

Can LLMs Reason Like Doctors? Exploring the Limits of Large Language Models in Complex Medical Reasoning

Flavio Merenda^{1,2}, José Manuel Gómez-Pérez¹, German Rigau²

¹ Language Technology Research Lab, Expert.ai

² HiTZ Center - Ixa, University of the Basque Country (UPV/EHU)

fmerenda@expert.ai, jmgomez@expert.ai, german.rigau@ehu.eus

Abstract

Large language models (LLMs) have shown remarkable progress in reasoning across multiple domains. However, it remains unclear whether their abilities reflect genuine reasoning or sophisticated pattern matching, a distinction critical in medical decision-making, where reliable multi-step problem-solving is required. Accordingly, we conduct one of the largest evaluations to date, assessing 77 LLMs with diverse fine-tuning approaches, ranging from 1 billion parameters to frontier models. Guided by medical problem-solving theory, we select three medical question answering (QA) benchmarks targeting key reasoning skills: reasoning processes, susceptibility to cognitive biases, and metacognitive abilities. Additionally, we manually annotate a subset of questions to assess the abduction, deduction, and induction capabilities of LLMs, offering detailed insight into the reasoning mechanisms followed by physicians, an aspect that has received relatively limited attention in this domain. Most models, particularly smaller ones, struggle even with specialized fine-tuning or advanced prompting. Larger models perform better but still show clear limitations in complex medical reasoning. Our findings highlight the need to improve specific reasoning strategies to better reflect medical decision-making. The datasets and code used in this study are publicly available at: <https://github.com/expertailab/Can-LLMs-Reason-Like-Doctors>.

1 Introduction

The advancement of LLMs has revealed remarkable capabilities, substantially enhancing natural language processing and understanding beyond previous expectations (Zhao et al., 2023). Notably, their emergent multi-step reasoning ability (Wei et al., 2022a) has been enabled by chain-of-thought prompting (Wei et al., 2022b) and further enhanced by dedicated reasoning models (Xu et al., 2025).

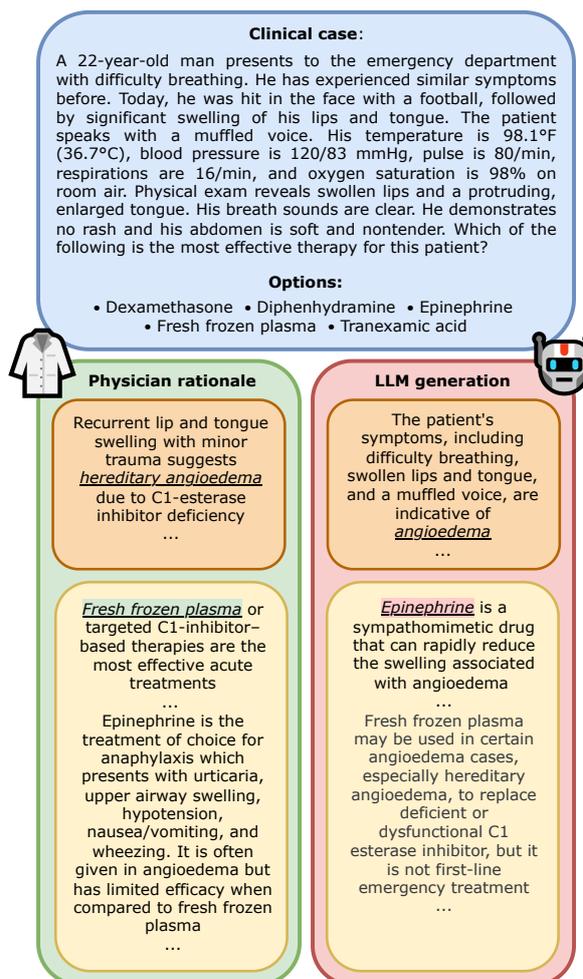


Figure 1: Reasoning comparison between a physician and llama-3.3-70b-instruct on a MedBullets clinical case (Chen et al., 2025a). This example highlights that current LLMs may be unreliable when solving complex problems that involve multi-step reasoning processes, such as abduction and deduction.

These breakthroughs have led to remarkable performance across diverse tasks and benchmarks, with some researchers suggesting that LLMs approach human-level performance and may even show early signs of artificial general intelligence

(Bubeck et al., 2023). The reasoning abilities of LLMs span across domains, including medicine, where the application of LLMs has demonstrated promising outcomes (Yuan et al., 2023; Eriksen et al., 2023; McDuff et al., 2023; Kanjee et al., 2023) and improved explainability (Savage et al., 2023) in tasks involving complex medical diagnosis. Recent research has released open-source models, including medical-specific (Zhou et al., 2023) and reasoning-focused (Li et al., 2025b), as well as quantized or distilled smaller models that rival the performance of their larger counterparts (Srivastava et al., 2025).

