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

Anonymous ACL submission

Abstract

Large Language Models (LLMs) show substantial medical potential, yet their capacity for systematic, verifiable reasoning remains limited. This limitation drives a shift from single-step answer generation toward architectures explicitly engineered for clinical deduction. We propose a taxonomy of reasoning enhancement methodologies, categorizing training-time strategies (supervised fine-tuning, reinforcement learning) and test-time mechanisms (structured prompting, multi-agent orchestration). We analyze these methods across text, image, and code modalities within core applications like diagnostic formulation and treatment planning. Furthermore, we trace the evaluation paradigm shift from static accuracy toward dynamic reasoning quality and multimodal interpretability assessments. Synthesizing foundational studies from 2022 to 2026, this review critically examines persistent barriers, including the faithfulness-plausibility gap, epistemic tension between parametric knowledge and dynamic evidence, the need for natively multimodal reasoning, and efficiency-performance trade-offs. Ultimately, we establish actionable trajectories for constructing robust, interpretable, and responsible medical AI.

1 Introduction

Specialized foundation models demonstrate utility in processing biomedical knowledge (Singhal et al., 2025; Wu et al., 2023; Luo et al., 2022; Zhou et al., 2023), yet struggle executing the multi-step inference required for high-stakes clinical decisions (Sim and Chen, 2025; Kim et al., 2025a). Medical diagnosis requires deep causal inference, integrating fragmented symptoms, history, and multimodal test outcomes into a cohesive picture (Richens et al., 2020; Xue et al., 2024; Mattingly, 1991). Diagnostic inaccuracies remain a major source of malpractice claims, underscoring risks associated with inadequate deduction (Schaffer et al., 2017).

Driven by structured elicitation techniques extracting intermediate inferential stages (Wei et al., 2023; Nachane et al., 2024), researchers formulate architectures simulating diagnostic workflows, substantiating conclusions, and adapting to intricate scenarios (Savage et al., 2024; Kim et al., 2024b). These systems function as advanced instruments for decision support and medical education, as AI-driven feedback elevates diagnostic proficiencies (Brügge et al., 2024; Schaye et al., 2025). Progress emphasizes embedding verifiable logic through reinforcement learning and knowledge graph constraints, transitioning models from passive repositories to active deductive agents (Pan et al., 2025b; Wu et al., 2025).

We queried PubMed, Scopus, and arXiv for manuscripts published between 2022 and 2026. Utilizing keywords encompassing clinical reasoning, multimodal agents, and deductive enhancement, our search produced over two hundred articles. After rigorous screening for explicit deductive mechanisms and clinical alignment, we established a core collection of foundational investigations. This review builds upon early capability surveys (Zhou et al., 2023; Shool et al., 2025) by focusing explicitly on the technical mechanics of reasoning enhancement, alongside associated epistemological challenges (Kim and Thorne, 2024a; Soffer et al., 2025).

We systematically evaluate these advancements. We first analyze reasoning technique adaptation across text, image, and code modalities (Section 2). We then propose a dual taxonomy categorizing enhancements into training-time and test-time interventions (Section 3), complemented by a clinical cognitive stage mapping in Appendix C. Following an exploration of clinical applications (Section 4), we trace the evaluation paradigm shift from static accuracy toward process-oriented reasoning assessment (Section 5). The review concludes by dissecting remaining translational barriers, specif-

ically addressing epistemic tension between consensus and novelty, native multimodal architecture demands, the efficiency-performance frontier, and prerequisites for responsible sociotechnical integration (Section 6).

2 Medical Reasoning Under Various Modalities

The specific data modality (text, image, or code) dictates the medical reasoning challenge, necessitating tailored deductive strategies. This section analyzes how reasoning techniques adapt across these three modalities.

2.1 Reasoning over Text

Clinical texts lack formalized logic, challenging correct inferential generation. Training renders reasoning explicit, producing rationales aligned with clinical patterns (Xu et al., 2024) or knowledge graphs (Wu et al., 2025). Pathway validity is enforced via first-order logic (Zafar et al., 2025) or reinforcement learning, rewarding correctness and penalizing hallucinations (Zhang et al., 2025a; Liu et al., 2025b). Advanced research deepens inference via test-time scaling (Huang et al., 2025a; Yu et al., 2025; Shi et al., 2024a) and broadens it using multi-agent debate for consensus (Hong et al., 2024; Tang et al., 2024; Zhu and Wu, 2025).

2.2 Reasoning over Image

Medical imaging requires bridging pixel data and clinical abstractions. Deductive processes rely on visual grounding to maintain trustworthiness, driving tighter vision-language integration. Early cross-modal alignment utilizes reinforcement learning, teaching Vision-Language Models (VLMs) to associate visual findings with diagnostic labels (Pan et al., 2025a; Lai et al., 2025; Su et al., 2025; Zhang et al., 2025b). Sophisticated strategies coordinate deduction across modalities. Frameworks like MC-CoT (Wei et al., 2024b) and Elicit and Enhance (Mu et al., 2025) utilize an orchestrator-perceiver dynamic, where language models generate plans dictating VLM visual analysis. Advanced methodologies pursue deep spatial grounding by integrating fine-grained anatomical knowledge, using advanced segmentation for precise localization (Trinh et al., 2025) or anatomical ontologies to formalize visual interpretations (Li et al., 2025b).

2.3 Reasoning over Code

Code-based deduction supports procedural, verifiable workflows, relying on an ecosystem of data, environments, and platforms. The data foundation, validated by BioCoder (Tang et al., 2023), confirms sufficient public biomedical code exists for training. Operational environments like MedAgentGym (Xu et al., 2025b) provide standardized sandboxes for rigorous agent training. Integrated platforms, such as BioDSBench (Wang et al., 2024b), transition validated agents toward practical application by embedding them within data science pipelines for medical professionals. This ecosystem establishes an operational lifecycle, separating controlled agent development from end-user clinical interfaces.

3 A Taxonomy of Medical Reasoning Enhancement Techniques

To equip LLMs with robust medical inference capabilities, investigators have formulated methodologies that separate into two primary phases. Training-time techniques modify a model's internal weights to construct foundational deductive abilities. Test-time techniques direct and optimize the model's output during inference without altering the core architecture. This section systematically examines these approaches (Figure 1).

3.1 Training-time Techniques: Building the Foundation

Training-time methodologies are high-cost, high-impact interventions embedding clinical logic directly into model parameters, though they frequently face data scalability limits.

3.1.1 Supervised Fine-tuning (SFT)

Supervised fine-tuning (SFT) shifts models from surface correlations to procedural knowledge. Multi-stage fine-tuning applies curriculum learning to decompose reasoning, improving reliability. Task abstraction staging builds hierarchies from concrete facts to abstract inference, as FineMedLM-o1 (Yu et al., 2025) demonstrates by progressing from factual data to complex causal reasoning. Modality integration staging supports multimodal tasks, advancing from perception to interpretation. AOR (Li et al., 2025b) exemplifies this by sequentially identifying anatomy, mapping terminology, and formulating diagnoses.

Chain-aware fine-tuning faces a supervision bottleneck due to high acquisition costs. While human annotation provides authoritative rationales

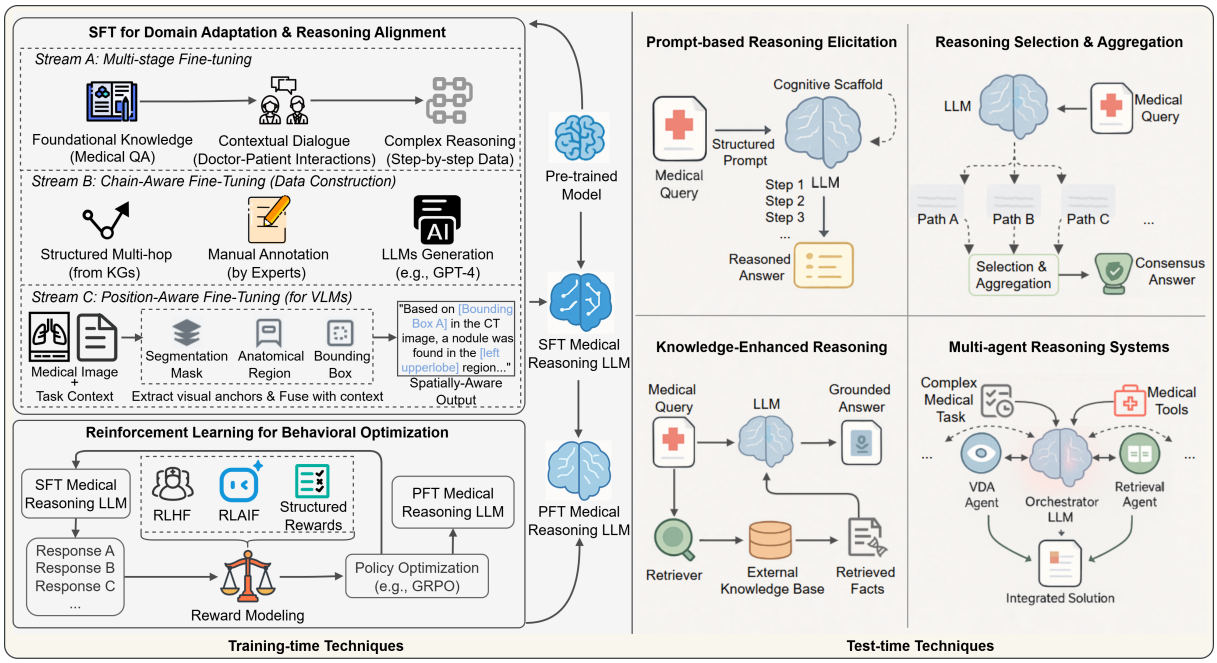

Figure 1: An overarching framework of techniques to enhance medical reasoning in LLMs. Training-time Techniques instill models with deductive capabilities by adjusting internal weights via methods such as Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL). Test-time Techniques enhance deduction at inference, encompassing Prompt-based Elicitation, Reasoning Selection and Aggregation, Knowledge-Enhanced Reasoning, and Multi-agent Systems.

(Singhal et al., 2025), AI-generated sequences offer scalability. Frameworks like HuatuoGPT-o1 (Chen et al., 2024a) employ teacher models via generation and self-correction cycles, risking inherited biases. Verifiable methods impose structural constraints, with MedReason (Wu et al., 2025) restricting generation to knowledge graph traversals for auditability. Other approaches refine datasets using data-centric filtering (Thapa et al., 2025) or stylistic alignment (Xu et al., 2024).

