Medical Reasoning in the Era of LLMs: A Systematic Review of Enhancement Techniques and Applications

Wenxuan Wang[1*] **Zizhan Ma**[2*] **Meidan Ding**[3] **Shiyi Zheng**[3] **Shengyuan Liu**[2]
Jie Liu[4] **Jiaming Ji**[5] **Wenting Chen**[4] **Xiang Li**[6] **Linlin Shen**[3] **Yixuan Yuan**[2]
[1]Renmin University of China [2]The Chinese University of Hong Kong
[3] Shenzhen University [4] City University of Hong Kong [5] Peking University
[6] Massachusetts General Hospital and Harvard Medical School
[1]wangwenxuan@ruc.edu.cn [2]zzma2@cse.cuhk.edu.hk [3]wentichen7-c@my.cityu.edu.hk

Abstract

The proliferation of Large Language Models (LLMs) in medicine has enabled impressive capabilities, yet a critical gap remains in their ability to perform systematic, transparent, and verifiable reasoning—a cornerstone of clinical practice. This has catalyzed a shift from single-step answer generation to the development of LLMs explicitly designed for medical reasoning. This paper provides the first systematic review of this emerging field. We propose a taxonomy of reasoning enhancement techniques, categorized into training-time strategies (e.g., supervised fine-tuning, reinforcement learning) and test-time mechanisms (e.g., prompt engineering, multi-agent systems). We analyze how these techniques are applied across different data modalities (text, image, code) and in key clinical applications such as diagnosis, education, and treatment planning. Furthermore, we survey the evolution of evaluation benchmarks from simple accuracy metrics to sophisticated assessments of reasoning quality and visual interpretability. Based on an analysis of 60 seminal studies from 2022-2025, we conclude by identifying critical challenges, including the faithfulness-plausibility gap and the need for native multimodal reasoning, and outlining future directions toward building efficient, robust, and sociotechnically responsible medical AI.

1 Introduction

The emergence of Large Language Models (LLMs) has catalyzed remarkable progress in the medical domain, with specialized models like Med-PaLM (Singhal et al., 2025), PMC-LLaMA (Wu et al., 2023), and BioGPT (Luo et al., 2022) demonstrating significant capabilities. However, these models, which often generate answers directly, struggle with the complex, multi-step inference crucial for high-stakes clinical decision-making. The process of medical diagnosis is not one of simple pattern matching but of deep, causal reasoning, where clinicians synthesize symptoms, patient history, and test results to form a coherent explanation (Richens et al., 2020; Xue et al., 2024). Diagnostic errors, a leading cause of medical malpractice claims, underscore the profound risks of inadequate reasoning (Schaffer et al., 2017).

This critical need has spurred research into reasoning LLMs. Inspired by breakthroughs like Chain-of-Thought (CoT) prompting (Wei et al., 2023), which elicits intermediate inferential steps, the field is developing models that can simulate clinical workflows, justify conclusions, and adapt to complex diagnostic challenges. These models are not only vital for decision support but also as educational tools, where AI-driven structured feedback has been shown to significantly enhance students' clinical skills (Brügge et al., 2024).

This paper presents the first systematic review of this emerging field. To ensure a comprehensive analysis, we conducted a structured literature search across major academic databases, including PubMed, Scopus, Google Scholar, and arXiv, for papers published between 2022 and 2025. Using keywords such as "LLM," "medical reasoning," "clinical reasoning," and "complex medical tasks," our initial query yielded over 200 articles. These were screened by title and abstract, followed by a full-text review for relevance to explicit reasoning mechanisms. This process resulted in a final corpus of 60 core studies that form the basis of our review. We have also released a companion GitHub repository hosting all curated papers, benchmarks, and resources for medical reasoning with LLMs and MLLMs at https://github.com/zzma2/medical-llm-reasoning-survey.

The remainder of this paper is structured to guide the reader from fundamental concepts to future challenges. Section 3 examines how reasoning techniques are adapted across different data modalities:

[*] Wenxuan Wang and Zizhan Ma equally contribute to this paper.

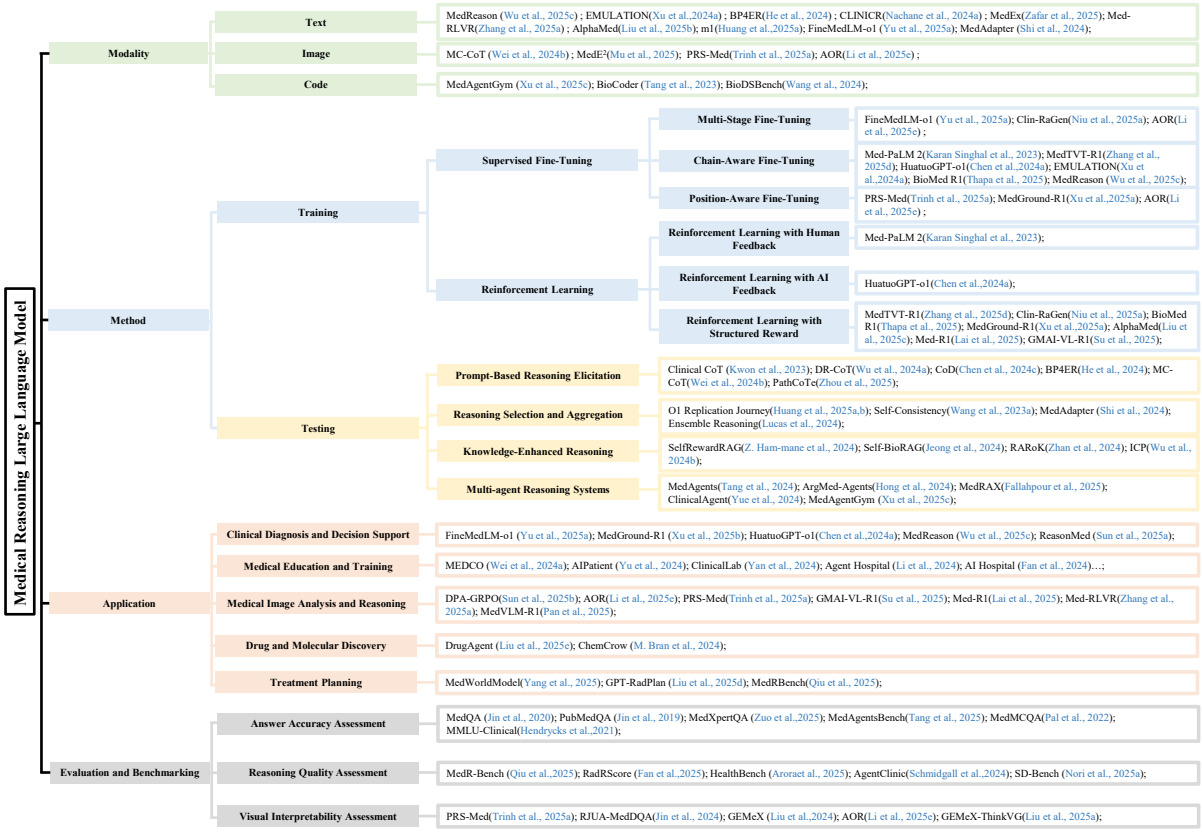

Figure 1: A taxonomy of medical reasoning (LLMs). This figure provides a visual summary of the topics discussed in this review, outlining the primary data modalities (Section 3), the core reasoning enhancement techniques (Section 4), key medical applications (Section 5), and the evolving paradigms for evaluation (Section 6).

text, image, and code. Section 4 presents our core contribution, a taxonomy of reasoning enhancement techniques organized into training-time and test-time strategies. Subsequently, Section 5 surveys the key clinical applications of these models, while Section 6 details the evolution of benchmarks for assessing their performance. Finally, Section 7 discusses critical challenges and future research directions before we offer concluding remarks.

