# From Medical LLMs to Versatile Medical Agents: A Comprehensive Survey

Yucheng Zhou, Huan Zheng, Dubing Chen, Hongji Yang, Wencheng Han, and Jianbing Shen

**Abstract**—The integration of Large Language Models (LLMs) into healthcare has catalyzed a significant technological leap, evolving from text-based Medical LLMs to Multimodal Medical LLMs (MLLMs) capable of interpreting complex clinical imaging. Despite these advancements, current models predominantly function as passive knowledge engines, proficient in answering queries but lacking the autonomy to navigate the dynamic, longitudinal nature of real-world patient care. This limitation has spurred a paradigm shift toward *Medical Agents*: proactive systems engineered to sense, reason, plan, and execute actions within clinical environments. In this survey, we provide a comprehensive roadmap of this evolutionary trajectory. We first review the foundational architectures and training strategies of state-of-the-art Medical LLMs and MLLMs. Subsequently, we formalize the construction of *Medical Agentic Systems*, distinguishing between the cognitive frameworks required for independent *Single-Agent Systems* and the collaborative paradigms of *Multi-Agent Systems* that simulate multidisciplinary clinical teams. Central to our analysis is the evolution of medical reasoning, which we categorize into three distinct stages: *Core Reasoning* for internal deliberation, *Augmented Reasoning* for tool-mediated and multimodal grounding, and *Collective Reasoning* for distributed medical intelligence. Finally, the survey examines the necessary transition in evaluation methodologies, from static benchmarks to interactive simulations, and discusses pressing open challenges, offering a forward-looking perspective on building reliable, safe, and clinically impactful medical AI. Project sources: https://github.com/yczhou001/Awesome-Medical-LLM-Agent.

**Index Terms**—Medical Large Language Models, Medical Agents, LLM Reasoning, Multimodal AI for Healthcare, Medical Multi-Agents

## 1 Introduction

The integration of Artificial Intelligence into healthcare has undergone a transformative evolution. Initially, the emergence of Large Language Models (LLMs) marked a significant leap in processing biomedical knowledge. Models such as Med-PaLM [1], [2] and PMC-LLaMA [3] demonstrated expert-level proficiency in medical question answering by leveraging vast textual corpora. Recognizing that clinical practice is inherently multimodal, relying on radiology, pathology, and vital signs, this paradigm expanded into Multimodal Medical LLMs (MLLMs). By fusing vision and language encoders, systems like LLaVA-Med [4] and Med-Flamingo [5] enabled automated report generation and visual diagnostic assistance, bridging the gap between textual records and pixel-level evidence.

Despite these capabilities, existing LLMs and MLLMs largely remain *passive information processors* [6], [7]. They excel at answering queries given a static context ("What is the diagnosis?") but lack the *agency* to navigate the dynamic, longitudinal, and interactive nature of real-world clinical care. A physician's role extends beyond knowledge retrieval; it involves active information seeking, tracking long-term patient states, utilizing tools (e.g., ordering labs, prescribing drugs), and coordinating with other specialists. The inability of standard models to plan, execute, and self-correct limits their utility to that of a sophisticated search engine rather than an autonomous clinical partner.

This limitation has catalyzed a paradigm shift from responsive models to proactive ***Medical Agents***, intelligent systems endowed with the capacity to sense, reason, and act. Unlike their predecessors, Medical Agents are architected with a perception-action loop, maintaining a memory of patient history, formulating multi-step plans, and executing actions through external tools [8], [9]. Furthermore, this evolution is not limited to individual cognition. Inspired by Multidisciplinary Teams (MDTs) in hospitals, recent research has pivoted towards ***Medical Multi-Agent Systems (MAS)***, where specialized agents (e.g., a Diagnostician, a Pharmacist, and a Radiologist) collaborate, debate, and negotiate to solve complex medical problems, creating a form of collective clinical intelligence [10], [11].

The engine driving this transition from static models to autonomous agents is the evolution of *medical reasoning*. We observe a clear trajectory in how reasoning is enhanced: from *Core Reasoning* that improves internal deliberation via techniques like Chain-of-Thought [12]; to *Augmented Reasoning* that grounds agents in reality through Retrieval-Augmented Generation (RAG) and tool use [13]; and finally to *Collective Reasoning*, where intelligence emerges from the interaction and consensus of multiple agents [14]. This progression is critical for transforming AI from a black-box predictor into a transparent, reliable, and verifiable participant in the medical workflow.

**Contributions.** To the best of our knowledge, this is the first

- *Yucheng Zhou, Dubing Chen, Huan Zheng, Hongji Yang, Wencheng Han, and Jianbing Shen are with the SKL-IOTSC, Department of Computer and Information Science, University of Macau, Macau, China. (Email: yucheng.zhou@connect.um.edu.mo)*
- *Corresponding author: Jianbing Shen.*

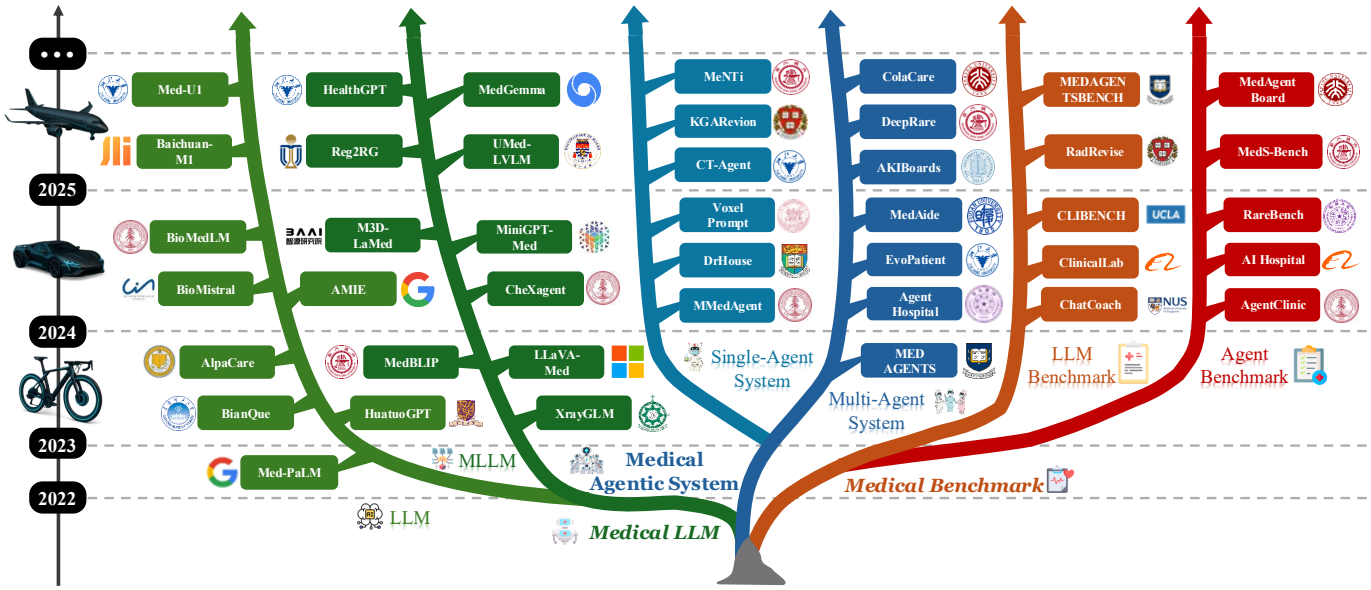


Fig. 1. The evolutionary landscape of Medical AI. This figure illustrates the comprehensive progression from foundational **Medical LLMs** and **Multimodal LLMs (VLMs)** (green branches) to proactive **Medical Agentic Systems** (blue branches), categorized into single-agent and multi-agent architectures. The rightmost branches (orange) depict the concurrent development of specialized benchmarks and simulation environments essential for evaluating these advanced capabilities. Representative models and frameworks are mapped onto their respective evolutionary branches, highlighting the rapid expansion of the field.

survey that systematically bridges the gap between Medical LLMs and Medical Agentic Systems. Our key contributions are summarized as follows:

- We chart the evolutionary path from static Medical LLMs and MLLMs to autonomous Medical Agents, identifying the limitations of passive models that necessitated this paradigm shift.
- We propose a structured classification for Medical Agentic Systems, distinguishing between the cognitive architectures of Single-Agent systems and the collaboration paradigms of Multi-Agent frameworks.
- We introduce a three-stage taxonomy for medical reasoning, *Core*, *Augmented*, and *Collective*, providing a theoretical basis for understanding how AI intelligence is scaling in healthcare.
- Beyond models, we extensively survey the enabling ecosystem, including tool-use environments, simulation platforms, and the next-generation evaluation benchmarks required for agentic capabilities.

**Paper Organization.** The remainder of this survey is organized as follows. Section 2 reviews the foundational Medical LLMs and Multimodal models. Section 3 formalizes the construction of Medical Agentic Systems, detailing both single- and multi-agent architectures. Section 4 explores the evolution of reasoning techniques that power these agents. Section 5 discusses clinical applications and the shift in evaluation paradigms. Finally, Section 6 outlines challenges and future research directions.

## 2 Medical Large Language Models

The adaptation of Large Language Models (LLMs) to the medical domain represents a significant research frontier, driven by the potential to transform clinical workflows, augment diagnostic processes, and democratize access to medical knowledge [20]. This endeavor is non-trivial, as the medical field is characterized by its distinct epistemological complexity, stringent requirements for factual accuracy, and the imperative for safety and reliability [1]. Consequently, the development of medical LLMs necessitates specialized methodologies that extend beyond the direct application of general-purpose models, a challenge that has spurred a wealth of innovation [36].

This section provides a systematic overview of the progress in this domain. We first examine the development of unimodal LLMs, which focus on the interpretation and generation of clinical text. We then transition to the more recent and complex domain of multimodal LLMs, which integrate visual information from medical imaging with textual understanding. Our analysis will focus on the evolution of model architectures, training paradigms, and the increasing specialization required for clinical viability, thereby establishing the technical foundations for the medical agents discussed subsequently.

### 2.1 Medical LLMs

The foundational efforts in medical LLMs have centered on leveraging their inherent capacity for processing text. The primary objective is to develop models that can comprehend and generate nuanced clinical language as found in electronic health records (EHRs), biomedical literature, and clinical

TABLE 1
Representative medical LLMs and their characteristics.

| Model Name | Base Model | Parameter (B) | Training Dataset | Training Method | Release |
|---|---|---|---|---|---|
| Med-PaLM [1] | PaLM | 540 | Medical QA | SFT | 22/12 |
| ChatDoctor [15] | LLaMA | 7 | Patient–doctor Dialogues | SFT | 23/03 |
| Baize [16] | LLaMA | 7 | Quora, MedQuAD | SFT | 23/04 |
| MedAlpaca [17] | LLaMA | 7 / 13 | Medical QA and Dialogues | SFT | 23/04 |
| PMC-LLaMA [3] | LLaMA | 7 / 13 | Biomedical academic papers | PT, SFT | 23/04 |
| DoctorGLM [18] | ChatGLM | 6 | Chinese Medical Dialogues | SFT | 23/04 |
| Med-PaLM 2 [2] | PaLM 2 | 340 | Medical QA | SFT | 23/05 |
| HuatuoGPT [19] | BLOOMZ | 7 | Conversation data and Instruction | SFT, RL | 23/05 |
| GatorTronGPT [20] | GPT-3 | 5 / 20 | Clinical Text from UF Health, Pile | PT | 23/06 |
| ClinicalGPT [21] | BLOOM | 7 | Three MedQA, MD-EHR, MedDialog | SFT, RL | 23/06 |
| Zhongjing [22] | Ziya-LLaMA | 13 | Medical books, health records, clinical reports | PT, SFT, RL | 23/08 |
| CPLLM [23] | LLaMA 2 | 13 | eICU-CRD, MIMIC-IV | SFT | 23/09 |
| BianQue [24] | ChatGLM | 6 | BianQueCorpus | SFT | 23/10 |
| Qilin-Med [25] | Baichuan | 7 | Medical QA, plain texts, knowledge graphs | PT, SFT, RL | 23/10 |
| AlpaCare [26] | LLaMA | 7 / 13 | MedInstruct-52k | SFT | 23/10 |
| HuatuoGPT-II [27] | Baichuan 2 | 7 / 13 | Web Corpus, Books, Literature, Encyclopedia | SFT | 23/11 |
| MEDITRON [28] | LLaMA 2 | 7 / 70 | GAP-Replay, MedMCQA, PubMedQA, MedQA | PT, SFT | 23/11 |
| AMIE [29] | PaLM 2 | 340 | MedQA, MultiMedBench, MIMIC-III, RealWorld Dialogue | SFT | 24/01 |
| OncoGPT [30] | LLaMA | 7 | Oncology conversations | SFT | 24/02 |
| BioMistral [31] | Mistral | 7 | PubMed Central | PT, SFT | 24/02 |
| BiMediX [32] | Mistral | 8x7 | 1.3 million English-Arabic dataset | SFT | 24/02 |
| Me-LLaMA [33] | LLaMA 2 | 13 / 70 | Pile, MIMIC-III, MIMIC-IV, MIMIC-CXR, RedPajama | PT, SFT | 24/02 |
| Apollo [34] | Qwen | 7 | ApolloCorpora | PT, SFT | 24/03 |
| BioMedLM [35] | Transformer | 2.7 | PubMed Center, Pile | PT, SFT | 24/03 |
| Aloe-Alpha [36] | LLaMA-3 | 8 | Medical QA, CoT, synthetic data | PT, SFT, RL | 24/05 |
| MedGo [37] | Qwen2 | 72 | 14B tokens unsupervised, CMCQA, drug-instruction | PT, SFT, RL | 24/10 |
| FineMedLM-o1 [38] | LLaMA 3.1 | 8 | FineMed | SFT, RL | 25/01 |
| MedS$^3$ [39] | LLaMA 3.1 | 8 | Hybrid Data | SFT, RL | 25/01 |
| Hengqin-RA-v1 [40] | LLaMA | 7 | HQ-GCM-RA-C1 | SFT | 25/01 |
| Hypnos [41] | LLaMA | 7 | Medical QA, Synthetic QA | SFT | 25/01 |
| Citrus [42] | LLaMA 3 / Qwen 2.5 | 70 | Expert reasoning corpus, diagnostic dialogue (JMED) | PT, SFT, RL | 25/02 |
| ELMTEX [43] | LLaMA 3.1 / 3.2 | 1 / 3 / 8 | English clinical summaries (60k), German translations (24k) | SFT | 25/02 |
| Baichuan-M1 [44] | Baichuan | 14 | 20T general + medical tokens | SFT, RL | 25/02 |
| MedicalGLM [45] | ChatGLM | 6 | Pediatric QA, MKP dataset | SFT | 25/03 |
| Med-U1 [46] | Qwen2.5 | 3 / 7 | Hybrid Data | RL | 25/06 |
| MIRIAD [47] | LLaMA 3.1 | 8 | 5.82M medical QA pairs | SFT | 25/06 |

dialogues. The evolution of these models can be categorized by their core development strategies and the advanced techniques used to refine their clinical capabilities.

### 2.1.1 From Generalists to Specialists

Three principal strategies have emerged for adapting LLMs to the medical domain, differing in their computational requirements and the depth of domain-specific integration.

**Supervised Fine-Tuning (SFT).** A primary paradigm for domain adaptation is the supervised fine-tuning of general-purpose foundation models. This approach capitalizes on the extensive world knowledge and linguistic capabilities of pre-trained models, such as LLaMA or ChatGLM, and specializes them using curated medical datasets. This has been effective for developing conversational systems fine-tuned on patient-doctor dialogues [15], [17], [24], including computationally efficient methods for non-English languages [18]. The quality of the tuning data is paramount; for instance, high-quality multi-turn dialogue corpora have been generated through automated self-chat protocols to train models like Baize [16].

**Domain-Specific Continued Pre-training.** To imbue models with a more profound and comprehensive domain-specific lexicon and knowledge base, a second strategy involves continued pre-training (CPT) on large-scale medical corpora before SFT. This methodology has been successfully applied by further training foundational models on extensive biomedical literature, such as PubMed, resulting in specialized base models like PMC-LLaMA [3], MEDITRON [28], and BioMistral [31]. This intensive exposure to domain literature establishes a robust knowledge foundation, which has been shown to improve performance on downstream clinical tasks. The culmination of this approach can be seen in large-scale efforts such as Me-LLaMA [33], which utilized a 129-billion-token medical corpus for pre-training.

**De Novo Training.** A third, albeit resource-intensive, pathway involves training models from scratch (de novo) on domain-specific corpora. This approach aims to create models

whose architectures and learned representations are optimally tailored to the unique characteristics of medical data. Notable examples include GatorTronGPT [20], trained on a massive clinical and general text corpus, and BioMedLM [35], which demonstrated the effectiveness of this approach even at a smaller scale (2.7B parameters). The state-of-the-art in this category is represented by Baichuan-M1 [44], trained on approximately 20 trillion tokens to achieve a balance of specialized medical expertise and broad general capabilities.

