medical applications, but future progress will depend on more than further gains in model capability. The next stage of development should place greater emphasis on clinically meaningful evidence and validation, grounded and adaptive Med-LLMs, human-centered clinical integration, and deployment across diverse settings. These priorities reflect a broader shift from isolated performance gains toward systems that can be evaluated, updated, supervised, and used responsibly in clinical practice. Future work should focus on making these systems more reliable, better aligned with clinical needs, and suitable for sustained use in real clinical settings.

### Clinical evidence and validation

Future work on Med-LLMs should focus less on extending benchmark coverage and more on building clinically meaningful evidence. Benchmark evaluation and retrospective assessment remain useful, especially for early comparison across models and tasks, but they should not be treated as the main measure of progress [197]. High performance on medical examination questions or isolated task benchmarks does not

mean that a model is ready for clinical use [326]. Clinical adoption requires evidence that is closer to real tasks, real users, and real care processes. Recent studies already suggest this shift [279,319,326]. Some use blinded physician review to assess complex clinical outputs, while others examine model performance in practical clinical tasks such as documentation and decision support [279].

A more mature validation pathway for Med-LLMs should combine several forms of evidence, rather than rely on any single level of assessment. Task-level evaluation will remain necessary, including benchmark results, scenario-based testing, and expert review, but these should be treated as an entry point rather than the final basis for adoption. They can show whether a model performs well on a defined task, but they are less able to show how that performance will translate into routine clinical use. Future work should also examine how models perform within clinical workflows such as documentation, summarization, triage, referral, and decision support, where value depends on timing, interpretability, and fit with existing care processes [31,326]. This level of evidence is especially important because many Med-LLM applications are intended to support clinical work rather than to function as isolated prediction tools. A further step is to build stronger prospective and real-world evidence through multi-center validation, post-deployment monitoring, and outcome-aware assessment when feasible [19]. Clinical validation should be understood not simply as improved model testing, but as the generation of evidence needed to support safe, effective, and sustained use in clinical practice.

**Grounded and adaptive medical large language models**

Future Med-LLMs should be developed as systems that are both grounded in reliable medical evidence and able to adapt to changing clinical knowledge [119]. In medical settings, model outputs should not rely only on internal parametric memory [306]. Medical knowledge changes over time, and clinical practice may also vary across different institutions. A more grounded system can link its responses to retrievable and reviewable sources, which makes clinical checking easier and helps reduce unsupported outputs. At the same time, future systems also need to remain adaptive by staying aligned with updated evidence, institutional protocols, and task requirements over time. Recent reviews of RAG in healthcare have highlighted the same need, showing that Med-LLMs require stronger support from external knowledge sources if they are to remain reliable and usable in practice [92,273]. Related work on dynamic deployment has likewise shown that medical AI systems, especially LLM-based systems, should not be treated as fixed tools once they enter clinical use [306,307,312].

Several future directions follow from this shift. One priority is better integration of external medical knowledge, including clinical guidelines, curated institutional resources, and other high-quality reference sources [92]. Retrieval-augmented approaches are already moving in this direction. Recent medical RAG studies suggest that their value extends beyond factual accuracy to source traceability and long-term maintainability [119,271]. Another priority is to make Med-LLMs easier to update safely. Model adaptation should support version control, revalidation, and traceable changes as knowledge and practice evolve, rather than relying on static model releases [308]. In practice, the timing of such updates should be guided by changes in clinical guidelines, institutional protocols, and local governance requirements rather than by a fixed technical schedule. The third priority is to extend grounding beyond text-only retrieval. Future systems will likely need stronger task-specific grounding for documentation, decision support, and multimodal clinical reasoning, especially when outputs depend on structured records or image-based evidence. Recent work on iterative medical RAG and hybrid systems that combine fine-tuning with RAG also points in the same direction [123]. Future progress may depend less on any single technique than on system designs that better integrate external knowledge, task adaptation, and ongoing validation. In this sense, the goal is not simply to build stronger Med-LLMs, but to build systems that remain clinically reliable as evidence, workflows, and care environments continue to change.

**Human-centered clinical integration**

Clinical integration will be a major determinant of whether Med-LLMs become useful in real medical settings. In many medical tasks, the real value of Med-LLMs depends on how its output integrates into clinical workflows, rather than just the capabilities of the model itself [326]. A system may perform well in individual evaluations, but its clinical value remains limited if it reaches the wrong users, enters the workflow at the wrong stage, or creates an additional review burden. Future research should address how Med-LLMs can support clinical roles, integrate into routine workflows, and operate under appropriate human oversight. However, recent studies suggest that this type of clinical integration remains limited in practice [316,319,327]. For example, in tasks such as clinical documentation or decision support, the usefulness of Med-LLMs often depends on how well they fit existing workflows rather than on model performance alone [40].

The more useful direction for future development is to

design Med-LLMs around specific forms of clinical collaboration rather than around generic model capability. Systems designed for different clinical tasks often require different outputs, interface structures, and levels of autonomy [328]. In clinical use, clinicians need to recognize uncertainty, revise model suggestions when necessary, and retain control over final decisions. This also means that workflow design should be treated as a central part of model development rather than as a later implementation step [329]. Med-LLMs are more likely to be adopted when they support review, revision, handoff, and traceability within existing clinical processes, instead of adding a separate layer of work. In this sense, future progress will depend on developing forms of human-AI collaboration that can operate safely and responsibly in clinical practice.

**Deployment across diverse clinical settings**

Deployment of Med-LLMs will require more than evidence from a single institution or a small set of standardized settings [330]. Real clinical environments differ in patient populations, languages, workflows, infrastructure, and organizational capacity. As a result, a system that appears useful in one setting may not transfer smoothly to another. Therefore, future deployment of Med-LLM systems should account for heterogeneity rather than assuming that clinical environments are broadly uniform. Recent implementation reviews show that real-world integration of LLMs in clinical workflows remains limited and uneven, while work on global health equity has highlighted how current development and deployment remain concentrated in high-income settings [316,326,331]. This observation also highlights an important issue for future research. Effective deployment depends on model quality as well as careful evaluation and adaptation across different clinical settings.

Several issues should be addressed to support deployment across diverse clinical settings. Broader external validation will be essential before large-scale adoption, particularly across different institutions and patient populations [93]. Deployment planning should also incorporate fairness and equity from the outset rather than treating them as secondary considerations. Future systems should be assessed for uneven performance across demographic groups, languages, and care environments, especially in settings that are underrepresented in current development pipelines [332]. Deployment infrastructure should account for privacy, security, and governance requirements [291]. In practice, implementation often depends as much on organizational capacity as on technical performance. In remote military or other austere settings, future deployment may also require lightweight Med-LLMs that can be integrated with edge computing infrastructure to support real-time applications under constrained connectivity and hardware conditions. Such designs may be particularly important for field hospitals, deployed units, and other operational environments where low latency, local processing, and rapid clinical response are necessary. In such environments, continual adaptation may need to rely less on frequent full-model retraining and more on lower-cost strategies such as retrieval-based knowledge updating, parameter-efficient adaptation, and centralized revalidation, so that edge-side systems can remain lightweight while still benefiting from controlled model maintenance. In addition, Med-LLMs intended for real-world use should remain maintainable after deployment. Grounded systems need continued monitoring and updating if they are to remain reliable as evidence, workflows, and local practice continue to change. The long-term goal is to support context-aware deployment that remains reliable and appropriate across diverse clinical settings, rather than assuming that systems can be applied uniformly.

## Conclusions

Med-LLMs have shown broad potential across a range of clinical tasks, including decision support, documentation, and patient communication. At the same time, their clinical value cannot be determined by model capability alone. Current evidence suggests that reliable use of Med-LLMs in practice depends on more than performance in controlled settings. It relies on the strength of supporting evidence, alignment with current medical knowledge, and effective integration into clinical workflows. By bringing together recent work on applications, challenges, and emerging directions, this review highlights the key conditions for meaningful use of Med-LLMs in clinical practice.

The development of Med-LLMs is increasingly centered on clinical validity, system design, and effective use in practice, rather than on model performance itself. Future advances will rely on stronger clinical evidence, better alignment with evolving knowledge, and more effective integration into routine care. It will also require approaches to deployment that account for differences across institutions, patient populations, and resource conditions. The focus should be on supporting systems that remain reliable in clinical practice while adapting over time and being used under appropriate oversight. Med-LLMs will have clinical impact only if they can function as part of a broader clinical system that remains accountable, maintainable, and responsive to real clinical needs.

### Abbreviations

AI: Artificial intelligence
EHR: Electronic health record
LLMs: Large language models
Med-LLMs: Medical large language models
MoE: Mixture of experts
NLP: Natural language processing
PEFT: Parameter-efficient fine-tuning
RAG: Retrieval-augmented generation
SFT: Supervised fine-tuning
SSMs: State space models
TCM: Traditional Chinese medicine

### Acknowledgements

GPT-5.4 and Gemini 3.1-Pro were used solely to improve the readability and grammar of the text. All scientific content, interpretations, and conclusions were developed and verified by the authors.

### Authors' contributions

YS, PC, TT, and KFL conceived and designed the topic of this review. YS and LY drafted the manuscript. YS, LY, ZL, XZ, AK, LSW, and YY performed the literature search and review. YS prepared the figures and tables. PC, TT, KFL, YL, TZ, SSG, ZW, RH, KL, and SKI reviewed and revised the manuscript. All authors contributed to the interpretation and final preparation of the manuscript. All authors read and approved the final manuscript.

### Funding

This work was supported by the Macao Polytechnic University (RP/FCA-14/2023), the Science and Technology Development Funds (FDCT) of Macao (0033/2023/RIB2), the Joint Research Funding Program between FDCT and the Department of Science and Technology of Guangdong Province (FDCT-GDST) (0009/2024/AGJ), the Joint Research Fund between FDCT and the Ministry of Science and Technology of China (FDCT-MOST) (0106/2025/AMJ) and the Anusandhan National Research Foundation (ANRF) under the Partnerships for Accelerated Innovation and Research (PAIR) programme under the sanction order (ANRF/PAIR/2025/000029/PAIR).

### Availability of data and materials

Not applicable.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[1]Faculty of Applied Sciences, Macao Polytechnic University, Macau 999078, Macao SAR, China. [2]Department of Oncology, Shengjing Hospital of China Medical University, Shenyang 110004, China. [3]Department of Computer Science and Technology, School of Engineering and Technology, Central University of Punjab, Bathinda, Punjab 151401, India. [4]Division of Hematology and Oncology, Mayo Clinic, Jacksonville, FL 32224, USA. [5]Department of Computer Science, School of Mathematics and Computing Sciences, Rani Channamma University, Belagavi, Karnataka 590016, India. [6]Department of Radiology, the Netherlands Cancer Institute, Amsterdam 1066 CX, the Netherlands. [7]Faculty of Health and Life Sciences, INTI International University, Nilai, Negeri Sembilan 71800, Malaysia. [8]School of Nursing, Peking University, Beijing 100191, China. [9]Guangdong-Hong-Kong-Macao University Joint Laboratory of Interventional Medicine, the Fifth Affiliated Hospital, Sun Yat-Sen University, Zhuhai 519000, Guangdong, China. [10]Department of General Surgery and Institute of Precision Diagnosis and Treatment of Digestive System Tumors, Carson International Cancer Center, Shenzhen University General Hospital, Shenzhen University, Shenzhen 518055, Guangdong, China. [11]Medicine and Engineering Interdisciplinary Research Laboratory of Nursing and Materials, West China Hospital, Sichuan University, Chengdu 610041, China. [12]Institute of Biomedical Engineering, College of Medicine, Southwest Jiaotong University, Chengdu 610031, China. [13]Informatics Engineering Department, Faculty of Science and Technology, University of Coimbra, 3030-290 Coimbra, Portugal.