Nevertheless, the impressive performance of these models has been questioned in several recent studies (Reese et al., 2023; Mehandru et al., 2024; Hager et al., 2024; Chen et al., 2024a; Ranji, 2024; Li et al., 2024; Liu et al., 2024; Qiu et al., 2025; Tang et al., 2025; Yan et al., 2025a), raising concerns that LLMs may act primarily as advanced information retrievers rather than systems capable of genuine medical reasoning (Kim et al., 2025; Yan et al., 2025b). Initial optimism tied reasoning skills to model scale (Singhal et al., 2022), but recent studies reveal that even advanced models like ChatGPT-o1 rely on pattern-matching and intuitive responses, limiting their effectiveness in nuanced medical reasoning (Soffer et al., 2024; Kim et al., 2025). This suggests that LLMs may continue to struggle with cases requiring complex multi-step reasoning, as illustrated in Figure 1, thereby motivating the following research questions:

- **RQ1:** How do open-source LLMs of varying sizes and tuning perform on complex medical reasoning, and how are results affected by different prompting strategies?
- **RQ2:** How does the performance of small to medium-sized open-source LLMs compare to frontier models in complex medical reasoning?
- **RQ3:** What types of reasoning do physicians apply in medical decision-making, and how accurately can LLMs replicate or support each reasoning type?

To address these questions, we adopted a theoretical framework based in medical problem-solving literature. We benchmarked 77 open-source and frontier LLMs on complex medical tasks that require advanced knowledge and reasoning, robustness against cognitive biases, and metacognitive

skills, making this one of the largest evaluations of its kind to date. We further analyze specific reasoning types applied by physicians, as described in the literature, by evaluating LLMs' abduction, deduction, and induction competences. To this end, we manually annotate a subset of the data, enabling the first fine-grained assessment of such reasoning within this domain. Our results show that many open-source models fail to surpass a random baseline, and even frontier models face significant limitations. This detailed analysis highlights the specific limitations of LLMs and indicates the need to enhance task-specific reasoning to more accurately reflect human medical decision-making.

2 Related Work

Recent studies have systematically evaluated the reasoning capabilities of LLMs in medical contexts. MedAgentsBench, building on prior datasets, introduced a challenging subset designed to test advanced medical knowledge and multi-step reasoning (Tang et al., 2025). While datasets such as JAMA and Medbullets QA reveal shortcomings in LLMs' ability to generate explainable medical answers (Chen et al., 2025b), MedR-Bench shows that these models frequently omit critical reasoning steps in complex tasks like treatment planning (Qiu et al., 2025). GPBench evaluations further highlight persistent deficiencies in areas such as medication use and complication recognition, indicating that LLMs are not yet safe for unsupervised deployment in general practice (Li et al., 2025a). Moreover, LLMs demonstrate critical limitations in real-world clinical scenarios, including misdiagnosis, guideline violations, poor integration (Hager et al., 2024), sensitivity to subtle diagnostic cues (Yan et al., 2025a), and insufficient numerical reasoning (Malghan, 2025), undermining their reliability in practice. Researchers are exploring iterative refinement, agent-based methods (Tang et al., 2025; Kim et al., 2024a; Chen et al., 2024c; Wang et al., 2024a), hybrid symbolic-argumentation systems (Hong et al., 2024), and multi-agent collaboration to reduce diagnostic errors across specialties (Es-teitieh et al., 2025). Recent studies have explored the cognitive processes involved in medical reasoning. MetaMedQA incorporates metacognitive tasks and reveals that models often overconfidently answer questions despite lacking sufficient knowledge (Griot et al., 2025). The Med-ARC benchmark shows that models struggle with cognitive

