frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 18 June 2025
DOI 10.3389/frai.2025.1616145

OPEN ACCESS

EDITED BY
Tim Hulsen,
Rotterdam University of Applied Sciences, Netherlands

REVIEWED BY
Hasi Hays,
University of Arkansas, United States
Hossein Motahari-Nezhad,
Óbuda University, Hungary

*CORRESPONDENCE
Birger Moëll
✉ bmoell@kth.se

RECEIVED 22 April 2025
ACCEPTED 29 May 2025
PUBLISHED 18 June 2025

CITATION
Moëll B, Sand Aronsson F and Akbar S (2025) Medical reasoning in LLMs: an in-depth analysis of DeepSeek R1.
Front. Artif. Intell. 8:1616145.
doi: 10.3389/frai.2025.1616145

Medical reasoning in LLMs: an in-depth analysis of DeepSeek R1

Birger Moëll[1]*, Fredrik Sand Aronsson[2,3] and Sanian Akbar[4]

[1]Division of Speech, Music and Hearing, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden, [2]Division of Speech and Language Pathology, Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Stockholm, Sweden, [3]Theme Womens Health and Allied Health Professionals, Section of Speech and Language Pathology, Karolinska University Hospital, Stockholm, Sweden, [4]Stockholm Health Care Services, Stockholm, Sweden

Introduction: The integration of large language models (LLMs) into healthcare holds immense promise, but also raises critical challenges, particularly regarding the interpretability and reliability of their reasoning processes. While models like DeepSeek R1-which incorporates explicit reasoning steps-show promise in enhancing performance and explainability, their alignment with domain-specific expert reasoning remains understudied.

Methods: This paper evaluates the medical reasoning capabilities of DeepSeek R1, comparing its outputs to the reasoning patterns of medical domain experts.

Results: Through qualitative and quantitative analyses of 100 diverse clinical cases from the MedQA dataset, we demonstrate that DeepSeek R1 achieves 93% diagnostic accuracy and shows patterns of medical reasoning. Analysis of the seven error cases revealed several recurring errors: anchoring bias, difficulty integrating conflicting data, limited consideration of alternative diagnoses, overthinking, incomplete knowledge, and prioritizing definitive treatment over crucial intermediate steps.

Discussion: These findings highlight areas for improvement in LLM reasoning for medical applications. Notably the length of reasoning was important with longer responses having a higher probability for error. The marked disparity in reasoning length suggests that extended explanations may signal uncertainty or reflect attempts to rationalize incorrect conclusions. Shorter responses (e.g., under 5,000 characters) were strongly associated with accuracy, providing a practical threshold for assessing confidence in model-generated answers. Beyond observed reasoning errors, the LLM demonstrated sound clinical judgment by systematically evaluating patient information, forming a differential diagnosis, and selecting appropriate treatment based on established guidelines, drug efficacy, resistance patterns, and patient-specific factors. This ability to integrate complex information and apply clinical knowledge highlights the potential of LLMs for supporting medical decision-making through artificial medical reasoning.

KEYWORDS

LLM, medical reasoning, DeepSeek R1, AI in medicine, reasoning models, medical benchmarking

1 Introduction

The accelerating adoption of artificial intelligence (AI) in healthcare, particularly large language models (LLMs), presents unprecedented opportunities to augment clinical decision-making and potentially improve patient outcomes. Clinical reasoning, the cornerstone of medical practice, is a complex cognitive process where practitioners integrate heterogeneous data streams, apply specialized knowledge frameworks, and navigate uncertainty to arrive at diagnostic and therapeutic decisions (Jay et al., 2024; Sudacka et al., 2023). This high-stakes process remains vulnerable to systemic failures, as evidenced by research suggesting medical errors contribute to over 250,000 deaths annually

in the US, making it the third leading cause of death. Medical error includes unintended acts, execution failures, planning errors, or deviations from care processes that may cause harm (Makary and Daniel, 2016).

These challenges are exacerbated as healthcare systems worldwide face mounting pressures from workforce shortages (World Health Organization, 2023) and increasing diagnostic complexity. In this strained environment, LLMs have emerged as potential aids to support clinical decision-making by potentially reducing cognitive burdens and mitigating error risks. However, the integration of these systems into medical workflows demands rigorous examination of their reasoning capabilities—not just their factual knowledge, but their ability to emulate the nuanced cognitive processes of expert clinicians while addressing systemic vulnerabilities in care delivery.

1.1 Clinical reasoning in healthcare

Clinical reasoning is an essential skill for healthcare professionals, particularly physicians (Crescitelli et al., 2019; Durning et al., 2024). It encompasses all aspects of clinical practice, including patient management, treatment decisions, and ongoing care (Crescitelli et al., 2019). While extensive research has focused on this area, challenges remain in understanding and implementing effective clinical reasoning (Yazdani and Abardeh, 2019).

A tension exists between explicit, quantitative approaches and the inherent limitations of human cognition, leading to the recognition that clinical reasoning involves both analytical and non-analytical processes, as described in dual-process theory (Pelaccia et al., 2011; Ferreira et al., 2010). Understanding how clinicians utilize both System 1 (intuitive) and System 2 (analytical) reasoning is crucial for evaluating whether LLMs can replicate this nuanced cognitive process.

1.1.1 Theoretical models and cognitive processes

Several theoretical frameworks have shaped our understanding of clinical reasoning:

- **Hypothetico-deductive reasoning:** Clinicians generate and test hypotheses using clinical data (Nierenberg, 2020). This model, while foundational, has been refined as research indicates clinical reasoning is more domain-specific and knowledge-dependent than initially thought.
- **Script theory:** Medical knowledge is organized into "illness scripts"–cognitive frameworks that integrate clinical findings, risk factors, and pathophysiology (Gee et al., 2017; Charlin et al., 2000). Evaluating LLMs requires assessing their ability to form and utilize analogous script-like structures.
- **Dual process theory:** This influential framework describes two systems of thinking: a fast, intuitive system (Type 1) and a slower, analytical system (Type 2) (Gold et al., 2022; Custers, 2013). Clinicians flexibly switch between these modes based on experience and situation (Boushehri et al., 2015). This highlights the need to evaluate LLMs on both rapid, pattern-recognition tasks and more complex, analytical scenarios.
- **Situated and distributed cognition:** Clinical reasoning is influenced by environmental factors, patient interactions, and team dynamics (Gold et al., 2022; Durning and Artino, 2011). Factors like fatigue and time pressure can impact the process (Torre et al., 2020). This suggests that evaluating LLMs should consider their performance under various contextual constraints.

Clinical reasoning operates through both rapid, intuitive (System 1) and slower, analytical (System 2) cognitive processes. System 1 relies on pattern recognition and experience to generate immediate diagnostic hypotheses, while System 2 involves deliberate, systematic evaluation of information (Shimozono et al., 2020; Barbosa Chaves et al., 2022). Clinicians flexibly switch between these modes depending on case complexity (Shimizu and Tokuda, 2012; Olupeliyawa, 2017).

1.1.2 Development of expertise

The development of clinical reasoning expertise involves a progression from deductive reasoning to the refinement of illness scripts, enabling more efficient diagnostic processes (Shin, 2019; Radović et al., 2022; Lubarsky et al., 2015). This involves mastering data gathering, hypothesis generation, differential diagnosis, and management planning (Weinstein et al., 2017). Assessing an LLM's ability to simulate this developmental trajectory could provide insights into its potential for clinical reasoning.

1.1.3 Diagnostic errors

Diagnostic errors, often linked to reasoning failures, contribute significantly to preventable adverse events (Mettarikanon and Tawanwongsri, 2024; Zwaan et al., 2010). Cognitive errors, particularly biases in information processing, are implicated in a majority of diagnostic errors (Graber et al., 2005; Mukhopadhyay and Choudhari, 2024; Schiff et al., 2013). Common biases include representative heuristic, availability heuristic, and anchoring (Kim and Lee, 2018). This underscores the importance of evaluating LLMs for susceptibility to similar cognitive biases.

Structured reflection and deliberate analysis can improve diagnostic accuracy (Moroz, 2017). However, the optimal balance between intuitive and analytical reasoning depends on various factors (Welch et al., 2017). This suggests that evaluating LLMs should involve tasks that require both rapid, intuitive responses and more deliberate, analytical reasoning.

The theoretical frameworks of clinical reasoning will inform the evaluation of DeepSeek R1 by providing a basis for analyzing its reasoning chains, identifying potential cognitive biases, and assessing its ability to navigate complex clinical scenarios analogous to human experts.