2 Background

Medical reasoning, the cognitive process of synthesizing patient data to formulate diagnoses and treatment plans, is fundamental to medical practice. While Large Language Models (LLMs) excel at processing medical text, their standard probabilistic architecture is not inherently suited for the structured, multi-step inference required in high-stakes clinical decision-making. This limitation has spurred the development of specialized **Reasoning LLMs**, which are architected to produce transparent, verifiable, and robust inferential pathways, addressing a critical need for trustworthy AI in medicine (Savage et al., 2024; Kim et al., 2024; Yu et al., 2025).

2.1 From Generalist LLMs to Specialized Reasoners

Conventional LLMs, operating as probabilistic sequence models, are proficient at knowledge retrieval but often struggle with complex logical operations, such as distinguishing correlation from causation or managing uncertainty (Wei et al., 2023). A key breakthrough was Chain-of-Thought (CoT) prompting, which demonstrated that by instructing a model to generate step-by-step reasoning, its latent inferential capabilities could be elicited, significantly improving performance on logical tasks (Wei et al., 2023; Nachane et al., 2024).

Building on this insight, the focus has shifted from prompting techniques to architecting models where reasoning is a primary design objective. State-of-the-art models, such as OpenAI's o1, now integrate supervised fine-tuning on explicit reasoning traces and reinforcement learning from human feedback (RLHF) to reward logically sound processes (OpenAI et al., 2024; Pan et al., 2025b). Recent advances have demonstrated that structured

clinical reasoning approaches significantly enhance diagnostic accuracy (Sonoda et al., 2025). Therefore, Reasoning LLMs are defined not just by their performance but by their designed capacity for transparent inference—a critical feature for their safe application in high-stakes domains (Wu et al., 2025).

2.2 The Imperative for Robust Reasoning in Medical Practice

The need for robust LLM reasoning is particularly acute in medicine, where inferential quality directly impacts patient safety and outcomes. Diagnostic errors, often stemming from flawed clinical reasoning, are a leading cause of preventable harm, contributing to an estimated 31.8% of medical malpractice claims, with a significant portion resulting in patient death (Schaffer et al., 2017).

Reasoning LLMs offer a promising approach to mitigate these challenges. In medical education, AI-driven tutors are already being used to improve students' diagnostic and history-taking skills through structured feedback (Brügge et al., 2024). For practicing clinicians, these models can act as cognitive partners. For instance, recent systems can analyze diagnostic reasoning documented in electronic health records to provide real-time feedback on potential cognitive biases or logical gaps (Schaye et al., 2025). By augmenting clinical reasoning in both training and practice, these advanced models represent a new frontier for improving the quality, safety, and consistency of patient care.

3 Medical Reasoning Under Various Modalities

The data modality—whether text, image, or code—fundamentally shapes the medical reasoning challenge. Consequently, the strategies to imbue LLMs with reasoning capabilities are tailored to the unique constraints and affordances of each data type. This section analyzes how reasoning techniques are adapted across these three primary modalities.

3.1 Reasoning over Text

Textual data, found in clinical notes, dialogues, and medical literature, is information-dense but lacks inherent logical structure. The primary challenge is to guide the model's generative process along a factually correct and clinically valid inferential path. Research has coalesced around three main strategies. First, to impose structure, models are trained to make their reasoning explicit. Techniques like **explicit path generation** guide models to produce step-by-step rationales, either by aligning with clinical inference patterns (Xu et al., 2024) or by grounding each step in a structured knowledge graph (Wu et al., 2025). Second, to ensure the validity of these paths, researchers focus on **enforcing logical consistency**. This is achieved by incorporating formal methods like first-order logic (FOL) to verify claims (Zafar et al., 2025) or by using reinforcement learning to reward factual correctness and penalize hallucinations (Zhang et al., 2025a; Liu et al., 2025b). Third, to move beyond single, linear paths, other work explores **deepening and broadening inference**. This includes test-time scaling (TTS) to allocate more computation for deeper reasoning on a single problem (Huang et al., 2025a; Yu et al., 2025; Shi et al., 2024), and multi-agent systems that simulate collaborative debate or dialogue to explore diverse perspectives and build a more robust, explainable consensus (Hong et al., 2024; Tang et al., 2024; Zhu and Wu, 2025).

3.2 Reasoning over Image

In medical imaging, the central challenge is bridging the "modality gap" between low-level pixel data and high-level clinical concepts. Reasoning must be visually grounded to be trustworthy. The research landscape reflects a progression toward tighter integration of vision and language. An initial approach focuses on **cross-modal alignment**, often using reinforcement learning to teach Vision-Language Models (VLMs) to associate visual findings with correct diagnostic labels, rewarding the model for making clinically sound connections (Pan et al., 2025a; Lai et al., 2025; Su et al., 2025; Zhang et al., 2025b). A more sophisticated strategy involves **coordinating the reasoning process** between modalities. Frameworks like MC-CoT (Wei et al., 2024b) and the two-phase paradigm of Elicit and Enhance (Mu et al., 2025) establish an "orchestrator-perceiver" dynamic, where a language model generates a high-level reasoning plan that explicitly directs the VLM's visual analysis. The most advanced methods aim for deep **clinical and spatial grounding**. These models move beyond simple object detection to incorporate fine-grained anatomical knowledge, either by integrating segmentation capabilities to reason about precise spatial locations (Trinh et al., 2025) or by using anatomical ontologies to structure the interpreta-