Benchmark	Size	Avg Lens	Options	Description
MedAgentsBench (Tang et al., 2025)	862	120	3-10	Collection of hard subsets from multiple datasets requiring deep clinical reasoning
MedBullets (Chen et al., 2025a)	89	210	5	USMLE Step 2&3 questions from an online platform requiring clinical reasoning
MedExQA (Kim et al., 2024b)	100	20	4	Covers questions from five underrepresented medical specialties
MedMCQA (Pal et al., 2022)	100	17	4	Questions from real-world entrance exams AIIMS & NEET PG
MedQA (Jin et al., 2021)	100	186	4	Medical board exam questions requiring clinical reasoning and decision-making
MedXpert-R (Zuo et al., 2025)	100	270	10	Advanced medical questions designed to assess expert-level reasoning capabilities
MedXpert-U (Zuo et al., 2025)	100	199	10	Advanced medical questions designed to assess expert-level knowledge
MLLU (Hendrycks et al., 2020)	73	51	4	Biomedical section of a multitask dataset spanning multiple academic domains
MLLU-Pro (Wang et al., 2024b)	100	46	10	Extends MMLU with harder reasoning-based questions and expands choices from 4 to 10
PubMedQA (Jin et al., 2019)	100	326	3	Biomedical research questions from PubMed abstracts with yes/no/maybe answers
MedARC-QA (Kim et al., 2025)	100	195	5-7	Manually curated and designed to exploit memorization (Einstellung effect)
MetaMedQA ¹ (Griot et al., 2025)	100	186	6	Extends MedQA to assess metacognitive capabilities (self-evaluation and knowledge gap recognition)

¹ MetaMedQA was downsampled to align with the MedQA subset used in MedAgentsBench.

Table 1: Overview of clinical question-answering benchmarks.

biases like the Einstellung effect (Binz and Schulz, 2021), underperforming compared to physicians and frequently hallucinating or displaying overconfidence (Kim et al., 2025).

While existing studies provide insights into LLMs in medical reasoning, a comprehensive evaluation across models of varying scales and fine-tuning strategies is still lacking. In addition, LLM competency in specific medical reasoning strategies used by physicians, as documented in the literature, remains underexplored.

3 Methodology

This study aims to evaluate the capabilities of LLMs in complex medical reasoning and to explore their potential limitations by adopting a structured framework based on the medical problem-solving literature. We selected challenging medical benchmarks specifically designed to evaluate three core aspects of reasoning: reasoning processes, susceptibility to cognitive biases, and metacognitive capabilities (Ledley and Lusted, 1959; Fieschi and Gouvernet, 1999; Corrao and Argano, 2022). To enable a fine-grained analysis of reasoning processes, we further constructed a manually annotated subset that captures the reasoning types physicians apply in medical problem-solving: abduction, deduction, and induction (Pietarinen and Stanley, 2025). Detailed information on the selected data and annotation process is provided in Section 4.1.

4 Experimental Setup

4.1 Data

To assess LLMs in complex medical reasoning, existing QA benchmarks were used, and a manually annotated subset was created to categorize questions according to the type of reasoning required:

Benchmarks. The QA benchmarks selected test three essential components of medical problem-solving: advanced knowledge and reasoning, susceptibility to cognitive biases, and metacognitive skills. The specifics of the different benchmarks are shown in Table 1.

MedAgentsBench (Tang et al., 2025) is a comprehensive benchmark that aggregates challenging subsets from eight widely used datasets, targeting advanced knowledge and multi-step reasoning. The MedBullets (Chen et al., 2025a) and PubMedQA (Jin et al., 2019) subsets come with expert-provided rationals to support the evaluation of complex problem-solving. The benchmark was constructed using difficulty-based filtering, selecting questions that fewer than 50% of systems answered correctly (see Table 1 for subset details), and a thorough contamination analysis was conducted to ensure data integrity. **MedARC-QA** (Kim et al., 2025), manually curated by the authors, assesses clinical reasoning through scenarios designed to expose the *Einstellung effect* (Luchins, 1942), highlighting LLMs’ bias toward rigid pattern matching over flexible reasoning. To prevent memorization, it incorporates adversarial answer choices and emphasizes long-tail reasoning. **MetaMedQA** (Griot et al., 2025) extends the *MedQA* dataset (Jin et al., 2021) to evaluate LLMs’ metacognitive abilities, specifically their capacity for self-evaluation and recognition of knowledge gaps. It introduces malformed questions lacking critical information, fictional questions to assess uncertainty detection, and modified items with options such as "none of the above" and "I don’t know or cannot answer". To further increase difficulty, our evaluation is limited to the hardest subset of *MedQA*, shared with *MedAgentsBench*.

To mitigate the known bias toward lettered options in multiple-choice QA (Zheng et al., 2023),

Dataset	Abduction-Deduction	Abduction	Deduction	Induction
MedBullets	27	7	13	0
PubMedQA	0	0	0	50
MedQA	3	4	2	0
MedXpertQA-R	18	9	5	0
MedXpertQA-U	0	2	0	0
MMLU-PRO	2	1	0	0
MMLU	0	2	5	0
Total	50	25	25	50

Table 2: Number of annotated samples per reasoning category across datasets.

we removed option labels in data samples and prompted models to generate full answer texts across all benchmarks. Additionally, we used the MELD (Memorization Affects Levenshtein Detector) framework (Nori et al., 2023), following prior work (Tang et al., 2025), to estimate potential data contamination between the evaluation sets and the models used (see Appendix B).