#### 2.1.2 Enhancing Clinical Acumen: Training Techniques

Beyond initial adaptation, significant research has focused on methodologies to elevate model performance from simple knowledge recall to sophisticated clinical reasoning and safe application.

**Instruction Tuning and Data-Centric Paradigms.** Instruction tuning has proven critical for aligning model behavior with the specific formats and expectations of clinical tasks. Seminal work on Med-PaLM and its successor, Med-PaLM 2, demonstrated substantial performance improvements on standardized medical qualifying exams through this technique [1], [2]. The efficacy of this method is highly dependent on the quality of the instruction dataset, a principle demonstrated by AlpaCare [26], which achieved significant improvements using the MedInstruct-52k dataset. Data-centric innovations, such as the unified one-stage training protocol of HuatuoGPT-II [27], have also emerged to improve the efficiency and stability of knowledge integration.

**Knowledge Grounding and Reasoning Enhancement.** A fundamental limitation of LLMs is their reliance on static, parametric knowledge, which can lead to factual inaccuracies or "hallucinations" [48]. To address this, Retrieval-Augmented Generation (RAG) has been adopted, equipping models with mechanisms for real-time information retrieval from external knowledge sources, a technique utilized by models like ChatDoctor [15] and Qilin-Med [25]. The development of large-scale, high-quality knowledge sources, such as the MIRIAD dataset of 5.82 million question-response pairs derived from peer-reviewed literature, is instrumental for grounding model outputs in evidence-based medicine [47].

**Alignment with Human Preferences and Reasoning.** Ensuring that model outputs are not only factually correct but also safe and exhibit sound reasoning is a paramount concern. To this end, alignment techniques such as Direct Preference Optimization (DPO) and Reinforcement Learning from Human/AI Feedback (RLHF/RLAIF) have been employed [38], [36]. Models like HuatuoGPT [19] and Zhongjing [22] specifically utilized RL to align with expert preferences. More advanced paradigms seek to explicitly structure the model's reasoning process. These include frameworks that incentivize unified reasoning across diverse tasks through reinforcement learning [46], or employ structured search methods to construct verifiable, step-by-step logical chains, a concept termed "Slow Thinking" [39]. Other research aims to close the gap between AI and expert cognition by training models on synthetic data that explicitly mimics physicians' cognitive pathways [42].

#### 2.1.3 Specialization and Application Frontiers

The maturation of the field is marked by a trend towards specialization, with models being developed for specific linguistic contexts, medical disciplines, and clinical functions.

**Linguistic and Regional Specialization.** To ensure global equity and clinical relevance, models are being developed for diverse linguistic and cultural contexts. This includes systems tailored for Chinese medical dialogue [18], [37], Traditional Chinese Medicine [22], [40], and the Arabic language through bilingual Mixture-of-Experts architectures [32]. Large-scale initiatives like the Apollo project aim to democratize medical AI by supporting multiple major world languages [34].

**Sub-Discipline Specialization.** Models are increasingly being focused on specific medical sub-disciplines to provide deeper, more relevant expertise. Examples include systems designed for oncology conversations [30], anesthesiology decision support [41], and pediatric question answering [45]. This specialization allows for training on highly specific data, leading to greater accuracy and utility within a defined clinical scope.

**Functional Specialization.** Beyond conversational roles, LLMs are being functionally specialized for structured clinical tasks. Research has demonstrated their efficacy in clinical prediction, such as forecasting disease onset or hospital readmission [23]. Another critical function is structured information extraction from unstructured clinical notes, for which dedicated workflows are being developed to enhance data interoperability and support clinical validation [43].

### 2.2 Multimodal Medical LLMs

A significant portion of diagnostic data in medicine is visual. Multimodal Large Language Models (MLLMs), which integrate vision and language processing [79], [80], represent a critical evolution towards a more comprehensive clinical AI. These models aim to reason jointly over medical images and associated textual information.

#### 2.2.1 Architectural Blueprint: Vision-Language Fusion

The design of medical MLLMs has converged on a modular architecture that effectively combines vision and language processing components.

**The De Facto Architecture.** The predominant architecture for medical MLLMs is a tripartite design consisting of: (1) a Vision Encoder for feature extraction; (2) an LLM for reasoning; and (3) a Connector module to align the two modalities. This modularity allows for leveraging powerful, independently pre-trained components. This design is instantiated in numerous works [4], [54], [57], including systems optimized for efficiency by freezing LLM parameters during training [56] or by adopting a linear-attention design [81].

**Adaptation to Diverse Imaging Modalities.** A key architectural challenge lies in adapting to the wide array of medical imaging modalities. For 2D images, specialized models have been developed for radiography [50], digital pathology [52], [62], and dermatology [65], [67]. The analysis of volumetric

TABLE 2
Overview of medical vision-language models.

| Model Name | Vision Encoder | LLM Backbone | Training Dataset | Release |
|---|---|---|---|---|
| MedBLIP [49] | EVA-CLIP-ViT | BioMedLM | MedNLI, NACC, OASIS | 23/05 |
| XrayGLM [50] | ViT-G | ChatGLM | MIMIC-CXR, OpenI | 23/05 |
| MedVInT [51] | PMC-CLIP-ViT | PMC-LLaMA | PMC-VQA | 23/05 |
| PathAsst [52] | PathCLIP-ViT | Vicuna | PathCap, PathInstruct | 23/05 |
| PCLMed [53] | EVA-CLIP-ViT | ChatGLM | ImageCLEF 2023 caption prediction | 23/06 |
| LLaVA-Med [4] | CLIP-ViT | LLaMA | PMC-15M, VQA-RAD, SLAKE, PathVQA | 23/06 |
| XrayGPT [54] | MedCLIP-ViT | Vicuna | MIMIC-CXR, OpenI | 23/06 |
| Med-Flamingo [5] | CLIP-ViT | LLaMA | MTB, PMC-OA | 23/07 |
| Med-PaLM M [29] | ViT-e, ViT-22B | PaLM | MultiMedBench | 23/07 |
| RadFM [55] | 3D ViT | Med-LLaMA-13B | RadMD, RadMD | 23/08 |
| R2GenGPT [56] | Swin Transformer | LLaMA 2 | IU Xray, MIMIC-CXR | 23/09 |
| Qilin-Med-VL [57] | ViT | LLaMA-2-Chinese | Med-VL | 23/10 |
| MAIRA-1 [58] | RAD-DINO | Vicuna | MIMIC-CXR | 23/11 |
| PeFoM-Med [59] | EVA-CLIP-ViT | LLaMA-2 | PMC-OA, VQA-RAD | 24/01 |
| CheXagent [60] | EVA-CLIP-ViT | Mistral | KInstruct | 24/01 |
| M3D-La-Med [61] | 3D ViT | LLaMA-2 | M3D-Data | 24/03 |
| PathChat [62] | UNI | LLaMA 2 | Med, WSI | 24/06 |
| HuatuoGPT-Vision [63] | CLIP-ViT | Yi-1.5 | MedVision, HuatuoGPT-II | 24/06 |
| miniGPT-Med [64] | EVA-CLIP-ViT | LLaMA 2 | MIMIC, NLST, SLAKE, RSNA, RadVQA | 24/07 |
| SkinGPT-4 [65] | ViT | LLaMA 2 | ISICON, Dermnet | 24/07 |
| LLaVA-Med++ [66] | CLIP-ViT | LLaMA 2 | PMC-Trinity-25M, VQA-RAD, SLAKE, PathVQA | 24/08 |
| MpoxVLM [67] | CLIP + ViT | LLaMA-2-7B | Mpox skin lesion VQA | 24/11 |
| UMed-LVLM [68] | PMC-CLIP-ViT | MedVInT | MAU dataset | 25/01 |
| MedVLM-R1 [69] | ViT (Qwen2-VL-2B) | Qwen2-VL-2B | Radiology VQA | 25/02 |
| HealthGPT [70] | CLIP-L/14 | phi-4 | Unified comprehension & generation dataset | 25/02 |
| RetinalGPT [71] | ViT | LLaMA | Retinal image analysis | 25/03 |
| Med-R1 [72] | ViT (Qwen2-VL-2B) | Qwen2-VL-2B | Multi-modal medical VQA | 25/04 |
| PathVLM-R1 [73] | ViT (Qwen2.5-VL-7B) | Qwen2.5-VL-7B | Pathology VQA | 25/04 |
| QoQ-Med [74] | ViT (Qwen2.5-VL) + ECG-JEPA | Qwen2.5-VL | 1D ECG, 2D/3D medical images, text | 25/05 |
| Reg2RG [75] | ViT3D + Mask Encoder | LLaMA2-7B | Region-grounded CT report generation | 25/05 |
| ChestGPT [76] | ViT (EVA-ViT) | LLaMA-2 | Chest X-ray disease detection | 25/07 |
| MCA-RG [77] | ResNet-50 | LLaMA2-7B | Radiology Report | 25/07 |
| MedGemma [78] | MedSigLIP | Gemma 3 4B / 27B | Multimodal medical reasoning data | 25/07 |

3D data, such as CT and MRI, requires more complex vision backbones. Solutions have included novel modules to bridge 3D data with 2D encoders [49] or the direct integration of 3D-aware vision transformers [55], [61].

**Towards a Generalist Biomedical AI.** The ultimate objective is a unified model capable of processing and reasoning over the full spectrum of biomedical data. Pioneering work produced systems that can jointly process text, medical images, and genomics within a single framework [29]. More recent architectures are expanding this to include other data types, such as time-series ECG signals, demonstrating the feasibility of a truly generalist, all-modality clinical reasoning model [74].

### *2.2.2 Core Capabilities and Clinical Applications*

The integration of vision enables MLLMs to perform a range of clinically relevant tasks that are inaccessible to their unimodal counterparts.

**Automated Report Generation.** A principal application is the automation of radiology report generation. Initial approaches focused on generating descriptive text from global image features [53], [58]. More sophisticated methods now enhance clinical accuracy by explicitly aligning visual features with a medical concept library [77] or by grounding the generated text in specific anatomical regions of interest, thereby improving the report's fidelity and interpretability [75].

**Interactive Diagnostic Assistance.** MLLMs can function as interactive assistants for clinicians, facilitating visual question answering (VQA). The development of large-scale VQA datasets has been instrumental for training such models [51]. This has led to the creation of specialized conversational assistants for disciplines like pathology [62] and ophthalmology [71], as well as few-shot learners capable of generating rationales for their answers [5].

**Abnormality Detection and Localization.** For diagnostic utility, identifying and localizing abnormalities is crucial. MLLMs are being designed to output not just a diagnosis but also spatial coordinates, often as bounding boxes, correspond-

ing to pathological findings [64], [76]. Training paradigms are also being refined to explicitly improve this capability, for instance, by incorporating an "abnormal-aware" feedback mechanism that rewards the model for correctly attending to and identifying pathological regions [68].

#### 2.2.3 Data Imperative and Training Paradigms

Progress in medical MLLMs is fundamentally coupled with advancements in data curation and the sophistication of training strategies.

**The Role of Large-Scale Annotated Datasets.** The performance of MLLMs is contingent upon the availability of large-scale, high-quality, paired image-text data. The field has progressed from leveraging existing figure-caption pairs in biomedical literature to the systematic construction of massive datasets. An important innovation is the use of powerful MLLMs themselves to re-format and de-noise web-crawled data, creating high-quality VQA pairs at scale [63]. The current state-of-the-art in data curation involves building datasets with multi-granular annotations, providing both image-level labels and fine-grained, region-of-interest annotations, which enables the training of more precise and versatile foundation models [66].

**Evolution of Training Strategies.** Training paradigms have evolved in sophistication. While a two-stage process of vision-language pre-training followed by instruction fine-tuning remains a common and effective strategy [4], [60], Parameter-Efficient Fine-Tuning (PEFT) methods are increasingly used to reduce the substantial computational burden [82]. A significant recent trend is the adoption of reinforcement learning to move beyond simple pattern matching towards explicit reasoning. RL-based frameworks are being used to guide models to generate interpretable, human-aligned reasoning paths, which is considered a critical step for enhancing the trustworthiness and clinical adoption of these systems [69], [73], [72].

### 2.3 From Foundational Models to Autonomous Agents: A Necessary Leap

The rapid development of unimodal and multimodal medical LLMs has equipped the field with powerful tools for knowledge synthesis and data interpretation. Through advanced training paradigms, these models can achieve expert-level performance on a range of specialized tasks. However, their role remains that of a sophisticated assistant, not an autonomous partner. They can answer complex questions about a given clinical context, but they cannot independently formulate a diagnostic plan, actively seek out new information using external tools, or manage a patient's care over time. This fundamental limitation in *agency*, the capacity to sense, reason, and act proactively, motivates the critical evolutionary leap from knowledge-rich models to goal-oriented Medical Agents, which we will explore in the next section.

## 3 Medical Agentic System Construction

This section provides a comprehensive overview of Medical Agentic Systems, detailing their fundamental concepts, the core components of an individual agent, and the prevalent architectural designs that enable complex, collaborative behaviors.

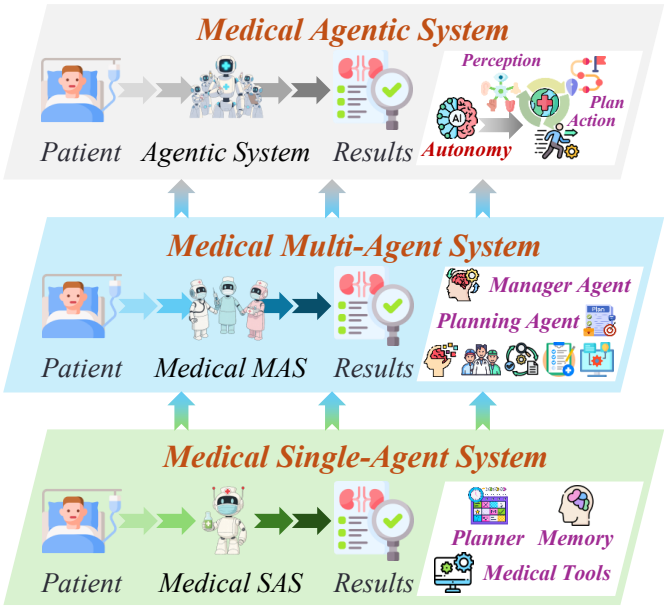


Fig. 2. **The architectural progression toward autonomous Medical Agents.** The diagram charts the paradigm shift from basic components to complex systems. It begins with the fundamental **Medical Single-Agent System**, which integrates planning, memory, and tools. This evolves into the more sophisticated **Medical Multi-Agent System**, where multiple agents collaborate under coordination to handle complex workflows. Both architectures are implementations of the overarching **Medical Agentic System**, defined by its autonomous perception-plan-act loop.

### 3.1 Conceptual Foundations: From Language Models to Agentic Systems

The paradigm shift from passive, information-processing models to proactive, autonomous agents represents a fundamental advancement in medical AI. This evolution necessitates a clear conceptual framework to systematically analyze the capabilities and architectures of these emerging systems. This subsection delineates the foundational concepts, establishing a clear distinction between Medical Large Language Models (LLMs) as knowledge engines and Medical Agents as goal-oriented actors within broader agentic systems.

#### 3.1.1 Definition

**Medical LLM.** The foundation of this field is the **Medical Large Language Model (Medical LLM)**, a sophisticated AI model, typically based on a transformer architecture, that has been trained or fine-tuned on vast corpora of medical literature and clinical data. Its primary strength lies in understanding and generating medical text, enabling it to answer questions, summarize records, and assist in documentation [83], [84]. Some

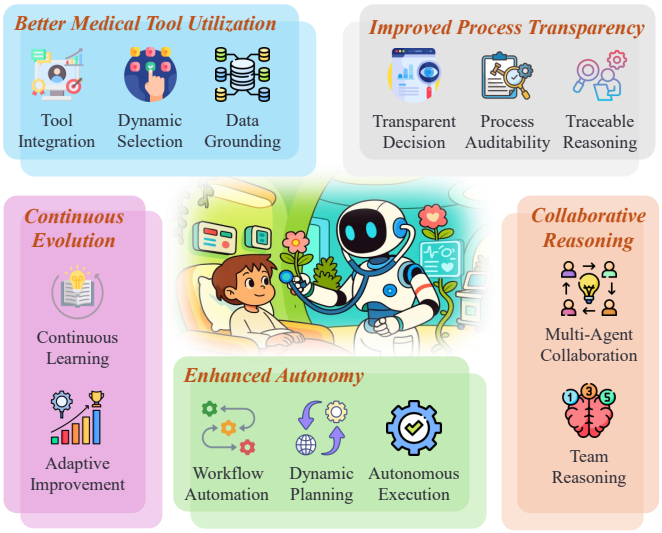


Fig. 3. Core capabilities unlocked by the paradigm shift to autonomous Medical Agents, centered on the principle of **Enhanced Autonomy**. This is supported by four key advancements: **(1) Better Medical Tool Utilization** for dynamic integration with external tools and data; **(2) Improved Process Transparency** ensuring traceable and auditable reasoning; **(3) Continuous Evolution** through ongoing learning and adaptation; and **(4) Collaborative Reasoning** in multi-agent systems to solve complex clinical problems.