1.2 Bridging clinical-cognitive theory and LLM computation

- **Hypothetico-deductive reasoning:** Similar to the step-wise "chain-of-thought prompting (Wei et al., 2022) now used to force models into enumerating intermediate inferences before

committing to an answer; tokens in the hidden state act as provisional hypotheses that are pruned or strengthened as new context is ingested.

- **Illness-script theory**: Maps onto the Platonic representation hypothesis arguing that large language modules appear to learn the same representations independent of model (Huh et al., 2024): every clinical vignette is pulled toward a stable, cross-modal "ideal" embedding representing the prototypical presentation and guideline-recommended next steps. In this sense, both LLM knowledge and human knowledge of the same concept can be argued to be stored in the same conceptual place.
- **Dual-process theory**: Mirrored in the coexistence of fast, implicit completions (Type 1: zero-shot or few-shot inference) and slow, explicit reasoning traces (Type 2: deliberate chains or tree-of-thought sampling).
- **Situated/distributed cognition:** Corresponds to retrieval-augmented generation (RAG) (Lewis et al., 2020) and mixture-of-experts (MoE) (Li and Zhou, 2024) systems, where external knowledge bases or specialist subnetworks are dynamically routed in–much like clinicians consult colleagues, guidelines, or point-of-care tests when confronted with diagnostic uncertainty.

1.3 Clinical reasoning by LLMs

The rapid evolution of LLMs presents both unprecedented opportunities and profound challenges for healthcare applications. While models like GPT-4 demonstrate remarkable performance on medical licensing examinations, achieving 87.6% accuracy on USMLE-style questions (Nori et al., 2023), performance metrics alone provide insufficient evidence for clinical deployment. Modern medicine requires reasoning that extends beyond factual recall to encompass contextual adaptation, probabilistic weighting of competing hypotheses, and adherence to evolving clinical guidelines (Rajpurkar et al., 2022). A critical gap persists between LLMs' capacity to generate clinically plausible text and their ability to replicate the disciplined reasoning processes that underlie safe patient care (Singhal et al., 2023). Reasoning models such as DeepSeek R1 (DeepSeek-AI et al., 2025) output reasoning tokens, a chain of thought process of thinking in text before giving a text response. By evaluating reasoning tokens we can evaluate whether DeepSeek R1's (DeepSeek-AI et al., 2025) reasoning aligns with that of medical experts, particularly in complex clinical scenarios. DeepSeek R1 is designed to generate explicit inference chains through chain-of-thought prompting (DeepSeek-AI et al., 2025), offering a degree of interpretability that is crucial for medical applications. This paper focuses on DeepSeek R1 because its architecture, which emphasizes explicit reasoning steps, provides a unique opportunity to analyze the fidelity of its medical reasoning in comparison to human experts. The model is available open source which makes it possible to deploy on site for potential handling of sensitive clinical data.

The urgency of this research stems from the accelerating real-world deployment of medical LLMs despite unresolved limitations. A 2023 survey found 38% of U.S. health systems piloting LLM-based tools (Healthcare Information and Management Systems Society , HIMSS), while regulatory approvals for AI diagnostics increased 127% annually since 2020 (Benjamens et al., 2020). The potential risks of deploying LLMs without a thorough understanding of their reasoning abilities underscore the need for this research. Our work bridges critical gaps by:

- **Establishing validity metrics beyond answer correctness, focusing on medical reasoning ability.** We evaluate not just *what* the LLM answers, but *how* it arrives at that answer, analyzing the steps in its reasoning process. This goes beyond simple accuracy metrics to assess the quality and appropriateness of the reasoning itself.
- **Identifying high-risk error patterns requiring mitigation, such as anchoring bias, protocol violations, and misinterpretations of lab values.** Our analysis of DeepSeek R1's errors reveals specific cognitive biases and knowledge gaps that could lead to patient harm. Identifying these patterns is crucial for developing mitigation strategies.
- **Providing a foundation for medically-grounded architectures and training paradigms.** By understanding the strengths and weaknesses of current LLM reasoning, we can inform the design of future models that better align with clinical reasoning processes. This includes exploring techniques like retrieval augmented generation (RAG) and fine-tuning on medical reasoning data.

As LLMs transition from experimental tools to clinical assets, it is imperative for reasoning transparency equivalent to human practitioners. Through systematic evaluation of reasoning chain fidelity, we lay the groundwork for AI systems that complement rather than conflict with clinical judgment, harnessing LLMs' potential while safeguarding evidence-based medicine.

One key benefit of reasoning models over previous LLMs is the reasoning as a solution to the black box problem of LLM outputs (Wang Y. et al., 2024). By following the models reasoning we can evaluate their solutions and see what errors in thinking or knowledge led to incorrect outcomes. This has great potential both from a medical and a technical perspective. From a medical perspective, the information can be valuable if common LLM reasoning errors mimic errors that humans make. If so we can use LLM reasoning errors to understand how we can better train physicians to have robust medical reasoning skills. From a technical perspective, medical reasoning outputs and medical reasoning errors can be used for reasoning fine-tuning and reinforcement learning training (DeepSeek-AI et al., 2025) as well as understanding what data sources might need to be added to the model to improve performance.

By evaluation reasoning we get a more granular understanding of both what the model knows and doesn't know and its reasoning process and the errors within that reasoning process.

1.4 Current research on medical reasoning by LLMs

Research has looked into techniques for improving medical reasoning in LLMs. Lucas et al. (2024) showed that prompt techniques could improve the reasoning capabilities of LLMs in

the medical domain while, Wang J. et al. (2024) showed that RAG joint training techniques reduced hallucinations and improved reasoning capabilities. Maharana et al. (2025) found misaligned between prediction and reasoning with LLMs predicting correctly with faulty reasoning. Li et al. (2024) built a multi turn system for medical evaluation of LLMs helpful for assessing clinical reasoning ability. Recently Open AI releases HealthBench (Arora et al., 2025) a structured evaluation of LLMs in the medical domain focusing on the quality of outputs. Lai et al. (2025) trained and evaluated a Medical Reasoning Vision Language models in a similar style to DeepSeek R1. Similarly, Yu et al. (2025) fine tuned a LLM model specifically for Medical Reasoning and Wu et al. (2025) created a dataset for medical reasoning and fine-tuned a LLM for improved medical reasoning.

On the relationship between use of LLMs and reasoning skills (Goh et al., 2024) found no improvement in reasoning skills by physicians when using LLMs as a tool in comparison to conventional resources. In contrast, Borg et al. (2024) found that a social robot powered by an LLM was useful in clinical training.

2 Methodology

2.1 Dataset

2.1.1 Evaluation corpus

The study utilized 100 clinically diverse questions from the MedQA benchmark (Jin et al., 2021), a rigorously validated dataset derived from professional medical board examinations across multiple countries. MedQA's questions follow the United States Medical Licensing Examination (USMLE) format, testing diagnostic reasoning through clinical vignettes requiring:

- Interpretation of patient histories and physical findings.
- Selection of appropriate diagnostic tests.
- Application of therapeutic guidelines.
- Integration of pathophysiology knowledge.

Questions were selected through random sampling to ensure a cover of a range of specialties within medicine. The amount of questions (n = 100) was selected to facilitate human analysis of reasoning outputs.

2.2 Model implementation

We evaluated DeepSeek-R1 (DeepSeek-AI et al., 2025), a 671B parameter mixture of expert reasoning-enhanced language model built through a novel multi-stage training pipeline that combines reinforcement learning and fine-tuning on reasoning data. We used the DeepSeek-Reasoner model available through the DeepSeek API with default params. The code used for calling the model including data used and model outputs is available open source on Github.[1]

1 https://github.com/BirgerMoell/medical-reasoning

2.2.1 System prompt

Please analyze this medical question carefully. Consider the relevant medical knowledge, clinical guidelines, and logical reasoning needed. Then select the single most appropriate answer choice. Provide your answer as just the letter (A, B, C, or D).

2.3 Error classification protocol

- **Step 1**: Ground truth alignment check
  - Compare final answer to MedQA reference
- **Step 2**: Reasoning chain decomposition
  - Break down into diagnostic/treatment decision points
  - Map to clinical reasoning taxonomy
- **Step 3**: Expert validation
  - Clinician review all errors and compared them to medical reasoning best practice.

3 Results

Author S.A who is a active medical professional performed analysis of the medical reasoning of the model. Additional analysis focused on model performance and cognitive errors was done by authors B.M and F.S. The model achieved an overall accuracy of 93% on the 100 MedQA questions. Our analysis focused on the seven cases where the model made an error to identify patterns and mechanisms of reasoning failures.