Fine-grained dataset. Guided by the literature on medical problem-solving, we created a manually annotated subset of the three selected benchmarks to enable a novel, fine-grained analysis of abductive, deductive, and inductive reasoning. For the annotation, we selected tasks that existing literature associates with each reasoning type: i) abduction for hypothesis generation, ii) deduction for exploring the implications of hypotheses, and iii) induction for identifying patterns in observations and measurements to infer generalized rules (Sooknanan and Seemungal, 2019). We attempted to validate the identified reasoning processes using the portion of the data that includes physicians’ manual explanations. Manual inspection of samples and their rationales identified questions that involve implicit abductive reasoning followed by deduction, which we annotated as abduction-deduction to capture the complete reasoning chain required. Table 2 reports the annotation statistics for the resulting dataset, with a full description of the procedure provided in Appendix A.

4.2 Models

The experimental design aims to evaluate and compare a diverse set of open-source and frontier models of varying sizes and tuning strategies, resulting in a collection of 77 models. The primary objective is to assess the impact of model size, reasoning ability, and domain-specific fine-tuning on performance in complex medical reasoning tasks. The study includes multiple LLM families ranging

from 1 to 72 billion parameters, as well as selected Mixture of Experts (MoE) architectures (Cai et al., 2024) and frontier models. For consistency, different sizes of the same model version were tested when available.

General-purpose models include widely used instruction-tuned models such as Llama 3 (Dubey et al., 2024), Gemma 3 (Kamath et al., 2025), Phi 4 (Abdin et al., 2024), OLMo 2 (OLMo et al., 2024), as well as the non-reasoning variants of Qwen 3 (Yang et al., 2025) and DeepCogito V1 Preview (DeepCogito Research, 2025). **Reasoning models** consist of distilled DeepSeek R1 models (DeepSeek-AI et al., 2025), MiMo (Xia et al., 2025), and reasoning-enhanced versions of Phi 4 (Abdin et al., 2025), Qwen 3, and DeepCogito. **Domain-specific models** are fine-tuned on specialized datasets and include Aloe Beta (Garcia-Gasulla et al., 2025), Meditron (Chen et al., 2023), and MedGemma (Google DeepMind, 2025). **Reasoning and domain-specific models** are designed for long-chain-of-thought reasoning on domain-specific tasks, exemplified by Huatuo (Chen et al., 2024b). **Frontier models** include GPT-4o (Hurst et al., 2024), OpenAI o3 (OpenAI, 2025), DeepSeek-V3 (DeepSeek-AI et al., 2024), DeepSeek-R1 (DeepSeek-AI et al., 2025), Claude Sonnet 4 and Llama 4 Maverick.

To assess benchmark complexity, we implemented a **random baseline** for comparison, which, to our knowledge, has not been explored previously. This baseline was generated using Python’s built-in random choice function. For details on computational resources and software settings, see Appendix C.

4.3 Prompting strategies

To evaluate the impact of model prompting, we utilized diverse strategies from prior literature on problem-solving and clinical reasoning, including zero-shot, few-shot, and agent-based prompting.

Zero-shot prompting encompassed several strategies: a direct-answer baseline; domain-agnostic Chain-of-Thought (CoT) prompting (Wei et al., 2022b); Intuitive Reasoning (IR), a clinical reasoning approach (Savage et al., 2023); and Analogical Prompting (AP), a cognitive-inspired method where the model generates relevant examples from parametric memory (Yasunaga et al., 2023). **Few-shot prompting** (FS) leveraged three question–answer–explanation triplets from the *CasiMedicos* dataset (Goenaga et al., 2023), which provides medical exam questions with correct answers and detailed rationales. **Agent-based prompting** utilized a Multi-Persona (MP) reasoning framework, in which specialized agents embodying distinct medical roles collaboratively analyze each question through structured dialogue to reach a consensus answer (Wang et al., 2023).

A detailed presentation of the prompts used in our experimentation is provided in Appendix D.