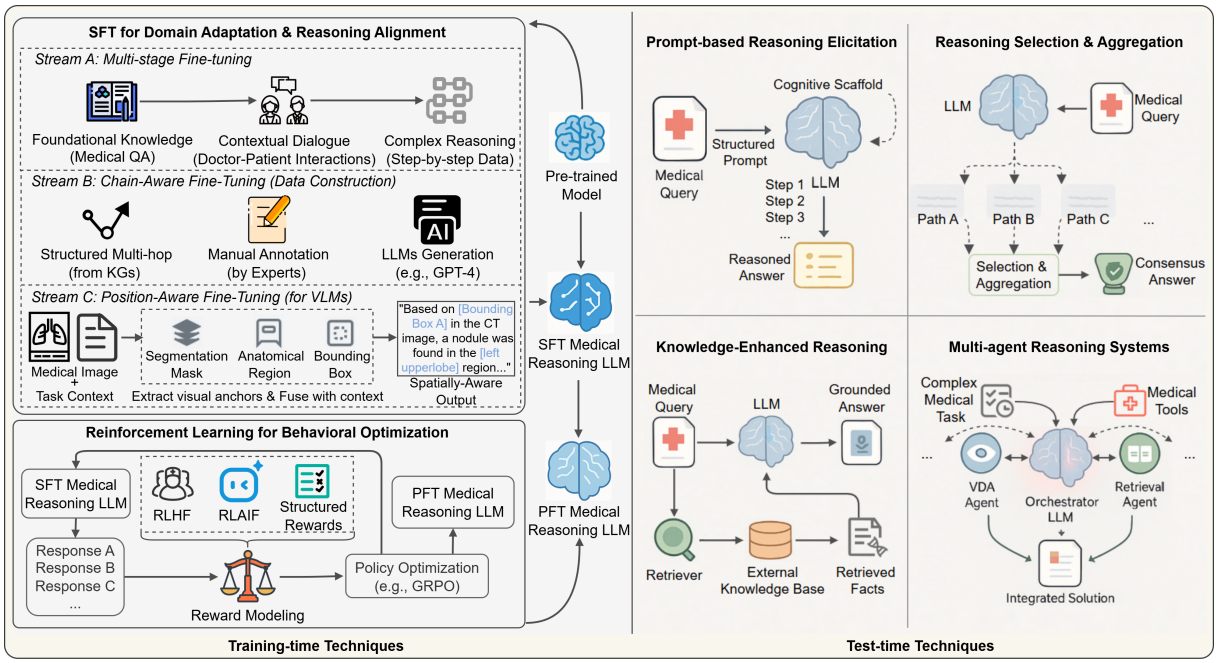

Figure 2: An overarching framework of techniques to enhance reasoning in Medical LLMs, divided into two primary stages. **(a) Training-time Techniques** fundamentally imbue models with reasoning capabilities by modifying their internal weights through methods like Supervised Fine-tuning (SFT) on reasoning-aware data and Reinforcement Learning (RL) with expert feedback. **(b) Test-time Techniques** improve reasoning at the moment of inference. These on-the-fly strategies include Prompt-based Elicitation to guide thought processes, Reasoning Selection & Aggregation for robustness, Knowledge-Enhanced Reasoning to ground responses in facts, and Multi-agent Systems that decompose complex problems for collaborative solving.

tion of findings in a clinically coherent manner (Li et al., 2025c).

3.3 Reasoning over Code

Code as a modality for medical reasoning is a new frontier, enabling procedural, verifiable, and automatable workflows. Unlike text and images, the primary challenge has been to build the foundational ecosystem for this type of research. The narrative of progress can be seen as constructing three essential pillars. The first pillar is the **environment**: MedAgentGym (Xu et al., 2025b) provides a standardized, extensible, and verifiable training and evaluation "gym" for medical agents, solving the need for a reproducible setting. The second pillar is the **data**: studies like BioCoder (Tang et al., 2023) have been crucial in analyzing and validating the richness of biomedical code available in public repositories, confirming that a sufficient data foundation exists to train capable models. The third pillar is the **platform**: with an environment and data, BioDSBench (Wang et al., 2024) represents the integration of these ideas into a usable platform, embedding LLMs within a data science pipeline where code serves as the direct interface for medical professionals to perform complex computational tasks.

4 A Taxonomy of Medical Reasoning Enhancement Techniques

To endow LLMs with robust medical reasoning, researchers have developed a suite of techniques that can be broadly categorized into two main stages: **training-time techniques**, which fundamentally alter a model's internal weights to build foundational reasoning capabilities, and **test-time techniques**, which steer and refine the model's output at the moment of inference without modifying the model itself. This section provides a systematic overview of these methods, illustrated in Figure 2.

4.1 Training-time Techniques: Building the Foundation

Training-time methods are high-cost, high-impact interventions that aim to bake clinical logic directly into the model's parameters. They represent the "heavy lifting" of creating a domain-specialized reasoner but face significant challenges related to data scalability.

4.1.1 Supervised Fine-tuning (SFT)

SFT marks a crucial epistemological shift from learning mere correlations to learning clinical *processes*. By training on data containing explicit reasoning chains, the model is forced to learn the "how" and "why" of a diagnosis. The innovation lies in the design of this data and the training strategy.

Multi-stage Fine-tuning: This approach applies the principle of curriculum learning, recognizing that complex clinical reasoning cannot be learned monolithically. The core idea is to break the skill into a sequence of manageable stages. We identify two primary strategies for this. The first, **staging by task abstraction**, is common for conceptual reasoning. It builds a hierarchy from concrete knowledge to abstract inference. FineMedLM-o1 (Yu et al., 2025) exemplifies this by first training on factual medical knowledge, then on interactive dialogues, and finally on complex causal reasoning. The second strategy, **staging by modality integration**, is crucial for multimodal tasks. It builds skills from perception to interpretation. AOR (Li et al., 2025c) is a canonical example, sequentially training the model to first recognize anatomical structures, then ground them to linguistic terms, and finally synthesize this information into a diagnostic conclusion.

Chain-Aware Fine-tuning: This paradigm's challenge is the 'supervision bottleneck'—the cost and difficulty of obtaining high-quality reasoning chains. Researchers have developed four distinct strategies to address this. The *gold standard* is to use **human expert annotation**, where clinicians provide the consensus and rationale for reasoning paths, as done for Med-PaLM 2 (Singhal et al., 2025). While authoritative, this is not scalable. To overcome this, the most common strategy is using **AI-generated chains**. Frameworks like HuatuoGPT-o1 (Chen et al., 2024a) leverage powerful teacher models (e.g., GPT-4) in a sophisticated cycle of generation, verification, and self-correction to create vast datasets at scale, though this risks inheriting the teacher's biases. A third, more verifiable approach is to impose **external structure**. MedReason (Wu et al., 2025), for example, constrains generation by forcing the reasoning path to be a valid traversal of a medical knowledge graph, making each step auditable. Finally, some methods focus on **refining existing data**. This includes data-centric approaches like BioMed-R1 (Thapa et al., 2025), which filters multiple benchmarks to curate a dataset of only the most reasoning-intensive samples, and stylistic approaches like EMULATION (Xu et al., 2024), which fine-tunes the model to ensure its reasoning style authentically mimics the abductive and deductive thought processes of clinicians.

Position-Aware Fine-tuning: For multimodal reasoning to be clinically useful, a diagnostic claim must be grounded to a specific visual location. This technique directly tackles this critical "grounding problem" by training models on data that enforces spatial correspondence. The strategies vary by the granularity of the spatial information provided. The most foundational approach uses **coarse-grained grounding** with bounding boxes, which serve as explicit intermediate reasoning steps in models like MedGround-R1 (Xu et al., 2025a). For higher clinical precision, **fine-grained grounding** with pixel-level segmentation masks is employed. PRS-MED (Trinh et al., 2025), for instance, trains on precise masks and requires the model to answer questions about these specific regions. The most sophisticated strategy is **semantically-rich grounding**, where visual regions are linked to a formal medical vocabulary. AOR (Li et al., 2025c) does this by aligning image areas with concepts from an anatomical ontology, allowing the model to reason not just about "where" a finding is, but also "what" it is in a structured, clinically meaningful way.