3.1 Reasoning analysis by medical professional

3.1.1 Error case 1: neonatal bilious vomiting

The model's reasoning is hampered by anchoring bias, difficulty integrating conflicting data, limited consideration of alternative diagnoses, overthinking, and a somewhat incomplete understanding of the embryology involved. It struggles to efficiently process the information and prioritize the most relevant clues, hindering its ability to confidently reach the correct diagnosis.

3.1.2 Error case 2: respiratory failure

The model correctly identifies key information such as age, risk factors, recent surgery and findings in the pulmonary artery. It excessively focuses on histological composition and fibrous remodeling, leading it to weighing other options as more likely.

3.1.3 Error case 3: acute limb ischemia

Limb ischemia is correctly identified. The model recognizes atrial fibrillation as a key risk factor for arterial emboli, and discusses Rutherford classifications and possible interventions

(surgery vs. thrombolysis). It emphasizes the urgency of revascularization and reasons that surgical thrombectomy should be done because the patient's presentation suggests an embolic source and immediate threat to the limb. It incorrectly weighs the definitive treatment as the answer and skips the important "next" step of heparin drip.

3.1.4 Error case 4: porphyria cutanea tarda (PCT)

Recognizes porphyria cutanea tarda (PCT) based on photosensitive blistering, dark urine, and hyperpigmentation. It explains that treatment typically involves phlebotomy or low-dose hydroxychloroquine. It dismisses invasive or less relevant options (liver transplantation, thalidomide) and incorrectly concludes that hydroxychloroquine (alternative first line treatment) is the best next step, largely because the patient's ferritin level is normal. Normally, a professional would reason that phlebomoty (first-line treatment) can induce remission even with normal iron stores and hydroxychloroquine is used if patient cannot tolerate phlebotomy.

3.1.5 Error case 5: enzyme kinetics

Recognizes hexokinase and glucokinase properties as candidates for an enzyme found in most tissues that phosphorylates glucose. It also correctly identifies it as hexokinase rather than glucokinase, noting that hexokinase has a low Km (high affinity). However, it concludes that this enzyme also has a high Vmax, leading it to pick the incorrect answer ("Low X and high Y"). The LLM's final reasoning step confuses hexokinase's lower capacity (lower Vmax) with a higher capacity, thereby arriving at the wrong choice.

3.1.6 Error case 6: preterm PDA management

It rightly identifies the continuous murmur as PDA-related and distinguishes between drugs that keep the ductus open (prostaglandin E1) and those that close it (indomethacin). However, it overestimates how age limits indomethacin's use, leading it prematurely to favor surgical ligation. In actual clinical practice, a stable 5-week-old would still warrant a trial of pharmacologic closure before considering surgical options.

3.1.7 Error case 7: niacin flushing

Correctly identifies that the patient experiences niacin-induced flushing after statin intolerance. It recognizes niacin as a likely cause of her evening flushing and pruritus, and appropriately considers–but rules out–alternative explanations such as carcinoid syndrome and pheochromocytoma, given hints of cancer in the patient's history. However, it departs from a typical medical approach by concluding that switching to fenofibrate (which primarily targets elevated triglycerides rather than LDL) is the best next step, rather than attempting to mitigate the flushing (for example, with NSAIDs) while maintaining niacin therapy. This oversight highlights a gap in its reasoning compared to standard clinical practice, where controlling niacin's side effects is usually preferred before abandoning a therapy that addresses the patient's elevated LDL cholesterol.

3.1.7.1 Risk scale

High: Foreseeable life- or limb-threat within hours-days. **Moderate**: Appreciable morbidity or accelerated disease progression, but sub-acute. **Low**: Negligible immediate harm; effects felt only over the long term or not at all, forensic question.

3.2 Detailed error analysis

3.2.1 Error case 1: neonatal bilious vomiting

- **Pathway of reasoning**:

$$\text{Bilious Vomit} \rightarrow \underbrace{\text{Duodenal Atresia}}_{\text{Model's Focus}} \rightarrow \text{Emergency Laparotomy} \leftarrow \text{Annular Pancreas}$$
$$\leftarrow \text{Delayed Presentation + Normal Prenatal US}$$

- **Critical failure**: Anchoring bias on classic duodenal obstruction pattern while ignoring:
  1. 3-week delayed presentation (incompatible with complete atresia)
  2. Absence of prenatal ultrasound findings
- **Clinical impact**: Risk of delayed annular pancreas diagnosis (24–48 h window for surgical intervention)

3.2.2 Error case 2: respiratory failure

- **Pathway of reasoning**:

$$\text{DVT} \rightarrow \text{PE} \rightarrow \text{Fibrosis} \rightarrow \text{Actual Cause} \rightarrow \underbrace{\text{CTEPH}}_{\text{Model's Focus}}$$

- **Critical failure**: Attributed wall remodeling (effect) as primary pathology
- **Risk amplification**: Increased mortality from missed vasculitis diagnosis

3.2.3 Error case 3: acute limb ischemia

- **Pathway of reasoning**:

$$\text{Ischemic Limb} \rightarrow \underbrace{\text{Direct Surgery}}_{\text{Model's Focus}} \rightarrow \text{Reperfusion Injury}$$
$$\leftarrow \text{Heparin Bridge} \leftarrow \text{Imaging Guidance}$$

- **Critical failure**: Bypassed essential anticoagulation and imaging steps
- **Risk amplification**: Increased limb loss probability with delayed anticoagulation

3.2.4 Error case 4: porphyria cutanea tarda (PCT)

- **Pathway of reasoning**:

$$\text{PCT} \rightarrow \text{Phlebotomy Required} \rightarrow \underbrace{\text{Normal Iron Stores}}_{\text{Model's Focus}}$$
$$\rightarrow \text{Hydroxychloroquinine}$$

- **Critical failure**: Equated serum ferritin with total body iron stores
- **Risk amplification**: Increased risk of cirrhosis from persistent iron overload

3.2.5 Error case 5: enzyme kinetics

- **Pathway of reasoning**:

$$\text{Tissue Distribution} \rightarrow \underbrace{\text{Low Vmax Assumption}}_{\text{Model's Focus}} \rightarrow \text{Metabolic Dysregulation}$$
$$\leftarrow \text{Hexokinase Signature} \leftarrow \text{Low Km/High Vmax}$$

- **Critical failure**: Confused hexokinase (high-affinity/high-capacity) with glucokinase kinetics
- **Risk amplification**: Error in predicting glucose utilization rates

3.2.6 Error case 6: preterm PDA management

- **Pathway of reasoning**:

$$\text{Preterm Birth} \rightarrow \text{PDA} \rightarrow \underbrace{\text{Surgical Ligation}}_{\text{Model's Focus}}$$
$$\leftarrow \text{Indomethacin Window} \leftarrow \text{5-Week Age}$$

- **Critical failure**: Overestimated surgical urgency in stable infant
- **Risk amplification**: Higher complication rate vs medical management

3.2.7 Error case 7: niacin flushing

- **Pathway of reasoning**:

$$\text{Niacin Use} \rightarrow \text{Flushing} \rightarrow \underbrace{\text{Fenofibrate Switch}}_{\text{Model's Focus}}$$
$$\xleftarrow{\text{PGD2 Pathway}} \text{Aspirin Prophylaxis}$$

- **Critical failure**: Misattributed prostaglandin-mediated flushing to rare neoplasms
- **Risk amplification**: Reduced lipid control efficacy with unnecessary agent switch

3.3 Analysis of diagnostic reasoning errors

We found recurring patterns of diagnostic reasoning errors. A key finding across multiple cases was **anchoring bias**, with fixation on an initial diagnosis (e.g., duodenal atresia in Case 1, CTEPH in Case 2) and subsequently failed to adequately incorporate conflicting evidence. This was often compounded by **confirmation bias**, with selectively attending to information supporting the initial impression while dismissing contradictory data (e.g., normal ferritin in the context of suspected PCT in Case 4).

Several cases demonstrated errors related to disease pathway understanding. In Case 2, **feature binding** led to misattributing wall remodeling as the primary pathology rather than recognizing it as a consequence of another underlying condition (vasculitis). A similar error in Case 5 involved confusing enzyme kinetics, misidentifying hexokinase as glucokinase, highlighting a lack of understanding of the specific biochemical pathways.

Omission bias was evident in Case 3, where crucial steps like anticoagulation and imaging were bypassed in the rush to surgery for acute limb ischemia. This suggests a failure to consider all necessary elements of the diagnostic and treatment pathway. In contrast, Case 6 demonstrated potential **commission bias** with the overestimation of surgical urgency in a stable infant with a PDA, potentially exposing the patient to unnecessary risk.