### Reference

1. Van Noorden R, Webb R. ChatGPT and science: the AI system was a force in 2023 – for good and bad. Nature. 2023;624(7992):509. https://doi.org/10.1038/d41586-023-03930-6.
2. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, *et al*. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9. https://doi.org/10.1038/s41586-021-03819-2.
3. Chen Q, Hu Y, Peng X, Xie Q, Jin Q, Gilson A, *et al*. Benchmarking large language models for biomedical natural language processing applications and recommendations. Nat Commun. 2025;16(1):3280. https://doi.org/10.1038/s41467-025-56989-2.
4. Liu F, Zhou H, Gu B, Zou X, Huang J, Wu J, *et al*. Application of large language models in medicine. Nat Rev Bioeng. 2025;3(6):445-64. https://doi.org/10.1038/s44222-025-00279-5.
5. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. Lancet Digit Health. 2024;6(9):e662-72. https://doi.org/10.1016/S2589-7500(24)00124-9.
6. Team G, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, *et al*. Gemini: a family of highly capable multimodal models. Preprint. arXiv; 2023. https://doi.org/10.48550/arXiv.2312.11805.
7. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, *et al*. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature. 2025;645(8081):633-8. https://doi.org/10.1038/s41586-025-09422-z.
8. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, *et al*. LLaMA: open and efficient foundation language models. Preprint. arXiv; 2023. https://doi.org/10.48550/arXiv.2302.13971.
9. Yang A, Li A, Yang B, Zhang B, Hui B, Zheng B, *et al*. Qwen3 technical report. Preprint. arXiv; 2025. https://doi.org/10.48550/arXiv.2505.09388.
10. Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, *et al*. Training compute-optimal large language models. Preprint. arXiv; 2022. https://doi.org/10.48550/arXiv.2203.15556.
11. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, *et al*. LLaMA 2: open foundation and fine-tuned chat models. Preprint. arXiv; 2023. https://doi.org/10.48550/arXiv.2307.09288.

12. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, *et al*. The LLaMA 3 herd of models. Preprint. arXiv; 2024. https://doi.org/10.48550/arXiv.2407.21783.
13. Liu A, Feng B, Xue B, Wang B, Wu B, Lu C, *et al*. DeepSeek-V3 technical report. Preprint. arXiv; 2024. https://doi.org/10.48550/arXiv.2412.19437.
14. Fedus W, Zoph B, Shazeer N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. J Mach Learn Res. 2022;23(1):5232-70. https://jmlr.org/papers/v23/21-0998.html.
15. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, *et al*. Retrieval-augmented generation for knowledge-intensive NLP tasks. Proceedings of Advances in Neural Information Processing Systems. 2020. p. 9459-74. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.
16. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, *et al*. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. https://doi.org/10.1038/s41586-023-06291-2.
17. Riedemann L, Labonne M, Gilbert S. The path forward for large language models in medicine is open. NPJ Digit Med. 2024;7(1):339. https://doi.org/10.1038/s41746-024-01344-w.
18. de Hond A, Leeuwenberg T, Bartels R, van Buchem M, Kant I, Moons KG, *et al*. From text to treatment: the crucial role of validation for generative large language models in health care. Lancet Digit Health. 2024;6(7):e441-3. https://doi.org/10.1016/S2589-7500(24)00111-0.
19. Chen SF, Alyakin A, Seas A, Yang E, Choi JJ, Lee JV, *et al*. LLM-assisted systematic review of large language models in clinical medicine. Nat Med. 2026;32(3):1152-9. https://doi.org/10.1038/s41591-026-04229-5.
20. Li H, Fu JF, Python A. Implementing large language models in health care: clinician-focused review with interactive guideline. J Med Internet Res. 2025;27:e71916. https://doi.org/10.2196/71916.
21. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. J Am Med Inform Assoc. 2024;31(9):1833-43. https://doi.org/10.1093/jamia/ocae045.
22. Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BB, Wang B. Clinical camel: an open expert-level medical language model with dialogue-based knowledge encoding. Preprint. arXiv; 2023. https://doi.org/10.48550/arXiv.2305.12031.
23. Wang F, Wan X, Sun R, Chen J, Arik SO. Astute RAG: overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2025. p. 30553-71. https://doi.org/10.18653/v1/2025.acl-long.1476.
24. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. Cureus. 2023;15(6):e40895. https://doi.org/10.7759/cureus.40895.
25. Han T, Adams LC, Papaioannou JM, Grundmann P, Oberhauser T, Figueroa A, *et al*. MedAlpaca-an open-source collection of medical conversational AI models and training data. Preprint. arXiv; 2023. https://doi.org/10.48550/arXiv.2304.08247.
26. Xie Q, Chen Q, Chen A, Peng C, Hu Y, Lin F, *et al*. Medical foundation large language models for comprehensive text analysis and beyond. NPJ Digit Med. 2025;8(1):141. https://doi.org/10.1038/s41746-025-01533-1.
27. Jiang S, Wang Y, Song S, Hu T, Zhou C, Pu B, *et al*. Hulu-med: a transparent generalist model towards holistic medical vision-language understanding. Preprint. arXiv; 2025. https://arxiv.org/abs/2510.08668.
28. Sellergren A, Kazemzadeh S, Jaroensri T, Kiraly A, Traverse M, Kohlberger T, *et al*. MedGemma technical report. Preprint. arXiv; 2025. https://doi.org/10.48550/arXiv.2507.05201.
29. Wang X, Cui Y, Wang J, Zhang F, Wang Y, Zhang X, *et al*. Multimodal learning with next-token prediction for large multimodal models. Nature. 2026;650(8101):327-33. https://doi.org/10.1038/s41586-025-10041-x.
30. Chen M, Wu Y, Ma J, Jia X, Gao C, Zhao F, *et al*. Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. NPJ Digit Med. 2026;9(1):222. https://doi.org/10.1038/s41746-026-02409-8.
31. Agrawal M, Chen IY, Gulamali F, Joshi S. The evaluation illusion of large language models in medicine. NPJ Digit Med. 2025;8(1):600. https://doi.org/10.1038/s41746-025-01963-x.
32. Mello MM, Cohen IG. Regulation of health and health care artificial intelligence. JAMA. 2025;333(20):1769-70. https://doi.org/10.1001/jama.2025.3308.
33. Mathes S, Ferber D, Dreyer T, Borm KJ, Modersohn L, Willem T, *et al*. Collaborative framework on responsible AI in LLM-driven CDSS for precision oncology leveraging real-world patient data. NPJ Precis Oncol. 2025;10(1):15. https://doi.org/10.1038/s41698-025-01180-5.
34. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40. https://doi.org/10.1038/s41591-023-02448-8.
35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al*. Attention is all you need. Proceedings of Advances in Neural Information Processing Systems. 2017. p. 6000-10. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
36. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, *et al*. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-65. https://doi.org/10.1038/s41586-023-05881-4.
37. Hussein R, Zink A, Ramadan B, Howard FM, Hightower M, Shah S, *et al*. Advancing healthcare AI governance through a comprehensive maturity model based on systematic review. NPJ Digit Med. 2026;9(1):236. https://doi.org/10.1038/s41746-026-02418-7.
38. Bean AM, Payne RE, Parsons G, Kirk HR, Ciro J, Mosquera-Gómez R, *et al*. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. Nat Med. 2026;32(2):609-15. https://doi.org/10.1038/s41591-025-04074-y.
39. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, *et al*. Scaling laws for neural language models. Preprint. arXiv; 2020. https://doi.org/10.48550/arXiv.2001.08361.
40. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, *et al*. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(9):2613-22. https://doi.org/10.1038/s41591-024-03097-1.
41. Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. NPJ Digit Med. 2025;8(1):143. https://doi.org/10.1038/s41746-025-01476-7.
42. Team G, Georgiev P, Lei VI, Burnell R, Bai L, Gulati A, *et al*. Gemini 1.5: unlocking multimodal understanding across millions of