Position-Aware Fine-tuning (PFT) connects diagnostic assertions to visual regions, addressing grounding via strict spatial correspondence. Coarse-grained anchoring uses bounding boxes as intermediate stages (Xu et al., 2025a). Fine-grained anchoring employs pixel-level masks for precision. PRS-Med (Trinh et al., 2025) trains on exact masks for localized responses. Semantically rich anchoring links visual regions to medical vocabularies. AOR (Li et al., 2025b) correlates image sections with ontology concepts to deduce location and pathology.

SFT methodologies exhibit high sensitivity to annotation quality, making manual labeling prohibitively expensive. Research frequently utilizes Chain-of-Thought (CoT) supervision from general-purpose architectures (Thapa et al., 2025; Yu et al., 2025). These sequences can appear formally coherent while remaining logically superficial, a problematic limitation in medicine. Developing high-quality CoT construction and filtering pipelines remains an active investigational focus.

3.1.2 Reinforcement Learning (RL)

RL aligns SFT capabilities with clinical safety and accuracy. Med-PaLM 2 (Singhal et al., 2025) utilized physician rankings to internalize diagnostic prudence via Human Feedback. Addressing scalability limits, RL with AI Feedback offers a pragmatic alternative, with HuatuoGPT-o1 (Chen et al., 2024a) employing GPT-4o for scalable reward indicators. Objective optimization uses Structured Rewards for reproducibility via the GRPO algorithm:

$$R_{total} = \alpha \cdot R_{accuracy} + \beta \cdot R_{logic} + \gamma \cdot R_{grounding} \quad (1)$$

MedGround-R1 (Xu et al., 2025a) utilizes accuracy rewards for correct diagnoses and grounding rewards via IoU scores measuring visual attention overlap with anatomical regions. These parameters incorporate rationale quality (Niu et al., 2025), spatial consistency (Xu et al., 2025a), and task-specific performance (Lai et al., 2025; Su et al., 2025; Liu et al., 2025c).

Reinforcement learning on challenging problems allows complex reasoning patterns to develop implicitly. AlphaMed (Liu et al., 2025b) and BioMed-R1 (Thapa et al., 2025) illustrate that applying objective rewards like multiple-choice accuracy enables sophisticated inferential capabilities to materialize without explicit CoT distillation. RL remains susceptible to reward architecture variations, necessitating intermediate rewards and auxiliary verifiable objectives to enhance reliability.

3.2 Test-time Techniques: Achieving Agility and Verifiability

Distinct from expensive retraining, test-time methodologies provide an adaptable, economical method to direct a pre-trained model's deductive process during inference. These strategies advance from fundamental input modification toward intricate, multi-agent system orchestration.

3.2.1 Prompt-based Reasoning Elicitation

Structured inputs externalize latent reasoning, progressing from generalized Chain-of-Thought (Nachane et al., 2024) to domain-specific clinical variations. Specialized versions include Clinical CoT (Kwon et al., 2023), Diagnostic Reasoning CoT (Wu et al., 2024a), and Chain of Diagnosis (Chen et al., 2024b). For complex scenarios, least-to-most prompting separates tasks into subproblems (He et al., 2024). Alternative methodologies apply iterative inquiry validating assertions (Vladika et al., 2025), manage cross-modal analysis (Wei et al., 2024b; Zhou et al., 2025a), or utilize bidirectional inference for transparent differential diagnoses (Zhou et al., 2025b). Prompt-focused approaches encounter consistency complications and logical hallucinations, while increasing input complexity restricts context windows. Empirically, domain-specific structuring, like mandating explicit symptom enumeration (Kwon et al., 2023; Chen et al., 2024b), produces superior diagnostic accuracy over generalized triggers. Recent structured approaches enforce multi-step workflows from history taking to differential diagnosis, enhancing explainability and generalization (Ayoub et al., 2026). Similarly, evidence-driven prompting integrates retrieval with multidimensional evidence evaluation, constraining reasoning strictly to authoritative sources and suppressing hallucinations (Xu, 2026).

3.2.2 Reasoning Selection and Aggregation

Generating and evaluating multiple deductive pathways mitigates stochastic variability in LLM outputs. 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, relying on a simple majority vote aggregation function to select the most frequent answer. Relying purely on brute-force sampling for majority consensus yields diminishing returns and is computationally prohibitive for routine clinical workflows. To mitigate this validity-cost trade-off, test-time adaptation offers a more efficient aggregation strategy. Rather than relying on exhaustive sampling from a massive base model, frameworks like MedAdapter (Shi et al., 2024a) use a small, lightweight model as a post-hoc ranker. This lightweight adapter is explicitly trained to score and select the most clinically plausible solution, acting as an efficient alternative to massive-scale majority voting.

3.2.3 Knowledge-Enhanced Reasoning

Unstructured Knowledge Contextualization Standard RAG methodologies prove inadequate for complex medical problem-solving. Iterative retrieval systems dynamically acquire data during logical phases (Yu et al., 2023), mimicking clinical workflows where models retrieve targeted evidence to confirm generated hypotheses. Advanced frameworks like SelfRewardRAG (Hammane, 2024) incorporate internal correction cycles, evaluating acquired medical literature relevance to refine subsequent queries, minimizing extraneous data causing hallucinations in conventional RAG. Recent systems also introduce privacy-aware multimodal retrieval frameworks, allowing users to toggle personal health record inclusion during reasoning, directly addressing compliance and explainability (Anonymous et al., 2026).

Structured Knowledge Guidance Structured Knowledge Graphs (KGs) impose logical frameworks to mitigate physiologically improbable hallucinations. Techniques like In-Context Padding (Wu et al., 2024b) embed foundational knowledge elements, specifically structured triplets extracted from formal ontologies like UMLS, directly into the active processing window. This methodology restricts generation strictly to validated biomedical pathways. Rigorous approaches, such as MedReason (Wu et al., 2025), deploy explicit trajectory

validation, ensuring each generated deductive stage undergoes real-time verification against the established knowledge graph. Hybrid configurations (Cabello et al., 2024) alongside MedEx (Zafar et al., 2025) incorporate First-Order Logic (FOL) or established medical consensus regulations (Staniek et al., 2025) to formally validate consistency between retrieved guidelines and the model's diagnostic assertions, promoting elevated reliability for clinical decision support systems.

These strategies separate into two complementary classifications. Retrieval-Augmented Generation (RAG) provides context using unstructured text sourced from medical corpora, functioning as accessible repositories for timely data (Jeong et al., 2024; Zhan et al., 2024), while advanced frameworks apply RAG facilitating step-by-step verification (Yun et al., 2025). The subsequent category employs structured knowledge elements (KGs) supplying logical frameworks directing generation along causally sound trajectories (Wu et al., 2024b, 2025). Integrating a broad factual corpus provided by RAG with a definitive logical framework supplied by KGs establishes a primary frontier in developing trustworthy systems. A detailed breakdown of these methodologies is available in Appendix D. Methodologically, these architectures experience noise sensitivity from irrelevant data retrieval (Jeong et al., 2024) and scalability complications maintaining KGs. Empirical evidence indicates purely retrieval-based procedures frequently lack necessary logical rigor. Optimal practices heavily favor hybrid architectures, demonstrated by MedReason (Wu et al., 2025), harnessing RAG for informational breadth and utilizing KGs ensuring structural verification. Refer to Appendix D.2 for an analysis detailing conflict resolution mechanisms.

3.2.4 Multi-agent Reasoning Systems

Clinical Reasoning-Centric Agents The primary objective is enhancing the deductive process to ensure accuracy, robustness, and interpretability. ArgMed-Agents (Hong et al., 2024) maps hypotheses, counter-arguments, priors, and posteriors, applying argumentation to optimize diagnoses across conversational iterations. AMANDA (Wang et al., 2025) and MultiMedRes (Gu et al., 2024) similarly occupy this space. Their Explorer and Learner components iteratively optimize questions and assimilate external knowledge to direct Med-VQA operations toward logically consistent determinations.

Clinical Workflow-Centric Agents These specialized systems manage complex operational workflows integrating imaging models, extensive databases, and predictive instruments. MedRAX (Fallahpour et al., 2025) and Medical AI Consensus (Elboardy et al., 2025) represent archetypal workflow-focused systems applied within radiology. MedChain (Liu et al., 2026) models clinical reasoning as an interactive sequential task, providing a comprehensive framework to benchmark agent performance within these workflows. They direct individual cases through modality detection, segmentation, classification, and report composition, attempting to support correct sequential submodule activation. ClinicalAgent (Yue et al., 2024) executes a comparable function regarding clinical trial design by directing queries through dedicated enrollment, safety, and efficacy components. MedAgentGym (Xu et al., 2025b) operates at a foundational level, functioning as a comprehensive training and evaluation environment for these workflow agents by exposing them to an expansive spectrum of code-based medical deduction tasks.

Clinical Evaluation-Centric Agents A third classification prioritizes assessment procedures rather than direct content generation. The evaluation components integrated within Medical AI Consensus and MedAgentGym operate as automated judging elements assessing reports, code solutions, or specific diagnostic decisions based on correctness, completeness, and clinical relevance. Within AMANDA and MultiMedRes, dedicated Evaluator components estimate overall answer quality and determine when sufficient information has been accumulated, functioning as an acquired termination and quality control protocol.

Multi-agent frameworks introduce complexity and stability risks, including the potential for unconstructive operational loops. Extensively validated best practices suggest that unconstrained collaborative efforts prove less effective than strict functional decomposition accompanied by rigid role assignments. Frameworks similar to MedAgents (Tang et al., 2024) and ArgMed-Agents (Hong et al., 2024) indicate that allocating distinct proposer and critic responsibilities helps the deductive process reliably converge toward a verifiable conclusion.

4 Applications in Clinical Workflow

4.1 History Taking and Patient Representation

Medical reasoning models process unstructured Electronic Health Records (EHRs) creating comprehensive patient representations. EHRAgent (Shi et al., 2024b) models tabular reasoning as tool-use planning decomposing intricate tasks. MedCopilot (Xu et al., 2025b) achieves accurate extraction through code-based retrieval logic. AIPatient (Yu et al., 2024) integrates EHRs with knowledge graphs, simulating clinical scenarios for medical education.