5 Results

5.1 Influence of model size, tuning, and prompting strategies (RQ1)

Table 3 indicates that achieving better performance necessitates the incorporation of explicit reasoning mechanisms or domain-specific fine-tuning. Overall, the highest performance is achieved by models that integrate both reasoning capabilities and medical domain-specific fine-tuning. Model size plays a critical role, particularly when combined with these enhancements. Notably, all top performers have around 70 billion parameters, highlighting the importance of both scale and targeted training.

As shown in Table 4, while it is challenging to isolate the effect of a single prompt across different strategies, the models seem to derive some, though limited, benefit from this approach. To better understand the results, we generated a plot shown in Figure 2. This box and whiskers plot displays the aggregated accuracy values for each model–prompt pair across all benchmarks. It is evident that models with shorter boxes and whiskers, such as OLMo and MiMo, generally perform worse, as reflected in their lower median accuracy and limited variability across runs. Looking at Table 4, these models particularly underperform on the MedARC-QA benchmark, which is specifically designed to detect the Einstellung effect. This suggests that, even when prompted, they tend to converge on frequent but incorrect answers, failing to explore alternative

reasoning paths. In contrast, models with larger boxes and whiskers, such as Huatuo, tend to show higher median accuracy along with greater variability. This pattern indicates a stronger ability to explore a wider range of responses, occasionally achieving significantly better outcomes. While this increased variability implies less consistent performance, it also reflects a higher potential for correct reasoning beyond common but misleading patterns.

Overall, the low performance across all benchmarks, with only occasional results slightly above a simple random baseline, highlights the challenges these models face in reasoning processes (MedAgentsBench), robustness to cognitive biases (MedARC-QA), and metacognitive capabilities (MetaMedQA).

5.2 Comparison with frontier models (RQ2)

Comparing the best open-source medium-sized models with frontier models in Table 5 reveals a familiar pattern. General-purpose frontier models perform similarly to medium-sized reasoning models, while frontier models fine-tuned for reasoning achieve higher performance, with OpenAI’s o3 leading. These results suggest that model size contributes to performance, but substantial gains require reasoning-focused fine-tuning. Nevertheless, even the top models in this benchmark have difficulty surpassing 50% accuracy, with notably poor performance on PubMedQA, where o3 also falls behind the baseline.

5.3 Fine-grained reasoning analysis (RQ3)

To further investigate the generally low model performances, we provide a fine-grained evaluation of reasoning capabilities, reporting results on the dataset annotated for the specific reasoning types used in medical problem-solving (Appendix A). The results (Figure 3) show that model size alone does not ensure success: larger models like DeepSeek V3 and GPT-4o can underperform, sometimes failing to exceed the random baseline. Smaller models fine-tuned for reasoning (Phi-4 Reasoning) or the medical domain (MedGemma) show improved performance, with models combining both (Huatuo) achieving the greatest gains. Frontier reasoning-focused models (DeepSeek R1, Gemini 2.5 Pro, and OpenAI o3) generally lead, though only o3 surpasses 50% accuracy across three tasks. All models struggle with induction, highlighting the need for targeted fine-tuning. Representative failures are detailed in Appendix E.