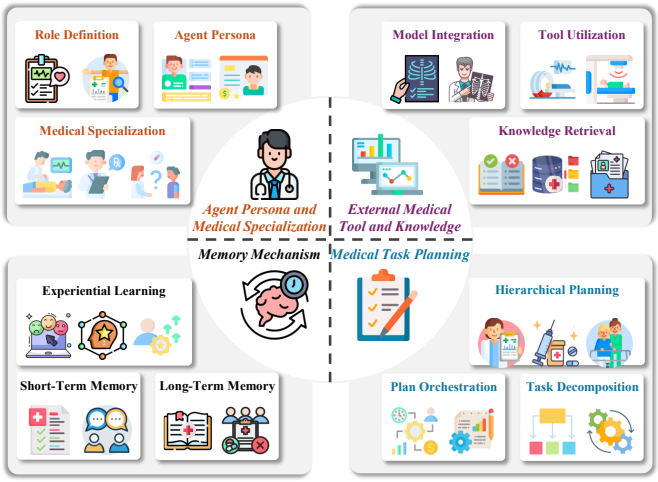


Fig. 4. Architectural blueprint of a Medical **Single-Agent System (SAS)**. The core components include a cognitive engine for **Medical Task Planning** and a dual **Memory Mechanism** for both short-term context and long-term learning. The agent's behavior is defined by its **Persona** and **Medical Specialization**, and its capabilities are extended via **Tool Utilization** to interact with external data and knowledge sources.

are enhanced with vision capabilities, becoming multimodal models [85]. However, a Medical LLM is fundamentally a reactive tool; it processes inputs and generates outputs but requires explicit, step-by-step human guidance for complex tasks. Even when augmented with retrieval capabilities (RAG), which grounds its knowledge, it primarily functions as a powerful reasoning engine on provided context, lacking true autonomy and the ability to execute multi-step plans [13].

**Medical Agent.** A **Medical Agent** represents a significant leap forward. It is an AI system that uses a Medical LLM as its core reasoning engine, but is further endowed with a framework for autonomous operation. Unlike a passive LLM that merely reacts to queries, a Medical Agent is proactive: it can perceive its clinical environment, autonomously formulate multi-step plans, maintain memory of past interactions, and actively leverage external tools to achieve specific objectives [83]. This shift from reactive information processing to proactive problem-solving is the defining characteristic of agency.

**Medical Agentic System.** This is the overarching term for any system built around one or more Medical Agents. A **Single-agent System** features one agent working autonomously to complete a task. A **Multi-agent System (MAS)** involves multiple agents collaborating, representing the most sophisticated architectural paradigm. The key distinction throughout this spectrum is the degree of **agency**, the capacity to operate proactively and autonomously to effect change or achieve goals within an environment.

### *3.1.2 Significance*

The shift towards agentic systems offers capabilities that are critical for real-world healthcare applications and addresses the limitations of passive LLMs:

**Enhanced Autonomy.** Agents can independently plan and execute multi-step clinical workflows, such as performing a differential diagnosis or managing a patient's care pathway, by dynamically interacting with their environment and tools [9], [86]. This moves beyond simple Q&A to actively participating in and driving clinical processes.

**Better Medical Tool Utilization.** Agents excel at dynamically selecting and utilizing a diverse array of external medical tools, from live databases and clinical calculators to specialized AI models for image analysis. This ensures their outputs are grounded in current, verifiable, and quantitative data, overcoming the static knowledge limitations of LLMs and reducing factual errors [87], [88], [89].

**Collaborative Reasoning.** Multi-agent systems simulate the collaborative intelligence of human multidisciplinary teams (MDTs). This architecture facilitates a richer reasoning process through structured debate, role-playing, and consensus-building, allowing for the exploration of diverse hypotheses and the mitigation of individual cognitive biases, a feat difficult for a monolithic LLM to achieve [90], [91], [11].

**Continuous Evolution.** Agentic frameworks can be designed for continuous learning. By interacting within simulated environments and reflecting on outcomes, agents can refine their strategies and improve performance over time. This is

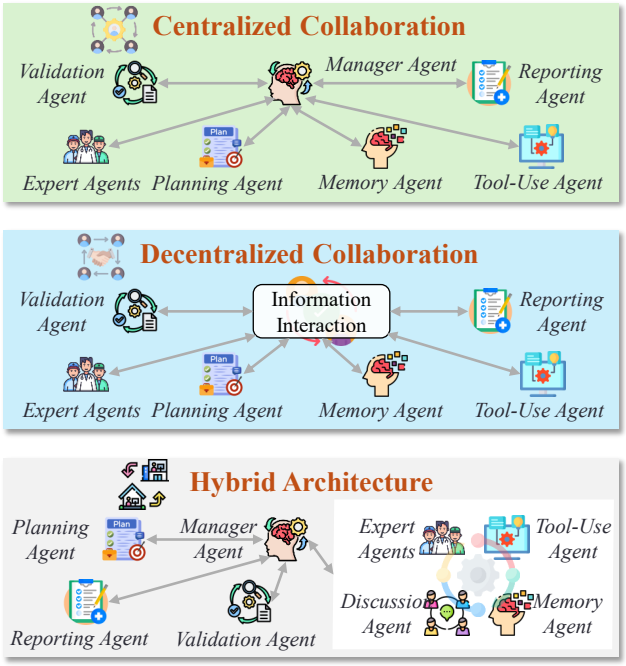


Fig. 5. Core collaboration paradigms for a Medical **Multi-Agent System (MAS)**. **(1) Centralized Collaboration** employs a 'Manager Agent' to orchestrate the workflow in a top-down manner. **(2) Decentralized Collaboration** enables emergent consensus through peer-to-peer agent interaction without a central controller. **(3) The Hybrid Architecture** strategically balances centralized coordination with flexible, direct inter-agent communication for enhanced adaptability.

a key feature for systems that must adapt to new medical knowledge and evolving clinical standards [10], [92].

**Improved Process Transparency.** By design, agents operate through explicit steps of planning and action. This creates a traceable log of their "thought process", offering greater transparency into the decision-making pipeline. This audit trail is crucial for building trust, enabling clinical scrutiny, and aligning with regulatory requirements in high-stakes medical applications [8], [93].

## 3.2 Single-agent System

To operationalize abstract reasoning capabilities into concrete clinical actions, a Medical Agent is typically constructed from several core cognitive components. These components collectively enable its autonomous and intelligent behavior.

### 3.2.1 Agent Persona and Medical Specialization

The foundation of a medical agent is its defined **persona** and degree of **medical specialization**. This involves establishing the agent's identity, such as a general practitioner, oncologist, radiologist, or geneticist, and endowing it with the requisite domain knowledge and communication style for that role [94], [95]. This is often achieved through tailored system prompts that define the agent's expertise and responsibilities, or by providing access to specific knowledge bases relevant to its designated function [90], [11]. For example, an agent's persona can determine whether it communicates in technical jargon for a specialist audience or in patient-friendly language, a critical aspect for applications like generating simplified medical reports [96]. Furthermore, systems like PIORS demonstrate the use of diverse, batch-generated patient personas to create more realistic and challenging training scenarios for conversational agents [97].

### 3.2.2 Medical Task Planning

A core capability is proficiency in **medical task planning and orchestration**, which serves as the direct operational manifestation of the agent's reasoning process. This entails decomposing complex medical objectives into a sequence of manageable sub-goals or executable actions. Frameworks like ReAct or Chain-of-Thought are often used to structure this process [98], [99]. For instance, ClinicalAgent uses a LEAST-TO-MOST strategy to break down the complex problem of clinical trial analysis into feasibility, safety, and efficacy assessments [8]. Some systems adopt hierarchical planning; MedAgent-Pro first generates a high-level, disease-specific diagnostic plan based on clinical guidelines, and then a patient-level reasoning module executes the relevant steps from that plan [100]. Orchestration involves managing this plan, which in multi-agent systems means delegating steps to the most appropriate specialized agents, such as a central "DDxDriver" coordinating knowledge retrieval and diagnosis agents in a dynamic, iterative loop to refine a differential diagnosis [14].

### 3.2.3 Memory Mechanism

To enable context retention and continuous improvement, agents incorporate sophisticated **memory mechanisms**, which are typically divided into two main types. **Short-term memory** is responsible for maintaining context within a single, ongoing consultation. **Long-term memory**, in contrast, is the mechanism crucial for **experiential learning**, allowing an agent to store, retrieve, and learn from past cases [9]. This is implemented in various ways to build a foundation for improvement. For example, MDTeamGPT maintains separate knowledge bases for correct consultations ("CorrectKB") and analyzed failures ("ChainKB") to enable reflection and prevent repeated errors [101]. Similarly, MedAgentSim uses an "Experience Records Buffer" to store corrected misdiagnoses, allowing the agent to learn from its mistakes [92]. In systems like Agent Hospital, agents evolve by expanding a "Medical Case Base" with successful treatments and an "Experience Base" with rules learned from failures, directly mimicking how human doctors gain expertise through experience [10]. This learning capacity, enabled by long-term memory, is vital for creating agents that can adapt and improve over time [102].

### 3.2.4 External Medical Tool and Knowledge

An agent's "agency" is most clearly demonstrated through its ability to interact with and leverage **external medical tools and knowledge**. This capability is enabled by a tool-use mechanism that defines the repertoire of actions an agent can perform. Seamless integration with these external resources

is critical, involving a process where the agent autonomously selects the appropriate tool for a sub-task, formats the input parameters, executes the action, and correctly interprets the output. This enables the agent to access a wide range of capabilities and information, such as querying web APIs and structured databases like DrugBank or OncoKB for up-to-date medical knowledge [87], [103]; calling specialized AI models for tasks like image segmentation (MedSAM) or bioinformatics analysis (Phenomizer) [88], [104], [90]; and generating code to perform clinical calculations, conduct statistical analysis, or interact with rich knowledge sources like EHRs [89], [9]. This tool-use capability is fundamental to an agent's problem-solving ability, extending its reach far beyond the static, inherent knowledge of its core LLM [93].

## 3.3 Medical Multi-agent System

While single-agent systems represent a significant step towards autonomous medical AI, their monolithic nature can struggle to replicate the collaborative intelligence essential to modern healthcare. Clinical decision-making, particularly for complex cases, and it usually relies on the converged expertise of multidisciplinary teams. This real-world paradigm motivates the shift towards multi-agent systems (MAS), which are architected to emulate the collaborative dynamics of a clinical team.

Compared to a single-agent system, an MAS introduces a fundamental evolution across its core functions. Instead of a single **persona**, an MAS comprises multiple agents with highly specialized roles (e.g., a radiologist, an oncologist), enabling deeper and more focused expertise. **Task planning** evolves from a linear decomposition of steps into a sophisticated process of coordination and delegation, often managed by a coordinator agent. **Memory** can be transformed into a shared context, akin to a patient's electronic health record, that all agents contribute to and reference. Finally, specialized agents can leverage more targeted **external tools**, enhancing both precision and efficiency. This collaborative architecture is designed to achieve a more robust, accurate, and transparent reasoning process by mitigating the cognitive biases of a single model and integrating diverse expert perspectives. These systems emulate real-world clinical workflows by structuring how agents interact and share information, with collaboration paradigms broadly categorized into centralized, decentralized, and hybrid architectures, each offering distinct advantages for specific medical scenarios.

### 3.3.1 Centralized Collaboration

In a centralized collaboration model, a single coordinator or manager agent orchestrates the workflow, directing tasks and synthesizing information from other specialized agents. This "hub-and-spoke" architecture ensures a controlled, interpretable, and often sequential reasoning process, making it ideal for tasks that require structured problem decomposition and integration of diverse data sources.

A common implementation involves a master agent that manages a team of subordinate specialist agents. For instance, the **ColaCare** framework utilizes a *MetaAgent* to collect initial reviews from multiple *DoctorAgents*, synthesize them into a preliminary report, and then moderate a multi-round debate to reach a consensus on EHR data analysis [105]. Similarly, **MAM** employs a *Director* agent to orchestrate a debate among specialists (e.g., Radiologist, General Practitioner), synthesize their opinions, and formulate the final diagnosis [106]. This approach is also evident in systems like ClinicalAgent, where a *Planning Agent* first decomposes a complex clinical trial problem into sub-tasks (feasibility, safety, efficacy) and a final *Reasoning Agent* aggregates the findings from specialized agents to produce a holistic conclusion [8]. The central orchestrator ensures that the final decision is grounded in a comprehensive analysis, mirroring the role of a lead physician or a case manager in a clinical setting.

### 3.3.2 Decentralized Collaboration

Decentralized collaboration models foster a more dynamic, peer-to-peer interaction among agents without a single, authoritative controller. This "round-table" approach emulates the collaborative reasoning of a Multidisciplinary Team (MDT), where consensus emerges from structured debate and discussion among specialists. This architecture is particularly effective for complex diagnostic challenges, such as rare diseases, where integrating multiple expert perspectives can mitigate individual model biases and lead to more accurate outcomes.

Many frameworks are explicitly designed to mirror an MDT case conference. In the MAC framework, multiple *Doctor Agents* and a *Supervisor Agent* engage in a multi-round conversation to diagnose complex and rare diseases, with the supervisor facilitating the debate rather than dictating the outcome [107]. Likewise, MedAgents and RareAgents assemble teams of specialist agents who analyze a case from their unique perspectives, debate hypotheses, and work towards a consensus diagnosis or treatment plan [11], [90]. Communication can be managed through structured protocols to enhance efficiency; for example, MDTeamGPT uses a "residual discussion structure" where agents only see summaries of previous rounds, reducing cognitive load while fostering a focused debate [101]. In these systems, the final decision is a product of collective intelligence, emerging from the dynamic interplay of expert opinions.

### 3.3.3 Hybrid Architecture

Hybrid architectures strategically combine elements of both centralized and decentralized models to balance control with flexibility. A prominent hybrid model is the sequential pipeline, where tasks are processed through a series of specialized agents in a predefined order. While the overall workflow is centrally controlled by the sequence, each agent operates with a degree of autonomy within its specific stage. This structure is ideal for standardizing and optimizing established clinical protocols.

For example, the Rx Strategist system verifies prescriptions through a rigid pipeline of agents for Feature Extraction, Indication Verification, and Dosage Verification, where each agent's output serves as the input for the next [93]. A more complex pipeline is demonstrated by PathFinder, which

emulates a pathologist's workflow for analyzing whole-slide images through a sequence of *Triage*, *Navigation*, *Description*, and *Diagnosis* agents [108]. Similarly, the Multi-Agent Inpatient Pathways (MAP) system simulates the entire inpatient journey with a sequence of specialized agents: a *Triage Agent*, a *Diagnosis Agent*, and a *Treatment Agent*, all overseen by a *Chief Agent* [109]. These pipeline-based systems enforce a logical progression of tasks while still leveraging the specialized expertise of individual agents at each step, providing a robust and structured approach to complex medical decision-making.

Finally, some of the most ambitious architectures use these collaborative principles to construct full-scale digital simulacra of healthcare environments. Systems like Agent Hospital [10], AI Hospital [110], and MedAgentSim [92] populate a virtual hospital with dozens of autonomous agents that collaborate to manage the entire patient lifecycle. Such environments serve as invaluable sandboxes for training and validating agent behaviors in a safe, controlled, and scalable manner.

### 3.4 From Architecture to Engine

Medical Agentic Systems, whether single or multi-agent, provide the necessary architectural "body" that allows AI to operate autonomously in dynamic clinical settings. They furnish the mechanisms for planning, remembering, acting, and collaborating. However, the intelligence and reliability of this body depend entirely on the sophistication of its core "engine": medical reasoning. Without advanced reasoning capabilities, even the most complex agentic architecture cannot effectively navigate the nuances of real-world medical decision-making. The following section will delve into the evolutionary trajectory of these reasoning techniques, exploring how they have advanced from simple prompt engineering to complex, self-evolving learning paradigms.

### 3.5 Medical Agentic System

A Medical Agentic System marks a fundamental paradigm shift, evolving AI from passive, information-retrieving Large Language Models (LLMs) into proactive, goal-oriented frameworks. The core innovation lies in **agency**: a capacity that extends beyond merely reacting to queries. Instead, agents can autonomously perceive their environment, formulate multi-step plans, maintain memory of past interactions, and execute actions by leveraging external tools to achieve specific objectives [83]. This proactive, self-directed nature is what allows agentic systems to effectively model and navigate the complexities of real-world clinical workflows, which are inherently dynamic, uncertain, and multi-faceted.

The value of this agency is demonstrated across critical clinical applications. For instance, agents can proactively gather information by dynamically optimizing questioning strategies in dialogues, as seen in *DoctorAgent-RL* [111]. They can also act as expert orchestrators that coordinate specialized tools, such as the image analysis models managed by *MMedAgent* and *ADAgent*, to form a cohesive diagnosis [88], [112]. Furthermore, agents like *MedAgent-Pro* can enforce adherence to evidence-based medicine by autonomously retrieving clinical guidelines and executing a procedural plan, ensuring their reasoning is both transparent and grounded in established standards [100].