4.1.2 Reinforcement Learning (RL)

If SFT provides raw capability, RL is the alignment phase sculpting this capability to fit the nuanced goals of clinical practice: safety, accuracy, and efficiency. The core challenge in applying RL is defining "good" clinical reasoning, which has led to a spectrum of feedback strategies, from holistic human judgment to granular, automated metrics.

At one end of this spectrum lies alignment with subjective, qualitative feedback. **RL with Human Feedback (RLHF)** directly captures complex clinical values by training on physician preferences. For instance, Med-PaLM 2 (Singhal et al., 2025) was optimized using preference rankings from a diverse panel of physicians, allowing it to learn intangible qualities like diagnostic prudence and safety. To address the significant cost and scalability limitations of RLHF, **RL with AI Feedback (RLAIF)** has emerged as a pragmatic alternative. HuatuoGPT-o1 (Chen et al., 2024a), for example, uses GPT-4o to provide scalable, binary reward sig-

nals on answer correctness, using a powerful AI as a proxy for human judgment.

At the other end of the spectrum lies optimization against objective, quantitative metrics using **Structured Rewards**. This engineering-driven approach offers scalability and reproducibility by defining explicit, measurable goals. This has become a powerful trend, with the policy optimization algorithm GRPO being widely used to train models on specific criteria. These include multi-faceted rationale quality (e.g., accuracy, coherence, and knowledge coverage in ClinRaGen (Niu et al., 2025)), precise multimodal grounding (e.g., spatial and semantic consistency in MedGround-R1 (Xu et al., 2025a)), and task-specific performance across diverse modalities and question types (e.g., Med-R1 (Lai et al., 2025), GMAI-VL-R1 (Su et al., 2025)).

Perhaps the most profound insight from this line of work is that RL can act as an "emergence engine" for complex reasoning. Studies like AlphaMed (Liu et al., 2025b) and BioMed-R1 (Thapa et al., 2025) demonstrate that by using simple, objective rewards (like multiple-choice accuracy) and focusing on a curated set of difficult problems, sophisticated reasoning capabilities can emerge without being explicitly taught via CoT distillation. This crucial finding challenges the "bigger is better" paradigm, suggesting a viable path toward creating smaller, more efficient, yet highly capable medical reasoning models.

4.2 Test-time Techniques: Achieving Agility and Verifiability

In contrast to costly retraining, test-time techniques offer a flexible, low-cost way to steer the reasoning of pre-trained models. These on-the-fly mechanisms represent a conceptual shift from viewing the LLM as a static oracle to a dynamic reasoning component. The strategies show a clear progression in sophistication, from simple input shaping to complex, multi-agent orchestration.

4.2.1 Prompt-based Reasoning Elicitation

This foundational technique uses structured prompts for "cognitive steering," compelling the model to externalize its latent thought process into an explicit, step-by-step format. The approach has evolved from generic Chain-of-Thought (CoT) prompting (Nachane et al., 2024) to domain-specific variants that emulate expert workflows. These include Clinical CoT (Kwon et al., 2023), Diagnostic Reasoning CoT (DR-CoT) (Wu et al., 2024a), and the formalized five-step Chain of Diagnosis (CoD) (Chen et al., 2024b), which breaks down diagnosis into explicit steps like symptom analysis and diagnostic testing. For more complex problems, techniques like least-to-most prompting decompose tasks into simpler sub-problems (He et al., 2024), while other methods use iterative questioning to verify claims (Vladika et al., 2025) or extend these concepts to orchestrate multimodal analysis (Wei et al., 2024b; Zhou et al., 2025).

4.2.2 Reasoning Selection and Aggregation

To mitigate the inherent stochasticity of LLM outputs, this pillar improves robustness by generating and evaluating multiple reasoning paths. The methods represent different points on a spectrum of computational cost versus performance gain. At the higher-cost end, **self-consistency** (Wang et al., 2023) and **ensemble reasoning** (Lucas et al., 2024) generate multiple candidate responses by introducing randomness during decoding and then select the most frequent or highest-quality answer via majority vote. Other methods invest more computation into a single, more exhaustive path through **test-time scaling** (Huang et al., 2025a,b). Another way, **test-time adaptation** uses a small, lightweight model like *MedAdapter* (Shi et al., 2024) as a post-hoc ranker to score and select the most clinically plausible solution from a pool of candidates generated by a much larger base model, achieving significant gains with minimal overhead.

4.2.3 Knowledge-Enhanced Reasoning

These techniques address the critical issues of hallucination and outdated knowledge by grounding the model's parametric memory in verifiable, external facts. The strategies fall into two main categories. The first is **"just-in-time" contextualization** via Retrieval-Augmented Generation (RAG). Before answering a question, the model first queries a medical database or text corpus for relevant information, then integrates this retrieved text into its context to generate a factually grounded answer (Hammane, 2024; Jeong et al., 2024; Zhan et al., 2024). The second strategy is **"just-in-place" guidance**, which uses structured knowledge to constrain the generation process. *In-Context Padding (ICP)* (Wu et al., 2024b), for instance, injects structured "knowledge seeds" (e.g., '(headache, is_symptom_of, migraine)') from a knowledge graph directly into the LLM's context,

guiding its generation along a logically sound and verifiable path.

4.2.4 Multi-agent Reasoning Systems

This frontier represents a paradigm shift from a monolithic intelligence to a distributed, specialized cognitive architecture where the LLM acts as an "orchestrator." This approach decomposes complex problems into tasks solved by multiple collaborating agents. We see two primary forms of collaboration. **Collaborative deliberation** frameworks simulate peer review; for example, one agent might act as a 'proposer' suggesting a diagnosis, while another acts as a 'critic,' challenging the evidence to force a more robust conclusion (Tang et al., 2024; Hong et al., 2024). **Functional decomposition** frameworks assign tasks to agents with specialized tools. This allows a central orchestrator to delegate sub-tasks to an 'imaging agent' that can call a segmentation model (Fallahpour et al., 2025), a 'data agent' that can execute database queries, or a 'trial agent' that can parse clinical trial documents (Yue et al., 2024). Supported by dedicated training environments (Xu et al., 2025b), this modular approach makes the entire reasoning process transparent and auditable by design (Gu et al., 2024; Zhu and Wu, 2025).

5 Applications & Use Cases

5.1 Clinical Diagnosis and Decision Support

Medical reasoning models enhance clinical diagnosis and decision support by delivering precise, evidence-based insights to optimize healthcare decisions. For example, FineMedLM-o1 (Yu et al., 2025) specializes in joint encoding of fine-grained symptoms, signs, and lab results; its candidate-ranking strategy markedly reduces misdiagnosis rates, MedGround-R1 (Xu et al., 2025a) integrates radiology annotations directly into language–vision alignment so that "read-the-scan, write-the-report" happens in a single forward pass, shortening reporting time for common imaging studies. Existing works like HuatuoGPT-o1 (Chen et al., 2024a), MedReason (Wu et al., 2025), and ReasonMed (Sun et al., 2025a) automatically generate a large batch of CoT reasoning drafts with verifiable mechanisms to strengthen the model's diagnostic capability.