tokens of context. Preprint. arXiv; 2024. https://doi.org/10.48550/arXiv.2403.05530.
43. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, *et al*. GPT-4 technical report. Preprint. arXiv; 2024. https://doi.org/10.48550/arXiv.2303.08774.
44. Nerella S, Bandyopadhyay S, Zhang J, Contreras M, Siegel S, Bumin A, *et al*. Transformers and large language models in healthcare: a review. Artif Intell Med. 2024;154:102900. https://doi.org/10.1016/j.artmed.2024.102900.
45. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, *et al*. A comprehensive overview of large language models. ACM Trans Intell Syst Technol. 2025;16(5):106:1-106:72. https://doi.org/10.1145/3744746.
46. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, *et al*. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60-9. https://doi.org/10.1038/s41591-024-03425-5.
47. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, Arx S von, *et al*. On the opportunities and risks of foundation models. Preprint. arXiv; 2022. https://doi.org/10.48550/arXiv.2108.07258.
48. Liao Z, Wang J, Yu H, Wei L, Li J, Wang J, *et al*. E2LLM: encoder elongated large language models for long-context understanding and reasoning. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2025. p. 19201-30. https://doi.org/10.18653/v1/2025.emnlp-main.970.
49. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, *et al*. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell. 2022;45(1): 87-110. https://doi.org/10.1109/TPAMI.2022.3152247.
50. Teo ZL, Thirunavukarasu AJ, Elangovan K, Cheng H, Moova P, Soetikno B, *et al*. Generative artificial intelligence in medicine. Nat Med. 2025;31(10):3270-82. https://doi.org/10.1038/s41591-025-03983-2.
51. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. 2019. p. 4171-86. https://doi.org/10.18653/v1/N19-1423.
52. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, *et al*. RoBERTa: a robustly optimized BERT pretraining approach. Preprint. arXiv; 2019. https://doi.org/10.48550/arXiv.1907.11692.
53. He P, Liu X, Gao J, Chen W. Deberta: decoding-enhanced bert with disentangled attention. Preprint. arXiv; 2020. https://doi.org/10.48550/arXiv.2006.03654.
54. Clark K, Luong MT, Le QV, Manning CD. ELECTRA: pre-training text encoders as discriminators rather than generators. Preprint. arXiv; 2020. https://doi.org/10.48550/arXiv.2003.10555.
55. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, *et al*. Language models are few-shot learners. Proceedings of Advances in Neural Information Processing Systems. 2020. p. 1877-901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.
56. Gemma Team, Kamath A, Ferret J, Pathak S, Vieillard N, Merhej R, *et al*. Gemma 3 technical report. Preprint. arXiv; 2025. https://doi.org/10.48550/arXiv.2503.19786.
57. GLM-5-Team, Zeng A, Lv X, Hou Z, Du Z, Zheng Q, *et al*. GLM-5: from vibe coding to agentic engineering. Preprint. arXiv; 2026. https://doi.org/10.48550/arXiv.2602.15763.
58. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, *et al*. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2020. p. 7871-80. https://doi.org/10.18653/v1/2020.acl-main.703.
59. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1-67. https://jmlr.org/papers/v21/20-074.html.
60. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, *et al*. Scaling instruction-finetuned language models. J Mach Learn Res. 2024; 25(70):1-53. https://jmlr.org/papers/v25/23-0870.html.
61. Tay Y, Dehghani M, Tran VQ, Garcia X, Wei J, Wang X, *et al*. UL2: unifying language learning paradigms. International Conference on Learning Representations. 2023. https://openreview.net/forum?id=6ruVLB727MC.
62. Warner B, Chaffin A, Clavié B, Weller O, Hallström O, Taghadouini S, *et al*. Smarter, better, faster, longer: a modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2025. p. 2526-47. https://doi.org/10.18653/v1/2025.acl-long.127.
63. Lee SA, Wu A, Chiang JN. Clinical ModernBERT: an efficient and long context encoder for biomedical text. Preprint. arXiv; 2025. https://doi.org/10.48550/arXiv.2504.03964.
64. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, *et al*. Current applications and challenges in large language models for patient care: a systematic review. Commun Med. 2025; 5(1):26. https://doi.org/10.1038/s43856-024-00717-2.
65. Rohekar RY, Gurwicz Y, Yu S, Aflalo E, Lal V. A causal world model underlying next token prediction: exploring GPT in a controlled environment. Proceedings of the International Conference on Machine Learning. 2025. p. 72196-209. https://proceedings.mlr.press/v267/yehezkel-rohekar25a.html.
66. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, *et al*. Toward expert-level medical question answering with large language models. Nat Med. 2025;31(3):943-50. https://doi.org/10.1038/s41591-024-03423-7.
67. Sha Y, Pan H, Meng W, Li K. Contrastive knowledge-guided large language models for medical report generation. Medical Image Computing and Computer Assisted Intervention. 2025. p. 111-20. https://doi.org/10.1007/978-3-032-04978-0_11.
68. Van Veen D, Van Uden C, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, *et al*. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30(4):1134-42. https://doi.org/10.1038/s41591-024-02855-5.
69. Guo M, Ainslie J, Uthus DC, Ontanon S, Ni J, Sung YH, *et al*. LongT5: efficient text-to-text transformer for long sequences. Findings of the Association for Computational Linguistics. 2022. p. 724-36. https://doi.org/10.18653/v1/2022.findings-naacl.55.
70. Yuan H, Yuan Z, Gan R, Zhang J, Xie Y, Yu S. BioBART: pretraining and evaluation of a biomedical generative language model. Proceedings of the Workshop on Biomedical Language Processing. 2022. p. 97-109. https://doi.org/10.18653/v1/2022.bionlp-1.9.
71. Phan LN, Anibal JT, Tran H, Chanana S, Bahadroglu E, Peltekian A, *et al*. SciFive: a text-to-text transformer model for biomedical literature. Preprint. arXiv; 2021. https://doi.org/10.48550/arXiv.2106.03598.
72. Xu R, Jiang P, Luo L, Xiao C, Cross A, Pan S, *et al*. A survey on unifying large language models and knowledge graphs for

biomedicine and healthcare. Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2025. p. 6195-205. https://doi.org/10.1145/3711896.3736556.

73. Bahri Y, Dyer E, Kaplan J, Lee J, Sharma U. Explaining neural scaling laws. Proc Natl Acad Sci U S A. 2024;121(27):e2311878121. https://doi.org/10.1073/pnas.2311878121.
74. Chen Z, Wang S, Xiao T, Wang Y, Chen S, Cai X, *et al*. Revisiting scaling laws for language models: the role of data quality and training strategies. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2025. p. 23881-99. https://doi.org/10.18653/v1/2025.acl-long.1163.
75. Luo J, Wu B, Luo X, Xiao Z, Jin Y, Tu RC, *et al*. A survey on efficient large language model training: from data-centric perspectives. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2025. p. 30904-20. https://doi.org/10.18653/v1/2025.acl-long.1493.
76. Cherti M, Beaumont R, Wightman R, Wortsman M, Ilharco G, Gordon C, *et al*. Reproducible scaling laws for contrastive language-image learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 2818-29. https://doi.org/10.1109/CVPR52729.2023.00276.
77. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, *et al*. Emergent abilities of large language models. Preprint. arXiv; 2022. https://doi.org/10.48550/arXiv.2206.07682.
78. Schaeffer R, Miranda B, Koyejo S. Are emergent abilities of large language models a mirage? Proceedings of Advances in Neural Information Processing Systems. 2023. p. 55565-81. https://proceedings.neurips.cc/paper_files/paper/2023/hash/adc98a266f45005c403b8311ca7e8bd7-Abstract-Conference.html.
79. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, *et al*. Training language models to follow instructions with human feedback. Proceedings of Advances in Neural Information Processing Systems. 2022;27730-44. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.
80. Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. Proceedings of Advances in Neural Information Processing Systems. 2023. p. 53728-41. https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.
81. Zhang S, Dong L, Li X, Zhang S, Sun X, Wang S, *et al*. Instruction tuning for large language models: a survey. ACM Comput Surv. 2026;58(7):1-36. https://doi.org/10.1145/3777411.
82. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, *et al*. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. 2024;7(1):258. https://doi.org/10.1038/s41746-024-01258-7.
83. Zhou J, Li H, Chen S, Chen Z, Han Z, Gao X. Large language models in biomedicine and healthcare. NPJ Artif Intell. 2025;1(1):44. https://doi.org/10.1038/s44387-025-00047-1.
84. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models and agents in healthcare: key challenges in clinical applications. Intell Med. 2025;5(02):151-63. https://doi.org/10.1016/j.imed.2025.03.002.
85. Patro BN, Agneeswaran VS. Mamba-360: survey of state space models as transformer alternative for long sequence modelling: methods, applications, and challenges. Eng Appl Artif Intell. 2025;159:111279. https://doi.org/10.1016/j.engappai.2025.111279.
86. Lenz B, Lieber O, Arazi A, Bergman A, Manevich A, Peleg B, *et al*. Jamba: hybrid transformer-mamba language models. International Conference on Learning Representations. 2025. p. 67959-84. https://openreview.net/forum?id=JFPaD7lpBD.
87. Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. Preprint. arXiv; 2023. https://doi.org/10.48550/arXiv.2312.00752.
88. Wang X, Wang S, Ding Y, Li Y, Wu W, Rong Y, *et al*. State space model for new-generation network alternative to transformers: a survey. Preprint. arXiv; 2024. https://doi.org/10.48550/arXiv.2404.09516.
89. Jiang Z, Xu F, Gao L, Sun Z, Liu Q, Dwivedi-Yu J, *et al*. Active retrieval augmented generation. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2023. p. 7969-92. https://doi.org/10.18653/v1/2023.emnlp-main.495.
90. Borgeaud S, Mensch A, Hoffmann J, Cai T, Rutherford E, Millican K, *et al*. Improving language models by retrieving from trillions of tokens. Proceedings of the International Conference on Machine Learning. 2022. p. 2206-40. https://proceedings.mlr.press/v162/borgeaud22a.html.
91. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. PLOS Digit Health. 2025;4(6):e0000877. https://doi.org/10.1371/journal.pdig.0000877.
92. Ke YH, Jin L, Elangovan K, Abdullah HR, Liu N, Sia ATH, *et al*. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med. 2025;8(1):187. https://doi.org/10.1038/s41746-025-01519-z.
93. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, *et al*. Mixtral of experts. Preprint. arXiv; 2024. https://doi.org/10.48550/arXiv.2401.04088.
94. Cai W, Jiang J, Wang F, Tang J, Kim S, Huang J. A survey on mixture of experts in large language models. IEEE Trans Knowl Data Eng. 2025;37(7):3896-915. https://doi.org/10.1109/TKDE.2025.3554028.
95. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. 2025;32(4):605-15. https://doi.org/10.1093/jamia/ocaf008.
96. Luo MJ, Pang J, Bi S, Lai Y, Zhao J, Shang Y, *et al*. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. JAMA Ophthalmol. 2024;142(9):798-805. https://doi.org/10.1001/jamaophthalmol.2024.2513.
97. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, *et al*. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. 2025;333(4):319-28. https://doi.org/10.1001/jama.2024.21700.
98. Jia M, Duan J, Song Y, Wang J. MedIKAL: integrating knowledge graphs as assistants of LLMs for enhanced clinical diagnosis on EMRs. Proceedings of the International Conference on Computational Linguistics. 2025. p. 9278-98. https://aclanthology.org/2025.coling-main.624.
99. Seinen TM, Kors JA, van Mulligen EM, Rijnbeek PR. Using structured codes and free-text notes to measure information

complementarity in electronic health records: feasibility and validation study. J Med Internet Res. 2025;27:e66910. https://doi.org/10.2196/66910.