## 4 Reasoning Evolution in Med Agents

The architectural frameworks described in the previous chapter provide the "body" for Medical Agents, enabling them to plan, act, and interact. This chapter focuses on the "engine" that drives this body, an engine specifically designed to perform one of the most complex cognitive tasks: medical reasoning. Traditionally, this is the process by which healthcare professionals diagnose, prognose, and manage patient health issues [115], [116], [117], [118], [119]. The reasoning paradigm in LLM-based agents is engineered to emulate this clinical workflow, advancing AI's reliability and interpretability in high-stakes medical applications by strategically decomposing complex problems into manageable, interconnected steps [120], [121], [122], [123], [124], [103], [125], [126].

This chapter charts this evolution across three distinct stages of increasing sophistication. Our exploration begins inwardly, with the foundational techniques that enhance an agent's *core reasoning* by improving its internal deliberation. From this foundation, the agent's capabilities expand outwardly in a pivotal leap to *augmented reasoning*, where it learns to connect with the external world by retrieving knowledge, using tools, and interpreting visual data. Finally, we examine the pinnacle of this evolution: a shift from individual cognition to *collective reasoning*, where collaborative intelligence emerges from the interaction of multiple specialized agents. These three evolutionary stages are summarized in Table 3.

### 4.1 Core Reasoning: Enhancing a Single Agent's Internal Deliberation

The foundation of an agent's intelligence lies in its core reasoning capability, the ability to process information and formulate a logical path to a conclusion without external assistance. These inference-time techniques enhance medical reasoning without updating model parameters by structuring the generation process to elicit more deliberate thought.

The foundational method is Chain-of-Thought (CoT) prompting, which encourages the model to generate intermediate steps before the final answer, mimicking a clinician's "thinking aloud" process [12], [127], [128], [129]. Generating this structured rationale before the conclusion, as opposed to providing a direct answer, has been shown to significantly improve performance on complex medical tasks [130], [131], [98], [113], [132], [133], [134], [42]. For instance, models like Med-PaLM 2 [113] and MedPrompt [133] leverage CoT to achieve expert-level performance on medical question-answering benchmarks. This reasoning is often guided by In-Context Learning (ICL), which adapts the model to a specific task by providing few-shot examples within the prompt [132]. For instance, by showing an LLM examples of how to extract structured information like symptoms or medication history from clinical narratives, its ability to perform that task on new notes is greatly improved [135], [136], [43], [137], and it can also be applied to tasks like radiology report generation [77].

TABLE 3
The Evolutionary Stages of Reasoning in Medical Agents. This taxonomy outlines the progression from enhancing an individual agent's internal thought processes to orchestrating complex, collaborative intelligence across multiple agents.

| Reasoning Stage | Core Principle | Key Techniques | Representative Work |
|---|---|---|---|
| **Core Reasoning** (Internal Deliberation) | Enhancing the deliberation process of a single agent by structuring its thought process, generating multiple reasoning paths, or iteratively refining its own outputs. | Chain-of-Thought, In-Context Learning, Self-Consistency, Self-Correction | Med-PaLM 2 [113], DrHouse [86], EnsReas [114] |
| **Augmented Reasoning** (External Grounding) | Grounding the agent's reasoning by connecting it to the external world, including dynamic knowledge bases, functional tools, and perceptual visual data. | Retrieval-Augmented Generation (RAG), Tool Use, Vision Augmentation | Almanac [13], EHRAgent [9], Med-Flamingo [5] |
| **Collective Reasoning** (Multi-Agent Collaboration) | Decomposing a complex problem to be solved by a collaborative team of specialized agents who communicate, debate, and deliberate to reach a consensus. | Multi-Agent Debate & Orchestration | MedAgents [11], Agent Hospital [10], MedDxAgent [14] |

To further bolster reliability, methods like Self-Consistency [138] and Ensemble Reasoning [114] generate multiple reasoning chains and select the most frequent or well-reasoned answer via voting or reconciliation. This approach has proven particularly effective in improving accuracy for complex diagnostic tasks [138], [136], [139]. Furthermore, iterative self-correction allows an agent to reflect on and refine its outputs, correcting errors before finalizing a decision, as exemplified in systems that iteratively refine diagnoses based on new information [140], [86].

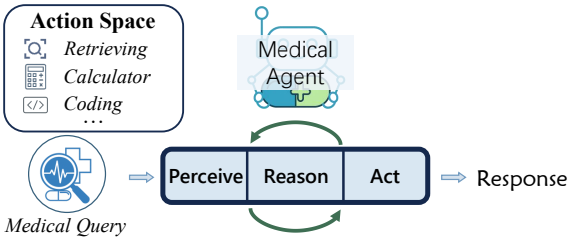


Fig. 6. Operational workflow of a tool-using medical agent.

## 4.2 Augmented Reasoning: Connecting to the External World

While core reasoning techniques refine an agent's internal thought process, their effectiveness is ultimately bound by the static knowledge within the LLM's parameters. The next crucial step in the evolution of reasoning is to empower agents to break free from these confines and interact with the external world. This augmented reasoning paradigm connects agents to external knowledge, tools, and perceptual data, dramatically expanding their capabilities and grounding their outputs in real-world, verifiable information.

### *4.2.1 Knowledge Augmentation via Retrieval*

A primary method to augment reasoning is through Retrieval-Augmented Generation (RAG), which expands agents' capabilities by grounding their outputs in verifiable, up-to-date evidence [13], [8], [86], [9]. RAG systems, such as Almanac [13] and Health-LLM [141], enable agents to retrieve information from vast medical literature, such as domain-specific information, clinical guidelines, or electronic health records (EHRs) [87], [142], [9]. This ensures responses are grounded in the latest evidence and reduces the risk of generating incorrect or outdated information, a critical safety feature for medical applications.

### *4.2.2 Functional Augmentation via Tool Use*

Beyond retrieving static information, agents can achieve a higher level of reasoning by using external tools to perform actions and calculations. Figure 6 illustrates the general workflow of such an agent, which operates in a cycle of perceiving, reasoning, and acting. This paradigm shift allows models to interact with knowledge bases, calculators, and clinical systems, enhancing factuality, timeliness, and reasoning capabilities [8], [9], [86], [143], [144], [145]. In the medical domain, agents can integrate with tools like medical calculators for clinical scoring [89], [146], [147], code interpreters for complex data analysis [9], [145], and web search to access the latest biomedical information [87], [86]. However, training an agent to reliably use tools introduces unique challenges that require specialized training paradigms.

**Training Paradigms for Tool-Using Medical Agents.** Training an LLM to effectively use tools requires specialized methods that teach the model to generate a sequence of thoughts and actions, a tool-use trajectory, that leads to a correct solution [148], [149]. Two primary strategies have become prominent for this purpose. The most direct method is through imitation learning, by fine-tuning the LLM on expert-demonstrated trajectories [150], [145]. This process teaches the model to generate specific tool-calling syntax, interpret the returned information, and synthesize it to address complex clinical queries [151], [152], [153].

A more advanced approach involves learning from the feedback generated by executing tool calls, allowing the agent to explore, adapt, and learn from its own mistakes. This process is often formalized using Reinforcement Learning (RL), where the agent's action (a tool call) is executed, and

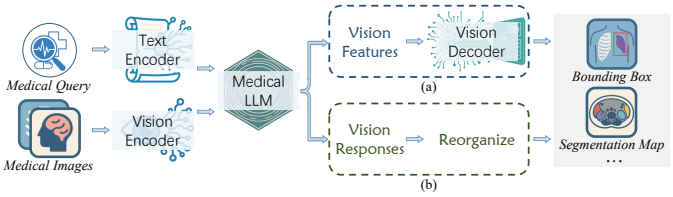


Fig. 7. Two primary strategies for medical visual reasoning. (a): Extracted features from the LLM are passed to an additional vision decoder for prediction. (b): The LLM directly generates predictions as part of its text sequence.

the outcome provides a learning signal. This enables the agent to develop more robust and generalizable problem-solving skills than imitation learning alone, as demonstrated in systems designed for complex reasoning and error recovery [154], [69], [155], [156], [111].

**Medical Data Curation for Tool-Using Agents.** Constructing datasets for tool-using agents is significantly more complex than for standard instruction tuning, as it requires capturing the entire reasoning and action process [157]. The gold standard is the manual creation of tool-use trajectories by clinical experts [147], [154]. To scale data creation, a common strategy is to use powerful teacher models (e.g., GPT-4) to generate synthetic tool-use examples, which are then rigorously reviewed and corrected by human experts. This human-in-the-loop approach, used to build datasets like those for MedAgentGym, effectively balances scalability with quality [145]. Another scalable method involves programmatically converting existing medical resources, like questions from medical benchmarks or problems from textbooks, into tool-use datasets [158], [159], [147].

#### *4.2.3 Perceptual Augmentation via Vision*

In medicine, reasoning is inherently multimodal, relying heavily on visual data. Vision augmentation equips agents with the ability to perceive and interpret medical images, a critical step towards emulating a clinician's diagnostic workflow. This includes 2D images such as X-rays and pathological slides, as well as 3D volumes like CT and MRI scans. The integration of visual data is paramount for enabling comprehensive diagnostic support and robust multimodal reasoning [160], [161], [162], with crucial applications in radiology [163], [162], [54], [64], ophthalmology [164], and pathology [108], [165].

**Grounded Medical Visual Reasoning.** While many multimodal models focus on textual outputs like VQA or report generation [54], [77], [64], [166], a significant advancement lies in grounded visual reasoning. This involves tasks like object detection or segmentation, which provide precise spatial information about clinical findings. Two primary strategies have emerged for this, as illustrated in Figure 7. The first involves training the LLM to directly generate spatial coordinates or mask representations as part of its text sequence [104], [88], [76]. A more common approach uses the LLM as a high-level reasoning engine while delegating the dense prediction task to a separate, specialized visual decoder (e.g., a UNet-style head), which allows each component to be optimized for its respective task [167], [168]. The training paradigm for these models typically involves a multi-task objective, combining a language loss with a vision-specific loss (e.g., Dice or IoU loss).

**Medical Visual Data Curation.** The collection and preparation of medical visual data present unique challenges. Data is typically sourced from hospital Picture Archiving and Communication Systems (PACS) and public datasets [169], [163], [54]. The primary bottleneck is annotation. While expert labeling by clinicians is the gold standard, it is exceptionally time-consuming and expensive [168], [165], [164]. To alleviate this, automated and hybrid methods are increasingly used, where foundational models generate initial annotations that are then refined by human experts in a "human-in-the-loop" approach [170], [171], [172], [168], [173]. A crucial final step is the accurate alignment of visual data with corresponding textual information, such as radiology reports, which is foundational for training powerful models like Med-Flamingo [5] and HuatuoGPT-Vision [161].

#### *4.2.4 The Data Imperative for Advanced Reasoning*

It is crucial to recognize that the evolution from core to augmented and collective reasoning is fundamentally driven by the sophistication of the underlying data. Each leap in reasoning capability requires a corresponding leap in data curation and construction methodology. Figure 8 provides a comprehensive overview of these data pipelines, synthesizing the distinct requirements for the major AI paradigms covered in this survey. The first two quadrants illustrate the data needed to build the foundational models themselves: Textual Data for unimodal LLMs (in Section 2) and Visual Medical Data for the multimodal capabilities central to augmented reasoning. The latter two quadrants show the shift towards more complex, process-oriented data required for advanced agentic behaviors: Tool-Use Trajectories for training tool-proficient agents and Collaborative Data for enabling the multi-agent systems we will discuss next.

### 4.3 Collective Reasoning: Collaborative Intelligence in Medical Multi-Agent Systems

The most advanced paradigm for medical reasoning transcends the capabilities of a single agent and moves towards orchestrating collaborative intelligence within multi-agent systems. By decomposing a complex medical problem into sub-problems handled by specialized agents, these systems emulate the multidisciplinary teams (MDTs) of human experts, aiming to achieve more robust, accurate, and interpretable medical decisions [103], [11], [10], [14], [94], [156], [106], [174]. Training these collaborative systems requires specialized paradigms that foster effective communication, coordination, and consensus-building [160].

#### *4.3.1 Training Paradigms for Collaborative Policies*

Teaching agents to collaborate effectively involves training them not just on task execution but also on interaction protocols. The most direct method is Supervised Fine-Tuning on Collaborative Trajectories, where the system learns to imitate

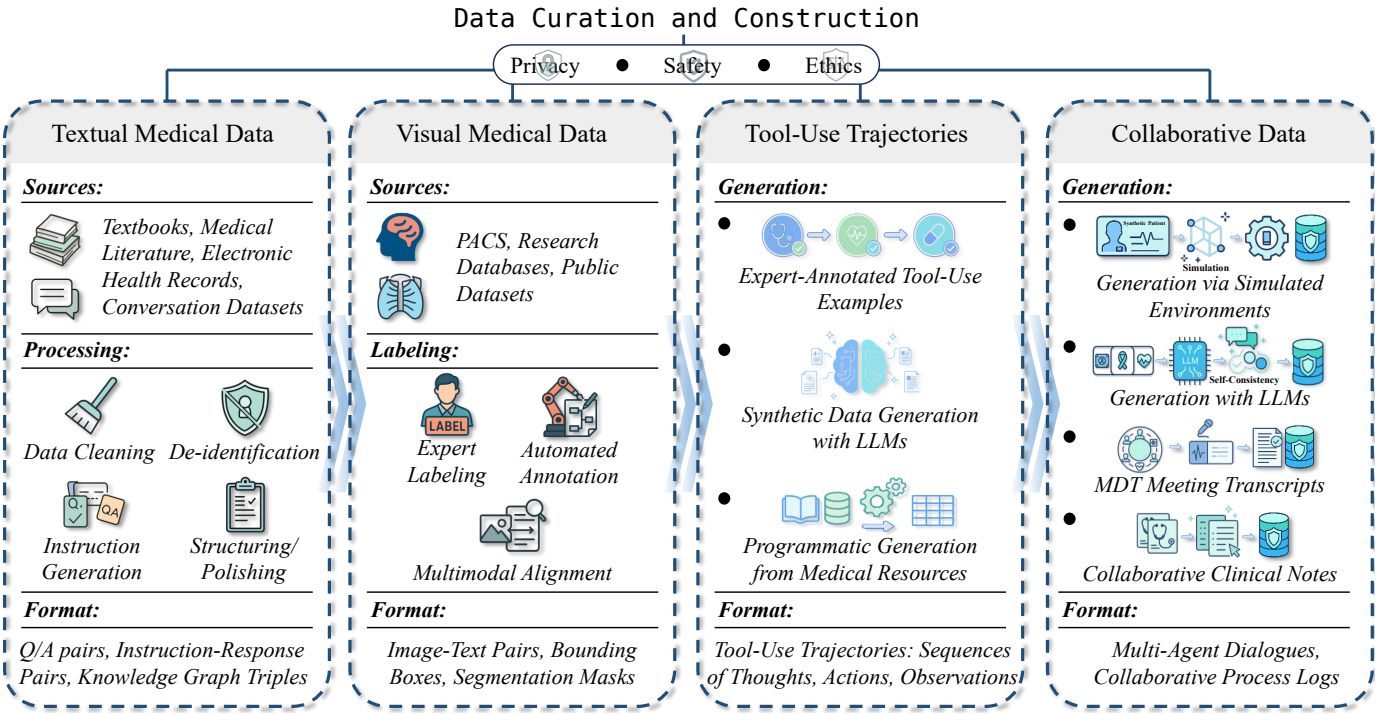


Fig. 8. Data Curation and Construction Pipelines for Medical AI Paradigms. This illustration showcases the distinct data sources, processing/generation methods, and data formats required for four key types of medical AI: (1) Textual LLMs, (2) Visual-Language Models, (3) Tool-Using Agents, and (4) Multi-Agent Systems. The overarching principles of privacy, safety, and ethics govern all data handling.

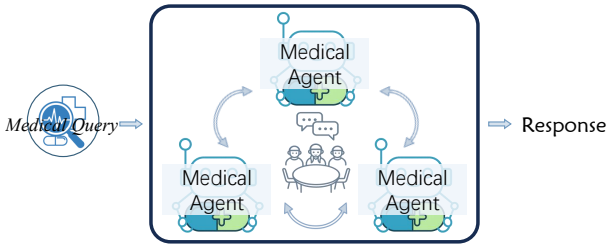


Fig. 9. Pipeline of orchestrating collaborative intelligence in multi-agent medical systems.

expert-demonstrated interactions from a dataset of collaborative dialogues. This is foundational for teaching agents specific roles and communication patterns [175], [145].

A more powerful paradigm is Multi-Agent Reinforcement Learning (MARL), where agents learn optimal behaviors through trial and error in a shared environment [156], [111]. The problem is often modeled as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP), and agents are trained to maximize a shared reward. This approach is exemplified in systems like DoctorAgent-RL [111] for optimizing clinical dialogues, enabling agents to develop more sophisticated and adaptive collaborative strategies.