Family	Model	MedAgentsBench								MedARC-QA	MetaMedQA	
		MedBullets	MedExQA	MedMCQA	MedQA	MedXpert-R	MedXpert-U	MLLU	MLLU-Pro			PubMedQA
random	random-baseline	17.2	24.3	29.7	23.0	11.0	9.7	29.2	11.7	30.0	23.0	18.7
Llama	Llama-3.2-1B-Instruct	9.0	12.0	15.3	16.0	6.3	8.7	16.9	6.7	23.3	7.3	14.3
	Llama-3.2-3B-Instruct	20.6	11.0	21.7	23.0	11.7	4.3	18.3	11.0	22.7	29.7	14.7
	Llama-3.1-8B-Instruct	17.2	16.7	23.0	17.7	11.7	5.3	18.7	17.3	17.3	27.3	11.0
	Llama-3.3-70B-Instruct	13.5	9.3	16.0	16.3	11.0	12.7	18.7	22.0	14.0	41.7	11.0
Gemma	gemma-3-1b-it	13.9	19.3	25.7	27.3	5.7	7.3	28.3	10.7	20.3	26.0	25.3
	gemma-3-4b-it	23.2	16.7	22.7	26.7	8.7	11.3	14.6	16.7	19.0	19.0	18.0
	gemma-3-12b-it	7.9	12.7	26.3	23.3	6.0	9.0	20.5	26.0	20.0	36.7	18.0
	gemma-3-27b-it	9.4	13.7	25.7	19.0	8.0	7.7	20.5	29.0	13.3	24.0	16.3
Phi	Phi-4-mini-instruct	11.6	12.3	22.0	15.3	9.3	8.7	17.4	14.0	23.3	34.7	11.7
	phi-4	13.9	7.0	20.7	15.3	8.0	11.0	22.8	30.3	21.3	18.3	12.3
	Phi-3.5-MoE-instruct	8.6	9.3	19.3	19.3	7.0	8.0	12.3	23.7	13.7	18.7	14.3
OLMo	OLMo-2-0425-1B-Instruct	21.3	19.3	21.7	31.3	13.0	11.0	28.3	9.0	22.3	7.7	22.0
	OLMo-2-1124-7B-Instruct	13.9	14.7	18.7	24.0	6.7	4.7	21.5	15.0	16.3	16.0	15.7
	OLMo-2-1124-13B-Instruct	11.6	12.7	17.0	18.3	7.0	4.7	17.4	15.3	12.0	23.3	14.7
	OLMo-2-0325-32B-Instruct	10.5	13.3	20.0	19.7	6.7	4.7	9.1	23.0	30.3	15.3	12.0
Qwen non-reasoning	Qwen3-0.6B	16.5	1.3	5.3	13.3	6.3	9.3	6.8	5.7	20.3	23.3	9.0
	Qwen3-1.7B	16.5	17.3	18.3	15.3	8.3	13.0	22.4	15.0	24.3	19.7	7.3
	Qwen3-4B	10.1	10.3	17.3	14.3	4.0	12.0	16.9	19.0	18.7	30.0	10.3
	Qwen3-8B	18.0	15.7	19.7	18.0	9.0	12.7	13.2	23.7	23.7	36.3	13.7
	Qwen3-14B	17.2	11.0	23.3	18.0	8.3	9.0	12.3	22.7	16.7	49.0	13.3
	Qwen3-32B	11.2	12.7	24.3	21.0	11.0	12.7	20.1	21.7	19.0	52.7	12.0
	Qwen3-30B-A3B	17.2	14.0	21.3	26.3	8.0	8.7	20.5	27.0	11.0	44.0	11.7
DeepCogito non-reasoning	cogito-v1-preview-llama-3B	17.2	12.3	20.0	18.3	13.0	9.7	28.8	22.3	18.0	24.3	14.3
	cogito-v1-preview-llama-8B	13.5	16.7	26.3	16.0	9.7	7.0	23.7	22.3	15.0	20.3	14.3
	cogito-v1-preview-qwen-14B	15.0	16.0	26.3	18.0	9.7	7.7	22.4	28.3	16.7	33.7	12.3
	cogito-v1-preview-qwen-32B	9.7	16.7	23.7	14.7	7.0	6.3	29.2	35.7	16.0	26.7	12.0
	cogito-v1-preview-llama-70B	15.4	16.3	21.0	21.3	11.0	9.0	35.6	34.0	12.7	33.0	18.0
Deepseek	DeepSeek-R1-Distill-Qwen-1.5B	14.6	12.0	12.7	23.0	6.7	6.3	19.2	6.3	19.0	26.3	12.7
	DeepSeek-R1-Distill-Qwen-7B	16.5	12.7	20.7	16.0	9.0	5.3	23.7	8.7	22.0	16.3	11.3
	DeepSeek-R1-Distill-Llama-8B	15.0	17.7	23.7	18.3	10.0	6.7	21.0	13.7	25.7	22.3	14.0
	DeepSeek-R1-0528-Qwen3-8B	19.1	17.0	19.7	22.3	7.3	6.7	25.1	23.7	14.7	24.0	18.0
	DeepSeek-R1-Distill-Qwen-14B	20.6	12.7	24.7	22.0	8.0	9.7	21.5	29.0	16.3	29.7	18.0
	DeepSeek-R1-Distill-Qwen-32B	15.4	12.7	20.3	25.0	13.0	8.7	24.7	30.7	13.3	31.0	22.3
	DeepSeek-R1-Distill-Llama-70B	24.0	13.7	24.7	33.3	14.7	12.7	28.8	26.3	13.7	36.0	19.0