100. Wang X, Xiong Z, Zou K, Srinivasan S, Lo TWS, Wu Y, *et al*. Reasoning-driven large language models in medicine: opportunities, challenges, and the road ahead. Lancet Digit Health. 2026;8(1):100931. https://doi.org/10.1016/j.landig.2025.100931.
101. Zhong X, Li S, Chen Z, Ge L, Yu D, Wang S, *et al*. Considerations for patient privacy of large language models in health care: scoping review. J Med Internet Res. 2025;27:e76571. https://doi.org/10.2196/76571.
102. Ahadian P, Xu W, Liu D, Guan Q. Ethics of trustworthy AI in healthcare: challenges, principles, and practical pathways. Neurocomputing. 2026;661:131942. https://doi.org/10.1016/j.neucom.2025.131942.
103. Savage T, P Ma S, Boukil A, Rangan E, Patel V, Lopez I, *et al*. Fine-tuning methods for large language models in clinical medicine by supervised fine-tuning and direct preference optimization: comparative evaluation. J Med Internet Res. 2025;27:e76048. https://doi.org/10.2196/76048.
104. Tran H, Yang Z, Yao Z, Yu H. BioInstruct: instruction tuning of large language models for biomedical natural language processing. J Am Med Inform Assoc. 2024;31(9):1821-32. https://doi.org/10.1093/jamia/ocae122.
105. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, *et al*. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022;23(6):bbac409. https://doi.org/10.1093/bib/bbac409.
106. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, *et al*. A large language model for electronic health records. NPJ Digit Med. 2022;5(1):194. https://doi.org/10.1038/s41746-022-00742-2.
107. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, *et al*. A study of generative large language model for medical research and healthcare. NPJ Digit Med. 2023;6(1):210. https://doi.org/10.1038/s41746-023-00958-w.
108. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, *et al*. MEDITRON-70B: scaling medical pretraining for large language models. Preprint. arXiv; 2023. https://doi.org/10.48550/arXiv.2311.16079.
109. Bolton E, Venigalla A, Yasunaga M, Hall D, Xiong B, Lee T, *et al*. BioMedLM: a 2.7B parameter language model trained on biomedical text. Preprint. arXiv; 2024. https://doi.org/10.48550/arXiv.2403.18421.
110. Zhang H, Chen J, Jiang F, Yu F, Chen Z, Chen G, *et al*. HuatuoGPT, towards taming language model to be a doctor. Findings of the Association for Computational Linguistics. 2023. p. 10859-85. https://doi.org/10.18653/v1/2023.findings-emnlp.725.
111. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. Preprint. arXiv; 2023. https://doi.org/10.48550/arXiv.2306.09968.
112. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, *et al*. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. Proceedings of Advances in Neural Information Processing Systems. 2023. p. 28541-64. https://proceedings.neurips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.
113. Bannur S, Bouzid K, Castro DC, Schwaighofer A, Thieme A, Bond-Taylor S, *et al*. MAIRA-2: grounded radiology report generation. Preprint. arXiv; 2024. https://doi.org/10.48550/arXiv.2406.04449.
114. Shi W, Xu R, Zhuang Y, Yu Y, Sun H, Wu H, *et al*. MedAdapter: efficient test-time adaptation of large language models towards medical reasoning. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2024. p. 22294-314. https://doi.org/10.18653/v1/2024.emnlp-main.1244.
115. Christophe C, Kanithi P, Munjal P, Raha T, Hayat N, Rajan R, *et al*. Med42-evaluating fine-tuning strategies for medical LLMs: full-parameter vs. Parameter-efficient approaches. Preprint. arXiv; 2024. https://arxiv.org/abs/2404.14779.
116. Liu K, Wen C. MedLoRA: exploring LoRA and quantization combinations for medical QA. Proceedings of the International Conference on Big Data & Artificial Intelligence & Software Engineering. 2025. p. 437-45. https://ieeexplore.ieee.org/document/11181297.
117. He J, Li P, Liu G, Zhong S. Parameter-efficient fine-tuning medical multimodal large language models for medical visual grounding. Proceedings of the IEEE International Symposium on Biomedical Imaging. 2025. p. 1-5. https://doi.org/10.1109/ISBI60581.2025.10981029.
118. He J, Wang Y, Wang L, Lu H, He JY, Lan JP, *et al*. Multi-modal instruction tuned LLMs with fine-grained visual perception. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 13980-90. https://doi.org/10.1109/CVPR52733.2024.01326.
119. Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, *et al*. Almanac – retrieval-augmented language models for clinical medicine. NEJM AI. 2024;1(2). https://doi.org/10.1056/aioa2300068.
120. Zhao X, Liu S, Yang SY, Miao C. MedRAG: enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. Preprint. arXiv; 2025. https://doi.org/10.48550/arXiv.2502.04413.
121. Lu Y, Zhao X, Wang J. ClinicalRAG: enhancing clinical decision support through heterogeneous knowledge retrieval. Proceedings of the Workshop on Towards Knowledgeable Language Models. 2024. p. 64-8. https://doi.org/10.18653/v1/2024.knowllm-1.6.
122. Xiong G, Jin Q, Wang X, Zhang M, Lu Z, Zhang A. Improving retrieval-augmented generation in medicine with iterative follow-up questions. Proceedings of the Pacific Symposium on Biocomputing. 2025. p. 199-214. https://doi.org/10.1142/9789819807024_0015.
123. Chen Z, Liao Y, Jiang S, Wang P, Guo Y, Wang Y, *et al*. Towards omni-RAG: comprehensive retrieval-augmented generation for large language models in medical applications. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2025. p. 15285-309. https://doi.org/10.18653/v1/2025.acl-long.742.
124. Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. BioMistral: a collection of open-source pretrained large language models for medical domains. Findings of the Association for Computational Linguistics. 2024. p. 5848-64. https://doi.org/10.18653/v1/2024.findings-acl.348.
125. Wu C, Qiu P, Liu J, Gu H, Li N, Zhang Y, *et al*. Towards evaluating and building versatile large language models for medicine. NPJ Digit Med. 2025;8(1):58. https://doi.org/10.1038/s41746-024-01390-4.
126. Nazar W, Nazar G, Kamińska A, Danilowicz-Szymanowicz L. How to design, create, and evaluate an instruction-tuning dataset

for large language model training in health care: tutorial from a clinical perspective. J Med Internet Res. 2025;27:e70481. https://doi.org/10.2196/70481.

127. Xu L, Xie H, Qin SJ, Tao X, Wang FL. Parameter-efficient fine-tuning methods for pretrained language models: a critical review and assessment. IEEE Trans Pattern Anal Mach Intell. 2026;46(6):6107-26. https://doi.org/10.1109/TPAMI.2026.3657354.
128. Liu H, Chen Z, Li P, Liu YZ, Liu X, Xu RX, *et al*. Resource-efficient instruction tuning of large language models for biomedical named entity recognition. J Biomed Inform. 2025;170:104896. https://doi.org/10.1016/j.jbi.2025.104896.
129. Zhang X, Zhao J, Yang Z, Zhong Y, Guan S, Cao L, *et al*. UORA: uniform orthogonal reinitialization adaptation in parameter efficient fine-tuning of large models. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2025. p. 11709-28. https://doi.org/10.18653/v1/2025.acl-long.575.
130. Wang B, Xia I, Zhang Y, Wang J, Ouyang F, Han S, *et al*. From scores to steps: diagnosing and improving LLM performance in evidence-based medical calculations. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2025. p. 10809-33. https://doi.org/10.18653/v1/2025.emnlp-main.548.
131. Chang Q, Chen F, Chen Y, Cheng L, Dong D, Dong J, *et al*. 2025 Expert consensus on retrospective evaluation of large language model applications in clinical scenarios. Intell Med. 2025;5(4):318-30. https://doi.org/10.1016/j.imed.2025.09.001.
132. Sim SZY, Chen T. Critique of impure reason: unveiling the reasoning behaviour of medical large language models. eLife. 2025;14:e106187. https://doi.org/10.7554/eLife.106187.
133. Savage T, Wang J, Gallo R, Boukil A, Patel V, Safavi-Naini SAA, *et al*. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. J Am Med Inform Assoc. 2025;32(1):139-49. https://doi.org/10.1093/jamia/ocae254.
134. Fast D, Adams LC, Busch F, Fallon C, Huppertz M, Siepmann R, *et al*. Autonomous medical evaluation for guideline adherence of large language models. NPJ Digit Med. 2024;7(1):358. https://doi.org/10.1038/s41746-024-01356-6.
135. Liu GY, Yu D, Fan MM, Zhang X, Jin ZY, Tang C, *et al*. Antimicrobial resistance crisis: could artificial intelligence be the solution?. Mil Med Res. 2024;11(1):7. https://doi.org/10.1186/s40779-024-00510-1.
136. Woo BFY, Cato K, Cho H, You SB, Song J. The use of large language models in clinical documentation: a scoping review. Int J Nurs Stud. 2026;176:105322. https://doi.org/10.1016/j.ijnurstu.2025.105322.
137. Das BC, Amini MH, Wu Y. Security and privacy challenges of large language models: a survey. ACM Comput Surv. 2025;57(6):1-39. https://doi.org/10.1145/3712001.
138. Su H, Sun Y, Li R, Zhang A, Yang Y, Xiao F, *et al*. Large language models in medical diagnostics: scoping review with bibliometric analysis. J Med Internet Res. 2025;27:e72062. https://doi.org/10.2196/72062.
139. Wang W, Jin YH, Liu M, He Q, Xu JY, Wang MQ, *et al*. Guidance of development, validation, and evaluation of algorithms for populating health status in observational studies of routinely collected data (DEVELOP-RCD). Mil Med Res. 2024;11(1):52. https://doi.org/10.1186/s40779-024-00559-y.
140. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci. 2021;11(14):6421. https://doi.org/10.3390/app11146421.
141. Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: a dataset for biomedical research question answering. Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. 2019. p. 2567-77. https://doi.org/10.18653/v1/D19-1259.
142. Pal A, Umapathi LK, Sankarasubbu M. Medmcqa: a large-scale multi-subject multi-choice dataset for medical domain question answering. Proceedings of the Conference on Health, Inference, and Learning. 2022. p. 248-60. https://proceedings.mlr.press/v174/pal22a.html.
143. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, *et al*. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035. https://doi.org/10.1038/sdata.2016.35.
144. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, *et al*. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203-9. https://doi.org/10.1038/s41586-018-0579-z.
145. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, *et al*. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019;6(1):317. https://doi.org/10.1038/s41597-019-0322-0.
146. Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, *et al*. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc. 2016;23(2):304-10. https://doi.org/10.1093/jamia/ocv080.
147. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, *et al*. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence. 2019. p. 590-7. https://doi.org/10.1609/aaai.v33i01.3301590.
148. Wang X, Wang F, Li Y, Ma Q, Wang S, Jiang B, *et al*. CXPMRG-Bench: pre-training and benchmarking for X-ray medical report generation on CheXpert plus dataset. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025. p. 5123-33. https://doi.org/10.1109/CVPR52734.2025.00483.
149. Lau JJ, Gayen S, Ben Abacha A, Demner-Fushman D. A dataset of clinically generated visual questions and answers about radiology images. Sci Data. 2018;5:180251. https://doi.org/10.1038/sdata.2018.251.
150. He X, Zhang Y, Mou L, Xing E, Xie P. PathVQA: 30000+ questions for medical visual question answering. Preprint. arXiv; 2020. https://doi.org/10.48550/arXiv.2003.10286.
151. Hu Y, Li T, Lu Q, Shao W, He J, Qiao Y, *et al*. OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 22170-83. https://doi.org/10.1109/CVPR52733.2024.02093.
152. Jin Q, Wang Z, Floudas CS, Chen F, Gong C, Bracken-Clarke D, *et al*. Matching patients to clinical trials with large language models. Nat Commun. 2024;15(1):9074. https://doi.org/10.1038/s41467-024-53081-z.
153. Lan W, Wang W, Ji C, Yang G, Zhang Y, Liu X, *et al*. ClinicalGPT-R1: pushing reasoning capability of generalist disease diagnosis with large language model. Preprint. arXiv; 2025. https://doi.org/10.48550/arXiv.2504.09421.
154. Small WR, Austrian J, O'Donnell L, Burk-Rafel J, Hochman KA, Goodman A, *et al*. Evaluating hospital course summarization by an electronic health record-based large language model.