Another emerging approach is Self-Evolving and Debate-Based Learning. Frameworks like Agent Hospital [10] propose "evolvable medical agents" that adapt through simulated practice and debate against previous versions of themselves. This iterative refinement forces agents to constantly improve their reasoning to outperform their predecessors, enhancing collective reasoning quality without explicit reward modeling.

### 4.3.2 Medical Data Curation for Collaborative Training

The primary bottleneck for training collaborative agents is acquiring high-quality datasets that reflect their interactions. To overcome the scarcity of real-world data, researchers have developed sophisticated simulation platforms. Frameworks like Agent Hospital [10], AI Hospital [110], and AgentClinic [178] create virtual clinical environments to programmatically generate complex cases and record the entire multi-agent interaction trajectory. This provides a scalable method for creating rich, structured training data suitable for both SFT and MARL [179], [180], [181].

Another potent strategy involves using powerful "teacher" LLMs to generate synthetic collaborative dialogues. By assigning different specialist personas to LLM instances, developers can create plausible multi-turn consultations that serve as training examples [175], [11]. These synthetic datasets are often combined with curated real-world data, such as MDT meeting transcripts. The development of comprehensive benchmarks like MedAgentBoard [182], ClinicalLab [183], and 3MD-Bench [184] is crucial for standardizing the evaluation of these complex, collaborative tasks, ensuring that multi-agent systems are robustly validated and aligned with the realities of clinical practice.

TABLE 4
A Taxonomy of Medical Agent Applications by Clinical Domain. This classification maps emerging agentic capabilities to specific healthcare workflows, highlighting the transition from technical proficiency to clinical utility.

| Clinical Domain | Agentic Function | Key Clinical Tasks & Enablers | Representative Work |
|---|---|---|---|
| **Clinical Diagnosis & Consultation** | **Active Diagnostician** | **Interactive Anamnesis:** Simulating dynamic patient-doctor inquiry to refine symptom intake.<br>**Multimodal Synthesis:** Correlating radiological/pathological findings with clinical history.<br>**Reasoning Trajectories:** Utilizing Chain-of-Thought for differential diagnosis. | ChatDoctor[15]<br>MedDxAgent[14]<br>ClinicalAgent[8]<br>LLaVA-Med[4]<br>PathFinder[108] |
| **Therapeutic Planning & Precision Medicine** | **Evidence-Based Planner** | **Safety Verification:** Leveraging external knowledge bases (e.g., DrugBank) to validate interactions.<br>**Genomic Targeting:** Querying live oncology databases to match profiles with therapies.<br>**Protocol Matching:** Automating clinical trial eligibility screening via logic reasoning. | GeneGPT[87]<br>ClinicalAgent[8]<br>Rx Strategist[93]<br>MedAgents[11] |
| **Administrative Workflow & Documentation** | **Intelligent Assistant** | **Automated Documentation:** Synthesizing structured discharge summaries from longitudinal records.<br>**Health Literacy Support:** Translating technical reports into patient-accessible narratives.<br>**Data Interoperability:** Parsing unstructured EHRs into standardized formats. | MeQSum[176]<br>Sudarshan et al.[96]<br>RadGraph[177]<br>EHRAgent[9] |
| **Medical Education & System Simulation** | **Simulation Environment** | **Pedagogical Simulation:** Deploying "Patient Agents" for risk-free diagnostic training.<br>**Operational Optimization:** Modeling hospital workflows (admission-to-discharge) to identify inefficiencies.<br>**In Silico Trials:** Validating clinical pathways within synthetic multi-agent ecosystems. | Agent Hospital[10]<br>AI Hospital[110]<br>AgentClinic[178]<br>MedAgentSim[92] |

TABLE 5
The Evolution of Evaluation Paradigms for Medical AI. This framework delineates the progression from assessing static knowledge competence to validating dynamic clinical performance and safety.

| Evaluation Domain | Assessment Methodology | Key Metrics & Benchmarks |
|---|---|---|
| **Domain Competence** *(Medical Knowledge)* | **Static Question Answering:** Assessing performance on standardized, multiple-choice licensing exams or text generation tasks from fixed datasets. | **Metrics:** Accuracy, F1-Score, ROUGE [185], BERTScore [186].<br>**Benchmarks:** MedQA [2], PubMedQA [187], MMLU (Medical) [188], MedMCQA [189], MMLU-Pro [190]. |
| **Clinical Utility** *(Reasoning & Action)* | **Interactive, Task-Oriented Simulation:** Measuring an agent's ability to navigate dynamic environments, utilize tools, and achieve complex clinical goals. | **Metrics:** Task Success Rate, Diagnostic Efficiency (e.g., turns to diagnosis), Reasoning Validity.<br>**Benchmarks:** CliBench [191], MedAgentBench [192], MedAgentBoard [182], AgentClinic [178], ClinicalBench [193], MedChain [179]. |
| **Safety & Viability** *(Real-World Validation)* | **Human-in-the-Loop (HITL) Validation:** Involving clinical experts to assess the safety, reliability, and ethical alignment of AI outputs and behaviors. | **Methods:** Expert Review (e.g., scoring rubrics), Clinical Turing Tests, LLM-as-a-Judge [194] (as a scalable proxy for preliminary screening). |

## 5 Medical Applications and Evaluation

The evolution from passive Medical Large Language Models to proactive Medical Agents signifies a fundamental paradigm shift in clinical artificial intelligence. While foundational models function primarily as static knowledge repositories, agentic systems are architected to navigate the complexities of real-world medical workflows autonomously. Consequently, to capture the practical implications of this technological leap, we delineate the landscape of applications through the lens of *clinical scenarios* rather than model architectures. By adopting this workflow-centric perspective, we elucidate how specific agentic capabilities, such as active information seeking and tool utilization, address persistent bottlenecks in diagnosis, treatment, and healthcare administration. This section details these clinical domains and subsequently examines the necessary evolution of evaluation paradigms required to validate such autonomous systems.

### 5.1 Clinical Applications of Medical Agents

We structure the application landscape into four cardinal domains: Diagnosis, Therapeutics, Administration, and Education. Table 4 provides a systematic overview of how agentic behaviors are deployed across these sectors.

#### 5.1.1 Clinical Diagnosis and Consultation

The diagnostic process is fundamentally iterative and multimodal, demanding capabilities that extend beyond static question answering. **Interactive Diagnostic Support:** In contrast to the reactive nature of standard LLMs, agents such as MedDxAgent [14] and DoctorAgent-RL [111] are engineered for *active inquiry*. These systems emulate the clinical anamnesis process, dynamically formulating follow-up questions to clarify ambiguity and narrow the differential diagnosis, thereby replicating the hypothesis-verification loop of a human clinician. **Multimodal Clinical Synthesis:** In specialized fields like pathology and radiology, agents serve as visual reasoning partners. Systems including PathFinder [108] and LLaVA-Med [4] demonstrate the capacity to identify regions of interest within imaging data and semantically align these visual features with textual patient history, facilitating a holistic diagnostic assessment.

#### 5.1.2 Therapeutic Planning and Precision Medicine

The formulation of treatment plans requires strict adherence to evidence-based protocols and real-time data verification, a requirement that necessitates **Tool-Augmented Reasoning**. **Evidence-Based Prescribing:** By integrating with external tools, agents can bridge the gap between parametric knowledge and dynamic pharmaceutical data. For instance, Rx Strategist [93] queries live databases (e.g., FDA guidelines) to verify drug-drug interactions and dosage appropriateness, mitigating the risk of hallucinations inherent in static models. **Precision Oncology and Genomics:** In data-intensive domains, agents like GeneGPT [87] utilize API calls to access specialized knowledge bases such as OncoKB. This capability allows for the precise interpretation of genomic variants and the recommendation of targeted therapies, effectively operationalizing vast biomedical databases for individual patient care. **Clinical Trial Optimization:** Furthermore, agents like ClinicalAgent [8] apply logical planning modules to screen patient records against complex trial eligibility criteria, streamlining the recruitment process for investigational treatments.

#### 5.1.3 Administrative Workflow and Documentation

The administrative burden on clinicians is a primary driver of professional burnout. Medical agents offer a solution by functioning as intelligent, context-aware assistants. **Automated Clinical Documentation:** Beyond generic summarization, agents are capable of generating domain-specific documentation, such as hospital discharge summaries and insurance pre-authorization forms, by synthesizing dispersed information from the Electronic Health Record (EHR) [176], [9]. **Enhancing Patient Communication:** Medical Agents are also deployed to bridge the communication gap between clinicians and patients. By translating complex radiological or pathological reports into layperson-accessible language, systems discussed by Sudarshan et al. [96] aim to improve health literacy and patient engagement.

#### 5.1.4 Medical Education and System Simulation

The advent of Multi-Agent Systems (MAS) enables the construction of sophisticated simulation environments for training and operations research. **Immersive Pedagogical Platforms:** Frameworks such as AgentClinic [178] and Agent Hospital [10] function as high-fidelity "sandboxes." These systems deploy diverse "Patient Agents" capable of simulating specific pathologies and personality traits, allowing medical trainees to practice diagnostic interviewing and decision-making in a risk-free environment. **Healthcare Systems Modeling:** At the macro level, MAS facilitates the digital twinning of healthcare facilities. By simulating the interactions between doctor, nurse, and patient agents, researchers can model patient flow and resource allocation [110], [195], providing an empirical basis for optimizing hospital operations before real-world implementation.

### 5.2 Evolution of Medical Evaluation Paradigms

As the scope of medical AI expands from passive information retrieval to autonomous decision-making, evaluation methodologies must undergo a parallel transformation. The traditional reliance on static accuracy metrics is increasingly insufficient for assessing medical agents designed to navigate dynamic, multi-step clinical workflows. Consequently, we propose a hierarchical evaluation framework that moves beyond assessing what the model *knows* to validating what the agent can *do* and ensuring it operates *safely*. This progression mirrors the medical training continuum: from standardized knowledge exams to simulated clinical practice, and ultimately, supervised real-world validation.

#### 5.2.1 Evaluating Foundational Medical Knowledge

The first layer of assessment focuses on the model's internal knowledge base, analogous to a medical licensing examination. Evaluation at this stage primarily relies on static benchmarks comprising multiple-choice questions sourced from professional boards. Datasets such as MedQA [2], MedMCQA [189], and the medical subdomains of MMLU [188] serve as the standard proxies for this capability. As illustrated in Table 6, leading models have recently achieved expert-level accuracy on these tests. However, while these metrics are necessary to verify domain competence, they are insufficient predictors of clinical utility. High performance on static QA demonstrates rote memorization and pattern matching but does not guarantee the ability to synthesize information or execute complex reasoning in a dynamic clinical environment, highlighting a critical gap between *competence* (knowing) and *performance* (doing).

#### 5.2.2 Evaluating Clinical Utility and Agentic Reasoning

To bridge the gap between static knowledge and practice, evaluation must shift towards interactive, task-oriented benchmarks that simulate real-world clinical workflows. This level of assessment focuses on *clinical utility*, the capacity of an agent to formulate plausible, actionable plans that positively influence patient outcomes. Benchmarks such as "CliBench" [191] represent this shift, moving beyond abstract questions to case-based decision-making. As shown in Table 7, the performance drop observed when models transition from

TABLE 6
The results of LLM/Agents in static QA. $*$ denotes the results come from [196], † denotes the results come from [197], ‡ denotes the results come from [2].

| Model | MedQA | PubMedQA | MedMCQA | MMLU-An | MMLU-CK | MMLU-CB | MMLU-CM | MMLU-MG | MMLU-PM |
|---|---|---|---|---|---|---|---|---|---|
| ***Close Source*** | | | | | | | | | |
| GPT-4* [198] | 78.9 | 75.2 | 69.5 | 80.0 | 86.0 | 95.1 | 76.9 | 91.0 | 93.0 |
| GPT-3.5* [132] | 50.8 | 71.6 | 50.1 | 56.3 | 69.8 | 72.2 | 61.3 | 70.0 | 70.2 |
| Gemini-Pro† [199] | 67.0 | 70.7 | 62.2 | 76.9 | 78.6 | 89.5 | 79.3 | 81.8 | 83.8 |
| Gemini-2.5-Pro [196] | 92.6 | 75.8 | 81.1 | 91.1 | 91.7 | 98.6 | 89.0 | 96.0 | 96.3 |
| Flan-PaLM [200] | 67.6 | 79.0 | 57.6 | 63.7 | 80.4 | 88.9 | 76.3 | 75.0 | 83.8 |
| Med-PaLM 2‡ [2] | 86.5 | 81.8 | 72.3 | 84.4 | 88.7 | 95.8 | 83.2 | 92.0 | 95.2 |
| ***Open Source*** | | | | | | | | | |
| Deepseek-R1§ [201] | 90.1 | 77.2 | 78.8 | 91.1 | 91.7 | 98.6 | 90.8 | 99.0 | 95.6 |
| Llama3-8b [202] | 59.7 | 74.8 | 57.5 | 68.9 | 74.7 | 78.5 | 61.9 | 83.0 | 70.2 |
| Llama3-Instruct-8b [202] | 60.7 | 74.6 | 56.9 | 62.2 | 70.9 | 73.6 | 65.3 | 82.0 | 74.6 |
| Gemma-7b [203] | 48.7 | 75.6 | 49.3 | 58.5 | 69.4 | 77.1 | 60.7 | 70.0 | 63.2 |
| Mistral-7b-v0.1 [204] | 50.8 | 75.4 | 48.2 | 55.6 | 68.7 | 68.1 | 59.5 | 71.0 | 68.4 |
| BioMistral-7b [31] | 46.1 | 71.0 | 41.5 | 51.1 | 63.8 | 61.1 | 53.8 | 66.0 | 52.9 |
| OpenBioLLM-8b [205] | 58.9 | 74.1 | 56.9 | 69.8 | 76.1 | 84.2 | 68.0 | 86.1 | 78.2 |
| OpenBioLLM-70b [205] | 78.2 | 78.9 | 74.0 | 83.9 | 92.9 | 93.8 | 85.8 | 93.3 | 93.8 |
| Apollo-7B [34] | 55.2 | 39.8 | 53.8 | 61.5 | 62.3 | 70.8 | 55.5 | 72.0 | 69.1 |
| Meditron-70b [28] | 57.1 | 76.6 | 46.9 | 53.3 | 66.8 | 76.4 | 63.0 | 69.0 | 71.7 |
| MedAlpaca-7b [17] | 41.7 | 72.8 | 37.5 | 57.0 | 57.4 | 65.3 | 54.3 | 69.0 | 67.3 |
| ClinicalGPT [21] | 26.1 | 63.8 | 28.2 | 30.4 | 30.6 | 25.0 | 24.3 | 27.0 | 19.5 |
| MedGemma-4b [78] | 64.4 | 73.4 | 55.7 | 59.3 | 71.3 | 70.8 | 65.3 | 83.0 | 76.8 |
| MedGemma-27b [78] | 87.7 | 76.8 | 74.2 | 83.7 | 86.0 | 96.5 | 86.1 | 97.0 | 93.4 |
| An = Anatomy, CK = Clinical Knowledge, CB = College Biology, CM = College Medicine, MG = Medical Genetics, PM = Professional Medicine | | | | | | | | | |

TABLE 7
Performance evaluation on CliBench (Values represent the **F1-Score** (%)). Results are sourced from [191].

| Model | Diagnosis | Procedures | Prescriptions |
|---|---|---|---|
| GPT-4o [206] | 46.02 | 13.15 | 63.55 |
| GPT-3.5 turbo [132] | 39.03 | 13.28 | 52.57 |
| Llama3-8b [202] | 15.62 | 12.71 | 46.69 |
| Llama3-Instruct-70b [202] | 41.03 | 12.48 | 63.08 |
| OpenBioLLM-8b [205] | 17.78 | 8.76 | 43.72 |
| Meditron-7b [28] | 9.17 | 5.84 | 23.68 |
| Asclepius-7b [207] | 7.38 | 9.30 | 18.49 |
| BioMistral-7b [31] | 14.41 | 9.68 | 36.76 |

QA to these tasks underscores the complexity of applying knowledge in context.

Furthermore, advanced evaluation platforms like "MedAgentBench" [192] and "MedAgentBoard" [182] introduce multi-turn scenarios that rigorously test specific agentic capabilities, including information gathering, tool utilization, and strategic planning. In these environments, simple accuracy metrics are replaced by process-oriented metrics such as "Task Success Rate", "Diagnostic Efficiency" (e.g., number of turns to correct diagnosis), and the logical coherence of the reasoning trajectory. This shift is essential for validating an agent's ability to function as an autonomous clinical partner rather than a mere search engine.

#### *5.2.3 The Gold Standard: Human-in-the-Loop Clinical Validation*

Ultimately, in the high-stakes domain of healthcare, algorithmic metrics cannot fully capture the nuances of clinical viability. "Human-in-the-Loop" (HITL) validation remains the indispensable gold standard for ensuring patient safety and system trustworthiness [208]. This process involves qualified medical professionals rigorously reviewing agent outputs not only for factual correctness but also for safety, ethical alignment, and empathetic communication, qualities that automated metrics struggle to quantify.