	DeepSeek-R1-Distill-Qwen-70B	24.0	13.7	24.7	33.3	14.7	12.7	28.8	26.3	13.7	36.0	19.0
MiMo	MiMo-7B-Base	9.4	9.3	8.3	11.0	6.0	2.3	5.0	8.3	21.7	24.0	8.7
	MiMo-7B-RL-Zero	8.6	6.3	8.0	14.0	4.7	5.0	9.1	15.3	19.0	10.7	6.3
	MiMo-7B-SFT	13.1	12.3	23.3	18.0	4.3	3.3	16.4	12.0	16.7	17.0	10.7
	MiMo-7B-RL	6.4	8.0	17.0	11.7	4.0	2.3	10.5	10.3	8.3	12.3	6.3
Phi	Phi-4-mini-reasoning	5.6	7.0	7.3	5.7	3.0	2.3	13.2	12.7	3.7	12.0	3.3
	Phi-4-reasoning	28.8	12.3	23.3	33.3	12.0	14.0	31.1	31.7	26.0	36.7	26.3
	Phi-4-reasoning-plus	22.5	13.0	15.0	27.0	8.7	10.7	23.7	23.7	24.7	28.0	23.3
Qwen reasoning	Qwen3-0.6B	11.2	8.3	19.0	19.7	5.3	5.7	12.8	4.0	25.0	22.7	11.0
	Qwen3-1.7B	19.5	9.3	25.0	19.0	6.7	8.3	16.4	17.3	19.7	14.7	10.7
	Qwen3-4B	12.7	13.0	19.0	21.0	5.3	8.0	22.4	23.0	20.3	24.3	18.0
	Qwen3-8B	17.2	14.7	20.7	17.3	9.0	9.3	28.8	30.3	15.7	24.0	16.7
	Qwen3-14B	20.6	14.0	22.3	29.3	11.7	6.7	26.0	31.3	13.0	35.7	18.7
	Qwen3-32B	23.2	12.7	24.3	30.3	19.0	9.3	26.0	31.3	18.0	45.0	25.0
	Qwen3-30B-A3B	20.2	12.7	21.7	27.0	9.7	8.3	21.0	30.3	17.3	31.0	20.0
DeepCogito reasoning	cogito-v1-preview-llama-3B	11.2	9.7	19.0	15.0	5.0	4.7	23.3	16.7	19.0	13.7	9.0
	cogito-v1-preview-llama-8B	13.9	14.3	23.0	17.3	5.3	7.7	21.9	26.3	16.3	20.3	15.7
	cogito-v1-preview-qwen-14B	15.7	13.7	22.0	23.0	8.0	7.7	31.1	34.3	20.0	27.3	13.0
	cogito-v1-preview-qwen-32B	18.0	12.3	25.0	27.7	13.3	8.0	32.9	31.7	17.0	28.7	23.7
	cogito-v1-preview-llama-70B	21.3	14.0	23.0	32.3	11.0	10.0	38.4	34.7	16.0	24.0	24.0
Aloe	Qwen2.5-Aloe-Beta-7B	14.6	19.7	27.3	18.0	8.7	6.0	19.2	17.0	21.7	30.7	14.0
	Llama3.1-Aloe-Beta-8B	15.7	15.0	25.3	18.7	7.7	6.3	16.0	21.3	16.3	21.3	17.3
	Llama3.1-Aloe-Beta-70B	19.5	12.0	23.3	16.7	10.7	17.7	12.3	23.0	14.3	34.3	12.3
	Qwen2.5-Aloe-Beta-72B	12.4	20.3	26.3	16.3	8.3	11.0	25.6	29.3	14.0	55.7	13.3
	Qwen2.5-Aloe-Beta-72B	12.4	20.3	26.3	16.3	8.3	11.0	25.6	29.3	14.0	55.7	13.3
Meditron	Meditron3-Gemma2-2B	16.9	12.3	18.0	16.7	8.3	11.0	16.9	12.3	25.0	13.3	10.7
	Meditron3-Qwen2.5-7B	20.6	21.0	26.0	15.3	10.3	8.7	21.5	14.0	19.0	38.7	14.0
	Meditron3-8B	12.0	15.0	19.0	18.0	9.0	7.7	17.4	18.0	18.3	31.0	13.7
	Meditron3-Gemma2-9B	12.7	10.3	16.3	14.3	5.7	10.3	14.6	14.0	20.3	24.0	13.3
	Meditron3-Qwen2.5-14B	13.1	14.3	25.0	17.3	10.0	8.3	18.7	22.7	15.3	29.0	10.0
	Meditron3-Phi4-14B	14.2	11.7	20.0	22.0	7.0	9.3	19.2	28.7	22.0	28.0	17.3
	Meditron3-70B	13.9	12.3	25.0	21.7	10.3	8.7	19.6	24.7	6.3	44.0	13.7
MedGemma	medgemma-4b-it	15.0	15.3	20.0	19.3	6.7	10.3	21.0	20.3	16.3	35.3	16.3
	medgemma-27b-text-it	22.8	15.7	27.0	32.7	8.0	8.0	22.4	32.0	18.3	36.0	28.7
Huatuotuo	HuatuotuoGPT-o1-7B	12.7	12.3	25.7	18.3	4.7	11.3	15.1	26.7	20.3	22.0	18.3
	HuatuotuoGPT-o1-8B	19.5	10.3	24.0	29.3	10.7	7.7	21.0	23.3	20.0	15.7	24.7
	HuatuotuoGPT-o1-70B	30.3	16.7	33.0	50.7	21.7	17.3	27.9	32.0	17.3	34.0	38.3
	HuatuotuoGPT-o1-72B	33.0	15.7	36.0	46.0	21.0	13.3	31.1	36.0	17.3	37.0	39.3