JAMA Netw Open. 2025;8(8):e2526339. https://doi.org/10.1001/jamanetworkopen.2025.26339.
155. Tanno R, Barrett DGT, Sellergren A, Ghaisas S, Dathathri S, See A, *et al*. Collaboration between clinicians and vision-language models in radiology report generation. Nat Med. 2025;31(2):599-608. https://doi.org/10.1038/s41591-024-03302-1.
156. Lopez I, Swaminathan A, Vedula K, Narayanan S, Nateghi Haredasht F, Ma SP, *et al*. Clinical entity augmented retrieval for clinical information extraction. NPJ Digit Med. 2025;8(1):45. https://doi.org/10.1038/s41746-024-01377-1.
157. Qiu P, Wu C, Zhang X, Lin W, Wang H, Zhang Y, *et al*. Towards building multilingual language model for medicine. Nat Commun. 2024;15(1):8384. https://doi.org/10.1038/s41467-024-52417-z.
158. Keles B, Gunay M, Caglar SI. LLMs-in-the-loop part-1: expert small AI models for bio-medical text translation. Preprint. arXiv; 2024. https://doi.org/10.48550/arXiv.2407.12126.
159. Salim MS, Fu L, Ramakrishnan AA, Yao Z, Yu H. MedCOD: enhancing English-to-Spanish medical translation of large language models using enriched chain-of-dictionary framework. Findings of the Association for Computational Linguistics. 2025. p. 6579-97. https://doi.org/10.18653/v1/2025.findings-emnlp.350.
160. Le-Duc K, Tran T, Tat BP, Bui NKH, Anh QD, Tran HP, *et al*. MultiMed-ST: large-scale many-to-many multilingual medical speech translation. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2025. p. 11838-963. https://doi.org/10.18653/v1/2025.emnlp-main.599.
161. Wang S, Zhao F, Bu D, Lu Y, Gong M, Liu H, *et al*. LINS: a general medical Q&A framework for enhancing the quality and credibility of LLM-generated responses. Nat Commun. 2025;16(1):9076. https://doi.org/10.1038/s41467-025-64142-2.
162. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, *et al*. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024;21(8):1470-80. https://doi.org/10.1038/s41592-024-02201-0.
163. Tripathi A, Waqas A, Schabath MB, Yilmaz Y, Rasool G. HONeYBEE: enabling scalable multimodal AI in oncology through foundation model-driven embeddings. NPJ Digit Med. 2025;8(1):622. https://doi.org/10.1038/s41746-025-02003-4.
164. Liu Z, Wu Y, Xu H, Wang M, Weng S, Pei D, *et al*. Multimodal fusion of radio-pathology and proteogenomics identify integrated glioma subtypes with prognostic and therapeutic opportunities. Nat Commun. 2025;16(1):3510. https://doi.org/10.1038/s41467-025-58675-9.
165. Yang K, Zhang T, Kuang Z, Xie Q, Huang J, Ananiadou S. MentaLLaMA: interpretable mental health analysis on social media with large language models. Proceedings of the ACM Web Conference. 2024. p. 4489-500. https://doi.org/10.1145/3589334.3648137.
166. Na H. CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation. 2024. p. 2930-40. https://aclanthology.org/2024.lrec-main.261.
167. Zhang W, Chen J, Zhu E, Cheng W, Li Y, Li Y, *et al*. MLlm-DR: towards explainable depression recognition with MultiModal large language models. ACM Trans Multimedia Comput Commun Appl. 2026;22(4):111:1-111:23. https://doi.org/10.1145/3796722.
168. Sha Y, Pan H, Luo G, Shi C, Chen W, Wang J, *et al*. MDD-thinker: a reasoning-enhanced large language model for diagnosis of major depressive disorder. J Affect Disord. 2026;403:121405. https://doi.org/10.1016/j.jad.2026.121405.
169. Jia Y, Ji X, Wang X, Zhang H, Meng Z, Zhang J, *et al*. Qibo: A large language model for traditional Chinese medicine. Expert Syst Appl. 2025;284:127672. https://doi.org/10.1016/j.eswa.2025.127672.
170. Yan Y, Ma T, Li R, Zheng X, Shan G, Li C. JingFang: a traditional Chinese medicine large language model of expert-level medical diagnosis and syndrome differentiation-based treatment. Preprint. arXiv; 2025. https://doi.org/10.48550/arXiv.2502.04345.
171. Chen J, Cai Z, Liu Z, Yang Y, Wang R, Xiao Q, *et al*. ShizhenGPT: towards multimodal LLMs for traditional Chinese medicine. Preprint. arXiv; 2025. https://doi.org/10.48550/arXiv.2508.14706.
172. Zhuang Y, Yu L, Jiang N, Ge Y. TCM-KLLaMA: intelligent generation model for traditional Chinese medicine prescriptions based on knowledge graph and large language model. Comput Biol Med. 2025;189:109887. https://doi.org/10.1016/j.compbiomed.2025.109887.
173. Hartman V, Zhang X, Poddar R, McCarty M, Fortenko A, Sholle E, *et al*. Developing and evaluating large language model-generated emergency medicine handoff notes. JAMA Netw Open. 2024;7(12):e2448723. https://doi.org/10.1001/jamanetworkopen.2024.48723.
174. Páleník J, Soták M, Černý M, Komarc M, Svoboda N, Hayu D, *et al*. Conversational AI in tactical combat casualty care: baseline GPT-4o improves medic decision-making. Clin Simul Nurs. 2025;107:101803. https://doi.org/10.1016/j.ecns.2025.101803.
175. Jiang Y, Black KC, Geng G, Park D, Zou J, Ng AY, *et al*. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. NEJM AI. 2025;2(9):AIdbp2500144. https://doi.org/10.1056/AIdbp2500144.
176. Yang W, Liu Z, Li Y, Yan B, Li L, He M, *et al*. Empowering locally deployable medical agent via state enhanced logical skills for FHIR-based clinical tasks. Preprint. arXiv; 2026. https://doi.org/10.48550/arXiv.2603.06902.
177. Che H, Jin H, Guo Z, Lin Y, Jin C, Chen H. LLM-driven medical report generation via communication-efficient heterogeneous federated learning. IEEE Trans Med Imaging. 2026;45(1):28-39. https://doi.org/10.1109/TMI.2025.3591185.
178. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, *et al*. Towards accurate differential diagnosis with large language models. Nature. 2025;642(8067):451-7. https://doi.org/10.1038/s41586-025-08869-4.
179. Omar M, Sorin V, Collins JD, Reich D, Freeman R, Gavin N, *et al*. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med. 2025;5(1):330. https://doi.org/10.1038/s43856-025-01021-3.
180. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, *et al*. Towards generalist biomedical AI. NEJM AI. 2024;1(3):23. https://doi.org/10.1056/AIoa2300138.
181. Ghosh S, Schneider M, Reinicke C, Eickhoff C. A survey on LLM-assisted clinical trial recruitment. Proceedings of the International Joint Conference on Natural Language Processing. 2025. p. 625-46. https://doi.org/10.18653/v1/2025.ijcnlp-long.35.
182. Wu J, Zhu J, Qi Y, Chen J, Xu M, Menolascina F, *et al*. Medical graph RAG: evidence-based medical large language model via graph retrieval-augmented generation. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2025. p. 28443-67. https://doi.org/10.18653/v1/2025.acl-long.1381.

183. Kim Y, Wu J, Abdulle Y, Wu H. MedExQA: medical question answering benchmark with multiple explanations. Proceedings of the Workshop on Biomedical Natural Language Processing. 2024. p. 167-81. https://doi.org/10.18653/v1/2024.bionlp-1.14.
184. Ghosh A, Acharya A, Jain R, Saha S, Chadha A, Sinha S. Clipsyntel: clip and LLM synergy for multimodal question summarization in healthcare. Proceedings of the AAAI Conference on Artificial Intelligence. 2024. p. 22031-9. https://ojs.aaai.org/index.php/AAAI/article/view/30206.
185. Song JW, Park J, Kim JH, You SC. Large language model assistant for emergency department discharge documentation. JAMA Netw Open. 2025;8(10):e2538427. https://doi.org/10.1001/jamanetworkopen.2025.38427.
186. Williams CYK, Subramanian CR, Ali SS, Apolinario M, Askin E, Barish P, *et al*. Physician- and large language model-generated hospital discharge summaries. JAMA Intern Med. 2025;185(7):818-25. https://doi.org/10.1001/jamainternmed.2025.0821.
187. Han T, Kumar A, Agarwal C, Lakkaraju H. MedSafetyBench: evaluating and improving the medical safety of large language models. Proceedings of Advances in Neural Information Processing Systems. 2024. p. 33423-54. https://doi.org/10.52202/079017-1054.
188. Rajashekar NC, Shin YE, Pu Y, Chung S, You K, Giuffre M, *et al*. Human-algorithmic interaction using a large language model-augmented artificial intelligence clinical decision support system. Proceedings of the CHI Conference on Human Factors in Computing Systems. 2024. p. 1-20. https://doi.org/10.1145/3613904.3642024.
189. Rodrigues T, Teixeira Lopes C. Harnessing large language models for clinical information extraction: a systematic literature review. ACM Trans Comput Healthc. 2025;6(4):1-35. https://doi.org/10.1145/3744660.
190. Builtjes L, Bosma J, Prokop M, van Ginneken B, Hering A. Leveraging open-source large language models for clinical information extraction in resource-constrained settings. JAMIA Open. 2025;8(5):ooaf109. https://doi.org/10.1093/jamiaopen/ooaf109.
191. Soroush A, Glicksberg BS, Zimlichman E, Barash Y, Freeman R, Charney AW, *et al*. Large language models are poor medical coders-benchmarking of medical code querying. NEJM AI. 2024;1(5):23. https://doi.org/10.1056/AIdbp2300040.
192. Kouzy R, Hong JC, Bitterman DS. One shot at trust: building credible evidence for medical artificial intelligence. Lancet Digit Health. 2025;7(7):100883. https://doi.org/10.1016/j.landig.2025.100883.
193. Li CY, Chang KJ, Yang CF, Wu HY, Chen W, Bansal H, *et al*. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. Nat Commun. 2025;16(1):2258. https://doi.org/10.1038/s41467-025-57426-0.
194. Yang J, Ding Q, Tian J, Lai P. Technical roadmap towards trustworthy large-scale models in medicine. Innov Med. 2024;2(1):100058-2. https://doi.org/10.59717/j.xinn-med.2024.100058.
195. Bednarczyk L, Reichenpfader D, Gaudet-Blavignac C, Ette AK, Zaghir J, Zheng Y, *et al*. Scientific evidence for clinical text summarization using large language models: scoping review. J Med Internet Res. 2025;27:e68998. https://doi.org/10.2196/68998.
196. Gu B, Shao V, Liao Z, Carducci V, Brufau SR, Yang J, *et al*. Scalable information extraction from free text electronic health records using large language models. BMC Med Res Methodol. 2025;25(1):23. https://doi.org/10.1186/s12874-025-02470-z.
197. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, *et al*. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med. 2025;31(1):77-86. https://doi.org/10.1038/s41591-024-03328-5.
198. Lin C-Y. ROUGE: a package for automatic evaluation of summaries. Text Summarization Branches Out. 2004. p. 74-81. https://aclanthology.org/W04-1013.
199. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2002. p. 311-8. https://doi.org/10.3115/1073083.1073135.
200. Carandang KAM, Arana JM, Casin ER, Monterola C, Tan DS, Valenzuela JFB, *et al*. Are LLMs reliable? An exploration of the reliability of large language models in clinical note generation. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2025. p. 1413-22. https://doi.org/10.18653/v1/2025.acl-industry.99.
201. Chen L, He R, Lu P, Jin Y, Zhou L, Li N, *et al*. Operationalizing large language models for clinical research data extraction: methods, quality control, and governance. J Med Syst. 2026;50(1):25. https://doi.org/10.1007/s10916-026-02353-w.
202. Hasani E, Richter S, Juratli TA, Buszello CH, Prem MG, Willkommen S, *et al*. Layperson-friendly AI translation of medical documents to improve doctor-patient communication: protocols for the AI-INFOCARE and AI-MEDTALK randomized controlled trials. JMIR Res Protoc. 2025;14:e77204. https://doi.org/10.2196/77204.
203. Kong M, Fernandez A, Bains J, Milisavljevic A, Brooks KC, Shanmugam A, *et al*. Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: a comparative analysis. BMJ Qual Saf. 2026;35(3):150-8. https://doi.org/10.1136/bmjqs-2024-018384.
204. Lopez I, Velasquez DE, Chen JH, Rodriguez JA. Operationalizing machine-assisted translation in healthcare. NPJ Digit Med. 2025;8(1):584. https://doi.org/10.1038/s41746-025-01944-0.
205. Gérardin C, Xiong Y, Wajsbürt P, Carrat F, Tannier X. Impact of translation on biomedical information extraction: experiment on real-life clinical notes. JMIR Med Inform. 2024;12:e49607. https://doi.org/10.2196/49607.
206. Ray M, Kats DJ, Moorkens J, Rai D, Shaar N, Quinones D, *et al*. Evaluating a large language model in translating patient instructions to Spanish using a standardized framework. JAMA Pediatr. 2025;179(9):1026-33. https://doi.org/10.1001/jamapediatrics.2025.1729.
207. Neves M, Grozea C, Thomas P, Roller R, Bawden R, Névéol A, *et al*. Findings of the WMT 2024 biomedical translation shared task: test sets on abstract level. Proceedings of the Conference on Machine Translation. 2024. p. 124-38. https://doi.org/10.18653/v1/2024.wmt-1.6.
208. Tiedemann J. News from OPUS-a collection of multilingual parallel corpora with tools and interfaces. Proceedings of Recent Advances in Natural Language Processing. 2009. p. 237-48. https://doi.org/10.1075/cilt.309.19tie.
209. Kors JA, Clematide S, Akhondi SA, van Mulligen EM, Rebholz-Schuhmann D. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. J Am Med Inform Assoc. 2015;22(5):948-56. https://doi.org/10.1093/jamia/ocv037.
210. Falcão J, Borg C, Aranberri N, Abela K. COMET for low-resource machine translation evaluation: a case study of English-Maltese