Finally, Case 7 illustrated an error in attribution, misattributing niacin-induced flushing to rare neoplasms instead of recognizing it as a prostaglandin-mediated effect. This misattribution led to an unnecessary and detrimental change in lipid-lowering medication.

These findings emphasize the importance of recognizing and mitigating cognitive biases and ensuring a thorough understanding of disease pathways to improve diagnostic accuracy and patient safety. The quantified risk amplifications associated with each error underscore the potential clinical impact of these reasoning flaws.

Another error we think is important to address is the one found in the first Case E1. If you follow the reasoning trace of the model it actually decides on A Abnormal migration of ventral pancreatic bud (correct) but outputs B, Complete failure of proximal duodenum to recanalize (false) . The model first reason and then outputs the answer. Although this only happened a single time, we want to highlight this because it shows that the reasoning might differ from the response. This means that in a clinical setting it is wise to have both model reasoning and model output in order to minimize the risk of errors. If a clinician would have access to both reasoning and output, the reasoning might help the clinician find the right diagnosis but having only access to the model output would lead to a potential misdiagnosis. This highlights the benefit or R1, which shows reasoning patterns, which are hidden in similar reasoning models such as O1 and O3 made by Open AI.

3.4 Statistical analysis of reasoning lengths in correct vs. incorrect responses

We conducted an independent two-sample Welch's t-test to compare the average reasoning length between correct and incorrect answers, as the groups exhibited unequal variances, correct answers (n = 93) averaged 3,648 characters (SD = 2,132; variance = 4.55×10), incorrect answers (n = 7) averaged 8,118 characters (SD = 4,277; variance = 1.83×10).

The analysis revealed a statistically significant difference (t = −2.74, p = 0.032), with incorrect answers containing substantially longer reasoning (mean = 8,118 characters) compared to correct answers (mean = 3,648 characters). The effect size was very large: Cohen's d = 1.93, indicating that an average incorrect explanation is nearly two pooled standard deviations longer than a correct one. The negative t-value reflects the directional difference, where incorrect responses were consistently lengthier.

The marked disparity in reasoning length suggests that extended explanations may signal uncertainty or reflect attempts to rationalize incorrect conclusions. Shorter responses (e.g., under

TABLE 1 Distribution of reasoning errors in 100 clinical cases.

Error type	Count	Percentage	Exemplar case
Protocol misapplication	2	2%	Acute limb ischemia management
Anchoring bias	1	1%	Neonatal bilious vomiting
Etiology-consequence confusion	1	1%	Pulmonary artery fibrosis
Lab value overinterpretation	1	1%	Porphyria cutanea tarda
Isoform misunderstanding	1	1%	Enzyme kinetics
Overinvestigation tendency	1	1%	Niacin-induced flushing

5,000 characters) were strongly associated with accuracy, providing a practical threshold for assessing confidence in model-generated answers. This metric could enhance user transparency by flagging verbose outputs as potential indicators of unreliability.

3.5 Analysis of reasoning success

Although our effort focused on reasoning errors in most cases the model was successful with 93% accuracy. In our analysis of the successful cases we found that the medical reasoning of the model was sound.

3.5.1 Classification as medical reasoning

The reasoning by the R1 model would likely qualify as medical reasoning. The thought process demonstrates key elements of clinical decision-making demonstrated here on case C1 (see Table 1):

3.5.2 Correct case 1: a 23-year-old pregnant women at 22 weeks gestation presents with burning upon urination

The model identifies that the patient is a pregnant woman at 22 weeks gestation with signs of a lower urinary tract infection. It systematically evaluates the safety and efficacy of each antibiotic option in pregnancy: it rules out ampicillin due to common resistance, ceftriaxone because it is overly broad for simple cystitis, and doxycycline because it is contraindicated in pregnancy. It concludes that nitrofurantoin is safe and effective in the second trimester, making option D the correct choice.

- **Data synthesis:** Systematically reviews the patient's history, symptoms, and exam findings.
- **Differential diagnosis:** Rules out pyelonephritis (absence of CVA tenderness) and narrows to cystitis.
- **Application of guidelines:** Considers pregnancy-specific risks and antibiotic safety profiles.
- **Critical appraisal of options:** Evaluates drug efficacy, resistance patterns, and contraindications.
- **Risk-benefit analysis:** Balances fetal safety (e.g., avoiding doxycycline) with maternal treatment efficacy.

3.5.3 Structured clinical approach

- **Begins with clinical context:** Identifies pregnancy as a critical factor influencing management.
- **Prioritizes diagnosis:** Distinguishes cystitis from pyelonephritis based on exam findings (no CVA tenderness).
- **Antibiotic stewardship:** Avoids overly broad agents (ceftriaxone) for uncomplicated cystitis and considers resistance patterns (ampicillin's limitations).
- **Guideline adherence:** Correctly applies recommendations for nitrofurantoin use in pregnancy (safe in second trimester, avoided in first/third).

3.5.4 Reasoning process

The reasoning follows a hypothetico-deductive model common in clinical medicine:

- **Information gathering:** Patient demographics, symptoms, vital signs, and exam findings.
- **Problem representation:** "Pregnant woman with dysuria, no systemic signs, likely cystitis."
- **Differential diagnosis:** Prioritizes cystitis over pyelonephritis.
- **Treatment selection:**
  - **Elimination:** Doxycycline (contraindicated).
  - **Comparison of remaining options:** Ampicillin (resistance), ceftriaxone (overly broad), nitrofurantoin (guideline-supported).
  - **Final decision:** Nitrofurantoin, justified by safety in the second trimester and efficacy for uncomplicated cystitis.

We believe that the structured reasoning approach with high accuracy shows the usefulness of DeepSeek R1 in the healthcare sector. The sound reasoning combines with an open source model gives a clear path forward for integrating this in the healthcare domain.

3.6 Specialty-level accuracy

Mapping the seven erroneous answers (E1–E7) to their respective clinical domains revealed no consistent clustering of mistakes. Only seven of the thirty specialties represented in the 100-item set recorded any error, and in five of those the domain contained a single question (Physiology, Neonatology) or two questions (Pharmacology), such that a lone miss reduced accuracy to 0% or 50%. Among larger categories the system remained highly reliable: **Pediatrics** 86% (6/7 correct), **Surgery** 80% (4/5), and **Pulmonology** 75% (3/4). All remaining 23 disciplines–including Neurology (12 items), Infectious Disease (10), and Cardiology

10.3389/frai.2025.1616145

TABLE 2 Examples of responses with a focus on incorrect responses and reasoning.