5.2 Medical Education and Training

Systems that prioritise the development of explicit clinical-reasoning pathways now underpin medical education. MEDCO (Wei et al., 2024a) guides students through structured differential-diagnosis chains and collaborative hypothesis building, while AIPatient (Yu et al., 2024) integrates electronic health records with knowledge graphs to simulate realistic clinical scenarios. In simulation-based training, medical reasoning models are stress-tested in sandboxed clinical scenarios before deployment (Wei et al., 2024a; Wu and colleagues, 2025; Yu et al., 2024). These platforms like ClinicalLab (Yan et al., 2024), Agent Hospital (Li et al., 2024), and AI Hospital (Fan et al., 2024) deliver interactive patient cases with real-time, adaptive feed, acting as personalized tutors that sharpen learners' diagnostic reasoning skills.

5.3 Medical Image Analysis and Reasoning

Multimodal medical reasoning models not only pinpoint pathological cues but also narrate their clinical relevance in plain language and link each observation to concrete next-step decisions—an evolution that promises more transparent, efficient, and trusted radiologic care (Sun et al., 2025b; Li et al., 2025c). For example, PRS-Med (Trinh et al., 2025) improves anatomical and pathological reasoning for precise diagnostics across diverse imaging modalities, and some recent studies (Su et al., 2025; Pan et al., 2025a; Lai et al., 2025; Zhang et al., 2025a) add RL-based methods such as GRPO (Shao et al., 2024) to boost the reasoning quality and traceability.

5.4 Drug and Molecular Discovery

Medical reasoning models are emerging as end-to-end engines that span both the design of novel therapeutics and the personalization of their clinical use. On the discovery side, DrugAgent (Liu et al., 2025d) treats drug-target interaction prediction as a sequential reasoning problem and reports a 4.92% ROC-AUC gain over strong baselines, while ChemCrow (M. Bran et al., 2024) stitches together 18 chemistry tools and exposes its chain of thought to autonomously plan multi-step syntheses and suggest new molecular scaffolds.

5.5 Treatment Planning

For the treatment planning, reasoning-centric models are being repurposed to navigate the high-dimensional design space of treatment planning (Rao et al., 2024; Yang et al., 2025). For example, GPT-RadPlan (Liu et al., 2025c) represents the first MLLM agent that mimics the be-

haviors of human planners in radiation oncology clinics, achieving promising results in automating the treatment planning process without the need for additional training. MedRBench(Qiu et al., 2025) offers a comprehensive benchmark that assesses LLMs on the factual accuracy, completeness, and computational efficiency of their treatment-planning rationales. Collectively, these advances position medical-reasoning LLMs as transparent, end-to-end copilots capable of accelerating drug discovery and delivering more precise, patient-specific therapies.

6 Evaluation & Benchmarking

6.1 Answer Accuracy Assessment

Traditional evaluation benchmarks (Jin et al., 2020; Pal et al., 2022; Jin et al., 2019; Hendrycks et al., 2021) use accuracy measures such as exact match or multiple-choice score on known-answer questions derived from medical exams. However, top-tier LLMs now achieve near-expert scores on several routine medical QA tests like MedQA (Jin et al., 2020) and PubMedQA (Jin et al., 2019). This success underscores the need for more difficult evaluation sets that move beyond straightforward recall of medical facts.

Recent studies have developed new benchmarks that emphasize complex, multi-step clinical reasoning and hard-to-solve questions (Gaber et al., 2025). For example, MedXpertQA (Zuo et al., 2025) incorporates specialty board review questions and performs multi-round expert reviews to build a high-quality benchmark. MedAgentsBench (Tang et al., 2025) focuses on challenging medical questions requiring multi-step clinical reasoning, diagnosis formulation, and treatment planning. By concentrating on truly difficult cases and standardizing evaluation, these benchmarks push models beyond rote knowledge retrieval, revealing performance gaps that were obscured by easier questions.

6.2 Reasoning Quality Assessment

In medical scenarios, particularly high-stakes situations, the quality of the reasoning process is just as crucial as reaching the correct conclusion (Nachane et al., 2024; Li et al., 2025b; Zhu et al., 2025). For example, MedR-Bench (Qiu et al., 2025) introduces a "Reasoning Evaluator", an automated tool that scores free-text clinical reasoning responses along multiple dimensions: efficiency, actuality, and completeness. RadRScore (Fan et al., 2025) is proposed to assess the factual correctness, completeness, and effectiveness of each step in a model's explanation, using clinically validated reasoning chains as references. HealthBench (Arora et al., 2025) is designed to capture realistic clinical reasoning—diagnostic triage, patient education, tailoring depth to user expertise, and safety-critical decision steps, while AgentClinic(Schmidgall et al., 2024) and SD-Bench (Nori et al., 2025) simulate the process of a doctor's clinical reasoning, including asking questions, arranging tests, and making the final diagnosis.

6.3 Visual Interpretability Assessment

The ability to visually interpret a model's decisions is essential for building trust, ensuring clinical acceptance, and safeguarding patient outcomes. Recent evaluation frameworks have specifically focused on testing a model's capacity to visually justify its reasoning and link its outputs to relevant image or textual data. For example, PRS-Med (Trinh et al., 2025) integrates vision-language models with segmentation capabilities to generate not only accurate segmentation masks but also corresponding spatial reasoning outputs. RJUA-MedDQA (Jin et al., 2024) evaluates whether a model can read visually complex medical documents, extract the correct evidence, and produce an answer that clearly cites or grounds its reasoning in the source data. In chest X-ray diagnosis, GEMeX (Liu et al., 2024, 2025a) and AOR (Li et al., 2025c) both focus on region-level, multi-step reasoning by evaluating the visual grounding and structured processes.

7 Discussion: Challenges & Future Directions

While progress in medical reasoning LLMs is accelerating, significant hurdles remain before they can be considered safe and effective clinical tools. Moving from promising research to widespread adoption requires confronting a series of distinct challenges in model capability, evaluation, and real-world implementation.

7.1 The Faithfulness-Plausibility Gap

A primary danger is 'plausible hallucination,' where models generate clinically plausible but factually incorrect explanations—a critical mismatch between rhetoric and ground truth (Sun et al., 2025a; Chen et al., 2024a). This is more perilous than a simple wrong answer; a model might invent

lab values that perfectly fit a diagnostic narrative, leading a clinician to a correct conclusion for the wrong reasons, thereby masking the model's flawed logic. Addressing this requires moving beyond surface-level explanations. One promising strategy is to impose external structure, for instance by constraining generation with a medical knowledge graph, which forces the reasoning path to be a sequence of verifiable '(subject, predicate, object)' triples (Wu et al., 2025). The ultimate goal, however, is to build models with intrinsic epistemic humility—the ability to express calibrated uncertainty and explicitly differentiate between evidence-backed claims and speculative inference.