4.2 Diagnostic Reasoning and Differential Diagnosis

Architectures like HuatuoGPT-o1 (Chen et al., 2024a), MedReason (Wu et al., 2025), ReasonMed (Sun et al., 2025a), and Dual-Inf (Zhou et al., 2025b) generate Chain-of-Thought drafts with verifiable mechanisms, strengthening diagnostic capabilities (Wang, 2025). FineMedLM-o1's ranking strategy reduces misdiagnosis rates (Yu et al., 2025). PRS-Med (Trinh et al., 2025) improves pathological reasoning across imaging modalities, while studies (Su et al., 2025; Pan et al., 2025a; Lai et al., 2025; Zhang et al., 2025a) utilize RL methods like GRPO (Shao et al., 2024) boosting reasoning quality. Concurrently, MedGround-R1 (Xu et al., 2025a) integrates radiology annotations into language-vision alignment.

4.3 Personalized Treatment Planning

Reasoning models navigate the high-dimensional design space of medical interventions (Rao et al., 2024; Yang et al., 2025). GPT-RadPlan (Liu et al., 2025d) mimics human planners in radiation oncology clinics, automating treatment planning without additional training. MedRBench (Qiu et al., 2025) assesses LLMs on factual accuracy, completeness, and efficiency of treatment planning rationales. These advances position reasoning LLMs as transparent copilots delivering precise therapies.

4.4 Prognosis Prediction and Risk Stratification

Pinpointing pathological cues provides foundations for accurate prognosis prediction and risk stratification (Han et al., 2023; Kim et al., 2024a). CardioCoT (Rui et al., 2025) proposes a hierarchical reasoning-enhanced survival analysis framework, improving interpretability and predictive performance for major adverse cardiovascular events (MACE) recurrence in acute myocardial infarction patients. Linking observations to concrete next-step decisions increases risk assessment reliability.

4.5 Accelerating Drug Discovery and Development

These models show potential as integrated engines designing novel therapeutics. DrugAgent (Liu et al., 2025f) frames drug target interaction prediction as a sequential reasoning problem, reporting considerable ROC-AUC gains. Systems like ChemCrow (M. Bran et al., 2024) and AgentMol (Karabowicz et al., 2025) function as autonomous agents integrating chemistry tools planning syntheses and proposing molecular scaffolds, accelerating the drug discovery pipeline.

5 Evaluation & Benchmarking

5.1 Answer Accuracy Assessment

Early benchmarks (Jin et al., 2020; Pal et al., 2022; Jin et al., 2019; Hendrycks et al., 2021) used multiple-choice accuracy on medical exams. With top-tier LLMs attaining near-expert scores on routine QA tests (Jin et al., 2020, 2019), recent studies developed benchmarks emphasizing complex clinical reasoning (Gaber et al., 2025). Recent initiatives prioritize ecologically valid assessments over static leaderboards (Chen et al., 2025). Frameworks like MR-Bench (Ren et al., 2026) utilize hospital data exposing disparities between exam performance and authentic clinical accuracy. Specialized datasets evaluate models within high-acuity settings, such as ER-Reason (Mehandru et al., 2025) for emergency rooms, and mARC (Shidara et al., 2026), assessing cognitive flexibility against heuristic biases. MedXpertQA (Zuo et al., 2025) incorporates specialty board questions and multi-round expert reviews evaluating advanced knowledge. MedAgentsBench (Tang et al., 2025) focuses on challenging questions requiring diagnosis formulation and treatment planning. Concentrating on difficult cases pushes models beyond rote knowledge retrieval, revealing performance gaps obscured by easier questions. Furthermore, large-scale studies evaluating frontier models against physicians on standardized and real-world emergency cases reveal that while LLMs match or exceed human performance in diagnostic accuracy, they still fall short in early-stage differential diag-

nosis formulation, exhibiting distinct error patterns in multimodal contexts (Jose et al., 2026; Siam et al., 2025; Rao et al., 2026).

5.2 Reasoning Quality Assessment

Reasoning process integrity is vital in clinical environments (Nachane et al., 2024; Li et al., 2025a; Zhu et al., 2025). MedR-Bench (Qiu et al., 2025) introduces a Reasoning Evaluator quantifying free-text rationales across efficiency, actuality, and completeness. RadRScore (Fan et al., 2025) assesses factual correctness of intermediate steps using validated clinical chains. Beyond static QA, HealthBench (Arora et al., 2025) operationalizes workflows spanning diagnostic triage and safety-critical decisions. AgentClinic (Schmidgall et al., 2024) and SD-Bench (Nori et al., 2025) emulate longitudinal encounters, requiring models to actively query patients. MedChain (Liu et al., 2026) establishes a rigorous benchmarking environment quantifying agent performance within sequential clinical encounters, advancing the granularity of process-oriented evaluations. MedThink-Bench (Zhou et al., 2025c) advances this paradigm with expert-annotated rationales for scalable evaluation. Comprehensive evaluations confirm that despite high accuracy, LLMs' reasoning trajectories lag behind human doctors' structured logic (Merenda et al., 2026). To scale evaluation, novel frameworks increasingly utilize expert-aligned LLMs as automated judges scoring reasoning chains (Zhou et al., 2025d; Rouillard et al., 2026). MedRECT (Iwase et al., 2025) targets reasoning modalities for error correction and interpretable decision pathways.

5.3 Visual Interpretability Assessment

Visually interpreting model decisions builds trust and safeguards patient outcomes. Evaluation frameworks increasingly test a model's capacity to visually justify its reasoning. PRS-Med (Trinh et al., 2025) integrates vision-language models with segmentation capabilities generating accurate masks and corresponding spatial reasoning outputs. RJUA-MedDQA (Jin et al., 2024) evaluates whether models can read complex medical documents, locate key information, and produce answers explicitly grounded in source data. Within chest X-ray diagnostics, GEMeX (Liu et al., 2024, 2025a) and AOR (Li et al., 2025b) concentrate on region-level, multi-step inference evaluating visual grounding and structured mechanisms.

6 Discussion

Despite accelerating progress, hurdles in capability, evaluation, and real-world implementation remain before medical reasoning LLMs become safe clinical tools.

6.1 The Faithfulness-Plausibility Gap

Plausible hallucinations—producing clinically probable yet erroneous explanations—present significant hazards (Sun et al., 2025a; Chen et al., 2024a; Soffer et al., 2025). Models might fabricate lab values supporting diagnostic narratives, directing clinicians toward conclusions derived from incorrect justifications. Addressing this requires imposing rigorous external structure. Restricting generation via medical knowledge graphs mandates the deductive pathway manifest as verifiable semantic combinations (Wu et al., 2025) or logic trees (Zang et al., 2025). Calibrated uncertainty scores compel models to report confidence, providing clinicians a metric for output reliability (Chen et al., 2024b). To dynamically mitigate this, learning-based adaptive frameworks explicitly trade off novelty and consistency during reasoning, controlling hallucination risks under distribution shifts (Sahane et al., 2026).

Epistemological studies suggest medical LLMs lack holistic knowledge structures, struggling to distinguish core principles from peripheral evidence (Kim and Thorne, 2024a). In high-sensitivity domains, they exhibit overconfidence and insufficient abstention (Atif et al., 2025). Consequently, LLMs function as statistical cognitive tools rather than autonomous epistemic agents, supporting their role as decision support systems where humans retain ultimate responsibility (Kim and Thorne, 2024b). Iterative prompting for epistemic uncertainty estimation aids hallucination detection in high-stakes scenarios (Yadkori et al., 2024). Trustworthy medical LLMs require explicit epistemic boundaries, controlled refusal mechanisms, and uncertainty-aware safety guards.

6.2 Toward Native Multimodal Reasoning

Vision-language models often employ loosely coupled architectures integrating representations late in processing (Pan et al., 2025a; Lai et al., 2025; Zhang et al., 2025a). This fails to capture iterative clinical reasoning, as models cannot dynamically re-interrogate visual data using text. Building natively multimodal architectures interleaving visual and textual tokens in a shared reasoning space

is the next frontier, necessitating iterative cross-modal attention and image-anchored counterfactual analysis evaluating alternative diagnoses. Systems like Emu3 (Wang et al., 2024a) and Kimi 2.5 (Team et al., 2026) embed signals into a shared space. Latent visual reasoning studies (Qin et al., 2025; Wu et al., 2026) optimize intrinsic reasoning through continuous visual token inference, overcoming language-centric constraints underutilizing spatial cues. Such innovations are highly instructive for medicine; unified representations seamlessly integrate diverse data (radiology, pathology, guidelines), while enhanced latent reasoning supports precise lesion detection and 3D anatomical analysis, advancing reliable medical multimodal models.

6.3 The Efficiency-Performance Frontier

A persistent tension exists between reasoning quality and computational footprint. Powerful reasoning strategies (Wei et al., 2023; Wang et al., 2023; Tang et al., 2024; Hong et al., 2024) demand significant resources and introduce latency, making them impractical for real-time clinical settings and accelerating research optimizing deductive efficiency. One trajectory involves lightweight post-hoc adaptation, where a small model ranks a larger model's outputs, functioning as results distillation (Shi et al., 2024a). Alternative approaches optimize inference density through meta-cognitive regulation improving efficiency (Zhao et al., 2026). Furthermore, complex logic emerges independently through goal-oriented reinforcement learning on challenging problems rather than imitation via SFT (Zhang et al., 2025a; Liu et al., 2025b), charting a path toward smaller, highly capable models. Evaluations of frontier systems (Bajwa et al., 2026; He et al., 2026) confirm that optimizing reasoning paths remains essential for acute internal medicine scenarios (Wei et al.) regardless of parameter scale.

6.4 Evaluation Beyond Task Accuracy

A restrictive focus on final accuracy precipitates an evaluation crisis. As architectures saturate static benchmarks (Jin et al., 2019; Tang et al., 2025), these metrics obscure logical deficiencies. Progress requires dynamic, longitudinal benchmarks simulating patient progression via multimodal datasets (Li et al., 2025b; Liu et al., 2024). This demands granular assessment of the deductive process, scrutinizing factual correctness, logical coherence, and evidence adherence (Fan et al., 2025; Zhou et al., 2025c). Multimodal evaluation must incorporate quantifiable visual grounding and interpretability metrics. Recognizing automated scores' limitations, robust standards integrate qualitative clinical expert reviews assessing genuine utility alongside stress testing against exceptionally rare pathologies.