Question	Strengths	Weaknesses	Diagnosis	R1 answer
C1. 23-year-old pregnant woman with UTI	- Identifies cystitis based on symptoms. - Recognizes need for treatment. - Rules out inappropriate options. - Selects Nitrofurantoin.	- Spends time on Cephalexin. - Could be more concise.	Cystitis	Cystitis **Correct**
C2. 3-month-old with SIDS	- Correctly identifies SIDS. - Recalls prevention strategies. - Evaluates answer choices. - Recognizes "Back to Sleep" campaign.	- None significant.	SIDS	SIDS **Correct**
C3. 20-year-old woman with menorrhagia	- Interprets lab results. - Considers differentials. - Recognizes family history. - Identifies vWD.	- Briefly considers Hemophilia A. - Mentions bleeding time.	Von Willebrand disease	Von Willebrand disease **Correct**
C4. 40-year-old zookeeper with pancreatitis	- Recalls causes of pancreatitis. - Identifies scorpion sting. - Considers other options.	- None significant.	Scorpion sting	Scorpion sting **Correct**
E1. 3-week-old with bilious vomiting	- Recognizes bilious vomiting as obstruction. - Considers relevant differentials. - Understands embryology.	- Initially rules out duodenal atresia. - Fixates on "complete" in option B. - Overemphasizes malrotation. - Repetitive explanation.	Abnormal migration of ventral pancreatic bud	Duodenal atresia **Incorrect** The models reason correctly but gives out the wrong response
E2. 58-year-old woman post-surgery	- Identifies risk factors. - Initially leans toward thromboembolism. - Considers each option. - Understands CTEPH.	- Gets fixated on histological composition. - Repetitive reasoning.	Thromboembolism	Pulmonary Hypertension **Incorrect**
E3. 68-year-old man with leg pain	- Correctly identifies acute limb ischemia. - Recognizes atrial fibrillation as a risk factor. - Applies Rutherford classifications to evaluate severity. - Understands that urgent management is needed to salvage limb.	- Incorrectly prioritizes definitive treatment over immediate anticoagulation with heparin. - Incorrectly states that thrombolysis is contraindicated in embolic events.	Heparin drip	Surgical thrombectomy **Incorrect**
E4. 48-year-old woman with photosensitive rash	- Correctly identifies porphyria cutanea tarda (PCT) as the most likely diagnosis. - Recognizes the significance of family history, dark urine, and photosensitivity. - Considers other porphyrias (variegate porphyria). - Appropriately rules out liver transplantion and thalomide as standard therapies, understands the role of phlebotomy and hydroxychloroquine in PCT treatment.	-Places excessive emphasis on normal ferritin levels, overlooking that phlebotomy can still induce remission even with normal iron stores. - Briefly considers unrelated conditions (epidermolysis bullosa, pseudoporphyria). - Incorrectly states that thalidomide is used in refractory cases of PCT.	Begin phlebotomy therapy	Begin oral hydroxychloroquine therapy **Incorrect**
E5. Enzyme Kinetics	- Correctly relates X to Km and Y to Vmax. - Correctly identifies the enzyme as hexokinase. - Understands the properties of hexokinase (low Km). - Correctly identifies that the enzyme in question phosphorylates glucose.	- Overthinks the Vmax, failing to definitively conclude whether it's high or low, causing confusion in the final step. -Confuses the concepts of Vmax and Km, incorrectly stating that a low Km indicates a high Vmax. - Incorrectly states that hexokinase has a higher Vmax than glucokinase and incorrectly states that hexokinase is inhibited by glucose-6-phosphate under these experimental conditions. - It overthinks minor details and loses track of the simpler hallmark difference	Low X and low Y	Low X and high Y **Incorrect**
E6. 5-week-old infant with a murmur	- Correctly identifies PDA as the most likely diagnosis. - Recognizes the significance of preterm birth. - Understands the implications of the continuous murmur. - Considers the infant's age and feeding changes. - Knows the general management options for PDA (Indomethacin, surgery).	- Incorrectly dismisses indomethacin as an option based on age alone without considering the full clinical picture - Overthinks the feeding changes and weight gain. - Overthinks age and arrives at the wrong first-line treatment in an otherwise stable infant.	Indomethacin infusion	Surgical ligation **Incorrect**

(Continued)

TABLE 2 (Continued)

Question	Strengths	Weaknesses	Diagnosis	R1 answer
E7. 53-year-old woman with flushing and itching	- Correctly identifies niacin-induced flushing as the most likely cause. - Considers other possibilities (carcinoid, pheochromocytoma, allergy). - Understands the limitations of statins and fibrates. - Recognizes the need for LDL management.	- Incorrectly prioritizes switching to fenofibrate over managing niacin side effects. - Overly focuses on the possibility of carcinoid syndrome despite the low likelihood. - Fails to recognize that taking aspirin 30 minutes before niacin can significantly reduce flushing.	Administer ibuprofen	Switch niacin to fenofibrate **Incorrect**

TABLE 3 Summary of reasoning errors across cases.

Case	Error type	Model answer	Key reasoning flaw
E1. Neonatal vomiting	Anchoring bias	B (duodenal atresia)	Overprioritized textbook presentation despite incompatible timeline
E2. Respiratory failure	Etiology confusion	C (Pulmonary hypertension)	Misattributed vascular remodeling to primary disease
E3. Limb ischemia	Protocol violation	C (surgery)	Skipped anticoagulation step in Rutherford IIb
E4. PCT management	Lab misinterpretation	D (hydroxychloroquine)	Overvalued serum ferritin over hepatic iron
E5. Enzyme kinetics	Isoform confusion	C (High Vmax)	Confused hexokinase/ glucokinase kinetic profiles
E6. PDA management	Therapeutic window error	C (Surgery)	Misjudged indomethacin efficacy in preterms
E7. Niacin flushing	Overinvestigation	D (Fenofibrate)	Ignored temporal drug-effect relationship

(7)–achieved 100% accuracy. Accuracy and errors are detailed in Table 2.

To test whether the observed distribution deviated from a uniform 7% error rate, we applied a χ^2 goodness-of-fit test, obtaining $\chi^2 = 4.8$ with $p = 0.31$; however, the result is tentative because $>$ 75% of cells contained zero errors, violating standard χ^2 assumptions. Overall, the data provide no convincing evidence of a discipline-specific weakness. The apparent dips are compatible with random variation in a small sample, and larger, domain-targeted test sets will be required to identify any genuine specialty-level performance gaps.

3.7 *Post-hoc* audit of ChatGPT-4o

To benchmark against a non-research-grade, commercially deployed LLM, we queried ChatGPT-4o via the public web interface on 28 May 2025, against the seven error vignettes. The system was instructed to reveal its chain of thought inside `<thinking>` tags and then commit to a final answer. This was the full prompt:

Please answer this question using a format where you first reason inside <thinking> tags, after thinking you give an output. Reason through chain of thought.

3.7.1 Method

A single prompt was issued per vignette; no temperature or system-level modifications were possible in the consumer UI. A physician reviewer scored the disclosed reasoning for (i) presence of clinical problem representation, (ii) generation of a pathophysiology-grounded differential, and (iii) adherence to guideline logic when selecting a management step.

3.7.2 Findings

Six of seven chains satisfied all three criteria, demonstrating recognizable medical reasoning. The sole exception (niacin flushing) exhibited a sound diagnostic path but erred at the final therapeutic choice, mirroring the pattern seen in DeepSeek R1 (Table 3).

3.7.3 Implications

Prompting was sufficient to give a non reasoning model reasoning steps that could be evaluated for medical reasoning. This is promising since it could be a step forward for explainability for non-reasoning LLMs. The model solved 6/7 questions which gives an accuracy of 85.71%. This accuracy is hard to compare since the subset is based on questions that DeepSeek R1 failed however it is below the average accuracy for the entire dataset for DeepSeek R1 (93%).

The close qualitative parity between ChatGPT-4o and DeepSeek R1 in terms of medical reasoning suggests that our error taxonomy captures generalizable failure modes of contemporary LLMs.

4 Discussion

This study provides a detailed analysis of the medical reasoning capabilities of DeepSeek R1, revealing both its strengths and limitations in handling complex clinical scenarios. While the model demonstrates high overall diagnostic accuracy (93%), our in-depth error analysis highlights specific areas where its reasoning leads to errors in clinical assessment see Figure 1 and Tables 3–6. These findings have several important implications for the development and deployment of LLMs in healthcare.

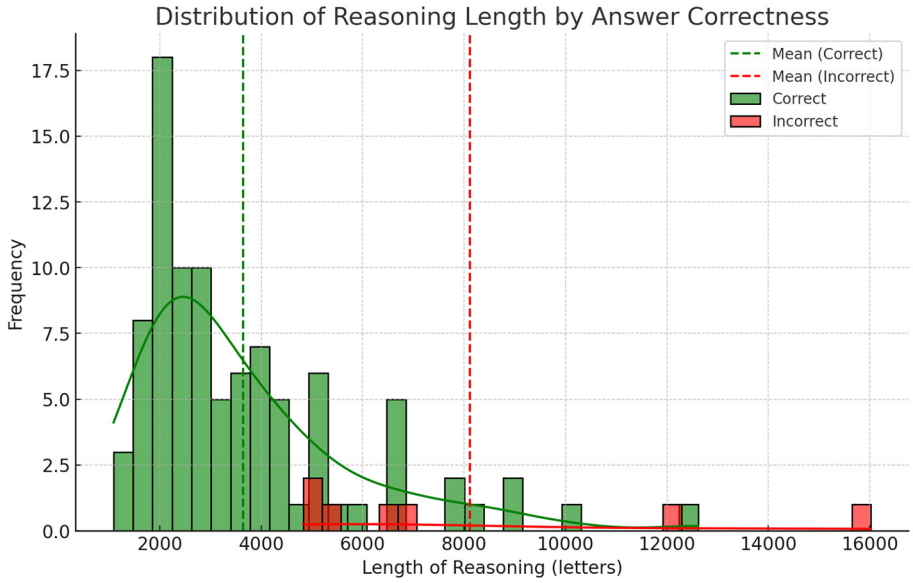

FIGURE 1
Length of reasoning and correctness.

4.1 A note on anthropomorphization of LLMs

In this work we evaluated the reasoning of LLMs and highlighted cognitive errors in its reasoning. There is a speculative nature to this since we assign human error mechanism to an LLM system. We want to be clear that the bias we found in reasoning is dependent on the analysis of the reasoning text and we provide all model reasoning outputs as supplementary material. Thinking about how we reason and how LLMs reason can be fruitful to improve our own reasoning process even though it might lead to bias and potential misunderstanding of the technology. The language we use to describe the reasoning and errors is made to help human understanding and we hope that this does not lead to anthropomorphization of these systems. We believe that LLMs should be viewed as tools but language regarding human cognition can help increase our understanding of their functioning.