7.2 Toward Native Multimodal Reasoning

Current vision-language models often use a loosely coupled architecture, fusing static image and text representations late in the process (Pan et al., 2025a; Lai et al., 2025; Zhang et al., 2025a). This fails to capture the dynamic, iterative nature of clinical reasoning. For example, a model might correctly identify "cardiomegaly" from an image and "shortness of breath" from text, but fail to infer the crucial causal link that the former is causing the latter because it cannot re-interrogate the visual data in light of the textual data. The next frontier is to build natively multimodal architectures that can interleave visual and textual tokens in a shared reasoning process. This involves developing techniques like iterative cross-modal attention, the ability to edit visual tokens during a chain-of-thought process, and image-grounded counterfactual analysis ("what if this shadow were not present?").

7.3 The Efficiency-Performance Frontier

A persistent tension exists between a model's reasoning quality and its computational footprint. The most powerful reasoning strategies, such as multi-step CoT (Wei et al., 2023; Wang et al., 2023) or multi-agent debate (Tang et al., 2024; Hong et al., 2024), demand significant computational resources and introduce latency, making them impractical for many real-time clinical settings. This has catalyzed vital research into "reasoning smarter, not harder." One direction is **lightweight post-hoc adaptation**, which is a form of results distillation; a small, efficient model is trained not on knowledge itself, but on the task of ranking the outputs of a larger, more powerful model (Shi et al., 2024). An even more profound insight comes from **eliciting emergent reasoning**, where studies show that complex reasoning can be "discovered" through goal-oriented reinforcement learning on hard problems, rather than purely "imitated" through SFT (Zhang et al., 2025a; Liu et al., 2025b). This suggests a path toward smaller, more efficient models that possess powerful reasoning capabilities.

7.4 Evaluation Beyond Task Accuracy

The field faces an evaluation crisis: as models saturate static benchmarks like MedQA (Jin et al., 2019) and MedAgentsBench (Tang et al., 2025), their scores mask real-world reasoning deficits. Progress requires a paradigm shift in evaluation. This entails moving to **dynamic, longitudinal benchmarks** that simulate a full patient journey with multimodal data (Li et al., 2025c; Liu et al., 2024). More importantly, it demands a focus beyond final-answer accuracy to a granular assessment of the **reasoning process** itself—scrutinizing its factual correctness, logical coherence, and evidence adherence, as pioneered by frameworks like ChestX-Reasoner (Fan et al., 2025). For multimodal models, this must also include quantifiable metrics for visual grounding and interpretability. Ultimately, automated scores are insufficient; the gold standard for validation must incorporate qualitative review by clinical experts to assess true clinical utility and rigorous stress-testing against rare "edge case" diseases where models are most likely to fail.

7.5 Prerequisites for Responsible Clinical Adoption

Even a technically perfect and rigorously evaluated model will fail if it cannot navigate the complex human and regulatory environment of healthcare. Responsible adoption hinges on a foundation of sociotechnical trust. This begins with the non-negotiable requirement of **patient privacy**; under regulations like HIPAA and GDPR, privacy-preserving techniques like Federated Learning (FL) are a critical architectural prerequisite (Jahan et al., 2025; Abbas et al., 2024; Li et al., 2025a). Beyond data handling, responsible deployment demands addressing the model's potential for **algorithmic bias**. This includes mitigating both demographic biases, which can worsen health inequities (Sandi, 2025), and cognitive biases, which replicate known patterns of human diagnostic error (Kim et al., 2025). Ultimately, both privacy and fairness are components of the largest challenge: establishing clear **accountability and trust**. Closing the "accountabil-

ity gap" (Habli et al., 2020) requires a robust framework built on shared responsibility policies for developers and institutions (Information Technology Industry Council, 2024), inherently auditable and explainable AI systems (Hong et al., 2024), and effective Human-in-the-Loop (HITL) workflows that empower clinicians as informed arbiters, not passive users of a black box.

8 Conclusion

This systematic review analyzes the crucial evolution of Large Language Models toward complex medical reasoning, presenting a core taxonomy of the training-time and test-time techniques enabling this capability. We survey the application of these reasoning techniques in LLMs across diverse medical modalities and clinical domains, and track the parallel shift in evaluation from measuring accuracy to validating the reasoning process itself. We identify formidable remaining challenges in model faithfulness, multimodal integration, efficiency, and responsible sociotechnical adoption. Overcoming these hurdles is the critical path to realizing the promise of Medical LLMs as trustworthy and interpretable reasoning partners in healthcare.

Limitations

While this paper provides a comprehensive systematic review, several limitations should be acknowledged. Our analysis is inherently constrained by the scope of publicly available literature, excluding proprietary models from industrial labs and potentially missing very recent pre-prints due to the field's rapid evolution. Our keyword-based search, while systematic, might also have inadvertently omitted papers using alternative terminologies. Furthermore, the taxonomy we propose is an interpretative lens; other valid frameworks could exist, and in prioritizing breadth to provide a panoramic overview, we could not delve into the deepest technical nuances of every method. Ultimately, this review serves as a snapshot of a rapidly moving target, and future breakthroughs may necessitate revisions to our framework.

References

Haider Abbas, Omar M. El-Gayar, and Areej Al-Malaise Al-Ghamdi. 2024. A comprehensive survey on federated learning in the healthcare area: Concept and applications. *Computer Modeling in Engineering & Sciences*, 140(3):2239–2273.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health. *Preprint*, arXiv:2505.08775.

Emilia Brügge, Sarah Ricchizzi, Malin Arenbeck, Marius Niklas Keller, Lina Schur, Walter Stummer, Markus Holling, Max Hao Lu, and Dogus Darici. 2024. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC medical education*, 24(1):1391.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024a. Huatuogpt-o1, towards medical complex reasoning with llms. *Preprint*, arXiv:2412.18925.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024b. CoD, towards an interpretable medical agent using chain of diagnosis. *Preprint*, arXiv:2407.13301.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. 2025. MedRAX: Medical reasoning agent for chest x-ray. *Preprint*, arXiv:2502.02673.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *Preprint*, arXiv:2402.09742.

Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. *Preprint*, arXiv:2504.20930.

Farieda Gaber, Maqsood Shaik, Fabio Allega, Agnes Julia Bilecz, Felix Busch, Kelsey Goon, Vedran Franke, and Altuna Akalin. 2025. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8(1):263.

Zishan Gu, Fenglin Liu, Changchang Yin, and Ping Zhang. 2024. Inquire, interact, and integrate: A proactive agent collaborative framework for zero-shot multimodal medical reasoning. *Preprint*, arXiv:2405.11640.

Ibrahim Habli, Tom Lawton, and Zoe Porter. 2020. Artificial intelligence in health care: accountability and safety. *Bulletin of the World Health Organization*, 98(4):251.

Z. Hammane. 2024. SelfRewardRAG: Enhancing medical reasoning with retrieval-augmented generation and self-evaluation. *IEEE Access*.