6.5 Prerequisites for Responsible Clinical Adoption

Successful clinical integration depends on navigating complex human and regulatory environments. Responsible implementation requires a solid foundation of sociotechnical trust (detailed in Appendix E). Ensuring patient privacy under regulations like HIPAA and GDPR makes privacy-preserving techniques like Federated Learning (FL) an architectural prerequisite (Jahan et al., 2025; Abbas et al., 2024; Zhao et al., 2025). Deployment necessitates addressing algorithmic bias, mitigating demographic disparities, and correcting cognitive biases replicating human errors (Kim et al., 2025b). Cultivating trust involves resolving the accountability gap (Habli et al., 2020) via shared responsibility policies (Information Technology Industry Council, 2024), human-centered explainable AI (Thirunavukarasu et al., 2023; Hong et al., 2024), and Human-in-the-Loop workflows supported by agile regulatory oversight (U.S. Food and Drug Administration, 2024). Moreover, recent human-AI interaction studies emphasize that physicians' trust in LLM outputs does not strictly correlate with benchmark performance. Models with high accuracy may still face mistrust if their reasoning lacks clinical alignment, highlighting the need for transparent interfaces (Xu et al., 2026).

7 Conclusion

This review examines LLM evolution toward complex medical inference, proposing a taxonomy of training-time and test-time enhancement methodologies. The application of these techniques is analyzed across diverse medical modalities and clinical domains, tracking the assessment shift from basic accuracy to validating underlying logical processes. Remaining challenges include model faithfulness, multimodal integration, operational efficiency, and responsible sociotechnical implementation. Overcoming these obstacles enables Medical LLMs to function as trustworthy analytical partners throughout the healthcare sector.

Limitations

Although this manuscript provides a systematic assessment, several limitations require acknowledgment. The analysis remains restricted by the scope of publicly accessible literature, intentionally excluding proprietary models developed by industrial laboratories and potentially omitting recent pre-prints due to rapid field evolution. The keyword-directed search strategy might have omitted relevant papers employing alternative terminologies. Furthermore, the taxonomy represents a specific interpretative framework, meaning alternative organizational structures could exist. Prioritizing breadth to supply a panoramic overview limited the capacity to investigate deep technical nuances for every method. This evaluation functions as a snapshot of a rapidly advancing target, implying future breakthroughs will necessitate continuous revisions.

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.

Anonymous et al. 2026. Medai+: A privacy-aware multimodal medical chat assistant using retrieval-augmented llm reasoning. In *IEEE*.

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.

Farah Atif et al. 2025. Sacred or synthetic? evaluating llm reliability and abstention for religious questions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8.

Muhammad Ayoub, Hai Zhao, Lifeng Li, Dongjie Yang, Shabir Hussain, and Junaid Abdul Wahid. 2026. Structured clinical approach to enable large language models to be used for improved clinical diagnosis and explainable reasoning. *Communications Medicine*.

Maria Bajwa, Robert Hoyt, Dacre Knight, and Maruf Haider. 2026. The performance of deepseek r1 and gemini 3 in complex medical scenarios: comparative study. *JMIRx Med*, 7:e76822.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *Preprint*, arXiv:2308.14346.

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.

Laura Cabello, Carmen Martin-Turrero, Uchenna Akujuobi, Anders Søgaard, and Carlos Bobed. 2024. Meg: Medical knowledge-augmented large language models for question answering. *arXiv preprint arXiv:2411.03883*.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2023. Medbench: A large-scale chinese benchmark for evaluating medical large language models. *Preprint*, arXiv:2312.12806.

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.

Wenting Chen, Guo Yu, Yiu-Fai Cheung, Meidan Ding, Jie Liu, Zizhan Ma, Wenxuan Wang, and Linlin Shen. 2025. Beyond the leaderboard: Rethinking medical benchmarks for large language models. *arXiv preprint arXiv:2508.04325*.

Ahmed T Elboardy, Ghada Khoriba, and Essam A Rashed. 2025. Medical ai consensus: A multi-agent framework for radiology report generation and evaluation. *arXiv preprint arXiv:2509.17353*.

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.

Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. 2025. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*, 4:e58670.

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*.

Changho Han, Dong Won Kim, Songsoo Kim, Seng Chan You, SungA Bae, and Dukyong Yoon. 2023. Large-language-model-based 10-year risk prediction of cardiovascular disease: insight from the uk biobank data. *medRxiv*, pages 2023–05.

Ning He, Lin Yang, Xinhong Hu, and Yuanfang He. 2026. Locally-deployed vs. cloud-based ai in healthcare: evaluating deepseek-r1: 8b, deepseek-r1, and chatgpt o3-mini-high for complex medical diagnostics. *Frontiers in Digital Health*, 8:1785443.

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.

Naoto Iwase, Hiroki Okuyama, and Junichiro Iwasawa. 2025. Medrect: A medical reasoning benchmark for error correction in clinical texts. *arXiv preprint arXiv:2511.00421*.

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.

P Jose, A Sai Manikanta, et al. 2026. Medai+: A privacy-aware multimodal medical chat assistant using retrieval-augmented llm reasoning. In *2026 International Conference on Electronics and Renewable Systems (ICEARS)*, pages 1831–1837. IEEE.

Maël Jullien, Marco Valentino, and André Freitas. 2024. Semeval-2024 task 2: Safe biomedical natural language inference for clinical trials. *arXiv preprint arXiv:2404.04963*.

Piotr Karabowicz, Radosław Charkiewicz, Alicja Charkiewicz, Anetta Sulewska, and Jacek Nikliński. 2025. Agentmol: Multi-model ai system for automatic drug-target identification and molecule development. *Methods and Protocols*, 8(6):143.

Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. 2025a. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific reports*, 15(1):39426.

Kyungwon Kim, Yongmoon Lee, Doohyun Park, Taejoon Eo, Daemyung Youn, Hyesang Lee, and Dosik Hwang. 2024a. Llm-guided multi-modal multiple instance learning for 5-year overall survival prediction of lung cancer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 239–249. Springer.

Minsu Kim and James Thorne. 2024a. Epistemology of language models: Do language models have holistic knowledge? In *Findings of the Association for Computational Linguistics: ACL 2024*.

Minsu Kim and James Thorne. 2024b. Epistemology of language models: Do language models have holistic knowledge? pages 12644–12669.

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. 2025b. 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. 2024b. 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.

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. 2025a. 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. 2025b. 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*.

Chi Liu, Derek Li, Yan Shu, Robin Chen, Derek Duan, Teng Fang, and Bryan Dai. 2025c. Fleming-r1: Toward expert-level medical reasoning via reinforcement learning. *arXiv preprint arXiv:2509.15279*.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang Su, Kao-Jung Chang, Haoliang Li, Linlin Shen, Michael R Lyu, and Wenting Chen. 2026. Medchain: Bridging the gap between llm agents and clinical practice with interactive sequence. *Advances in Neural Information Processing Systems*, 38.

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. 2025d. Automated radiotherapy treatment planning guided by gpt-4vision. *Preprint*, arXiv:2406.15609.

Shuyi Liu, Yuming Shang, and Xi Zhang. 2025e. Truthfulrag: Resolving factual-level conflicts in retrieval-augmented generation with knowledge graphs. *arXiv preprint arXiv:2511.10375*.

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. 2025f. 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.

Cheryl Mattingly. 1991. What is clinical reasoning? *The American Journal of Occupational Therapy*, 45(11):979–986.

Nikita Mehandru, Niloufar Golchini, David Bamman, Travis Zack, Melanie F Molina, and Ahmed Alaa. 2025. Er-reason: A benchmark dataset for llm-based clinical reasoning in the emergency room. *arXiv preprint arXiv:2505.22919*.

Flavio Merenda, José Manuel Gómez-Pérez, and German Rigau. 2026. Can llms reason like doctors? exploring the limits of large language models in complex medical reasoning. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2432–2452.

Mirja Mittermaier, Marium M Raza, and Joseph C Kvedar. 2023. Bias in ai-based models for medical applications: challenges and mitigation strategies. *NPJ Digital Medicine*, 6(1):113.

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.

Mahmud Omar, Shelly Soffer, Reem Agbareia, Nicola Luigi Bragazzi, Donald U Apakama, Carol R Horowitz, Alexander W Charney, Robert Freeman, Benjamin Kummer, Benjamin S Glicksberg, et al. 2025. Sociodemographic biases in medical decision making by large language models. *Nature Medicine*, pages 1–9.

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*.

Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, et al. 2024. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, 30(12):3590–3600.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. 2025. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*.

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.

Arya S Rao, Kaiz P Esmail, Richard S Lee, Sharon Jiang, Bianca Arraiza Carlo, Jasleen Gill, Praneet Khanna, Ezra Kalmowitz, Basile Montagnese, Kimia Heydari, et al. 2026. Large language model performance and clinical reasoning tasks. *JAMA Network Open*, 9(4):e264003.

Xiaohan Ren, Chenxiao Fan, Wenyin Ma, Hongliang He, Chongming Gao, Xiaoyan Zhao, and Fuli Feng. 2026. Medical reasoning with large language models: A survey and mr-bench. *arXiv preprint arXiv:2604.08559*.

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.

Amy Rouillard, Sitwala Mundiab, Linda Camarab, Michael Cameron Gramaniec, Ziyaad Dangorc, Ismail Kallad, Shabir A Madhic, Kajal Morarc, Marlvin T Ncubec, Haroon Saloojeee, et al. 2026. Can llms score medical diagnoses and clinical reasoning as well as expert panels? *arXiv preprint arXiv:2604.14892*.

Shaohao Rui, Haoyang Su, Jinyi Xiang, Lian-Ming Wu, and Xiaosong Wang. 2025. Cardiocot: Hierarchical reasoning for multimodal survival analysis. *Preprint*, arXiv:2505.19195.

Sudarshan Dnyaneshwar Sahane, Soma N Ghosh, and Manjari Gupta. 2026. Learning-based α–β guided pncd: An adaptive framework for hallucination mitigation in medical llms. In *2026 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 1–6. IEEE.

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. 2024a. 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.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May D Wang. 2024b. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2024, page 22315.

Kie Shidara, Preethi Prem, Jonathan Kim, Anna Podlasek, Feng Liu, Ahmed Alaa, and Danilo Bernardo. 2026. Advances in llm reasoning enable flexibility in clinical problem-solving. *arXiv preprint arXiv:2601.11866*.

Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. 2025. A systematic review of large language model (llm) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1):117.

Md Kamrul Siam, Angel Varela, Md Jobair Hossain Faruk, Jerry Q Cheng, Huanying Gu, Abdullah Al Maruf, and Zeyar Aung. 2025. Benchmarking large language models on the united states medical licensing examination for clinical reasoning and medical licensing scenarios. *Scientific Reports*.

Shamus Zi Yang Sim and Tyrone Chen. 2025. Critique of impure reason: Unveiling the reasoning behaviour of medical large language models. *Elife*, 14:e106187.

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

Shelly Soffer, Vera Sorin, Girish N Nadkarni, and Eyal Klang. 2025. Pitfalls of large language models in medical ethics reasoning. *NPJ Digital Medicine*, 8(1):461.

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.

Michael Staniek, Artem Sokolov, and Stefan Riezler. 2025. Training and evaluation of guideline-based medical reasoning in llms. *arXiv preprint arXiv:2512.03838*.

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*.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

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.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

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.

U.S. Food and Drug Administration. 2024. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. Technical report, FDA. Final Guidance for Industry and Food and Drug Administration Staff.

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*.

Mini Han Wang. 2025. An explainable ai framework for corneal imaging interpretation and refractive surgery decision support. *Bioengineering*, 12(11):1174.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. 2024a. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

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. 2024b. Can large language models replace data scientists in biomedical research? *arXiv preprint arXiv:2410.21591*.

Ziqing Wang, Chengsheng Mao, Xiaole Wen, Yuan Luo, and Kaize Ding. 2025. Amanda: Agentic medical knowledge augmentation for data-efficient medical visual question answering. *Preprint*, arXiv:2510.02328.

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.

Jintao Wei, Shou-Yin Jiang, Jiangchun Ma, Shanxiang Xu, Mao Zhang, et al. Performance evaluation of large language models in the diagnosis of emergency internal medicine diseases: A retrospective study. *Frontiers in Public Health*, 14:1780425.

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*.

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.

Linquan Wu, Tianxiang Jiang, Yifei Dong, Haoyu Yang, Fengji Zhang, Shichaang Meng, Ai Xuan, Linqi Song, and Jacky Keung. 2026. Lavit: Aligning latent visual thoughts for multi-modal reasoning. *arXiv preprint arXiv:2601.10129*.

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, Yuchen Zhuang, Yishan Zhong, Yue Yu, Xiangru Tang, Hang Wu, May D Wang, Peifeng Ruan, Donghan Yang, Tao Wang, et al. 2025b. Medagentgym: Training llm agents for code-based medical reasoning at scale. *arXiv preprint arXiv:2506.04405*.

Shan Xu, Zhaokun Yan, Chengxiao Dai, and Fan Wu. 2025c. Mega-rag: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of llms in public health. *Frontiers in Public Health*, 13:1635381.

Yuansong Xu, Yichao Zhu, Haokai Wang, Yuchen Wu, Yang Ouyang, Hanlu Li, Wenzhe Zhou, Xinyu Liu, Chang Jiang, and Quan Li. 2026. "do i trust the ai?" towards trustworthy ai-assisted diagnosis: Understanding user perception in llm-supported clinical reasoning. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–31.

Zezhou Xu. 2026. Evidence-driven chain-of-thought prompting with multi-dimensional evidence evaluation for medical question answering. In *2026 5th International Symposium on Computer Applications and Information Technology (ISCAIT)*, pages 1–7. IEEE.

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.

Yasin A Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari. 2024. To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty. volume 37, pages 58077–58117.

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.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*.

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.

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Yanjun Shao, Yonghoe Koo, Minhyeok Ko, Qingyu Chen, Mark Gerstein, et al. 2025. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. *arXiv preprint arXiv:2506.11474*.

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.

Zelin Zang, Wenyi Gu, Siqi Ma, Dan Yang, Yue Shen, Zhu Zhang, Guohui Fan, Wing-Kuen Ling, and Fuji Yang. 2025. A medical multimodal diagnostic framework integrating vision-language models and logic tree reasoning. *arXiv preprint arXiv:2512.21583*.

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*.

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

Huiya Zhao, Dehao Sui, Yasha Wang, Liantao Ma, and Ling Wang. 2025. Privacy-preserving federated learning framework for multi-source electronic health records prognosis prediction. *Sensors*, 25(8):2374.

Yu Zhao, Hao Guan, Yongcheng Jing, Ying Zhang, and Dacheng Tao. 2026. Medcog: Maximizing llm inference density in medical reasoning via meta-cognitive regulation. *arXiv preprint arXiv:2602.07905*.

Xianda Zheng, Zijian Huang, Meng-Fen Chiang, Michael J Witbrock, and Kaiqi Zhao. 2025. Kcr: Resolving long-context knowledge conflicts via reasoning in llms. *arXiv preprint arXiv:2508.01273*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

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

Shuang Zhou, Mingquan Lin, Sirui Ding, Jiashuo Wang, Canyu Chen, Genevieve B Melton, James Zou, and Rui Zhang. 2025b. Explainable differential diagnosis with dual-inference large language models. *npj Health Systems*, 2(1):12.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, et al. 2025c. Automating expert-level medical reasoning evaluation of large language models. *arXiv preprint arXiv:2507.07988*.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, et al. 2025d. Automating expert-level medical reasoning evaluation of large language models. *npj Digital Medicine*.

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.

Appendix

This appendix section supplies supplementary material supporting the primary manuscript. We commence by providing an extensive background regarding medical reasoning LLMs (Appendix A). This progression is succeeded by comprehensive taxonomy visualizations (Appendix B), a complementary mapping of clinical cognitive stages (Appendix C), an expanded investigation covering knowledge-enhanced deductive techniques (Appendix D), extended discussions on responsible clinical adoption and model biases (Appendix E and F), an analytical comparison of evaluation benchmarks (Appendix G), and highly detailed tables documenting model performance (Appendix I). We also introduce additional sections examining transfer gaps.

A 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 optimized for the structured, multi-step inference demanded during high-stakes clinical decision-making. This limitation has spurred the development of specialized Reasoning LLMs, which are structurally designed to generate transparent, highly verifiable, and definitively robust inferential pathways, resolving an essential requirement for trustworthy AI applications within medicine (Savage et al., 2024; Kim et al., 2024b; Yu et al., 2025).

A.1 From Generalist LLMs to Specialized Reasoners

Conventional LLMs, operating as probabilistic sequence models, are proficient regarding knowledge retrieval yet frequently encounter difficulties concerning complex logical operations, including distinguishing correlation from strict causation or effectively managing inherent uncertainty (Wei et al., 2023). A key breakthrough was Chain-of-Thought (CoT) prompting, which established that instructing an architecture to generate sequential logic successfully elicits its latent inferential capabilities, thereby significantly elevating overall performance across deductive 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. Contemporary advanced models currently integrate supervised fine-tuning utilizing explicit deductive traces alongside reinforcement learning from human feedback (RLHF) to systematically reward logically sound cognitive sequences (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). Consequently, Reasoning LLMs are characterized not merely by their raw performance metrics but primarily by their engineered capacity for transparent inference, representing an indispensable feature ensuring their safe deployment within high-stakes environments (Wu et al., 2025).

A.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 inaccuracies, frequently originating from flawed clinical deduction, constitute a primary source of entirely preventable harm, accounting for an estimated thirty-one percent of all medical malpractice claims, whereby a substantial portion unfortunately culminates in patient mortality (Schaffer et al., 2017). Reasoning LLMs offer a promising approach to mitigate these challenges. Within medical education, AI-powered tutoring systems are presently utilized to enhance students' diagnostic alongside history taking proficiencies via delivering highly structured feedback (Brügge et al., 2024). For practicing clinicians, these models can act as cognitive partners. For instance, modern configurations can systematically analyze diagnostic rationale documented within electronic health records to supply real-time feedback regarding potential cognitive biases or apparent logical inconsistencies (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.

B Taxonomy and Framework Visualizations

This section provides visual summaries of our proposed taxonomy. Figure 2 offers a high-level overview of the topics covered in this review. Ta-

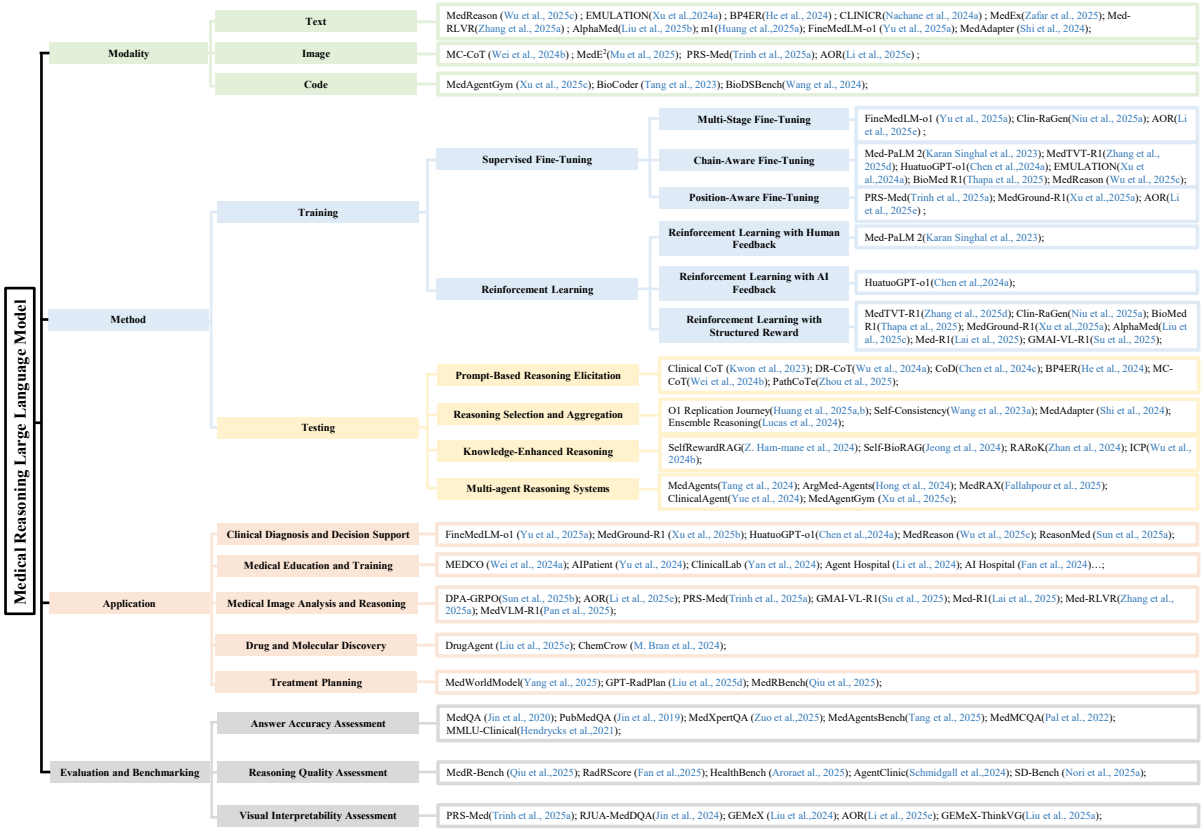

Figure 2: A taxonomy of medical reasoning with LLMs. This figure provides a visual summary of the topics discussed in this review, outlining the primary data modalities (Section 2), the core reasoning enhancement techniques (Section 3), key medical applications (Section 4), and the evolving paradigms for evaluation (Section 5).

ble 1 and Table 3 provide detailed breakdowns of representative works categorized by modality, enhancement method, application, and evaluation framework.