4.2 Opening the black box

Deep learning models including LLMs have been accused of being black box algorithms where the inner workings of the models are shielded from view (Wang Y. et al., 2024). This has limited their use in high risk areas such as healthcare where understanding of model outputs is essential for safe implementation. Open reasoning models such as R1 shows a path forwards by being transparent regarding reasoning which has the potential of making the model safer to use in a high risk setting.

4.3 Errors in medical reasoning

Errors that took place were overall a result of thinking errors where the model focused too much attention on details of a problem and lacked necessary understanding of medical protocols. These errors can be viewed similar to mistakes made by a human with medical knowledge and ability to reason about that knowledge making a mistake. That is, a doctor misdiagnosing a patient rather than a human without medical knowledge guessing the answer on a medical test. This is an important distinction because the difference between the two is years of clinical schooling and medical reasoning ability. As such we view these errors as promising and believe that training techniques and new reasoning models will enhance this already fairly adequate medical reasoning ability. Our findings that the length of reasoning was strongly linked to correctness is interesting and can be helpful for improving the usefulness of these models in a clinical setting. By simply using the length of reasoning as a reverse certainty score, we can help a clinician make sense of the models reasoning and even automate double checking, by rerunning long reasoning attempts with an added prompt that the reasoning is likely incorrect.

4.4 Lengths of reasoning and errors

One interesting finding was the strong correlation between the length of output and errors where longer results were more likely to be incorrect. We did a qualitative analysis and found that longer outputs seemed to suffer from “overthinking," where the model gets confused and uses additional tokens to think even though the thinking is not helpful to improve the quality of the results. A concrete finding from our paper is that showing the reasoning

TABLE 4 Distribution of medical questions by specialty.

Specialty	Number of questions	Percentage
Gynecology (OBGYN)	6	6%
Pediatrics	7	7%
Genetics	7	7%
Cardiology	7	7%
Neurology	12	12%
Hematology	7	7%
Gastroenterology	7	7%
Pulmonology	4	4%
Nephrology	6	6%
Urology	3	3%
Infectious disease	10	10%
Oncology	7	7%
Surgery	5	5%
Dermatology	3	3%
Endocrinology	5	5%
Psychiatry	3	3%
Orthopedics	2	2%
Emergency medicine	3	3%
Medical ethics	1	1%
Biostatistics/epidemiology	3	3%
Pharmacology	2	2%
ENT (otolaryngology)	4	4%
Pathology	2	2%
Immunology	1	1%
Toxicology	1	1%
Metabolic disorders	2	2%
Research methods	1	1%
Physiology	1	1%
Patient safety	1	1%
Neonatology	1	1%
Total	**100**	**100%**

length prominently could be a promising way to help clinician evaluate the quality of model reasoning.

4.5 Quality of medical reasoning

Overall we found that the model made few mistakes in its reasoning and the reasoning was medical in nature. The model could reason regarding medical scenarios and overall the reasoning of the model was excellent. This is promising because it shows that medical reasoning is possible through LLMs and that the reasoning is already functional and can be helpful in the healthcare sector if integrated in a safe way.

TABLE 5 Risk-ranked clinical impact of reasoning errors.

Case	Error (short label)	Risk level	Potential patient harm if followed
E1	Anchoring on duodenal atresia	**High**	Missed annular pancreas delayed surgery, bowel perforation, neonatal sepsis; mortality rises hour-to-hour.
E3	Skipping heparin step	**High**	Thrombus propagation during limb-ischaemia work-up irreversible limb loss or systemic emboli within hours.
E4	Overvaluing ferritin	**Moderate**	Persistent porphyria lesions and hepatic iron overload scarring, cirrhosis risk; morbidity high, mortality lower.
E6	Premature PDA ligation	**Moderate**	Avoidable surgical and anesthetic risk when indomethacin could suffice; potential vocal-cord palsy, bleeding.
E2	Treating effect as cause	**Low**	Vasculitis left untreated while managing "CTEPH" unchecked inflammation, right-heart failure, fatal pulmonary hemorrhage. Forensic question. Patient is already diseased.
E5	Hexokinase/ glucokinase mix-up	**Low**	Purely biochemical slip; no direct bedside decision tied to it, negligible immediate harm.
E7	Abandoning niacin instead of fixing flushing	**Low**	LDL undertreated for months-years incremental long-term CV risk; little short-term danger.

4.6 The future of LLMs in healthcare

As within other areas of healthcare, expert clinicians time become a bottleneck when evaluating LLMs. As models improve and show signs of medical reasoning it seems worthwhile to use LLMs to improve LLMs in healthcare. This seemingly paradoxical way of working is actually in line with how large AI labs work to improve LLMs (Anthropic, 2023). A capable LLM model can be used to refine and improve data that can be used to train another LLM and over time data quality improves as well as model performance. For larger medical datasets where human evaluation is simply unfeasible when thousand or millions of questions are evaluated this technique becomes necessary. Having a gold standard of human evaluation with lesser standards for evaluation using LLMs seems to be a possible way forward. As in other areas where LLMs are highly performant such as code generation, we should start to accustom ourself to a world where clinicians supervise AI systems that reason independently. In the future the job of the clinician might be to supervise an AI system that independently gives suggestions for diagnosis and treatment.

TABLE 6 Correct vs. incorrect responses by specialty and resulting accuracy.

Specialty	Questions (*n*)	Correct	Incorrect	Accuracy
Gynecology (OB/GYN)	6	6	0	100%
Pediatrics	7	6	1	86%
Genetics	7	7	0	100%
Cardiology	7	7	0	100%
Neurology	12	12	0	100%
Hematology	7	7	0	100%
Gastroenterology	7	7	0	100%
Pulmonology	4	3	1	75%
Nephrology	6	6	0	100%
Urology	3	3	0	100%
Infectious disease	10	10	0	100%
Oncology	7	7	0	100%
Surgery	5	4	1	80%
Dermatology	3	2	1	67%
Endocrinology	5	5	0	100%
Psychiatry	3	3	0	100%
Orthopedics	2	2	0	100%
Emergency medicine	3	3	0	100%
Medical ethics	1	1	0	100%
Biostatistics/ epidemiology	3	3	0	100%
Pharmacology	2	1	1	50%
ENT (Otolaryngology)	4	4	0	100%
Pathology	2	2	0	100%
Immunology	1	1	0	100%
Toxicology	1	1	0	100%
Metabolic Disorders	2	2	0	100%
Research Methods	1	1	0	100%
Physiology	1	0	1	0%
Patient Safety	1	1	0	100%
Neonatology	1	0	1	0%
Total*	**100**	**93**	**7**	**93%**

4.7 Improving human medical reasoning

Errors in medical reasoning by humans leads to thousands of deaths and injuries each year (Makary and Daniel, 2016). As such improving clinicians ability to reason might be one of the most important tasks for improving healthcare outcomes. The medical reasoning already available in the R1 model can take years for a clinician to acquire through medical training and mentorship and thus using models such as R1 to improve clinicians reasoning skills is one potential use of this technology. This is also in line with a human in the loop approach which improves safety while being aligned with regulatory bodies views on AI in healthcare (Parliament and of the European Union, 2024).

4.8 Improving clinical reasoning

The model was evaluated with a simple prompt and could likely improve through several methods.

1. Retrieval augmented generation (RAG) for improved clinical reasoning. By using a RAG system the performance of the system would likely improve by access to clinical guidelines and other medical texts.
2. Specialization in prompting and documents. In a clinical context, medical professionals usually reason about a smaller subset of clinical knowledge. By dividing the problem of medical reasoning by medical specialty; prompts and knowledge could be used to solve these subproblem more appropriately.
3. Fine tuning on medical reasoning. Improvements to medical reasoning would likely result from fine-tuning on medical reasoning data. Recent advancements in reinforcement learning training for text (DeepSeek-AI et al., 2025) could be useful in this regard.

4.9 Use in a clinical setting

Although the model had errors, overall the reasoning was sound from a medical perspective, as such we believe that these models can be useful in the medical domain and we think it is time for healthcare practitioner to start experimenting with these technologies. As long as healthcare workers are aware of limitations, we believe that use of these systems could help improve patient outcomes. For many clinicians especially in specialized care settings the work can be lonely and there might not be colleagues with similar experience to discuss medical diagnostics. Even though healthcare decisions should always be the responsibility of a human, we believe that reasoning models such as R1 can help clinicians in their diagnostic assessments.