Yuhong He, Yongqi Zhang, Shizhu He, and Jun Wan. 2024. BP4ER: Bootstrap prompting for explicit reasoning in medical dialogue generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Shengxin Hong, Liang Xiao, Xin Zhang, and Jianxia Chen. 2024. Argmed-agents: explainable clinical decision reasoning with llm disscusion via argumentation schemes. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5486–5493. IEEE.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025a. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *Preprint*, arXiv:2504.00869.

Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025b. O1 replication journey – part 3: Inference-time scaling for medical reasoning. *Preprint*, arXiv:2501.06458.

Information Technology Industry Council. 2024. ITI's AI Accountability Framework. https://www.itic.org/documents/artificial-intelligence/AIFIAIAccountabilityFrameworkFinal.pdf. Accessed: 2024-07-22.

Nusrat Jahan, Ratun Rahman, and Michel Wang. 2025. Federated learning: A survey on privacy-preserving collaborative intelligence. *Preprint*, arXiv:2504.17703.

Garam Jeong, Junseong Kim, Dongmin Bib, and Jaesik Choi. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented LLMs. In *Proceedings of the 32nd International Conference on Intelligent Systems for Molecular Biology (ISMB)*.

Congyun Jin, Ming Zhang, Xiaowei Ma, Li Yujiao, Yingbo Wang, Yabo Jia, Yuliang Du, Tao Sun, Haowen Wang, Cong Fan, Jinjie Gu, Chenfei Chi, Xiangguo Lv, Fangzhou Li, Wei Xue, and Yiran Huang. 2024. Rjua-meddqa: A multimodal benchmark for medical document question answering and clinical reasoning. *Preprint*, arXiv:2402.14840.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. 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)*, pages 2567–2577.

Su Hwan Kim, Sebastian Ziegelmayer, Felix Busch, Christian J. Mertens, Matthias Keicher, Lisa C. Adams, Keno K. Bressem, Rickmer Braren, Marcus R. Makowski, Jan S. Kirschke, Dennis M. Hedderich, and Benedikt Wiestler. 2025. LLM reasoning does not protect against clinical cognitive biases - an evaluation using BiasMedQA. *Preprint*, arXiv:2506.22405.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Yongsik Sim, Beomseok Sohn, Dongha Lee, and Jinyoung Yeo. 2023. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. *Preprint*, arXiv:2312.07401.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *Preprint*, arXiv:2503.13939.

Jiaqi Li, Zhen Yan, Xiaohui Liu, Xiao-Li Li, and Yang Liu. 2025a. Privacy-preserving federated learning framework for multi-source electronic health records prognosis prediction.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *Preprint*, arXiv:2405.02957.

Lei Li, Xiao Zhou, and Zheng Liu. 2025b. R2med: A benchmark for reasoning-driven medical retrieval. *Preprint*, arXiv:2505.14558.

Qingqiu Li, Zihang Cui, Seongsu Bae, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, Junjun He, and Shujun Wang. 2025c. Aor: Anatomical ontology-guided reasoning for medical large multimodal model in chest x-ray interpretation. *Preprint*, arXiv:2505.02830.

Bo Liu, Xiangyu Zhao, Along He, Yidi Chen, Huazhu Fu, and Xiao-Ming Wu. 2025a. Gemex-thinkvg: Towards thinking with visual grounding in medical vqa via reinforcement learning. *Preprint*, arXiv:2506.17939.

Bo Liu, Ke Zou, Liming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao,

Xiao-Ming Wu, and Huazhu Fu. 2024. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. *Preprint*, arXiv:2411.16778.

Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. 2025b. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. *arXiv preprint arXiv:2505.17952*.

Sheng Liu, Oscar Pastor-Serrano, Yizheng Chen, Matthew Gopaulchan, Weixing Liang, Mark Buyyounouski, Erqi Pollom, Quynh-Thu Le, Michael Gensheimer, Peng Dong, Yong Yang, James Zou, and Lei Xing. 2025c. Automated radiotherapy treatment planning guided by gpt-4vision. *Preprint*, arXiv:2406.15609.

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. 2025d. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *Preprint*, arXiv:2411.15692.

Mary M Lucas, Justin Yang, Jon K Pomeroy, and Christopher C Yang. 2024. Reasoning with large language models for medical question answering. *Journal of the American Medical Informatics Association*, 31(9):1964–1975.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.

Linjie Mu, Zhongzhen Huang, Yakun Zhu, Xiangyu Zhao, Shaoting Zhang, and Xiaofan Zhang. 2025. Elicit and enhance: Advancing multimodal reasoning in medical scenarios. *arXiv preprint arXiv:2505.23118*.

Saeel Sandeep Nachane, Ojas Gramopadhye, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. 2024. Few shot chain-of-thought driven reasoning to prompt LLMs for open-ended medical question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Shuai Niu, Jing Ma, Hongzhan Lin, Liang Bai, Zhihua Wang, Yida Xu, Yunya Song, and Xian Yang. 2025. Knowledge-augmented multimodal clinical rationale generation for disease diagnosis with small language models. *Preprint*, arXiv:2411.07611.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. 2025. Sequential diagnosis with language models. *Preprint*, arXiv:2506.22405.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, et al. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025a. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Qianjun Pan, Wenkai Ji, Yuyang Ding, Junsong Li, Shilian Chen, Junyi Wang, Jie Zhou, Qin Chen, Min Zhang, Yulan Wu, et al. 2025b. A survey of slow thinking-based reasoning llms using reinforced learning and inference-time scaling law. *arXiv preprint arXiv:2505.02665*.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Quantifying the reasoning abilities of llms on real-world clinical cases. *Preprint*, arXiv:2503.04691.

Arya Rao, John Kim, Winston Lie, Michael Pang, Lanting Fuh, Keith J Dreyer, and Marc D Succi. 2024. Proactive polypharmacy management using large language models: opportunities to enhance geriatric care. *Journal of medical systems*, 48(1):41.

Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. 2020. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):3923.

Fran Sandi. 2025. CareLens: Investigating LLM bias in healthcare. `https://www.fransandi.com/blog/llm-bias-in-healthcare`. Accessed: 2024-07-22.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20.

Adam C Schaffer, Anupam B Jena, Seth A Seabury, Harnam Singh, Venkat Chalasani, and Allen Kachalia. 2017. Rates and characteristics of paid malpractice claims among us physicians by specialty, 1992-2014. *JAMA internal medicine*, 177(5):710–718.

Verity Schaye, David DiTullio, Benedict Vincent Guzman, Scott Vennemeyer, Hanniel Shih, Ilan Reinstein, Danielle E Weber, Abbie Goodman, Danny TY Wu, Daniel J Sartori, et al. 2025. Large language model–based assessment of clinical reasoning documentation in the electronic health record across two institutions: Development and validation study. *Journal of Medical Internet Research*, 27:e67967.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *Preprint*, arXiv:2405.07960.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, and May D Wang. 2024. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2024, page 22294.

Karan Singhal, Tao Tu, and et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950.