C Cognitive Stages of Clinical Reasoning

We introduce a complementary taxonomy based on the cognitive stages of clinical reasoning itself. This framework, summarized in Table 5, divides reasoning techniques into three stages, encompassing Hypothesis Generation, Evidence Gathering, and Synthesis alongside Verification.

Hypothesis Generation & Differential Diagnosis To support this cognitive leap, models like Clinical CoT (Kwon et al., 2023) and PathCoT (Zhou et al., 2025a) employ Chain-of-Thought prompting to externalize the rationale behind hypothesis formation, ensuring initial diagnostic sets are grounded in logical deduction. To enhance differential diagnosis breadth, MedAgents (Tang et al., 2024) leverages a multi-agent framework where diverse roles propose and debate hypotheses, while Dr. Knows (Gao et al., 2025) integrates knowledge graphs to ground hypotheses in structured facts. Additionally, approaches like FineMedLM-o1 (Yu et al., 2025) optimize this via supervised fine-tuning on reasoning-intensive data, and Disc-MedLLM (Bao et al., 2023) employs counterfactual reasoning to test initial differential lists against confounders.

Evidence Gathering & Hypothesis Refinement Several models are explicitly designed for targeted information collection. BP4ER (He et al., 2024) focuses on evidence gathering through decomposition, using least-to-most prompting to break complex problems into sub-problems for systematic inquiry. MedEx (Zafar et al., 2025) centers on evidence verification, combining iterative questioning with First-Order Logic (FOL) to validate claims. In the visual domain, ChestX-Reasoner (Fan et al., 2025) adopts step-by-step verification of radiological findings for progressive evidence accumulation. Furthermore, Med-RLVR (Zhang et al., 2025a) utilizes reinforcement learning to reward accurate evidence interpretation, while BioCoder (Tang et al., 2023), MedAgentGym (Xu et al., 2025b), and BioDSBench (Wang et al., 2024b) leverage procedural workflows to enable verifiable data collection via code generation.

Synthesis & Verification Models in this category focus on integrating multisource evidence to verify the final conclusion. AOR (Li et al., 2025b) achieves high-level semantic synthesis through anatomical ontology alignment and multi-stage training, while MedGround-R1 (Xu et al., 2025a) incorporates bounding box annotations into cross-modal alignment to enable spatially grounded synthesis from scan interpretation to diagnostic generation. ReasonMed (Sun et al., 2025a) employs automated Chain-of-Thought reasoning to provide traceable and transparent diagnostic verification. Similarly, GEMeX (Liu et al., 2024) and GEMeX-ThinkVG (Liu et al., 2025a) emphasize explicit visual reasoning to support diagnostic synthesis. Despite their distinct technical approaches, these models converge on the essential goal of converting fragmented evidence into a reliable, verified diagnostic conclusion.

D A Deeper Dive into Knowledge-Enhanced Reasoning

This section provides a granular categorization of knowledge-enhanced reasoning methods, differentiated by the structure of the external knowledge source and the mechanism of integration.

D.1 Methodological Categorization

Unstructured Knowledge Contextualization This category is most prominently exemplified by Retrieval-Augmented Generation (RAG). In its standard form, this approach dynamically retrieves relevant passages from a medical text corpus at query time and integrates them into the model's context to generate a factually grounded answer (Jeong et al., 2024; Zhan et al., 2024). Advanced RAG frameworks optimize this for complex reasoning via iterative retrieval, where new information is fetched at each step of a reasoning chain (Yu et al., 2023), or self-correction mechanisms like SelfRewardRAG (Hammane, 2024).

Structured Knowledge Guidance This type of method utilizes formalisms like knowledge graphs (KGs) and ontologies to impose logical and causal constraints on the reasoning process. Mechanisms for integrating this structured knowledge include:
1. Constrained Generation: This strategy uses structured knowledge to directly guide the generation process. For example, In-Context Padding (ICP) (Wu et al., 2024b) injects highly structured knowledge elements directly into the LLM's context. This acts as a foundational framework, compelling the model's generation to adhere to valid

biomedical pathways.

2. Explicit Reasoning Path Validation: This stringent approach leverages KGs to construct and validate entire reasoning chains. While this principle can be instilled during fine-tuning, it can also be applied as a test-time verification layer, where each generated step is validated against the KG for factual correctness before proceeding.

3. Hybrid RAG and KG Integration: To leverage the strengths of both paradigms, hybrid models use RAG to retrieve a broad set of relevant facts from unstructured text, and then use a KG to reason over these entities and explore their connections (Cabello et al., 2024).

4. Formal Logic Integration: For the highest degree of rigor, methods like MedEx integrate First-Order Logic (FOL) to verify claims and inject structured knowledge, ensuring a high degree of logical reliability (Zafar et al., 2025).

D.2 Conflict Resolution in Knowledge Integration

Integrating external knowledge introduces a profound epistemic tension regarding how models should resolve direct conflicts between high-confidence, static Knowledge Graphs (representing established consensus) and novel, dynamic RAG retrievals (representing emerging, potentially unverified evidence). This is not merely a technical data conflict, but a fundamental challenge of clinical epistemology. Recent research highlights that managing this tension requires moving beyond simple conflict resolution toward safety-critical evidence prioritization.

Safety-first evidence hierarchies, involving source credibility and temporal weighting, dictate that in high-stakes clinical scenarios, reasoning chains must adopt a strict hierarchy of trust. Frameworks like TruthfulRAG (Liu et al., 2025e) propose that gold-standard sources, such as clinical guidelines encoded in KGs, must act as non-negotiable safety guardrails. A reasoning chain should default to the established KG to prevent hallucinated or isolated novel findings from overriding the standard of care. A dynamic retrieval should only override a static KG if the retrieval explicitly cites a high-impact, peer-reviewed update superseding the older consensus. To operationalize this hierarchy, methods such as MEGA-RAG (Xu et al., 2025c) and KCR (Zheng et al., 2025) utilize Natural Language Inference (NLI) modules to explicitly detect logical contradictions between the model's internal parametric knowledge, the KG, and the retrieved context. Escalation and epistemic uncertainty modeling are also essential; in life-threatening or high-stakes scenarios, the most responsible action for an LLM is to not autonomously resolve the conflict. Instead, robust reasoning models are designed to preserve the epistemic tension and escalate it. Evaluations on Safe Biomedical NLI tasks (Jullien et al., 2024) highlight the necessity of generating targeted uncertainty markers. This ensures that the ultimate resolution of conflicting medical evidence remains in the hands of the human clinician.

E Extended Discussion on Responsible Clinical Adoption

This section provides a deeper analysis of the sociotechnical prerequisites introduced in Section 6.5, focusing on specific technical and regulatory implementations.

Addressing Algorithmic and Cognitive Bias Risks While demographic fairness remains a primary concern, recent studies suggest that LLMs may also replicate more subtle human cognitive errors. For instance, an evaluation using the BiasMedQA benchmark (Kim et al., 2025b) indicates that reasoning models can exhibit clinical cognitive biases, such as anchoring, similar to those observed in human practice. If deployed without calibration, these models could potentially reinforce rather than correct existing diagnostic errors. This underscores the need for mitigation strategies that go beyond surface-level fairness to address the underlying reasoning processes.

Navigating Regulatory Frameworks and Accountability The integration of adaptive AI into clinical workflows introduces significant challenges regarding liability and safety monitoring. To address the accountability gap, regulatory bodies are moving toward dynamic oversight models. A notable example is the FDA guidance on Predetermined Change Control Plans (U.S. Food and Drug Administration, 2024), which outlines a pathway for software to undergo iterative improvements while adhering to pre-specified safety protocols.

Technical Approaches to Privacy and Data Security To mitigate the risks of data leakage associated with centralized model training, privacy-preserving architectures are becoming increasingly critical. Federated Learning (FL) has been proposed as a method to enable collaborative model

training across institutions without exchanging raw patient data. Recent work, such as the framework for multi-source electronic health records proposed by Zhao et al. (Zhao et al., 2025), demonstrates the technical feasibility of applying FL to complex prognosis prediction tasks while maintaining compliance with privacy regulations.

Responsible Deployment Mitigation Checklist To provide actionable guidance for clinical translation, we recommend a comprehensive mitigation strategy rather than isolated interventions. Continuous monitoring must establish a long-term mechanism for model performance in real clinical scenarios, tracking diagnostic accuracy, stability, and potential bias drift. Human oversight is essential, maintaining full human-in-the-loop involvement during clinical deployment to ensure that all model-assisted decisions are reviewed and confirmed by qualified clinicians. Risk escalation protocols should formulate a standardized procedure for abnormal model outputs, misdiagnosis risks, or system anomalies, with clear responsible subjects and response processes. Privacy preservation demands adopting technologies like Federated Learning in model training and inference to strictly comply with HIPAA, GDPR, and other regional privacy regulations. Bias evaluation must be conducted regularly for demographic factors and clinical cognitive biases, updating mitigation strategies accordingly. Accountability frameworks should be clearly established to match regional regulatory requirements (e.g., FDA Predetermined Change Control Plans) for adaptive AI model updates. Finally, interpretability features must provide accessible information for model outputs to assist clinicians in understanding AI-assisted decisions, supplemented by regular training for clinical users on model characteristics, usage norms, and risk awareness before and during deployment.