As clinicians we need to be creative in finding safe ways to use this technology in a clinical settings. Both for clinician facing and patient facing interfaces there are likely useful ways to use this technology in a way that is helpful for improving health outcomes.

4.10 Extending the evaluation framework: broader applicability and enhanced validity

Our reasoning analysis framework, applied to DeepSeek R1, is broadly adaptable to all reasoning models that contain reasoning

traces and all LLMs that can be prompted through chain-of-thought to show reasoning traces. Our small experiment with GPT-4o shows that a simple prompting technique can be used with a standard LLM to allow it to be evaluated in this way.

4.11 Limitations

This study has several limitations. First, the evaluation is based on a limited, albeit diverse, set of clinical cases from a single dataset. While MedQA provides a valuable benchmark, it may not fully capture the complexity of real-world clinical practice. Second, our analysis focuses on one specific LLM, DeepSeek R1. While this model represents a state-of-the-art approach to reasoning-enhanced LLMs, the findings may not be generalizable to all LLMs, especially those with different architectures or training methodologies. Third, the expert validation is still subject to the inherent limitations of human judgment and potential biases. Another limitation is that we only had a single medical expert evaluate the medical reasoning of the model.

4.12 Future research directions

Building on the reasoning failures identified in this study, we outline six targeted avenues to improve the clinical robustness of large language models (LLMs).

1. **Domain-Specific fine-tuning with curated medical reasoning corpora**
   Extending the work by Wu et al. (2025) on creating high-quality, explanation-rich datasets of medical reasoning, we advocate assembling case collections with chain-of-thought medical reasoning. Fine-tuning domain-specialized LLMs on such corpora should reduce diagnostic bias and increase factual completeness.
2. **Retrieval-augmented generation (RAG) over authoritative guidelines**
   Linking LLMs to continuously updated medical sources via lightweight RAG pipelines can ground outputs in best evidence, constrain hallucinations, and expose deviations from established care pathways.
3. **Multi-evaluator frameworks for reliable assessment**
   Adopting panels of at least three independent clinicians, standardized rubrics, and explicit inter-rater metrics (e.g. Cohen's κ) will yield more robust estimates of reasoning quality. Targeted training on LLM error taxonomies can further align evaluators.
4. **Prompt engineering for improvement of reasoning capabilities**
   Chain-of-thought prompting sparked the rise of reasoning-oriented models. Systematic exploration of additional prompting strategies–especially those requiring no retraining–could further enhance reasoning and should be rigorously evaluated.
5. **Hybrid human-AI workflows**
   An open research question is how best to integrate these tools into clinical practice. Future systems could auto-flag verbose or low-confidence chains of thought for clinician review, balancing automation with expert oversight.
6. **LLM-as-judge evaluation of reasoning ability**
   Human evaluation is costly, whereas text generation is relatively cheap. Leveraging LLM-as-judge techniques (Croxford et al., 2025) may provide scalable, low-cost assessment of reasoning quality–especially when tightly coupled to dataset creation and model training.

5 Conclusion

This study shows that DeepSeek R1 is capable of a form of medical reasoning as evaluated by analysis by human evaluation on a subset (n = 100) of the MedQA benchmark. The model had an accuracy of 93% and both correct and incorrect cases showed signs of medical reasoning. Using open reasoning models in healthcare improves explainability over non-reasoning models and we encourage continued investigation of how these models can be used to improve the future of healthcare.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://huggingface.co/datasets/birgermoell/medical-reasoning.

Author contributions

BM: Funding acquisition, Conceptualization, Validation, Resources, Writing – review & editing, Writing – original draft, Investigation, Supervision, Project administration, Formal analysis, Data curation, Methodology, Software, Visualization. FS: Data curation, Project administration, Writing – original draft, Methodology, Formal analysis, Writing – review & editing. SA: Formal analysis, Writing – review & editing, Methodology, Writing – original draft, Investigation.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. Generative AI was used to assist in layout of tables in the paper.

Publisher's note

References

Anthropic (2023). *Model Card and Evaluations for Claude Models. Technical Report*. San Francisco, CA: Anthropic PBC. Available online at: https://www-cdn.anthropic.com/files/4zrzovbb/website/Bd2a28d2535bfb0494cc8e2a3bf135d2e7523226.pdf

Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., et al. (2025). Healthbench: evaluating large language models towards improved human health. *arXiv* [Preprint]. arXiv:2505.08775. doi: 10.48550/arXiv.2505.08775

Barbosa Chaves, A., Sampaio Moura, A., de Faria, R. M. D., and Cayres Ribeiro, L. (2022). The use of deliberate reflection to reduce confirmation bias among orthopedic surgery residents. *Sci. Med.* 32:42216. doi: 10.15448/1980-6108.2022.1.42216

Benjamens, S., Dhunnoo, P., and Meskó, B. (2020). The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *npj Digit. Med.* 3:118. doi: 10.1038/s41746-020-00324-0

Borg, A., Jobs, B., Huss, V., Gentline, C., Espinosa, F., Ruiz, M., et al. (2024). Enhancing clinical reasoning skills for medical students: a qualitative comparison of LLM-powered social robotic versus computer-based virtual patients within rheumatology. *Rheumatol. Int.* 44, 3041–3051. doi: 10.1007/s00296-024-05731-0

Boushehri, E., Soltani Arabshahi, K., and Monajemi, A. (2015). Clinical reasoning assessment through medical expertise theories: past, present and future directions. *Med. J. Islamic Repub. Iran* 29:222.

Charlin, B., Tardif, J., and Boshuizen, H. (2000). Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Acad. Med.* 75, 182–190. doi: 10.1097/00001888-200002000-00020

Crescitelli, M. E. D., Ghirotto, L., Artioli, G., and Sarli, L. (2019). Opening the horizons of clinical reasoning to qualitative research. *Acta Biomed.* 90, 8–16.

Croxford, E., Gao, Y., First, E., Pellegrino, N., Schnier, M., Caskey, J., et al. (2025). Automating evaluation of AI text generation in healthcare with a large language model (LLM)-as-a-judge. *medRxiv*. doi: 10.1101/2025.04.22.25326219

Custers, E. (2013). Medical education and cognitive continuum theory: an alternative perspective on medical problem solving and clinical reasoning. *Acad. Med.* 88, 1074–1080. doi: 10.1097/ACM.0b013e31829a3b10

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., et al. (2025). Deepseek-r1: Incentivizing reasoning capability in LLMa via reinforcement learning. *arXiv* [Preprint]. arXiv:2501.12948. doi: 10.48550/arXiv.2501.12948

Durning, S., and Artino, A. (2011). Situativity theory: a perspective on how participants and the environment can interact: Amee guide no. 52. *Med. Teach.* 33, 188–199. doi: 10.3109/0142159X.2011.550965

Durning, S., Jung, E., Kim, D.-H., and Lee, Y.-M. (2024). Teaching clinical reasoning: principles from the literature to help improve instruction from the classroom to the bedside. *Korean J. Med. Educ.* 36, 145–155. doi: 10.3946/kjme.2024.292

European Parliament and Council of the European Union (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (EU) 2020/1828 (artificial intelligence act) (text with eea relevance)*. ELI. Available online at: http://data.europa.eu/eli/reg/2024/1689/oj (BG, ES, CS, DA, DE, ET, EL, EN, FR, GA, HR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV).