HITL methodologies range from structured reviews using standardized scoring rubrics to holistic assessments such as Clinical Turing Tests, where experts evaluate the indistinguishability of AI-generated plans from human-generated ones. While expert review is resource-intensive and challenging to scale, it is the only reliable mechanism to detect subtle errors that could lead to adverse events. Emerging research is exploring the use of "LLM-as-a-Judge" frameworks [194] as scalable surrogates for preliminary evaluation; however, these methods require careful calibration against human baselines and cannot yet replace the critical judgment of expert clinicians in the final validation phase.

## 6 Future Directions and Challenges

The evolution from Medical LLMs to Agentic Systems represents a paradigm shift from *passive knowledge retrieval* to

*active clinical management*. While current prototypes demonstrate technical feasibility, bridging the gap between *in silico* pilots and *bedside* utility requires addressing fundamental clinical needs. To transform these agents into reliable partners for precision medicine, the field must prioritize evidence-based validation, longitudinal continuity, and deep pathophysiological reasoning. We delineate three pivotal frontiers essential for the clinical translation and maturation of Medical Agentic Systems.

### 6.1 The Reliability Frontier: Evidence Grounding and Process Transparency

In clinical practice, the black-box nature of neural networks conflicts with the imperative for *explainability* and *accountability*. Future agents must transition from probabilistic generation to verifiable, evidence-based reasoning to mitigate iatrogenic risks.

#### *6.1.1 Evidence-Based Tool Augmentation*

Relying solely on internal parametric knowledge poses a significant risk of hallucination. To ensure safety, agents must adopt an Evidence-Based Tool Augmentation architecture [86], [87]. Instead of generating medical facts directly, agents should function as orchestrators that leverage verified external tools, such as querying live pharmacological databases (e.g., DrugBank) for contraindications or accessing up-to-date Clinical Practice Guidelines (CPGs). This separation of *reasoning* (the agent) from *knowledge* (the tool) ensures that every decision is grounded in retrievable, standard-of-care data sources, establishing a robust "safety guardrail" for clinical logic.

#### *6.1.2 Diagnostic Auditability and Traceability*

For an AI system to be integrated into the clinical workflow, its decision-making process must be fully auditable [8]. Future research should focus on agents that produce structured *reasoning trajectories* rather than simple outputs. This involves making the diagnostic logic transparent: explicitly citing the evidence for ruling in/out a differential diagnosis and mapping symptoms to specific diagnostic criteria. Such "Chain-of-Diagnosis" transparency enables clinicians to review, verify, and trust the agent's logic, satisfying the rigorous requirements for medical liability and ethical oversight.

### 6.2 The Ecological Validity Frontier: From Simulation to Bedside

A significant bottleneck in deploying medical agents is the "Sim-to-Real" gap. Current agents are often evaluated on sanitized, "textbook" cases, which fail to reflect the stochasticity and complexity of real-world hospital environments [10], [178].

#### *6.2.1 Handling Atypical and Comorbid Presentations*

Real-world patients rarely present with the clean, singular symptom profiles found in standardized exams. Future agents must be stress-tested on atypical presentations and multimorbid scenarios where symptoms of one condition mask another. Developing High-Fidelity Medical World Models that simulate noise, missing data, and conflicting lab results is essential. Agents must demonstrate the capability to navigate uncertainty and refine hypotheses dynamically, rather than forcing a fit to a standard pattern [94].

#### *6.2.2 Longitudinal Patient Trajectory Management*

Chronic disease management requires monitoring a patient's state over months or years, not just a single interaction. We must advance from episodic Q&A to Longitudinal Patient Trajectory Management [9], [209]. This requires agents to maintain a persistent memory of the patient's history, distinguishing between acute fluctuations and long-term deterioration. The challenge lies in identifying clinically significant trends within vast temporal EHR data, enabling proactive interventions for conditions like heart failure or diabetes progression.

### 6.3 The Cognitive Frontier: Precision Phenotyping and Etiological Depth

The transformative potential of Medical Agents lies in augmenting the depth of clinical inquiry. Future systems must transcend simple symptom matching to achieve Deep Pathophysiological Reasoning, moving from descriptive diagnosis to mechanistic understanding.

#### *6.3.1 Deep Pathophysiological Stratification*

Standard diagnosis often stops at a broad label (e.g., "Sepsis"). Future agents, particularly Multi-Agent Systems, should aim for Deep Stratification, dissecting diseases into specific phenotypes, clinical stages, and risk layers [11]. By integrating multi-omics data and clinical history, agents can facilitate *Precision Phenotyping*, guiding therapy based on specific biological subtypes rather than generic protocols.

#### *6.3.2 Multi-Scale Mechanistic Explanation*

True clinical understanding requires connecting symptoms to underlying biology. Future architectures must strive for Multi-Scale Reasoning, capable of explaining a diagnosis vertically from the *molecular level* (genetic variants) to the *organ level* (pathology) and finally to the *systemic level* (clinical presentation) [87], [88]. This integration allows for robust decision-making in complex cases, such as oncology or rare genetic disorders, where understanding the etiology, the pathway from gene to phenotype, is critical for selecting targeted therapies.

Taken together, the future of Medical Agentic Systems lies in their ability to reconcile the "messiness" of real-world biology with the rigor of evidence-based medicine. Their success will depend on becoming transparent, safe, and deeply knowledgeable partners in the preservation of human health.

## 7 Conclusion

In this survey, we have systematically charted the evolution from static Medical Large Language Models to dynamic Medical Agentic Systems. This transition represents a fundamental shift from the passive retrieval of biomedical knowledge to the active orchestration of clinical reasoning, multimodal

perception, and tool-mediated action. We formalized the architectural blueprints underpinning this evolution, ranging from the cognitive mechanisms of single-agent systems to the emergent collective intelligence of multi-agent frameworks. While foundation models have achieved expert-level proficiency in question answering, navigating the stochastic and high-stakes environment of real-world healthcare requires distinct agentic competencies. Essential capabilities identified include robust state tracking, symbolic guardrails for safety verification, and the synergy of heterogeneous inputs from text, imaging, and electronic health records. Looking forward, the realization of autonomous clinical partners hinges on bridging the gap between probabilistic generation and deterministic execution. Future research must rigorously address challenges in formal verification, simulation-to-real transfer, and cognitive alignment. Ultimately, the systems surveyed herein represent the precursors to a new era of machine intelligence, one that does not merely process medical data but actively collaborates in the preservation of human health.

## REFERENCES


[1] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[2] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis *et al.*, "Toward expert-level medical question answering with large language models," *Nature Medicine*, pp. 1–8, 2025.

[3] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, "Pmc-llama: toward building open-source language models for medicine," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1833–1843, 2024.

[4] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *Advances in Neural Information Processing Systems*, vol. 36, pp. 28 541–28 564, 2023.

[5] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar, "Med-flamingo: a multimodal medical few-shot learner," in *Machine Learning for Health (ML4H)*. PMLR, 2023, pp. 353–367.

[6] Y. Zhou, J. Shen, and Y. Cheng, "Weak to strong generalization for large language models with multi-capabilities," in *The Thirteenth International Conference on Learning Representations*, 2025.

[7] Y. Zhou, X. Li, Q. Wang, and J. Shen, "Visual in-context learning for large vision-language models," *arXiv preprint arXiv:2402.11574*, 2024.

[8] L. Yue, S. Xing, J. Chen, and T. Fu, "Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning," in *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2024, pp. 1–10.

[9] W. Shi, R. Xu, Y. Zhuang, Y. Yu, J. Zhang, H. Wu, Y. Zhu, J. Ho, C. Yang, and M. D. Wang, "Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records," *arXiv preprint arXiv:2401.07128*, 2024.

[10] J. Li, Y. Lai, W. Li, J. Ren, M. Zhang, X. Kang, S. Wang, P. Li, Y.-Q. Zhang, W. Ma *et al.*, "Agent hospital: A simulacrum of hospital with evolvable medical agents," *arXiv preprint arXiv:2405.02957*, 2024.

[11] X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan, and M. Gerstein, "Medagents: Large language models as collaborators for zero-shot medical reasoning," *arXiv preprint arXiv:2311.10537*, 2023.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[13] C. Zakka, R. Shad, A. Chaurasia, A. R. Dalal, J. L. Kim, M. Moor, R. Fong, C. Phillips, K. Alexander, E. Ashley *et al.*, "Almanac—retrieval-augmented language models for clinical medicine," *Nejm ai*, vol. 1, no. 2, p. AIoa2300068, 2024.

[14] D. Rose, C.-C. Hung, M. Lepri, I. Alqassem, K. Gashteovski, and C. Lawrence, "Meddxagent: A unified modular agent framework for explainable automatic differential diagnosis," *arXiv preprint arXiv:2502.19175*, 2025.

[15] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge," *Cureus*, vol. 15, no. 6, 2023.

[16] C. Xu, D. Guo, N. Duan, and J. McAuley, "Baize: An open-source chat model with parameter-efficient tuning on self-chat data," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6268–6278. [Online]. Available: https://aclanthology.org/2023.emnlp-main.385/

[17] T. Han, L. C. Adams, J.-M. Papaioannou, P. Grundmann, T. Oberhauser, A. Löser, D. Truhn, and K. K. Bressem, "Medalpaca–an open-source collection of medical conversational ai models and training data," *arXiv preprint arXiv:2304.08247*, 2023.

[18] H. Xiong, S. Wang, Y. Zhu, Z. Zhao, Y. Liu, L. Huang, Q. Wang, and D. Shen, "Doctorglm: Fine-tuning your chinese doctor is not a herculean task," *arXiv preprint arXiv:2304.01097*, 2023.

[19] H. Zhang, J. Chen, F. Jiang, F. Yu, Z. Chen, J. Li, G. Chen, X. Wu, Z. Zhang, Q. Xiao, X. Wan, B. Wang, and H. Li, "Huatuogpt, towards taming language models to be a doctor," *arXiv preprint arXiv:2305.15075*, 2023.

[20] C. Peng, X. Yang, A. Chen, K. E. Smith, N. PourNejatian, A. B. Costa, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, G. Lipori, D. A. Mitchell, N. S. Ospina, M. M. Ahmed, W. R. Hogan, E. A. Shenkman, Y. Guo, J. Bian, and Y. Wu, "A study of generative large language model for medical research and healthcare," *npj Digital Medicine*, vol. 6, no. 1, Nov. 2023. [Online]. Available: http://dx.doi.org/10.1038/s41746-023-00958-w

[21] G. Wang, G. Yang, Z. Du, L. Fan, and X. Li, "Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation," 2023. [Online]. Available: https://arxiv.org/abs/2306.09968

[22] S. Yang, H. Zhao, S. Zhu, G. Zhou, H. Xu, Y. Jia, and H. Zan, "Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 17, 2024, pp. 19 368–19 376.

[23] O. B. Shoham and N. Rappoport, "Cpllm: Clinical prediction with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2309.11295

[24] Y. Chen, Z. Wang, X. Xing, Z. Xu, K. Fang, J. Wang, S. Li, J. Wu, Q. Liu, X. Xu *et al.*, "Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt," *arXiv preprint arXiv:2310.15896*, 2023.

[25] Q. Ye, J. Liu, D. Chong, P. Zhou, Y. Hua, F. Liu, M. Cao, Z. Wang, X. Cheng, Z. Lei *et al.*, "Qilin-med: Multi-stage knowledge injection advanced medical large language model," *arXiv preprint arXiv:2310.09089*, 2023.

[26] X. Zhang, C. Tian, X. Yang, L. Chen, Z. Li, and L. R. Petzold, "Alpacare:instruction-tuned large language models for medical application," 2025. [Online]. Available: https://arxiv.org/abs/2310.14558

[27] J. Chen, X. Wang, K. Ji, A. Gao, F. Jiang, S. Chen, H. Zhang, D. Song, W. Xie, C. Kong *et al.*, "Huatuogpt-ii, one-stage training for medical adaption of llms," *arXiv preprint arXiv:2311.09774*, 2023.

[28] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami *et al.*, "Meditron-70b: Scaling medical pretraining for large language models," *arXiv preprint arXiv:2311.16079*, 2023.

[29] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena *et al.*, "Towards generalist biomedical ai," *Nejm Ai*, vol. 1, no. 3, p. AIoa2300138, 2024.

[30] F. Jia, X. Liu, L. Deng, J. Gu, C. Pu, T. Bai, M. Huang, Y. Lu, and K. Liu, "Oncogpt: A medical conversational model tailored with oncology domain expertise on a large language model meta-ai (llama)," *arXiv preprint arXiv:2402.16810*, 2024.

[31] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, "Biomistral: A collection of open-source pre-trained large language models for medical domains," *arXiv preprint arXiv:2402.10373*, 2024.

[32] S. Pieri, S. S. Mullappilly, F. S. Khan, R. M. Anwer, S. Khan, T. Baldwin, and H. Cholakkal, "Bimedix: Bilingual medical mixture of experts llm," *arXiv preprint arXiv:2402.13253*, 2024.

[33] Q. Xie, Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, X. Peng, J. Huang, J. Zhang, V. Keloth *et al.*, "Medical foundation large language models for comprehensive text analysis and beyond," *npj Digital Medicine*, vol. 8, no. 1, p. 141, 2025.

[34] X. Wang, N. Chen, J. Chen, Y. Wang, G. Zhen, C. Zhang, X. Wu, Y. Hu, A. Gao, X. Wan *et al.*, "Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people," *arXiv preprint arXiv:2403.03640*, 2024.

[35] E. Bolton, A. Venigalla, M. Yasunaga, D. Hall, B. Xiong, T. Lee, R. Daneshjou, J. Frankle, P. Liang, M. Carbin *et al.*, "Biomedlm: A 2.7 b parameter language model trained on biomedical text," *arXiv preprint arXiv:2403.18421*, 2024.

[36] A. K. Gururajan, E. Lopez-Cuena, J. Bayarri-Planas, A. Tormos, D. Hinjos, P. Bernabeu-Perez, A. Arias-Duart, P. A. Martin-Torres, L. Urcelay-Ganzabal, M. Gonzalez-Mallo *et al.*, "Aloe: A family of fine-tuned open healthcare llms," *arXiv preprint arXiv:2405.01886*, 2024.

[37] H. Zhang and B. An, "Medgo: A chinese medical large language model," *arXiv preprint arXiv:2410.20428*, 2024.

[38] H. Yu, T. Cheng, Y. Cheng, and R. Feng, "Finemedlm-o1: Enhancing the medical reasoning ability of llm from supervised fine-tuning to test-time training," *arXiv preprint arXiv:2501.09213*, 2025.

[39] S. Jiang, Y. Liao, Z. Chen, Y. Zhang, Y. Wang, and Y. Wang, "Meds3: Towards medical small language models with self-evolved slow thinking," *arXiv preprint arXiv:2501.12051*, 2025.

[40] Y. Liu, S. Luo, Z. Zhong, T. Wu, J. Zhang, P. Ou, Y. Liang, L. Liu, and H. Pan, "Hengqin-ra-v1: advanced large language model for diagnosis and treatment of rheumatoid arthritis with dataset based traditional chinese medicine," *arXiv preprint arXiv:2501.02471*, 2025.

[41] Z. Wang, J. Jiang, Y. Zhan, B. Zhou, Y. Li, C. Zhang, B. Yu, L. Ding, H. Jin, J. Peng *et al.*, "Hypnos: A domain-specific large language model for anesthesiology," *Neurocomputing*, vol. 624, p. 129389, 2025.

[42] G. Wang, M. Gao, S. Yang, Y. Zhang, L. He, L. Huang, H. Xiao, Y. Zhang, W. Li, L. Chen *et al.*, "Citrus: Leveraging expert cognitive pathways in a medical language model for advanced medical decision support," *arXiv preprint arXiv:2502.18274*, 2025.

[43] A. Guluzade, N. Heiba, Z. Boukhers, F. Hamiti, J. H. Polash, Y. Mohamad, and C. A. Velasco, "Elmtex: Fine-tuning large language models for structured clinical information extraction. a case study on clinical reports," *arXiv preprint arXiv:2502.05638*, 2025.

[44] B. Wang, H. Zhao, H. Zhou, L. Song, M. Xu, W. Cheng, X. Zeng, Y. Zhang, Y. Huo, Z. Wang *et al.*, "Baichuan-m1: Pushing the medical capability of large language models," *arXiv preprint arXiv:2502.12671*, 2025.

[45] X. Wang, Z. Sun, P. Wang, and B. Wei, "Medicalglm: A pediatric medical question answering model with a quality evaluation mechanism," *Journal of Biomedical Informatics*, vol. 165, p. 104793, 2025.

[46] X. Zhang, Y. Wang, Z. Feng, R. Chen, Z. Zhou, Y. Zhang, H. Xu, J. Wu, and Z. Liu, "Med-u1: Incentivizing unified medical reasoning in llms via large-scale reinforcement learning," *arXiv preprint arXiv:2506.12307*, 2025.