and Spanish-Basque. Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation. 2024. p. 3553-65. https://doi.org/10.63317/59mm6uhq2ihq.

211. Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T, *et al*. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open. 2025;8(2):e2457879. https://doi.org/10.1001/jamanetworkopen.2024.57879.
212. Ayoub M, Zhao H, Li L, Yang D, Hussain S, Wahid JA. Structured clinical approach to enable large language models to be used for improved clinical diagnosis and explainable reasoning. Commun Med. 2026;6(1):86. https://doi.org/10.1038/s43856-025-01348-x.
213. Wang L, Chen X, Deng X, Wen H, You M, Liu W, *et al*. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med. 2024;7(1):41. https://doi.org/10.1038/s41746-024-01029-4.
214. Lee RW, Jun TJ, Lee JM, Cho SI, Park HJ, Suh J. Vulnerability of large language models to prompt injection when providing medical advice. JAMA Netw Open. 2025;8(12):e2549963. https://doi.org/10.1001/jamanetworkopen.2025.49963.
215. Kim M, Kim Y, Kang HJ, Seo H, Choi H, Han J, *et al*. Fine-tuning LLMs with medical data: can safety be ensured?. NEJM AI. 2025;2(1):20. https://doi.org/10.1056/AIcs2400390.
216. Ji K, Guo Y, Zhang Z, Zhu X, Tian Y, Liu N. MedOmni-45: a safety-performance benchmark for reasoning-oriented LLMs in medicine. Proceedings of the AAAI Conference on Artificial Intelligence. 2026. p. 35536-44. https://doi.org/10.1609/aaai.v40i42.40864.
217. Raghu Subramanian C, Yang DA, Khanna R. Enhancing health care communication with large language models-the role, challenges, and future directions. JAMA Netw Open. 2024;7(3):e240347. https://doi.org/10.1001/jamanetworkopen.2024.0347.
218. Fahrner LJ, Chen E, Topol E, Rajpurkar P. The generative era of medical AI. Cell. 2025;188(14):3648-60. https://doi.org/10.1016/j.cell.2025.05.018.
219. Yang L, Xu S, Sellergren A, Kohlberger T, Zhou Y, Ktena I, *et al*. Advancing multimodal medical capabilities of Gemini. Preprint. arXiv; 2024. https://doi.org/10.48550/arXiv.2405.03162.
220. Tong Q, Lu Z, Liu J, Zheng Y, Lu ZM. MediSee: reasoning-based pixel-level perception in medical images. Proceedings of the ACM International Conference on Multimedia. 2025. p. 2742-51. https://doi.org/10.1145/3746027.3754736.
221. Xin Y, Ates GC, Gong K, Shao W. Med3DVLM: an efficient vision-language model for 3D medical image analysis. IEEE J Biomed Health Inform. 2026;30(3):2524-36. https://doi.org/10.1109/JBHI.2025.3604595.
222. Moor M, Huang Q, Wu S, Yasunaga M, Dalmia Y, Leskovec J, *et al*. Med-Flamingo: a multimodal medical few-shot learner. Proceedings of Machine Learning for Health. 2023. p. 353-67. https://proceedings.mlr.press/v225/moor23a.html.
223. Lee S, Kim WJ, Chang J, Ye JC. LLM-CXR: instruction-finetuned LLM for CXR image understanding and generation. International Conference on Learning Representations. 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/7f70331dbe58ad59d83941dfa7d975aa-Abstract-Conference.html.
224. Saab K, Tu T, Weng WH, Tanno R, Stutz D, Wulczyn E, *et al*. Capabilities of gemini models in medicine. Preprint. arXiv; 2024. https://doi.org/10.48550/arXiv.2404.18416.
225. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, *et al*. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616-24. https://doi.org/10.1038/s41586-023-06139-9.
226. Shu L, Tang J, Guan X, Zhang D. A comprehensive survey of genome language models in bioinformatics. Brief Bioinform. 2026;27(1):bbaf724. https://doi.org/10.1093/bib/bbaf724.
227. Brixi G, Durrant MG, Ku J, Naghipourfar M, Poli M, Sun G, *et al*. Genome modelling and design across all domains of life with Evo 2. Nature. 2026;652(8112):1349-61. https://doi.org/10.1038/s41586-026-10176-5.
228. Tang X, Tran A, Tan J, Gerstein MB. MolLM: a unified language model for integrating biomedical text with 2D and 3D molecular representations. Bioinformatics. 2024;40(Suppl 1):i357-68. https://doi.org/10.1093/bioinformatics/btae260.
229. Sha Y, Meng W, Luo G, Zhai X, Tong HHY, Wang Y, *et al*. MetDIT: transforming and analyzing clinical metabolomics data with convolutional neural networks. Anal Chem. 2024;97(7):2949-57. https://doi.org/10.1021/acs.analchem.3c04607.
230. Sha Y, Zhang Q, Zhai X, Hou M, Lu J, Meng W, *et al*. CerviFusionNet: a multi-modal, hybrid CNN-transformer-GRU model for enhanced cervical lesion multi-classification. iScience. 2024;27(12):111313. https://doi.org/10.1016/j.isci.2024.111313.
231. Lu MY, Chen B, Williamson DFK, Chen RJ, Liang I, Ding T, *et al*. A visual-language foundation model for computational pathology. Nat Med. 2024;30(3):863-74. https://doi.org/10.1038/s41591-024-02856-4.
232. Xiao J, Yao A, Li Y, Chua TS. Can I trust your answer? Visually grounded video question answering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 13204-14. https://doi.org/10.1109/CVPR52733.2024.01254.
233. Liu B, Zou K, Zhan LM, Lu Z, Dong X, Chen Y, *et al*. GEMeX: a large-scale, groundable, and explainable medical VQA benchmark for chest X-ray diagnosis. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025. p. 21310-20. https://doi.org/10.1109/ICCV51701.2025.01979.
234. Wang Z, Jin Q, Wei CH, Tian S, Lai PT, Zhu Q, *et al*. GeneAgent: self-verification language agent for gene-set analysis using domain databases. Nat Methods. 2025;22(8):1677-85. https://doi.org/10.1038/s41592-025-02748-6.
235. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, *et al*. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:e081554. https://doi.org/10.1136/bmj-2024-081554.
236. Hua Y, Na H, Li Z, Liu F, Fang X, Clifton D, *et al*. A scoping review of large language models for generative tasks in mental health care. NPJ Digit Med. 2025;8(1):230. https://doi.org/10.1038/s41746-025-01611-4.
237. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, *et al*. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. NPJ Ment Health Res. 2024;3(1):12. https://doi.org/10.1038/s44184-024-00056-z.
238. Malgaroli M, Schultebraucks K, Myrick KJ, Andrade Loch A, Ospina-Pinillos L, Choudhury T, *et al*. Large language models for the mental health community: framework for translating code to care. Lancet Digit Health. 2025;7(4):e282-5. https://doi.org/10.1016/S2589-7500(24)00255-3.
239. McCoy TH, Castro VM, Perlis RH. Estimating depression severity in narrative clinical notes using large language models. J Affect Disord. 2025;381:270-4. https://doi.org/10.1016/j.jad.2025.04.014.
240. Sha Y, Pan H, Xu W, Meng W, Luo G, Du X, *et al*. MDD-LLM:

towards accuracy large language models for major depressive disorder diagnosis. J Affect Disord. 2025;388:119774. https://doi.org/10.1016/j.jad.2025.119774.
241. De Duro ES, Improta R, Stella M. Introducing CounseLLMe: a dataset of simulated mental health dialogues for comparing LLMs like Haiku, LLaMAntino and ChatGPT against humans. Emerg Trends Drugs Addict Health. 2025;5:100170. https://doi.org/10.1016/j.etdah.2025.100170.
242. Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, *et al*. The distress analysis interview corpus of human and computer interviews. Proceedings of the International Conference on Language Resources and Evaluation. 2014. p. 3123-8. https://doi.org/10.63317/3o7bccg9xequ.
243. Sun H, Lin Z, Zheng C, Liu S, Huang M. PsyQA: a Chinese dataset for generating long counseling text for mental health support. Findings of the Association for Computational Linguistics. 2021. p. 1489-503. https://doi.org/10.18653/v1/2021.findings-acl.130.
244. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. 2024;11:e57400. https://doi.org/10.2196/57400.
245. Ong JCL, Seng BJJ, Law JZF, Low LL, Kwa ALH, Giacomini KM, *et al*. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: current state and future directions. Cell Rep Med. 2024;5(1):101356. https://doi.org/10.1016/j.xcrm.2023.101356.
246. Luo H, Dai S, Ni C, Li X, Zhang G, Wang K, *et al*. AgentAuditor: human-level safety and security evaluation for LLM agents. Preprint. arXiv; 2025. https://doi.org/10.48550/arXiv.2506.00641.
247. Ren Y, Luo X, Wang Y, Li H, Zhang H, Li Z, *et al*. Large language models in traditional Chinese medicine: a scoping review. J Evid Based Med. 2025;18(1):e12658. https://doi.org/10.1111/jebm.12658.
248. Wang Z, Hao M, Peng S, Huang Y, Lu Y, Yao K, *et al*. TCMEval-SDT: a benchmark dataset for syndrome differentiation thought of traditional Chinese medicine. Sci Data. 2025;12(1):437. https://doi.org/10.1038/s41597-025-04772-9.
249. Tan Y, Zhang Z, Li M, Pan F, Duan H, Huang Z, *et al*. MedChatZH: a tuning LLM for traditional Chinese medicine consultations. Comput Biol Med. 2024;172:108290. https://doi.org/10.1016/j.compbiomed.2024.108290.
250. Yue W, Wang X, Zhu W, Guan M, Zheng H, Wang P, *et al*. TCMBench: a comprehensive benchmark for evaluating large language models in traditional Chinese medicine. Preprint. arXiv; 2024. https://doi.org/10.48550/arXiv.2406.01126.
251. Dai J, Xu H, Chen T, Huang T, Liang W, Zhang R, *et al*. Artificial intelligence for medicine 2025: navigating the endless frontier. Innov Med. 2025;3(1):100120-1-100120-15. https://doi.org/10.59717/j.xinn-med.2025.100120.
252. Wang B, Lu Y, Wang Z, Ge P, Wang G, Yao K, *et al*. TCMEval-PA: a question-answering benchmark dataset for the prescription audit of Traditional Chinese Medicine. Sci Data. 2025;13(1):79. https://doi.org/10.1038/s41597-025-06387-6.
253. Long H, Deng Y, Guo Y, Shen Z, Zhang Y, Bao J, *et al*. Large language model evaluation in traditional Chinese medicine for stroke: quantitative benchmarking study. JMIR Form Res. 2025;9:e81545. https://doi.org/10.2196/81545.
254. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. Preprint. arXiv; 2020. https://doi.org/10.48550/arXiv.1904.09675.
255. Wei S, Peng X, Wang Y, Shen T, Si J, Zhang W, *et al*. BianCang: a traditional Chinese medicine large language model. IEEE J Biomed Health Inform. 2025;1-12. https://doi.org/10.1109/JBHI.2025.3612415.
256. Wang X, Wang Y, Dong W, Guo S, Wang K, He S, *et al*. TCM-Agent: advancing network pharmacology and herbal medicine discovery with LLM-based multi-agent systems. J Pharm Anal. 2026;101581. https://doi.org/10.1016/j.jpha.2026.101581.
257. Nettesheim N, Powell D, Vasios W, Mbuthia J, Davis K, Yourk D, *et al*. Telemedical support for military medicine. Mil Med. 2018;183(11-12):e462-70. https://doi.org/10.1093/milmed/usy127.
258. Pamplin JC, Davis KL, Mbuthia J, Cain S, Hipp SJ, Yourk DJ, *et al*. Military telehealth: a model for delivering expertise to the point of need in austere and operational environments. Health Aff. 2019;38(8):1386-92. https://doi.org/10.1377/hlthaff.2019.00273.
259. Maddry JK, Perez CA, Mora AG, Lear JD, Savell SC, Bebarta VS. Impact of prehospital medical evacuation (MEDEVAC) transport time on combat mortality in patients with non-compressible torso injury and traumatic amputations: a retrospective study. Mil Med Res. 2018;5(1):22. https://doi.org/10.1186/s40779-018-0169-2.
260. Bennett WN, Markelz AE, Kile MT, Pamplin JC, Barsoumian AE. Infectious Disease Teleconsultation to the Deployed U.S. Military From 2017–2022. Mil Med. 2023;188(7-8):e1990-5. https://doi.org/10.1093/milmed/usac308.
261. McLeroy RD, Kile MT, Yourk D, Hipp S, Pamplin JC. Advanced virtual support for operational forces: a 3-year summary. Mil Med. 2022;187(5):742-6. https://doi.org/10.1093/milmed/usab388.
262. Xie T, Liu XR, Chen GL, Qi L, Xu ZY, Liu XD. Development and application of triage and medical evacuation system for casualties at sea. Mil Med Res. 2014;1:12. https://doi.org/10.1186/2054-9369-1-12.
263. Wang W, Ma Z, Wang Z, Wu C, Ji J, Chen W, *et al*. A survey of LLM-based agents in medicine: how far are we from baymax?. Findings of the Association for Computational Linguistics. 2025. p. 10345-59. https://doi.org/10.18653/v1/2025.findings-acl.539.
264. He K, Wang SY, Ren J. Challenges, opportunities, and future perspectives of portable field endoscopy. Mil Med Res. 2025;12(1):80. https://doi.org/10.1186/s40779-025-00666-4.
265. Gottlieb S, Silvis L. How to safely integrate large language models into health care. JAMA Health Forum. 2023;4(9):e233909. https://doi.org/10.1001/jamahealthforum.2023.3909.
266. Verlingue L, Boyer C, Olgiati L, Brutti Mairesse C, Morel D, Blay JY. Artificial intelligence in oncology: ensuring safe and effective integration of language models in clinical practice. Lancet Reg Health Eur. 2024;46:101064. https://doi.org/10.1016/j.lanepe.2024.101064.
267. Sun G, Zeng JQ, Wu JL, Wang SF, Peng LH, Yan B, *et al*. Evaluation and application of an innovative portable field endoscope: addressing battlefield and biosafety concerns. Mil Med Res. 2025;12(1):57. https://doi.org/10.1186/s40779-025-00644-w.
268. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, *et al*. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med. 2025;8(1):274. https://doi.org/10.1038/s41746-025-01670-7.
269. Pandit S, Xu J, Hong J, Wang Z, Chen T, Xu K, *et al*. MedHallu: a comprehensive benchmark for detecting medical hallucinations in large language models. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2025. p. 2858-73. https://doi.org/10.18653/v1/2025.emnlp-main.143.

270. Wang S, Tang Z, Yang H, Gong Q, Gu T, Ma H, *et al*. A novel evaluation benchmark for medical LLMs illuminating safety and effectiveness in clinical domains. NPJ Digit Med. 2025;9(1):91. https://doi.org/10.1038/s41746-025-02277-8.
271. Li M, Zhan Z, Yang H, Xiao Y, Zhou H, Huang J, *et al*. Benchmarking retrieval-augmented large language models in biomedical NLP: application, robustness, and self-awareness. Sci Adv. 2025;11(47): eadr1443. https://doi.org/10.1126/sciadv.adr1443.
272. Moore J, Grabb D, Agnew W, Klyman K, Chancellor S, Ong DC, *et al*. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. 2025. p. 599-627. https://doi.org/10.1145/3715275.3732039.
273. Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking retrieval-augmented generation for medicine. Findings of the Association for Computational Linguistics. 2024. p. 6233-51. https://doi.org/10.18653/v1/2024.findings-acl.372.
274. Lu J, Liu J, Zheng X, Yang M, Wang J, Wang P, *et al*. MHB: medical hallucination benchmark for large language models in complex clinical tasks. Proceedings of the AAAI Conference on Artificial Intelligence. 2026. p. 38971-8. https://doi.org/10.1609/aaai.v40i45.41243.
275. McCoy LG, Swamy R, Sagar N, Wang M, Bacchi S, Fong JMN, *et al*. Assessment of large language models in clinical reasoning: a novel benchmarking study. NEJM AI. 2025;2(10):AIdbp2500120. https://doi.org/10.1056/AIdbp2500120.
276. Croxford E, Gao Y, First E, Pellegrino N, Schnier M, Caskey J, *et al*. Evaluating clinical AI summaries with large language models as judges. NPJ Digit Med. 2025;8(1):640. https://doi.org/10.1038/s41746-025-02005-2.
277. Griot M, Vanderdonckt J, Yuksel D, Hemptinne C. Pattern recognition or medical knowledge? The problem with multiple-choice questions in medicine. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2025. p. 5321-41. https://doi.org/10.18653/v1/2025.acl-long.266.
278. Mallinar N, Heydari AA, Liu X, Faranesh AZ, Winslow B, Hammerquist N, *et al*. A scalable framework for evaluating health language models. NPJ Digit Med. 2026; https://doi.org/10.1038/s41746-026-02492-x.
279. Kocaman V, Kaya MA, Feier AM, Talby D. Clinical large language model evaluation by expert review (CLEVER): framework development and validation. JMIR AI. 2025;4:e72153. https://doi.org/10.2196/72153.
280. Zhang M, Shen Y, Li Z, Sha H, Hu B, Wang Y, *et al*. LLMEval-Med: a real-world clinical benchmark for medical LLMs with physician validation. Findings of the Association for Computational Linguistics. 2025. p. 4888-914. https://doi.org/10.18653/v1/2025.findings-emnlp.263.
281. Bedi S, Jiang Y, Chung P, Koyejo S, Shah N. Fidelity of medical reasoning in large language models. JAMA Netw Open. 2025;8(8):e2526021. https://doi.org/10.1001/jamanetworkopen.2025.26021.
282. Dillon GF, Boulet JR, Hawkins RE, Swanson DB. Simulations in the United States Medical Licensing Examination (USMLE). Qual Saf Health Care. 2004;13 Suppl 1(Suppl 1):i41-45. https://doi.org/10.1136/qhc.13.suppl_1.i41.
283. Liu F, Wu X, Huang J, Yang B, Branson K, Schwab P, *et al*. Aligning, autoencoding and prompting large language models for novel disease reporting. IEEE Trans Pattern Anal Mach Intell. 2025;47(5):3332-43. https://doi.org/10.1109/TPAMI.2025.3534586.
284. Badawi A, Rahimi E, Laskar MTR, Grach S, Bertrand L, Danok L, *et al*. When can we trust LLMs in mental health? Large-scale benchmarks for reliable LLM evaluation. Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics. 2026. p. 3873-96. https://doi.org/10.18653/v1/2026.eacl-long.180.
285. Gu J, Jiang X, Shi Z, Tan H, Zhai X, Xu C, *et al*. A survey on LLM-as-a-judge. Innovation. 2026;101253. https://doi.org/10.1016/j.xinn.2025.101253.
286. Zhu Z, Zhang Y, Zhuang X, Zhang F, Wan Z, Chen Y, *et al*. Can we trust AI doctors? A survey of medical hallucination in large language and large vision-language models. Findings of the Association for Computational Linguistics. 2025. p. 6748-69. https://doi.org/10.18653/v1/2025.findings-acl.350.
287. Ong JCL, Chang SYH, William W, Butte AJ, Shah NH, Chew LST, *et al*. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health. 2024;6(6):e428-32. https://doi.org/10.1016/S2589-7500(24)00061-X.
288. Kusche I. Possible harms of artificial intelligence and the EU AI act: fundamental rights and risk. J Risk Res. 2024;1-14. https://doi.org/10.1080/13669877.2024.2350720.
289. Manakul P, Liusie A, Gales M. SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2023. p. 9004-17. https://doi.org/10.18653/v1/2023.emnlp-main.557.
290. Marks M, Haupt CE. AI chatbots, health privacy, and challenges to HIPAA compliance. JAMA. 2023;330(4):309-10. https://doi.org/10.1001/jama.2023.9458.
291. Jonnagaddala J, Wong ZSY. Privacy preserving strategies for electronic health records in the era of large language models. NPJ Digit Med. 2025;8(1):34. https://doi.org/10.1038/s41746-025-01429-0.
292. Jiang YL, Zhao G, Wang SH, Li N. Leveraging artificial intelligence for clinical decision support in personalized standard regimen recommendation for cancer. Mil Med Res. 2025;12(1):31. https://doi.org/10.1186/s40779-025-00617-z.
293. Li M, Xu P, Hu J, Tang Z, Yang G. From challenges and pitfalls to recommendations and opportunities: implementing federated learning in healthcare. Med Image Anal. 2025;101:103497. https://doi.org/10.1016/j.media.2025.103497.
294. Zong Y, Yang Y, Hospedales T. MEDFAIR: benchmarking fairness for medical imaging. International Conference on Learning Representations. 2023. https://openreview.net/forum?id=6ve2CkeQe5S.
295. van Kolfschooten H, van Oirschot J. The EU artificial intelligence act (2024): implications for healthcare. Health Policy. 2024;149: 105152. https://doi.org/10.1016/j.healthpol.2024.105152.
296. Vardas EP, Marketou M, Vardas PE. Medicine, healthcare and the AI act: gaps, challenges and future implications. Eur Heart J Digit Health. 2025;6(4):833-9. https://doi.org/10.1093/ehjdh/ztaf041.
297. Schmidt J, Schutte NM, Buttigieg S, Novillo-Ortiz D, Sutherland E, Anderson M, *et al*. Mapping the regulatory landscape for artificial intelligence in health within the European Union. NPJ Digit Med. 2024;7(1):229. https://doi.org/10.1038/s41746-024-01221-6.
298. Aboy M, Minssen T, Vayena E. Navigating the EU AI act: implications for regulated digital medical products. NPJ Digit Med. 2024;7(1):237. https://doi.org/10.1038/s41746-024-01232-3.
299. Balendran A, Beji C, Bouvier F, Khalifa O, Evgeniou T, Ravaud P,