Ferreira, A. P. R. B., Ferreira, R., Rajgor, D., Shah, J., Menezes, A., Pietrobon, R., et al. (2010). Clinical reasoning in the real world is mediated by bounded rationality: implications for diagnostic clinical practice guidelines. *PLoS ONE* 5:e10265. doi: 10.1371/journal.pone.0010265

Gee, W., Anakin, M., and Pinnock, R. (2017). Using theory to interpret how senior clinicians define, learn, and teach clinical reasoning. *MedEdPublish* 6:182. doi: 10.15694/mep.2017.000182

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., et al. (2024). Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw. Open* 7:e2440969. doi: 10.1001/jamanetworkopen.2024.40969

Gold, J., Knight, C. L., Christner, J., Mooney, C. E., Manthey, D., Lang, V. J., et al. (2022). Clinical reasoning education in the clerkship years: a cross-disciplinary national needs assessment. *PLoS ONE* 17:e0273250. doi: 10.1371/journal.pone.0273250

Graber, M., Franklin, N., and Gordon, R. (2005). Diagnostic error in internal medicine. *Arch. Intern. Med.* 165, 1493–1499. doi: 10.1001/archinte.165.13.1493

Healthcare Information and Management Systems Society (HIMSS) and Medscape (2024). *Medscape & HIMSS AI Adoption by Health Systems Report 2024*. Available online at: https://cdn.sanity.io/files/sqo8bpt9/production/68216fa5d161adebceb50b7add5b496138a78cdb.pdf

Huh, M., Cheung, B., Wang, T., and Isola, P. (2024). The platonic representation hypothesis. *arXiv* [Preprint]. arXiv:2405.07987. doi: 10.48550/arXiv.2405.07987

Jay, R., Davenport, C., and Patel, R. (2024). Clinical reasoning—the essentials for teaching medical students, trainees and non-medical healthcare professionals. *Br. J. Hosp. Med.* 85, 1–8. doi: 10.12968/hmed.2024.0052

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., Szolovits, P., et al. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* 11:6421. doi: 10.3390/app11146421

Kim, K., and Lee, Y. M. (2018). Understanding uncertainty in medicine: concepts and implications in medical education. *Korean J. Med. Educ.* 30, 181–188. doi: 10.3946/kjme.2018.92

Lai, Y., Zhong, J., Li, M., Zhao, S., and Yang, X. (2025). Med-r1: reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv* [Preprint]. arXiv:2503.13939. doi: 10.48550/arXiv.2503.13939

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*. 33, 9459–9474. doi: 10.48550/arXiv.2005.11401

Li, S., Balachandran, V., Feng, S., Ilgen, J., Pierson, E., Koh, P. W. W., et al. (2024). "Mediq: question-asking llms and a benchmark for reliable interactive clinical reasonin," in *Advances in Neural Information Processing Systems*. 37, 28858–28888. doi: 10.48550/arXiv.2406.00922

Li, Z., and Zhou, T. (2024). Your mixture-of-experts llm is secretly an embedding model for free. *arXiv* [Preprint]. arXiv:2410.10814. doi: 10.48550/arXiv.2410.10814

Lubarsky, S., Dory, V., Audétat, M., Custers, E., and Charlin, B. (2015). Using script theory to cultivate illness script formation and clinical reasoning in health professions education. *Can. Med. Educ. J.* 6, e61–e70. doi: 10.36834/cmej.36631

Lucas, M. M., Yang, J., Pomeroy, J. K., and Yang, C. C. (2024). Reasoning with large language models for medical question answering. *J. Am. Med. Inf. Assoc.* 31, 1964–1975. doi: 10.1093/jamia/ocae131

Maharana, U., Verma, S., Agarwal, A., Mruthyunjaya, P., Mahapatra, D., Ahmed, S., et al. (2025). Right prediction, wrong reasoning: uncovering llm misalignment in ra disease diagnosis. *arXiv* [Preprint]. arXiv:2504.06581. doi: 10.48550/arXiv.2504.06581

Makary, M. A., and Daniel, M. (2016). Medical error—the third leading cause of death in the us. *BMJ* 353:i2139. doi: 10.1136/bmj.i2139

Mettarikanon, D., and Tawanwongsri, W. (2024). Analysis of patient information and differential diagnosis with clinical reasoning in pre-clinical medical students. *Int. Med. Educ.* 3, 23–31. doi: 10.3390/ime3010003

Moroz, A. (2017). Clinical reasoning workshop: cervical spine and shoulder disorders. *MedEdPORTAL* 13:10560. doi: 10.15766/mep_2374-8265.10560

Mukhopadhyay, D., and Choudhari, S. G. (2024). Clinical reasoning skills among second-phase medical students in west bengal, india: an exploratory study. *Cureus* 16:e68839. doi: 10.7759/cureus.68839

Nierenberg, R. (2020). Using the chief complaint driven medical history: theoretical background and practical steps for student clinicians. *MedEdPublish* 9:17. doi: 10.15694/mep.2020.000017.1

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. *arXiv* [Preprint]. arXiv:2303.13375. doi: 10.48550/arXiv.2303.13375

Olupeliyawa, A. (2017). Clinical reasoning: implications and strategies for postgraduate medical education. *J. Postgrad. Inst. Med.* 4:59. doi: 10.4038/jpgim.8174

Pelaccia, T., Tardif, J., Triby, E., and Charlin, B. (2011). An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. *Med. Educ. Online* 16:5890 doi: 10.3402/meo.v16i0.5890

Radović, M., Petrovic, N., and Tosic, M. (2022). An ontology-driven learning assessment using the script concordance test. *Appl. Sci.* 12:1472. doi: 10.3390/app12031472

Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). AI in health and medicine. *Nat. Med.* 28, 31–38. doi: 10.1038/s41591-021-01614-0

Schiff, G., Puopolo, A., Huben-Kearney, A., Yu, W., Keohane, C. A., McDonough, P., et al. (2013). Primary care closed claims experience of Massachusetts malpractice insurers. *JAMA Inter. Med.* 173, 2063–2068. doi: 10.1001/jamainternmed.2013.11070

Shimizu, T., and Tokuda, Y. (2012). Pivot and cluster strategy: a preventive measure against diagnostic errors. *Int. J. Gen. Med.* 5, 917–921. doi: 10.2147/IJGM.S38805

Shimozono, H., Nawa, N., Takahashi, M., Tomita, M., and Tanaka, Y. (2020). A cognitive bias in diagnostic reasoning and its remediation by the "2-dimensional approach". *MedEdPublish* 9:123. doi: 10.15694/mep.2020.000123.1

Shin, H. S. (2019). Reasoning processes in clinical reasoning: from the perspective of cognitive psychology. *Korean J. Med. Educ.* 31, 299–308. doi: 10.3946/kjme.2019.140

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., et al. (2023). Large language models encode clinical knowledge. *Nature* 620, 172–180. doi: 10.1038/s41586-023-06291-2

Sudacka, M., Durning, S. J., Georg, C., Huwendiek, S., Kononowicz, A. A., Schlegel, C., et al. (2023). Clinical reasoning: what do nurses, physicians, and students reason about. *J. Interprof. Care* 37, 990–998. doi: 10.1080/13561820.2023.2208605

Torre, D., Durning, S., Rencic, J., Lang, V. J., Holmboe, E., Daniel, M., et al. (2020). Widening the lens on teaching and assessing clinical reasoning: from "in the head" to "out in the world". *Diagnosis* 7, 181–190. doi: 10.1515/dx-2019-0098

Wang, J., Yang, Z., Yao, Z., and Yu, H. (2024). JMLR: joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv* [Preprint]. arXiv:2402.17887. doi: 10.48550/arXiv.2402.17887

Wang, Y., Ma, X., and Chen, W. (2024). "Augmenting black-box llms with medical textbooks for biomedical question answering (published in findings of EMNLP 2024)," in *Findings of the Association for Computational Linguistics: EMNLP* (Miami, FL: ACL). doi: 10.18653/v1/2024.findings-emnlp.95

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* 35, 24824–24837.

Weinstein, A., Gupta, S., Pinto-Powell, R. C., Jackson, J., Appel, J., Roussel, D., et al. (2017). Diagnosing and remediating clinical reasoning difficulties: a faculty development workshop. *MedEdPORTAL* 13:10650. doi: 10.15766/mep_2374-8265.10650

Welch, P., Plummer, D., Young, L., Quirk, F., Larkins, S., Evans, R., et al. (2017). Grounded theory - a lens to understanding clinical reasoning. *MedEdPublish* 6:2. doi: 10.15694/mep.2017.000002

World Health Organization (2023). *Health workforce snapshot*. Geneva: World Health Organization.

Wu, J., Deng, W., Li, X., Liu, S., Mi, T., Peng, Y., et al. (2025). Medreason: eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv* [Preprint]. arXiv:2504.00993. doi: 10.48550/arXiv.2504.00993

Yazdani, S., and Abardeh, M. H. (2019). Five decades of research and theorization on clinical reasoning: a critical review. *Adv. Med. Educ. Pract.* 10, 703–716. doi: 10.2147/AMEP.S213492

Yu, H., Cheng, T., Cheng, Y., and Feng, R. (2025). Finemedlm-o1: enhancing the medical reasoning ability of llm from supervised fine-tuning to test-time training. *arXiv* [preprint] arXiv:2501.09213. doi: 10.48550/arXiv.2501.09213

Zwaan, L., de Bruijne, M. D., Wagner, C., Thijs, A., Smits, M., van der Wal, G., et al. (2010). Patient record review of the incidence, consequences, and causes of diagnostic adverse events. *Arch. Intern. Med.* 170, 1015–1021. doi: 10.1001/archinternmed.2010.146