[47] Q. Zheng, S. Abdullah, S. Rawal, C. Zakka, S. Ostmeier, M. Purk, E. Reis, E. J. Topol, J. Leskovec, and M. Moor, "Miriad: Augmenting llms with millions of medical query-response pairs," *arXiv preprint arXiv:2506.06091*, 2025.

[48] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[49] Q. Chen and Y. Hong, "Medblip: Bootstrapping language-image pre-training from 3d medical images and texts," in *Proceedings of the Asian conference on computer vision*, 2024, pp. 2404–2420.

[50] R. Wang, Y. Duan, J. Li, P. Pang, and T. Tan, "Xrayglm: The first chinese medical multimodal model that chest radiographs summarization," *arXiv preprint arXiv: 2408.12345*, 2023.

[51] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie, "Pmc-vqa: Visual instruction tuning for medical visual question answering," *arXiv preprint arXiv:2305.10415*, 2023.

[52] Y. Sun, C. Zhu, S. Zheng, K. Zhang, L. Sun, Z. Shui, Y. Zhang, H. Li, and L. Yang, "Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 5034–5042.

[53] B. Yang, A. Raza, Y. Zou, and T. Zhang, "Customizing general-purpose foundation models for medical report generation," *arXiv preprint arXiv:2306.05642*, 2023.

[54] O. C. Thawakar, A. M. Shaker, S. S. Mullappilly, H. Cholakkal, R. M. Anwer, S. Khan, J. Laaksonen, and F. Khan, "Xraygpt: Chest radiographs summarization using large medical vision-language models," in *Proceedings of the 23rd workshop on biomedical natural language processing*, 2024, pp. 440–448.

[55] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data," *arXiv preprint arXiv:2308.02463*, 2023.

[56] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2gengpt: Radiology report generation with frozen llms," *Meta-Radiology*, vol. 1, no. 3, p. 100033, 2023.

[57] J. Liu, Z. Wang, Q. Ye, D. Chong, P. Zhou, and Y. Hua, "Qilin-med-vl: Towards chinese large vision-language model for general healthcare," *arXiv preprint arXiv:2310.17956*, 2023.

[58] S. Srivastav, M. Ranjit, F. Pérez-García, K. Bouzid, S. Bannur, D. C. Castro, A. Schwaighofer, H. Sharma, M. Ilse, V. Salvatelli *et al.*, "Maira at rrg24: A specialised large multimodal model for radiology report generation," in *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 2024, pp. 597–602.

[59] G. Liu, J. He, P. Li, G. He, Z. Chen, and S. Zhong, "Pefomed: Parameter efficient fine-tuning of multimodal large language models for medical imaging," *arXiv preprint arXiv:2401.02797*, 2024.

[60] Z. Chen, M. Varma, J.-B. Delbrouck, M. Paschali, L. Blankemeier, D. Van Veen, J. M. J. Valanarasu, A. Youssef, J. P. Cohen, E. P. Reis *et al.*, "Chexagent: Towards a foundation model for chest x-ray interpretation," *arXiv preprint arXiv:2401.12208*, 2024.

[61] F. Bai, Y. Du, T. Huang, M. Q.-H. Meng, and B. Zhao, "M3d: Advancing 3d medical image analysis with multi-modal large language models," *arXiv preprint arXiv:2404.00578*, 2024.

[62] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, M. Zhao, A. K. Chow, K. Ikemura, A. Kim, D. Pouli, A. Patel *et al.*, "A multimodal generative ai copilot for human pathology," *Nature*, vol. 634, no. 8033, pp. 466–473, 2024.

[63] J. Chen, C. Gui, R. Ouyang, A. Gao, S. Chen, G. Chen, X. Wang, Z. Cai, K. Ji, X. Wan *et al.*, "Towards injecting medical visual knowledge into multimodal llms at scale," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 7346–7370.

[64] A. Alkhaldi, R. Alnajim, L. Alabdullatef, R. Alyahya, J. Chen, D. Zhu, A. Alsinan, and M. Elhoseiny, "Minigpt-med: Large language model as a general interface for radiology diagnosis," *arXiv preprint arXiv:2407.04106*, 2024.

[65] J. Zhou, X. He, L. Sun, J. Xu, X. Chen, Y. Chu, L. Zhou, X. Liao, B. Zhang, S. Afvari *et al.*, "Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4," *Nature Communications*, vol. 15, no. 1, p. 5649, 2024.

[66] Y. Xie, C. Zhou, L. Gao, J. Wu, X. Li, H.-Y. Zhou, S. Liu, L. Xing, J. Zou, C. Xie, and Y. Zhou, "Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=IwgmgidYPS

[67] X. Cao, W. Ye, K. Moise, and M. Coffee, "Mpoxvlm: A vision-language model for diagnosing skin lesions from mpox virus infection," *arXiv preprint arXiv:2411.10888*, 2024.

[68] Y. Zhou, L. Song, and J. Shen, "Improving medical large vision-language models with abnormal-aware feedback," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 12 994–13 011. [Online]. Available: https://aclanthology.org/2025.acl-long.636/

[69] J. Pan, C. Liu, J. Wu, F. Liu, J. Zhu, H. B. Li, C. Chen, C. Ouyang, and D. Rueckert, "Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning," *arXiv preprint arXiv:2502.19634*, 2025.

[70] T. Lin, W. Zhang, S. LI, Y. Yuan, B. Yu, H. Li, W. He, H. Jiang, M. Li, S. Tang *et al.*, "Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation," in *Forty-second International Conference on Machine Learning*, 2025.

[71] W. Zhu, X. Li, X. Chen, P. Qiu, V. K. Vasa, X. Dong, Y. Chen, N. Lepore, O. Dumitrascu, Y. Su *et al.*, "Retinalgpt: A retinal clinical preference conversational assistant powered by large vision-language models," *arXiv preprint arXiv:2503.03987*, 2025.

[72] Y. Lai, J. Zhong, M. Li, S. Zhao, and X. Yang, "Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models," *arXiv preprint arXiv:2503.13939*, 2025.

[73] J. Wu, H. Yang, X. Zeng, G. He, Z. Chen, Z. Li, X. Zhang, Y. Ma, R. Fang, and Y. Liu, "Pathvlm-r1: A reinforcement learning-driven reasoning model for pathology visual-language tasks," *arXiv preprint arXiv:2504.09258*, 2025.

[74] W. Dai, P. Chen, C. Ekbote, and P. P. Liang, "Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training," *arXiv preprint arXiv:2506.00711*, 2025.

[75] Z. Chen, Y. Bie, H. Jin, and H. Chen, "Large language model with region-guided referring and grounding for ct report generation," *IEEE Transactions on Medical Imaging*, 2025.

[76] S. S. Khan, P. Przulj, A. Ashraf, and A. Abedi, "Chestgpt: Integrating large language models and vision transformers for disease detection and localization in chest x-rays," *arXiv preprint arXiv:2507.03739*, 2025.

[77] Q. Xing, Z. Song, Y. Zhang, N. Feng, J. Yu, and W. Yang, "Mca-rg: Enhancing llms with medical concept alignment for radiology report generation," *arXiv preprint arXiv:2507.06992*, 2025.

[78] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau *et al.*, "Medgemma technical report," *arXiv preprint arXiv:2507.05201*, 2025.

[79] Y. Zhou, Z. Rao, J. Wan, and J. Shen, "Rethinking visual dependency in long-context reasoning for large vision-language models," *arXiv preprint arXiv:2410.19732*, 2024.

[80] Y. Zhou, J. Zhang, G. Chen, J. Shen, and Y. Cheng, "Less is more: Vision representation compression for efficient video generation with large language models," 2024.

[81] W. Sun, J. Hu, Y. Zhou, J. Du, D. Lan, K. Wang, T. Zhu, X. Qu, Y. Zhang, X. Mo *et al.*, "Speed always wins: A survey on efficient architectures for large language models," *arXiv preprint arXiv:2508.09834*, 2025.

[82] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature machine intelligence*, vol. 5, no. 3, pp. 220–235, 2023.

[83] W. Wang, Z. Ma, Z. Wang, C. Wu, W. Chen, X. Li, and Y. Yuan, "A survey of llm-based agents in medicine: How far are we from baymax?" *arXiv preprint*, vol. arXiv:2502.11211, 2025. [Online]. Available: https://arxiv.org/abs/2502.11211

[84] S. Xu, Y. Zhou, Z. Liu, Z. Wu, T. Zhong, H. Zhao, Y. Li, H. Jiang, Y. Pan, J. Chen *et al.*, "Towards next-generation medical agent: How o1 is reshaping decision-making in medical scenarios," *arXiv preprint arXiv:2411.14461*, 2024.

[85] J. Feng, Q. Zheng, C. Wu, Z. Zhao, Y. Zhang, Y. Wang, and W. Xie, "Mˆ 3builder: A multi-agent system for automated machine learning in medical imaging," *arXiv preprint arXiv:2502.20301*, 2025.

[86] B. Yang, S. Jiang, L. Xu, K. Liu, H. Li, G. Xing, H. Chen, X. Jiang, and Z. Yan, "Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, pp. 1–29, 2024.

[87] Q. Jin, Y. Yang, Q. Chen, and Z. Lu, "Genegpt: Augmenting large language models with domain tools for improved access to biomedical information," *Bioinformatics*, vol. 40, no. 2, p. btae075, 2024.

[88] B. Li, T. Yan, Y. Pan, J. Luo, R. Ji, J. Ding, Z. Xu, S. Liu, H. Dong, Z. Lin *et al.*, "Mmedagent: Learning to use medical tools with multimodal agent," *arXiv preprint arXiv:2407.02483*, 2024.

[89] A. J. Goodell, S. N. Chu, D. Rouholiman, and L. F. Chu, "Augmentation of chatgpt with clinician-informed tools improves performance on medical calculation tasks," *medRxiv*, pp. 2023–12, 2023.

[90] X. Chen, Y. Jin, X. Mao, L. Wang, S. Zhang, and T. Chen, "Rareagents: Advancing rare disease care through llm-empowered multi-disciplinary team," *arXiv preprint arXiv:2412.12475*, 2024.

[91] Y. H. Ke, R. Yang, S. A. Lie, T. X. Y. Lim, H. R. Abdullah, D. S. W. Ting, and N. Liu, "Enhancing diagnostic accuracy through multi-agent conversations: using large language models to mitigate cognitive bias," *arXiv preprint arXiv:2401.14589*, 2024.

[92] M. Almansoori, K. Kumar, and H. Cholakkal, "Self-evolving multi-agent simulations for realistic clinical interactions," *arXiv preprint arXiv:2503.22678*, 2025.

[93] P. P. Van, D. N. Minh, A. D. Ngoc, and H. P. Thanh, "Rx strategist: Prescription verification using llm agents system," *arXiv preprint arXiv:2409.03440*, 2024.

[94] Y. Kim, C. Park, H. Jeong, Y. S. Chan, X. Xu, D. McDuff, H. Lee, M. Ghassemi, C. Breazeal, and H. W. Park, "Mdagents: An adaptive collaboration of llms for medical decision-making," *Advances in Neural Information Processing Systems*, vol. 37, pp. 79 410–79 452, 2024.

[95] F. Zeng, Z. Lyu, Q. Li, and X. Li, "Enhancing llms for impression generation in radiology reports through a multi-agent system," *arXiv preprint arXiv:2412.06828*, 2024.

[96] M. Sudarshan, S. Shih, E. Yee, A. Yang, J. Zou, C. Chen, Q. Zhou, L. Chen, C. Singhal, and G. Shih, "Agentic llm workflows for generating patient-friendly medical reports," *arXiv preprint arXiv:2408.01112*, 2024.

[97] Z. Bao, Q. Liu, Y. Guo, Z. Ye, J. Shen, S. Xie, J. Peng, X. Huang, and Z. Wei, "Piors: Personalized intelligent outpatient reception based on large language model with multi-agents medical scenario simulation," *arXiv preprint arXiv:2411.13902*, 2024.

[98] J. Liu, Y. Wang, J. Du, J. T. Zhou, and Z. Liu, "Medcot: Medical chain of thought via hierarchical expert," *arXiv preprint arXiv:2412.13736*, 2024.

[99] L. Wei, W. Wang, X. Shen, Y. Xie, Z. Fan, X. Zhang, Z. Wei, and W. Chen, "Mc-cot: A modular collaborative cot framework for zero-shot medical-vqa with llm and mllm integration," *arXiv preprint arXiv:2410.04521*, 2024.

[100] Z. Wang, J. Wu, L. Cai, C. H. Low, X. Yang, Q. Li, and Y. Jin, "Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow," *arXiv preprint arXiv:2503.18968*, 2025.

[101] K. Chen, X. Li, T. Yang, H. Wang, W. Dong, and Y. Gao, "Mdteamgpt: A self-evolving llm-based multi-agent framework for multi-disciplinary team medical consultation," *arXiv preprint arXiv:2503.13856*, 2025.

[102] S. Schmidgall, R. Ziaei, C. Harris, E. Reis, J. Jopling, and M. Moor, "Agentclinic: A multimodal agent benchmark to evaluate ai in simulated clinical environments," *arXiv preprint*, vol. arXiv:2405.07960, 2024. [Online]. Available: https://arxiv.org/abs/2405.07960

[103] D. Ferber, O. S. El Nahhas, G. Wölflein, I. C. Wiest, J. Clusmann, M.-E. Leßman, S. Foersch, J. Lammert, M. Tschochohei, D. Jäger *et al.*, "Autonomous artificial intelligence agents for clinical decision making in oncology," *arXiv preprint arXiv:2404.04667*, 2024.

[104] A. Hoopes, V. I. Butoi, J. V. Guttag, and A. V. Dalca, "Voxelprompt: A vision-language agent for grounded medical image analysis," *arXiv preprint arXiv:2410.08397*, 2024.

[105] Z. Wang, Y. Zhu, H. Zhao, X. Zheng, D. Sui, T. Wang, W. Tang, Y. Wang, E. Harrison, C. Pan *et al.*, "Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 2250–2261.

[106] Y. Zhou, L. Song, and J. Shen, "Mam: Modular multi-agent framework for multi-modal medical diagnosis via role-specialized collaboration," *arXiv preprint arXiv:2506.19835*, 2025.

[107] X. Chen, H. Yi, M. You, W. Liu, L. Wang, H. Li, X. Zhang, Y. Guo, L. Fan, G. Chen *et al.*, "Enhancing diagnostic capability with multi-agents conversational large language models," *NPJ digital medicine*, vol. 8, no. 1, p. 159, 2025.

[108] F. Ghezloo, M. S. Seyfioglu, R. Soraki, W. O. Ikezogwo, B. Li, T. Vivekanandan, J. G. Elmore, R. Krishna, and L. Shapiro, "Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology," *arXiv preprint arXiv:2502.08916*, 2025.

[109] Z. Chen, Z. Peng, X. Liang, C. Wang, P. Liang, L. Zeng, M. Ju, and Y. Yuan, "Map: Evaluation and multi-agent enhancement of large language models for inpatient pathways," *arXiv preprint arXiv:2503.13205*, 2025.

[110] Z. Fan, J. Tang, W. Chen, S. Wang, Z. Wei, J. Xi, F. Huang, and J. Zhou, "Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator," *arXiv preprint arXiv:2402.09742*, 2024.

[111] Y. Feng, J. Wang, L. Zhou, and Y. Li, "Doctoragent-rl: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue," *arXiv preprint arXiv:2505.19630*, 2025.

[112] W. Hou, G. Yang, Y. Du, Y. Lau, L. Liu, J. He, L. Long, and S. Wang, "Adagent: Llm agent for alzheimer's disease analysis with collaborative coordinator," *arXiv preprint arXiv:2506.11150*, 2025.

[113] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu *et al.*, "Can generalist foundation models outcompete special-purpose tuning? case study in medicine," *arXiv preprint arXiv:2311.16452*, 2023.

[114] C.-H. Chang, M. M. Lucas, Y. Lee, C. C. Yang, and G. Lu-Yao, "Beyond self-consistency: Ensemble reasoning boosts consistency and accuracy of llms in cancer staging," in *International Conference on Artificial Intelligence in Medicine*. Springer, 2024, pp. 224–228.

[115] M. Graber, "Diagnostic errors in medicine: a case of neglect," *The joint commission journal on quality and patient safety*, vol. 31, no. 2, pp. 106–113, 2005.

[116] M. Abrandt Dahlgren, K. Valeskog, K. Johansson, and S. Edelbring, "Understanding clinical reasoning: A phenomenographic study with entry-level physiotherapy students," *Physiotherapy Theory and Practice*, vol. 38, no. 13, pp. 2817–2826, 2022.

[117] M. Young, A. Thomas, S. Lubarsky, T. Ballard, D. Gordon, L. D. Gruppen, E. Holmboe, T. Ratcliffe, J. Rencic, L. Schuwirth *et al.*, "Drawing boundaries: the difficulty in defining clinical reasoning," *Academic Medicine*, vol. 93, no. 7, pp. 990–995, 2018.