F Model Bias and Equity

Collectively, existing research (Mittermaier et al., 2023; Pfohl et al., 2024; Omar et al., 2025; Kim et al., 2025a) underscores that medical reasoning LLMs are plagued by interconnected biases. These biases are amplified by LLMs' inflexible pattern matching, leading to inconsistent care suggestions that risk exacerbating healthcare disparities. Future work must center bias mitigation and equity-aware reasoning to ensure fair care for all patient groups.

G Comparative Analysis of Evaluation Benchmarks

As discussed in Section 5, the field is moving beyond simple accuracy metrics toward more sophisticated assessments. Table 6 provides a structured overview and comparison of different benchmarks, now organized into three key categories: Answer Accuracy, Reasoning Quality, and Visual Interpretability.

When assessing real-world clinical readiness, it is necessary to distinguish between reliable and potentially misleading benchmarks. Static QA datasets (e.g., MedQA (Jin et al., 2020), PubMedQA (Jin et al., 2019)) are useful for measuring rote medical knowledge and textual reasoning. However, using them as direct proxies for clinical capability can be misleading, as they lack workflow realism and do not fully represent the case mix, demographic diversity, or distribution shifts present in actual clinical practice. To validate true clinical reasoning, researchers should prioritize dynamic, workflow-integrated benchmarks. Frameworks such as HealthBench (Arora et al., 2025), AgentClinic (Schmidgall et al., 2024), and SD-Bench (Nori et al., 2025) provide more reliable evaluation proxies by explicitly modeling longitudinal patient encounters, diagnostic triage, and sequential evidence gathering.

H Mechanistic Explanations for Transfer Gaps

A critical issue across the literature is why gains on general QA benchmarks fail to reliably transfer to harder clinical reasoning or multimodal visual grounding tasks. This disconnect stems from three core mismatches. First, task granularity misalignment occurs because general QA (e.g., MMLU) prioritizes factual recall, while clinical reasoning demands multi-step causal chains (e.g., synthesizing symptoms to link imaging to pathology) (Gaber et al., 2025; Chen et al., 2025). Second, a modal specificity gap exists where text-centric QA gains fail to transfer to multimodal tasks because current VLMs rely on late-fusion architectures that lack dynamic cross-modal reasoning (Pan et al., 2025a; Mu et al., 2025). Third, a clinical pragmatics deficit arises as general QA ignores clinical norms (e.g., diagnostic prudence) and noisy real-world data (e.g., incomplete EHRs), leading to factually correct but clinically irrelevant outputs (Sim and Chen, 2025; Kim et al., 2025a).

To bridge these gaps, actionable training paradigms are required. These include Clinical Reasoning-Centric Curriculum Fine-Tuning (Yu et al., 2025; Li et al., 2025b), which moves beyond generic SFT to prioritize cross-modal integration. Additionally, RL with Clinical-Specific Rewards (Zhang et al., 2025a; Niu et al., 2025) optimizes models for evidence-based claims and penalizes non-clinical reasoning paths. Finally, Knowledge-Enhanced Transfer with Clinical Scaffolds (Wu et al., 2025; Zafar et al., 2025) constrains reasoning paths to verifiable medical knowledge graphs to mitigate modality discrepancies.

I Detailed Performance on Benchmarks

This section contains detailed tables summarizing the performance of various medical reasoning models on established training and testing benchmarks. Table 8 lists results on benchmarks commonly used during the training and fine-tuning phases, while Table 9 shows performance on benchmarks used for evaluation at test-time.

I.1 Quantitative Performance Comparison on MMLU-Pro and MMLU

Table 10 presents a selected comparison to isolate the effects of data quality and parameter size on complex reasoning tasks (MMLU-Pro). By analyzing different versions of MedReason, we observe that for the same base model (LLaMA-3.1-8B), the quality of Chain-of-Thought (CoT) data is pivotal, given that using MedReason CoT yields a considerable improvement over Huatuo CoT. Furthermore, the comparison between HuatuoGPT-o1-8B and 70B highlights the scaling law in medical reasoning, wherein scaling to 70B combined with Reinforcement Learning (PPO) results in a substantial leap, confirming that larger parameter spaces remain essential for handling complex medical logic.

Table 11 illustrates how open-source models are narrowing the gap with proprietary models through advanced post-training techniques, using key model families as examples. Foundation Model Dominance remains evident, as Med-PaLM 2 remains the state-of-the-art, primarily due to the massive scale of the PaLM 2 foundation model and its alignment with high-quality, physician-curated exam data. The Post-Training Lift becomes apparent through the FineMedLM series, demonstrating that standard SFT is often insufficient for smaller models. However, introducing preference learning (DPO) boosts performance significantly. Adding Test-Time Training (TTT) further pushes the score, demonstrating that inference-time compute can effectively compensate for smaller parameter counts. Adversarial Training alongside GRPO results within BioMed-R1 confirm that while advanced reinforcement learning strategies and adversarial data improve performance, model capacity remains a hard constraint for medical knowledge retrieval.

Table 1: Taxonomy of Research on Medical Reasoning LLMs (Modalities & Methods)

Section	Category	Representative Works
Modality	Text	MedReason (Wu et al., 2025), EMULATION (Xu et al., 2024), BP4ER (He et al., 2024), CLINICR (Nachane et al., 2024), MedEx (Zafar et al., 2025), Med-RLVR (Zhang et al., 2025a), AlphaMed (Liu et al., 2025b), m1 (Huang et al., 2025a), FineMedLM-o1 (Yu et al., 2025), MedAdapter (Shi et al., 2024a).
	Code	MC-CoT (Wei et al., 2024b), MedE2 (Mu et al., 2025), PRS-Med (Trinh et al., 2025), AOR (Li et al., 2025b).
	Image	MedAgentGym (Xu et al., 2025b), BioCoder (Tang et al., 2023), BioDSBench (Wang et al., 2024b).
Method	**Training**	
	Supervised Fine-Tuning	
	Multi-stage fine-tuning	FineMedLM-o1 (Yu et al., 2025), Clin-RaGen (Niu et al., 2025), AOR (Li et al., 2025b).
	Chain-aware fine-tuning	Med-PaLM 2 (Singhal et al., 2025), MedTVT-R1 (Zhang et al., 2025c), HuatuoGPT-o1 (Chen et al., 2024a), EMULATION (Xu et al., 2024), BioMed-R1 (Thapa et al., 2025), MedReason (Wu et al., 2025).
	Position-aware fine-tuning	PRS-Med (Trinh et al., 2025), MedGround-R1 (Xu et al., 2025a), AOR (Li et al., 2025b).
	Reinforcement learning	
	Reinforcement Learning with Human Feedback	Med-PaLM 2 (Singhal et al., 2025).
	Reinforcement Learning with AI Feedback	HuatuoGPT-o1 (Chen et al., 2024a).
	Reinforcement Learning with Structured Reward	MedTVT-R1 (Zhang et al., 2025c), Clin-RaGen (Niu et al., 2025), BioMed-R1 (Thapa et al., 2025), MedGround-R1 (Xu et al., 2025a), AlphaMed (Liu et al., 2025b), Med-R1 (Lai et al., 2025), GMAI-VL-R1 (Su et al., 2025), α–β Guided PNCD (Sahane et al., 2026).
	Testing	
	Prompt-Based Reasoning Elicitation	Clinical CoT (Kwon et al., 2023), DR-CoT (Wu et al., 2024a), CoD (Chen et al., 2024b), BP4ER (He et al., 2024), MC-CoT (Wei et al., 2024b), PathCoT (Zhou et al., 2025a), Structured Clinical (Ayoub et al., 2026), Evidence-Driven CoT (Xu, 2026).
	Reasoning Selection and Aggregation	O1 Replication Journey (Huang et al., 2025b), Self-Consistency (Wang et al., 2023), MedAdapter (Shi et al., 2024a), Ensemble Reasoning (Lucas et al., 2024), MedCog (Zhao et al., 2026).
	Knowledge-enhanced reasoning	SelfRewardRAG (Hammane, 2024), Self-BioRAG (Jeong et al., 2024), RARoK (Zhan et al., 2024), ICP (Wu et al., 2024b), MedAI+ (Anonymous et al., 2026).
	Multi-agent Reasoning systems	MedAgents (Tang et al., 2024), ArgMed-Agents (Hong et al., 2024), MedRAX (Fallahpour et al., 2025), ClinicalAgent (Yue et al., 2024), MedAgentGym (Xu et al., 2025b).

Table 3: Taxonomy of Research on Medical Reasoning LLMs (Applications & Evaluation)

Section	Category	Representative Works
Applications	Clinical Diagnosis and Decision Support	FineMedLM-o1 (Yu et al., 2025), MedGround-R1 (Xu et al., 2025a), HuatuoGPT-o1 (Chen et al., 2024a), MedReason (Wu et al., 2025), ReasonMed (Sun et al., 2025a).
	Medical Education and Training	MEDCO (Wei et al., 2024a), AIPatient (Yu et al., 2024), ClinicalLab (Yan et al., 2024), Agent Hospital (Li et al., 2024), AI Hospital (Fan et al., 2024).
	Medical Image Analysis and Reasoning	DPA-GRPO (Sun et al., 2025b), AOR (Li et al., 2025b), PRS-Med (Trinh et al., 2025), GMAI-VL-R1 (Su et al., 2025), Med-R1 (Lai et al., 2025), Med-RLVR (Zhang et al., 2025a), MedVLM-R1 (Pan et al., 2025a).
	Drug and Molecular Discovery	DrugAgent (Liu et al., 2025f), ChemCrow (M. Bran et al., 2024).
	Treatment planning	MedWorldModel (Yang et al., 2025), GPT-RadPlan (Liu et al., 2025d), MedRBench (Qiu et al., 2025).
Evaluation	Answer accuracy evaluation	MedQA (Jin et al., 2020), PubMedQA (Jin et al., 2019), MedXpertQA (Zuo et al., 2025), MedAgentsBench (Tang et al., 2025), MedMCQA (Pal et al., 2022), MMLU-Clinical (Hendrycks et al., 2021), MR-Bench (Ren et al., 2026), ER-Reason (Mehandru et al., 2025), mARC (Shidara et al., 2026), USMLE Bench (Siam et al., 2025), Frontier Eval (Bajwa et al., 2026), Local AI Eval (He et al., 2026).
	Reasoning quality evaluation	MedR-Bench (Qiu et al., 2025), RadRScore (Fan et al., 2025), HealthBench (Arora et al., 2025), AgentClinic (Schmidgall et al., 2024), SD-Bench (Nori et al., 2025), PrIME-LLM (Rao et al., 2026), Reason Like Doctors (Merenda et al., 2026), Auto Eval (Zhou et al., 2025d), LLMJury (Rouillard et al., 2026), Trust AI (Xu et al., 2026).
	Visualization & interpretability evaluation	PRS-Med (Trinh et al., 2025), RJUA-MedDQA (Jin et al., 2024), GEMeX (Liu et al., 2024), AOR (Li et al., 2025b), GEMeX-ThinkVG (Liu et al., 2025a).