Yuki Sonoda, Ryo Kurokawa, Akifumi Hagiwara, Yusuke Asari, Takahiro Fukushima, Jun Kanzawa, Wataru Gonoi, and Osamu Abe. 2025. Structured clinical reasoning prompt enhances llm's diagnostic capabilities in diagnosis please quiz cases. *Japanese Journal of Radiology*, 43(4):586–592.

Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibo Ju, Jin Ye, Pengcheng Chen, Ming Hu, Shixiang Tang, Lihao Liu, Bin Fu, Wenqi Shao, Xiaowei Hu, Xiangwen Liao, Yuanfeng Ji, and Junjun He. 2025. Gmai-vl-r1: Harnessing reinforcement learning for multimodal medical reasoning. *Preprint*, arXiv:2504.01886.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Yu Rong, Wenbing Huang, Qifeng Bai, and Tingyang Xu. 2025a. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. *Preprint*, arXiv:2506.09513.

Zheng Sun, Yi Wei, and Long Yu. 2025b. Image aesthetic reasoning: A new benchmark for medical image screening with mllms. *Preprint*, arXiv:2505.23265.

Xiangru Tang, Bill Qian, Rick Gao, Jiakang Chen, Xinyun Chen, and Mark Gerstein. 2023. Biocoder: A benchmark for bioinformatics code generation with contextual pragmatic knowledge.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, Arman Cohan, and Mark Gerstein. 2025. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *Preprint*, arXiv:2503.07459.

Xiangru Tang, Anni Zou, Zhuoheng Li, Yilun Hao, Yiqi Wang, Yiming Wang, Boyang Liu, Chaoyi Wu, Zhaofeng He, and S. Kevin Zhou. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, Shriya Reddy, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. 2025. Disentangling reasoning and knowledge in medical large language models. *Preprint*, arXiv:2505.11462.

Quoc-Huy Trinh, Minh-Van Nguyen, Jung Peng, Ulas Bagci, and Debesh Jha. 2025. Prs-med: Position reasoning segmentation with vision-language model in medical imaging. *Preprint*, arXiv:2505.11872.

Juraj Vladika, Florian Barth, and Udo Kruschwitz. 2025. Step-by-step fact verification system for medical claims with explainable reasoning. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Zifeng Wang, Benjamin Danek, Ziwei Yang, Zheng Chen, and Jimeng Sun. 2024. Can large language models replace data scientists in biomedical research? *arXiv preprint arXiv:2410.21591*.

Hao Wei, Jianing Qiu, Haibao Yu, and Wu Yuan. 2024a. Medco: Medical education copilots based on a multi-agent framework. *Preprint*, arXiv:2408.12496.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Lai Wei, Wenkai Wang, Xiaoyu Shen, Yu Xie, Zhihao Fan, Xiaojin Zhang, Zhongyu Wei, and Wei Chen. 2024b. Mc-cot: A modular collaborative cot framework for zero-shot medical-vqa with llm and mllm integration. *arXiv preprint arXiv:2410.04521*.

C. Wu et al. 2023. Pmc-llama: An open-source language model for medical applications. *Preprint*, arXiv:2304.14454.

Cheng-Kuang Wu, Wei-Lin Chen, and Hsin-Hsi Chen. 2024a. Large language models perform diagnostic reasoning. In *The Twelfth International Conference on Learning Representations*.

Chiyuan Wu and colleagues. 2025. Generative ai for medical education: Insights from a case study with medical students and an ai tutor for clinical reasoning. https://research.google/pubs/generative-ai-for-medical-education-insights-from-a-case-study-with-medical-students-and-an-ai-tutor-for-clinical-reasoning/.

Jiageng Wu, Zixuan Liu, and Zhiyong Lu. 2024b. Guiding clinical reasoning with large language models via knowledge seeds. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *Preprint*, arXiv:2504.00993.

Huihui Xu, Yuanpeng Nie, Hualiang Wang, Ying Chen, Wei Li, Junzhi Ning, Lihao Liu, Hongqiu Wang, Lei Zhu, Jiyao Liu, Xiaomeng Li, and Junjun He. 2025a. Medground-r1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization. *Preprint*, arXiv:2507.02994.

Kaishuai Xu, Yi Cheng, Wenjun Hou, Qiaoyu Tan, and Wenjie Li. 2024. Reasoning like a doctor: Improving medical dialogue systems via diagnostic reasoning process alignment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6796–6814, Bangkok, Thailand. Association for Computational Linguistics.

Ran Xu et al. 2025b. MedAgentGym: Training LLM agents for code-based medical reasoning at scale. *Preprint*, arXiv:2506.04405.

Chonghua Xue, Sahana S Kowshik, Diala Lteif, Shreyas Puducheri, Varuna H Jasodanand, Olivia T Zhou, Anika S Walia, Osman B Guney, J Diana Zhang, Serena Poésy, et al. 2024. Ai-based differential diagnosis of dementia etiologies on multimodal data. *Nature Medicine*, 30(10):2977–2989.

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan Zhao, Yixin Zhang, Renjun Zhang, Li Zhu, and Xuandong Zhao. 2024. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. *Preprint*, arXiv:2406.13890.

Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, and Jieneng Chen. 2025. Medical world model: Generative simulation of tumor evolution for treatment planning. *Preprint*, arXiv:2506.02327.

Hongzhou Yu, Tianhao Cheng, Ying Cheng, and Rui Feng. 2025. Finemedlm-o1: Enhancing the medical reasoning ability of llm from supervised fine-tuning to test-time training. *Preprint*, arXiv:2501.09213.

Huizi Yu, Jiayan Zhou, Lingyao Li, et al. 2024. Aipatient: Simulating patients with ehrs and llm powered agentic workflow. *Preprint*, arXiv:2409.18924.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning. *Preprint*, arXiv:2404.14777.

Aizan Zafar, Kshitij Mishra, and Asif Ekbal. 2025. Medex: Enhancing medical question-answering with first-order logic based reasoning and knowledge injection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9701–9720.

Yin-Ying Zhan, Yi-Fan Li, and Wei-Wei Wang. 2024. RARoK: Retrieval-augmented reasoning on knowledge for medical question answering. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. 2025a. Med-rlvr: Emerging medical reasoning from a 3b base model via reinforcement learning. *Preprint*, arXiv:2502.19655.

Yuting Zhang, Kaishen Yuan, Hao Lu, Yutao Yue, Jintai Chen, and Kaishun Wu. 2025b. Medtvt-r1: A multimodal llm empowering medical reasoning and diagnosis. *arXiv preprint arXiv:2506.18512*.

Junjie Zhou, Yanyun Qu, Yamei Chen, Zhaoyang Wang, Jing-Doo Wang, Can-Jie Cao, and Yuan-Chih Tsai. 2025. PathCoT: Chain-of-thought prompting for zero-shot pathology visual reasoning. *Preprint*, arXiv:2507.01029.

Jiayuan Zhu and Junde Wu. 2025. Ask patients with patience: Enabling LLMs for human-centric medical dialogue with grounded reasoning. *Preprint*, arXiv:2502.07143.

Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. Diagnosisarena: Benchmarking diagnostic reasoning for large language models. *Preprint*, arXiv:2505.14107.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *Preprint*, arXiv:2501.18362.