[118] S. Mamede, T. van Gog, K. van den Berge, R. M. Rikers, J. L. van Saase, C. van Guldener, and H. G. Schmidt, "Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents," *Jama*, vol. 304, no. 11, pp. 1198–1203, 2010.

[119] J. Higgs, G. M. Jensen, S. Loftus, F. V. Trede, and S. Grace, *Clinical Reasoning in the Health Professions E-Book: clinical Reasoning in the Health Professions E-Book*. Elsevier Health Sciences, 2024.

[120] B. Moëll, F. Sand Aronsson, and S. Akbar, "Medical reasoning in llms: an in-depth analysis of deepseek r1," *Frontiers in Artificial Intelligence*, vol. 8, p. 1616145, 2025.

[121] H. Xu, Y. Wang, Y. Xun, R. Shao, and Y. Jiao, "Artificial intelligence for clinical reasoning: the reliability challenge and path to evidence-based practice," *QJM: An International Journal of Medicine*, p. hcaf114, 2025.

[122] M. Khosravi, Z. Zare, S. M. Mojtabaeian, and R. Izadi, "Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews," *Health services research and managerial epidemiology*, vol. 11, p. 23333928241234863, 2024.

[123] H. A. Shah and M. Househ, "Chain of thought strategy for smaller llms for medical reasoning," *Studies in health technology and informatics*, vol. 327, pp. 783–787, 2025.

[124] J. Miao, C. Thongprayoon, S. Suppadungsuk, P. Krisanapan, Y. Radhakrishnan, and W. Cheungpasitporn, "Chain of thought utilization in large language models and application in nephrology," *Medicina*, vol. 60, no. 1, p. 148, 2024.

[125] A. Singh, "Future prospects of agentic rag in medical research," *Available at SSRN 5212343*, 2025.

[126] D. Patel, G. Raut, S. N. Cheetirala, B. Glicksberg, M. A. Levin, G. Nadkarni, R. Freeman, E. Klang, and P. Timsina, "Ai agents in modern healthcare: From foundation to pioneer–a comprehensive review and implementation roadmap for impact and integration in clinical settings," 2025.

[127] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[128] Y. Zhou, X. Geng, T. Shen, C. Tao, G. Long, J.-G. Lou, and J. Shen, "Thread of thought unraveling chaotic contexts," *arXiv preprint arXiv:2311.08734*, 2023.

[129] H. Hu, Y. Zhou, J. Si, Q. Wang, H. Zhang, F. Ren, F. Ma, L. Cui, and Q. Tian, "Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling," *arXiv preprint arXiv:2505.15715*, 2025.

[130] V. Liévin, C. E. Hother, A. G. Motzfeldt, and O. Winther, "Can large language models reason about medical questions?" *Patterns*, vol. 5, no. 3, 2024.

[131] T. Savage, A. Nayak, R. Gallo, E. Rangan, and J. H. Chen, "Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine," *NPJ Digital Medicine*, vol. 7, no. 1, p. 20, 2024.

[132] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[133] H. Nori, N. Usuyama, N. King, S. M. McKinney, X. Fernandes, S. Zhang, and E. Horvitz, "From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond," *arXiv preprint arXiv:2411.03590*, 2024.

[134] S. Sandeep Nachane, O. Gramopadhye, P. Chanda, G. Ramakrishnan, K. Sharad Jadhav, Y. Nandwani, D. Raghu, and S. Joshi, "Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering," *arXiv e-prints*, pp. arXiv–2403, 2024.

[135] B. J. Gutierrez, N. McNeal, C. Washington, Y. Chen, L. Li, H. Sun, and Y. Su, "Thinking about gpt-3 in-context learning for biomedical ie? think again," *arXiv preprint arXiv:2203.08410*, 2022.

[136] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[137] Z. Wu, A. Hasan, J. Wu, Y. Kim, J. P. Cheung, T. Zhang, and H. Wu, "Chain-of-though (cot) prompting strategies for medical error detection and correction," *arXiv preprint arXiv:2406.09103*, 2024.

[138] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[139] M. M. Lucas, J. Yang, J. K. Pomeroy, and C. C. Yang, "Reasoning with large language models for medical question answering," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1964–1975, 2024.

[140] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang *et al.*, "Self-refine: Iterative refinement with self-feedback," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 534–46 594, 2023.

[141] Q. Yu, M. Jin, D. Shu, C. Zhang, L. Fan, W. Hua, S. Zhu, Y. Meng, Z. Wang, M. Du *et al.*, "Health-llm: Personalized retrieval-augmented disease prediction system," *arXiv preprint arXiv:2402.00746*, 2024.

[142] H. Wang, S. Zhao, Z. Qiang, Z. Li, C. Liu, N. Xi, Y. Du, B. Qin, and T. Liu, "Knowledge-tuning large language models with structured medical knowledge bases for trustworthy response generation in chinese," *ACM Transactions on Knowledge Discovery from Data*, vol. 19, no. 2, pp. 1–17, 2025.

[143] Y. Wang, X. Ma, and W. Chen, "Augmenting black-box llms with medical textbooks for biomedical question answering (published in findings of emnlp 2024)," *arXiv preprint arXiv:2309.02233*, 2023.

[144] X. Tang, D. Shao, J. Sohn, J. Chen, J. Zhang, J. Xiang, F. Wu, Y. Zhao, C. Wu, W. Shi *et al.*, "Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning," *arXiv preprint arXiv:2503.07459*, 2025.

[145] R. Xu, Y. Zhuang, Y. Zhong, Y. Yu, X. Tang, H. Wu, M. D. Wang, P. Ruan, D. Yang, T. Wang *et al.*, "Medagentgym: Training llm agents for code-based medical reasoning at scale," *arXiv preprint arXiv:2506.04405*, 2025.

[146] A. J. Goodell, S. N. Chu, D. Rouholiman, and L. F. Chu, "Large language model agents can use tools to perform clinical calculations," *npj Digital Medicine*, vol. 8, no. 1, p. 163, 2025.

[147] Y. Zhu, S. Wei, X. Wang, K. Xue, X. Zhang, and S. Zhang, "Menti: Bridging medical calculator and llm agent with nested tool calling," *arXiv preprint arXiv:2410.13610*, 2024.

[148] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *International Conference on Learning Representations (ICLR)*, 2023.

[149] Y. Cheng, C. Zhang, Z. Zhang, X. Meng, S. Hong, W. Li, Z. Wang, Z. Wang, F. Yin, J. Zhao *et al.*, "Exploring large language model based intelligent agents: Definitions, methods, and prospects," *arXiv preprint arXiv:2401.03428*, 2024.

[150] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian *et al.*, "Toolllm: Facilitating large language models to master 16000+ real-world apis," *arXiv preprint arXiv:2307.16789*, 2023.

[151] Y. Mao, W. Xu, Y. Qin, and Y. Gao, "Ct-agent: A multimodal-llm agent for 3d ct radiology question answering," *arXiv preprint arXiv:2505.16229*, 2025.

[152] X. Su, Y. Wang, S. Gao, X. Liu, V. Giunchiglia, D.-A. Clevert, and M. Zitnik, "Kgarevion: an ai agent for knowledge-intensive biomedical qa," *arXiv preprint arXiv:2410.04660*, 2024.

[153] J. Chen, C. Gui, A. Gao, K. Ji, X. Wang, X. Wan, and B. Wang, "Cod, towards an interpretable medical agent using chain of diagnosis," *arXiv preprint arXiv:2407.13301*, 2024.

[154] Z. Fan, C. Liang, C. Wu, Y. Zhang, Y. Wang, and W. Xie, "Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification," *arXiv preprint arXiv:2504.20930*, 2025.

[155] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 8634–8652, 2023.

[156] A. Dutta and Y.-C. Hsiao, "Adaptive reasoning and acting in medical language agents," *arXiv preprint arXiv:2410.10020*, 2024.

[157] X. Li, S. Wang, S. Zeng, Y. Wu, and Y. Yang, "A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges," *Vicinagearth*, vol. 1, no. 1, p. 9, 2024.

[158] K. Zuo, Z. Zhong, P. Huang, S. Tang, Y. Chen, and Y. Jiang, "Healkggen: A hierarchical multi-agent llm framework with knowledge

graph enhancement for genetic biomarker-based medical diagnosis," *bioRxiv*, pp. 2025–06, 2025.
[159] K. Zuo, Y. Jiang, F. Mo, and P. Lio, "Kg4diagnosis: A hierarchical multi-agent llm framework with knowledge graph enhancement for medical diagnosis," in *AAAI Bridge Program on AI for Medicine and Healthcare*. PMLR, 2025, pp. 195–204.
[160] K. Zhang, R. Zhou, E. Adhikarla, Z. Yan, Y. Liu, J. Yu, Z. Liu, X. Chen, B. D. Davison, H. Ren *et al.*, "A generalist vision–language foundation model for diverse biomedical tasks," *Nature Medicine*, vol. 30, no. 11, pp. 3129–3141, 2024.
[161] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, and B. Wang, "Huatuogpt-o1, towards medical complex reasoning with llms," *arXiv preprint arXiv:2412.18925*, 2024.
[162] J. Lei, X. Zhang, C. Wu, L. Dai, Y. Zhang, Y. Zhang, Y. Wang, W. Xie, and Y. Li, "Autorg-brain: Grounded report generation for brain mri," *arXiv preprint arXiv:2407.16684*, 2024.
[163] S. Lee, J. Youn, H. Kim, M. Kim, and S. H. Yoon, "Cxr-llava: a multimodal large language model for interpreting chest x-ray images," *European Radiology*, pp. 1–13, 2025.
[164] W. Gao, Z. Deng, Z. Niu, F. Rong, C. Chen, Z. Gong, W. Zhang, D. Xiao, F. Li, Z. Cao *et al.*, "Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue," *arXiv preprint arXiv:2306.12174*, 2023.
[165] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, K. Ikamura, G. Gerber, I. Liang, L. P. Le, T. Ding, A. V. Parwani *et al.*, "A foundational multimodal vision language ai assistant for human pathology," *arXiv preprint arXiv:2312.07814*, 2023.
[166] J. Dai, Q. Zhu, J. Zhan, B. Wang, and X. Qiu, "Moss-med: a family of multimodal models serving medical image analysis," *ACM Transactions on Management Information Systems*, vol. 16, no. 2, pp. 1–14, 2025.
[167] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9579–9589.
[168] Z. Zhao, L. Dai, Y. Zhang, Y. Wang, and W. Xie, "Rethinking whole-body ct image interpretation: An abnormality-centric approach," *arXiv preprint arXiv:2506.03238*, 2025.
[169] Radiopaedia.org, "Radiopaedia: The peer-reviewed collaborative radiology resource." [Online]. Available: https://radiopaedia.org/
[170] X. Zhang, C. Wu, Z. Zhao, J. Lei, Y. Zhang, Y. Wang, and W. Xie, "Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis," *arXiv preprint arXiv:2404.16754*, 2024.
[171] X. Zhou, L. Sun, D. He, W. Guan, R. Wang, L. Wang, X. Sun, K. Sun, Y. Zhang, Y. Wang *et al.*, "A knowledge-enhanced pathology vision-language foundation model for cancer diagnosis," *arXiv preprint arXiv:2412.13126*, 2024.
[172] H. Wu, Z. Zhao, Y. Zhang, Y. Wang, and W. Xie, "Mrgen: Segmentation data engine for underrepresented mri modalities," *arXiv preprint arXiv:2412.04106*, 2024.
[173] Y. Hu, T. Li, Q. Lu, W. Shao, J. He, Y. Qiao, and P. Luo, "Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 170–22 183.
[174] J. Qiu, K. Lam, G. Li, A. Acharya, T. Y. Wong, A. Darzi, W. Yuan, and E. J. Topol, "Llm-based agentic systems in medicine and healthcare," *Nature Machine Intelligence*, vol. 6, no. 12, pp. 1418–1420, 2024.
[175] H. Wang, S. Zhao, Z. Qiang, N. Xi, B. Qin, and T. Liu, "Beyond direct diagnosis: Llm-based multi-specialist agent consultation for automatic diagnosis," *arXiv preprint arXiv:2401.16107*, 2024.
[176] A. B. Abacha and D. Demner-Fushman, "On the summarization of consumer health questions," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2228–2234.
[177] S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng *et al.*, "Radgraph: Extracting clinical entities and relations from radiology reports," *arXiv preprint arXiv:2106.14463*, 2021.
[178] S. Schmidgall, R. Ziaei, C. Harris, E. Reis, J. Jopling, and M. Moor, "Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments," *arXiv preprint arXiv:2405.07960*, 2024.
[179] J. Liu, W. Wang, Z. Ma, G. Huang, Y. SU, K.-J. Chang, W. Chen, H. Li, L. Shen, and M. Lyu, "Medchain: Bridging the gap between llm agents and clinical practice through interactive sequential benchmarking," *arXiv preprint arXiv:2412.01605*, 2024.
[180] Z. Du, L. Zheng, R. Hu, Y. Xu, X. Li, Y. Sun, W. Chen, J. Wu, H. Cai, and H. Ying, "Llms can simulate standardized patients via agent coevolution," *arXiv preprint arXiv:2412.11716*, 2024.
[181] D. Kyung, H. Chung, S. Bae, J. Kim, J. H. Sohn, T. Kim, S. K. Kim, and E. Choi, "Patientsim: A persona-driven simulator for realistic doctor-patient interactions," *arXiv preprint arXiv:2505.17818*, 2025.
[182] Y. Zhu, Z. He, H. Hu, X. Zheng, X. Zhang, Z. Wang, J. Gao, L. Ma, and L. Yu, "Medagentboard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks," *arXiv preprint arXiv:2505.12371*, 2025.
[183] W. Yan, H. Liu, T. Wu, Q. Chen, W. Wang, H. Chai, J. Wang, W. Zhao, Y. Zhang, R. Zhang *et al.*, "Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world," *arXiv preprint arXiv:2406.13890*, 2024.
[184] I. Sviridov, A. Miftakhova, A. Tereshchenko, G. Zubkova, P. Blinov, and A. Savchenko, "3mdbench: Medical multimodal multi-agent dialogue benchmark," *arXiv preprint arXiv:2504.13861*, 2025.
[185] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
[186] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.
[187] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 2567–2577.
[188] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," in *International Conference on Learning Representations*, 2021.
[189] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," in *Conference on health, inference, and learning*. PMLR, 2022, pp. 248–260.
[190] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang *et al.*, "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark," *arXiv preprint arXiv:2406.01574*, 2024.
[191] M. D. Ma, C. Ye, Y. Yan, X. Wang, P. Ping, T. S. Chang, and W. Wang, "Clibench: A multifaceted and multigranular evaluation of large language models for clinical decision making," *arXiv preprint arXiv:2406.09923*, 2024.
[192] Y. Jiang, K. C. Black, G. Geng, D. Park, A. Y. Ng, and J. H. Chen, "Medagentbench: Dataset for benchmarking llms as agents in medical applications," *arXiv preprint arXiv:2501.14654*, 2025.
[193] C. Chen, J. Yu, S. Chen, C. Liu, Z. Wan, D. Bitterman, F. Wang, and K. Shu, "Clinicalbench: Can llms beat traditional ml models in clinical prediction?" *arXiv preprint arXiv:2411.06469*, 2024.
[194] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 595–46 623, 2023.
[195] S. Mukherjee, P. Gamble, M. S. Ausin, N. Kant, K. Aggarwal, N. Manjunath, D. Datta, Z. Liu, J. Ding, S. Busacca *et al.*, "Polaris: A safety-focused llm constellation architecture for healthcare," *arXiv preprint arXiv:2403.13313*, 2024.
[196] A. Pal and M. Sankarasubbu, "Gemini goes to med school: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations," in *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 2024, pp. 21–46.
[197] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.
[198] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[199] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.
[200] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.
[201] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[202] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[203] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière *et al.*, "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025.

[204] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[205] M. S. Ankit Pal, "Openbiollms: Advancing open-source large language models for healthcare and life sciences," https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024.

[206] OpenAI, "Hello gpt-4o." 2024, accessed: 2025-05-31. [Online]. Available: https://openai.com/index/hello-gpt-4o

[207] S. Kweon, J. Kim, J. Kim, S. Im, E. Cho, S. Bae, J. Oh, G. Lee, J. H. Moon, S. C. You *et al.*, "Publicly shareable clinical large language model built on synthetic clinical notes," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 5148–5168.

[208] T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, A. V. Stolyar, K. Polanska, K. R. McCarthy, H. Osterhoudt, X. Wu, S. Visweswaran, S. Fu *et al.*, "A framework for human evaluation of large language models in healthcare derived from literature review," *NPJ digital medicine*, vol. 7, no. 1, p. 258, 2024.

[209] H. Hu, Y. Zhou, C. Ma, Q. Wang, Z. Zhang, F. Ma, L. Cui, and Q. Tian, "Theramind: A strategic and adaptive agent for longitudinal psychological counseling," *arXiv preprint arXiv:2510.25758*, 2025.