*et al*. A scoping review of robustness concepts for machine learning in healthcare. NPJ Digit Med. 2025;8(1):38. https://doi.org/10.1038/s41746-024-01420-1.

300. Han T, Nebelung S, Khader F, Wang T, Müller-Franzes G, Kuhl C, *et al*. Medical large language models are susceptible to targeted misinformation attacks. NPJ Digit Med. 2024;7(1):288. https://doi.org/10.1038/s41746-024-01282-7.
301. Koch LM, Baumgartner CF, Berens P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. NPJ Digit Med. 2024;7(1):120. https://doi.org/10.1038/s41746-024-01085-w.
302. Subasri V, Krishnan A, Kore A, Dhalla A, Pandya D, Wang B, *et al*. Detecting and remediating harmful data shifts for the responsible deployment of clinical AI models. JAMA Netw Open. 2025;8(6):e2513685. https://doi.org/10.1001/jamanetworkopen.2025.13685.
303. Oh C, Fang Z, Im S, Du X, Li Y. Understanding multimodal LLMs under distribution shifts: an information-theoretic approach. Proceedings of the International Conference on Machine Learning. 2025. p. 46943-70. https://proceedings.mlr.press/v267/oh25a.html.
304. Ktena I, Wiles O, Albuquerque I, Rebuffi S-A, Tanno R, Roy AG, *et al*. Generative models improve fairness of medical classifiers under distribution shifts. Nat Med. 2024;30(4):1166-73. https://doi.org/10.1038/s41591-024-02838-6.
305. Chen S, Gao M, Sasse K, Hartvigsen T, Anthony B, Fan L, *et al*. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. NPJ Digit Med. 2025;8(1):605. https://doi.org/10.1038/s41746-025-02008-z.
306. Wu W, Xu X, Gao C, Diao X, Li S, Salas LA, *et al*. Assessing and mitigating medical knowledge drift and conflicts in large language models. Findings of the Association for Computational Linguistics. 2025. p. 707-30. https://doi.org/10.18653/v1/2025.findings-emnlp.38.
307. Rosenthal JT, Beecy A, Sabuncu MR. Rethinking clinical trials for medical AI with dynamic deployments of adaptive systems. NPJ Digit Med. 2025;8(1):252. https://doi.org/10.1038/s41746-025-01674-3.
308. Shi H, Xu Z, Wang H, Qin W, Wang W, Wang Y, *et al*. Continual learning of large language models: a comprehensive survey. ACM Comput Surv. 2025;58(5):1-42. https://doi.org/10.1145/3735633.
309. Masanneck L, Meuth SG, Pawlitzki M. Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology. NPJ Digit Med. 2025;8(1):137. https://doi.org/10.1038/s41746-025-01536-y.
310. McCoy LG, Manrai AK, Rodman A. Large language models and the degradation of the medical record. N Engl J Med. 2024;391(17):1561-4. https://doi.org/10.1056/NEJMp2405999.
311. Wang X, Tan M, Jin Q, Xiong G, Hu Y, Zhang A, *et al*. MedCite: can language models generate verifiable text for medicine? Findings of the Association for Computational Linguistics. 2025. p. 18891-913. https://doi.org/10.18653/v1/2025.findings-acl.967.
312. Solaiman B, Mekki YM, Qadir J, Ghaly M, Abdelkareem M, Al-Ansari A. A "true lifecycle approach" towards governing healthcare AI with the GCC as a global governance model. NPJ Digit Med. 2025;8(1):337. https://doi.org/10.1038/s41746-025-01614-1.
313. Wang P, Li Z, Zhang N, Xu Z, Yao Y, Jiang Y, *et al*. Wise: rethinking the knowledge memory for lifelong model editing of large language models. Proceedings of Advances in Neural Information Processing Systems. 2024;37:53764-97. https://papers.nips.cc/paper_files/paper/2024/hash/60960ad78868fce5c165295fbd895060-Abstract-Conference.html.
314. Wang L, Zhang X, Su H, Zhu J. A comprehensive survey of continual learning: theory, method and application. IEEE Trans Pattern Anal Mach Intell. 2024;46(8):5362-83. https://doi.org/10.1109/TPAMI.2024.3367329.
315. Li Z, Zhang N, Yao Y, Wang M, Chen X, Chen H. Unveiling the pitfalls of knowledge editing for large language models. International Conference on Learning Representations. 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/56f187f2b3e87da2fb103af6df611970-Abstract-Conference.html.
316. Wang G, Zhang K, Jiang J, Wang C, Bi H, Liang H, *et al*. Human-large language model collaboration in clinical medicine: a systematic review and meta-analysis. NPJ Digit Med. 2026;9(1):195. https://doi.org/10.1038/s41746-026-02382-2.
317. Goh E, Bunning B, Khoong EC, Gallo RJ, Milstein A, Centola D, *et al*. Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. Commun Med. 2025;5(1):59. https://doi.org/10.1038/s43856-025-00781-2.
318. Agweyu A, Mwaniki P, Musau W, Korom R, Isaaka L, Wanyama C, *et al*. Safety of a large language model-based clinical decision support system in African primary healthcare. Nat Health. 2026;1-12. https://doi.org/10.1038/s44360-026-00082-5.
319. Duggan MJ, Gervase J, Schoenbaum A, Hanson W, Howell JT, Sheinberg M, *et al*. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Netw Open. 2025;8(2):e2460637. https://doi.org/10.1001/jamanetworkopen.2024.60637.
320. Templin T, Fort S, Padmanabham P, Seshadri P, Rimal R, Oliva J, *et al*. Framework for bias evaluation in large language models in healthcare settings. NPJ Digit Med. 2025;8(1):414. https://doi.org/10.1038/s41746-025-01786-w.
321. Teferra BG, Johny N, Huang S, Rueda A, Kamaleddin MA, Dunlop K, *et al*. Assessing the impact of safety guardrails on large language models using irritability metrics. NPJ Digit Med. 2026;9(1):148. https://doi.org/10.1038/s41746-025-02333-3.
322. Lukac PJ, Turner W, Vangala S, Chin AT, Khalili J, Shih YCT, *et al*. Ambient AI scribes in clinical practice: a randomized trial. NEJM AI. 2025;2(12):10.1056/aioa2501000. https://doi.org/10.1056/aioa2501000.
323. Wang Z, Cao L, Jin Q, Chan J, Wan N, Afzali B, *et al*. A foundation model for human-AI collaboration in medical literature mining. Nat Commun. 2025;16(1):8361. https://doi.org/10.1038/s41467-025-62058-5.
324. Oliveira JD, Santos HDP, Ulbrich AHDPS, Couto JC, Arocha M, Santos J, *et al*. Development and evaluation of a clinical note summarization system using large language models. Commun Med. 2025;5(1):376. https://doi.org/10.1038/s43856-025-01091-3.
325. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Proceedings of Advances in Neural Information Processing Systems. 2017. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.
326. Gaber F, Shaik M, Allega F, Bilecz AJ, Busch F, Goon K, *et al*. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. NPJ Digit Med. 2025;8(1):263. https://doi.org/10.1038/s41746-025-01684-1.
327. Gomez C, Cho SM, Ke S, Huang CM, Unberath M. Human-AI collaboration is not very collaborative yet: a taxonomy of

interaction patterns in AI-assisted decision making from a systematic review. Front Comput Sci. 2025;6:1521066. https://doi.org/10.3389/fcomp.2024.1521066.

328. Kim Y, Park C, Jeong H, Chan YS, Xu X, McDuff D, *et al*. MDAgents: an adaptive collaboration of LLMs for medical decision-making. Proceedings of Advances in Neural Information Processing Systems. 2024;37:79410-52. https://proceedings.neurips.cc/paper_files/paper/2024/hash/90d1fc07f46e31387978b88e7e057a31-Abstract-Conference.html.
329. Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. NPJ Digit Med. 2024;7(1):84. https://doi.org/10.1038/s41746-024-01083-y.
330. Xia P, Chen Z, Tian J, Gong Y, Hou R, Xu Y, *et al*. CARES: a comprehensive benchmark of trustworthiness in medical vision language models. Proceedings of Advances in Neural Information Processing Systems. 2024;37:140334-65. https://proceedings.neurips.cc/paper_files/paper/2024/hash/fde7f40f8ced5735006810534dc66b33-Abstract-Datasets_and_Benchmarks_Track.html.
331. Chen H, Zeng D, Qin Y, Fan Z, Ng Yu Ci F, Klonoff DC, *et al*. Large language models and global health equity: a roadmap for equitable adoption in LMICs. Lancet Reg Health West Pac. 2025;63:101707. https://doi.org/10.1016/j.lanwpc.2025.101707.
332. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, *et al*. Sociodemographic biases in medical decision making by large language models. Nat Med. 2025;31(6):1873-81. https://doi.org/10.1038/s41591-025-03626-6.

**https:**//doi.org/10.1016/j.mmr.2026.100050
**Cite this article as:** Sha YY, Yu L, Lin ZH, Kaur A, Lou YY, Gornale SS, *et al*. A roadmap for medical large language models: a review of foundations, applications, and challenges. *Mil Med Res*. 2026;13(1):100050.