Table 5: Taxonomy of Medical Reasoning Techniques based on Clinical Cognitive Stages.

Division	Definition
Hypothesis Generation & Differential Diagnosis (Kwon et al., 2023; Tang et al., 2024)	Triggered by initial problem representation, this phase involves formulating a prioritized list of potential diseases. Corresponding LLM techniques activate relevant illness scripts and generate a wide range of potential answers.
Evidence Gathering & Hypothesis Refinement (He et al., 2024; Zafar et al., 2025)	An iterative, cyclical process where new information is sought to discriminate between competing hypotheses. Models in this category focus on targeted information retrieval (RAG) and Chain-of-Thought (CoT) methods to guide sequential inquiry.
Synthesis & Verification (Li et al., 2025b; Sun et al., 2025a)	All accumulated evidence is integrated to form a coherent, verified narrative explaining the patient's condition. This involves higher-order cognitive skills and integrating multi-source evidence to confirm the final working diagnosis.

Table 6: A comparative analysis of key benchmarks for medical reasoning evaluation. This table is formatted to ensure readability and prevent column overlap by allowing text wrapping in headers and content.

Type	Benchmark	#Size	Multi-turn Conversation	Clinical Scenarios	Diverse Specialties	Reasoning Step Eval.	Multi-dim. Eval.
Answer Accuracy	MedQA (Jin et al., 2020)	1,273	×		×	×	×
	PubMedQA (Jin et al., 2019)	1,000	×	×	×	×	×
	MedXpertQA (Zuo et al., 2025)	4,450	×			×	×
	MedAgentsBench (Tang et al., 2025)	862	×		×	×	×
Reasoning Quality	MedBench (Cai et al., 2023)	1,453	×				
	HealthBench (Arora et al., 2025)	5,000		×	×	×	
	AgentClinic (Schmidgall et al., 2024)	/				×	×
	SD-Bench (Nori et al., 2025)	304			×	×	
Visual Interpretability	PRS-Med (Trinh et al., 2025)	/	×	×		×	
	RJUA-MedDQA (Jin et al., 2024)	2,000	×		×		×

Table 8: Results on Training Benchmarks

Model	Size	Dataset	Metric / Acc.
FineMedLM (Yu et al., 2025)	8B	MMLU	71.4
FineMedLM-o1 (Yu et al., 2025)	8B	MMLU	78.2
FineMedLM-o1 (TTT) (Yu et al., 2025)	8B	MMLU	81.36
HuatuoGPT-o1 (Chen et al., 2024a)	8B	MMLU-Pro	58.7
HuatuoGPT-o1 (Chen et al., 2024a)	70B	MMLU-Pro	71.0
HuatuoGPT-II (Chen et al., 2024a)	7B	CMMLU	59.1
HuatuoGPT-II (Chen et al., 2024a)	13B	CMMLU	61.5
BioMed-R1 (Thapa et al., 2025)	8B	MMLU	71.1
BioMed-R1 (Thapa et al., 2025)	32B	MMLU	80.7
Med-PaLM 2 (ER) (Singhal et al., 2025)	-	MMLU	92
Med-PaLM 2 (best) (Singhal et al., 2025)	-	MMLU	92
MedReason (LLaMA) (Wu et al., 2025)	8B	MMLU-Pro	63.1
MedReason (Mistral) (Wu et al., 2025)	7B	MMLU-Pro	50.8
AlphaMed (Liu et al., 2025b)	7B	MMLU	66.8
Clin-RaGen (Niu et al., 2025)	793M	MIMIC-III (Micro)	0.6501 (F1)
Clin-RaGen (Niu et al., 2025)	793M	MIMIC-IV (Micro)	0.7127 (F1)
AOR (Li et al., 2025b)	-	CheXpert (closed)	71.58
AOR (Li et al., 2025b)	-	CheXpert (open)	53.85
MedTVT-R1 (Zhang et al., 2025c)	-	ECG-QA	0.0831 (BLEU)
PRS-Med (Trinh et al., 2025)	-	Lung X-ray	0.942 (mIoU)
PRS-Med (Trinh et al., 2025)	-	Lung X-ray	0.969 (mDice)
MedGround-R1 (Xu et al., 2025a)	-	MS-CXR	83.12
AlphaMed (LLaMA3.1) (Liu et al., 2025b)	8B	MMLU-ProM	59.3
Med-R1 (Qwen2-VL) (Lai et al., 2025)	3B	OmniMedVQA	72.57
GMAI-VL-R1 (Qwen2.5-VL) (Su et al., 2025)	7B	GMAI-MMBench (test)	43.84

Table 9: Results on Testing Benchmarks

Model	Size	Dataset	Metric / Acc.
CoD (Chen et al., 2024b)	65B	MMLU	63.5
MedAdapter (GPT-4) (Shi et al., 2024a)	-	MMLU	87.42
MedAdapter (GPT-3.5) (Shi et al., 2024a)	-	MMLU	70.55
MedAdapter (LLaMA-2) (Shi et al., 2024a)	13B	MMLU	33.96
BioRAG (Jeong et al., 2024)	7B	MMLU	53.9
BioRAG (Jeong et al., 2024)	13B	MMLU	57.2
Clinical CoT (GPT-4) (Kwon et al., 2023)	-	ADNI	68.4
DR-CoT (Qwen-Instruct) (Wu et al., 2024a)	7B	HotpotQA	72.29 (F1)
BP4ER (He et al., 2024)	-	MedDG	44.78 (BLEU)
MC-CoT (Wei et al., 2024b)	7B	PATH-VQA	45.07
PathCoT (Zhou et al., 2025a)	-	PubMed	40.9
Self-Consistency (GPT-3.5) (Wang et al., 2023)	-	TruthfulQA	67.7
MedAgents (GPT-3.5) (Tang et al., 2024)	-	MedQA	64.1
RARoK (LLaMA2-Chat) (Zhan et al., 2024)	7B	GenMedGPT-5k	70.1 (F1)
ArgMed-Agents (GPT-3.5) (Hong et al., 2024)	-	MedQA	62.1
DR-CoT (Wu et al., 2024a)	-	DDXPlus	33.5
O1 Replication Journey (Huang et al., 2025b)	72B	MedQA	86.48
SelfRewardRAG (GPT-3.5) (Hammane, 2024)	-	PubMedQA	81.1
ClinicalAgent (Yue et al., 2024)	-	TOP	0.7

Table 10: Representative Comparison of Reasoning Performance on MMLU-Pro

Model Name	Base Model	Data Size	Data Sources	Training Method	MMLU-Pro
MedReason (Wu et al., 2025)	LLaMA-3.1-8B-Instruct	SFT 32k	MedQA, MedMCQA, MMLU, PubMedQA, MedXpert, HLE	SFT	63.1
MedReason (Wu et al., 2025)	LLaMA-3.1-8B-Instruct	SFT 32k	MedQA, MedMCQA, MMLU, PubMedQA, MedXpert, HLE	SFT	59.9
MedReason (Wu et al., 2025)	Mistral-Instruct-7B	SFT 32k	MedQA, MedMCQA, MMLU, PubMedQA, MedXpert, HLE	SFT	50.8
MedReason (Wu et al., 2025)	Mistral-Instruct-7B	SFT 32k	MedQA, MedMCQA, MMLU, PubMedQA, MedXpert, HLE	SFT	47.6
HuatuoGPT-o1-8B (Chen et al., 2024a)	LLaMA-3.1-8B-Instruct	SFT 29k	GPT-4o verifiable problems, MedQA, MedMCQA, MMLU-Pro	SFT	58.7
HuatuoGPT-o1-70B (Chen et al., 2024a)	LLaMA-3.1-70B-Instruct	SFT 29k + PPO 20k	GPT-4o verifiable problems, MedQA, MedMCQA, MMLU-Pro	SFT + PPO	**71.0**

Table 11: Performance Progression via Advanced Training Paradigms on MMLU

Model Name	Base Model	Data Size	Data Sources	Training Method	MMLU
Med-PaLM 2 (Singhal et al., 2025)	PaLM 2	SFT 193k	MedMCQA	SFT	**92.0**
FineMedLM (Yu et al., 2025)	LLaMA-3.1-8B-Instruct	SFT 264k	FineFineWeb	SFT	**71.44**
FineMedLM-o1 (Yu et al., 2025)	LLaMA-3.1-8B-Instruct	SFT 264k + DPO 33k	FineFineWeb, QwQ-32B positive and self-negative	SFT + DPO	**78.24**
FineMedLM-o1 (+TTT) (Yu et al., 2025)	LLaMA-3.1-8B-Instruct	SFT 264k + DPO 33k + TTT	FineFineWeb, QwQ-32B positive and self-negative	SFT + DPO + TTT	**81.4**
BioMed-R1 (8B) (Thapa et al., 2025)	LLaMA-3.1-8B-Instruct	SFT 23k + GRPO 7,627 + adversarial 25%	m23k reasoning, DeepSeek-R1 (MedQA, MCQA, HeadQA)	SFT + GRPO + adversarial	**71.1**
BioMed-R1 (32B) (Thapa et al., 2025)	Qwen2.5-32B-Instruct	SFT 23k + GRPO 7,627 + adversarial 25%	m23k reasoning, DeepSeek-R1 (MedQA, MCQA, HeadQA)	SFT + GRPO + adversarial	**80.7**