Table 3: Performance across benchmarks, showing the accuracy of models compared to a random baseline (values below and above the baseline). The model categories are general-purpose, reasoning, domain-specific, and domain-specific reasoning. Results are averaged over three runs with different random seeds to ensure robustness.

6 Discussion

Experimental results indicate that strong performance on complex medical reasoning tasks requires more than model scale or advanced prompting. While larger models generally perform better, success critically depends on reasoning-specific fine-tuning (see Section 5.1 and Table 3). Even frontier models without such adaptation, like DeepSeek V3 and GPT-4o, often fail to exceed ran-

dom baselines, demonstrating that scale alone is insufficient (see Section 5.2 and Table 5). Similarly, prompting strategies provide only limited benefit in encouraging exploration of alternative reasoning paths. Models exhibiting low variability, including OLMo and MiMo, frequently converge on common but incorrect answers, reflecting susceptibility to cognitive biases. Models exhibiting higher variability, such as Huatuotuo, can occasionally achieve improved outcomes, but their overall performance

Family	Prompt	MedAgentsBench									MedARC-QA	MetaMedQA
		MedBullets	MedExQA	MedMCQA	MedQA	MedXpert-R	MedXpert-U	MMLU	MMLU-Pro	PubMedQA		
random-baseline		17.2	24.3	29.7	23.0	11.0	9.7	29.2	11.7	30.0	23.0	18.7
Llama-3.3-70B-Instruct	Base	13.5	9.3	16.0	16.3	11.0	12.7	18.7	22.0	14.0	41.7	11.0
	CoT	16.5	14.7	26.7	27.7	12.7	13.3	23.7	29.7	15.3	37.0	19.7
	IR	21.7	14.0	22.3	28.0	10.3	14.0	23.7	32.0	15.7	37.0	22.7
	AP	15.0	12.3	19.7	20.7	8.7	12.0	18.3	30.0	13.7	34.0	18.0
	FS	23.6	11.0	21.0	22.0	12.0	16.0	17.8	28.0	16.7	42.7	14.7
gemma-3-27b-it	MP	18.7	14.0	21.3	25.7	12.7	11.7	19.6	30.7	16.0	39.7	20.3
	Base	9.4	13.7	25.7	19.0	8.0	7.7	20.5	29.0	13.3	24.0	16.3
	CoT	12.4	13.3	21.3	24.3	6.7	6.3	21.0	34.7	16.3	27.7	20.3
	IR	6.7	12.0	20.0	25.3	6.3	6.0	22.4	33.7	11.3	27.3	20.0
	AP	11.6	14.0	22.7	23.7	7.7	5.7	19.6	29.0	11.0	28.0	16.0
phi-4	FS	13.9	11.7	15.7	31.0	7.3	8.7	21.0	26.3	10.7	26.7	24.7
	MP	16.1	13.7	24.0	24.3	6.7	6.7	23.3	32.0	13.3	29.3	22.3
	Base	13.9	7.0	20.7	15.3	8.0	11.0	22.8	30.3	21.3	18.3	12.3
	CoT	16.1	13.0	27.7	22.3	8.7	8.7	22.8	28.3	18.7	24.0	20.7
	IR	19.1	9.3	22.0	17.7	9.7	7.0	21.5	27.7	19.0	16.3	14.0
OLMo-2-0325-32B-Instruct	AP	13.1	12.3	22.3	22.3	11.7	10.0	24.2	26.3	16.0	17.0	17.0
	FS	16.5	11.0	24.7	19.3	8.3	11.3	19.6	29.3	17.7	23.3	18.0
	MP	13.5	13.3	20.0	24.0	9.7	9.3	25.1	27.7	16.7	22.7	21.0
	Base	10.5	13.3	20.0	19.7	6.7	4.7	9.1	23.0	30.3	15.3	12.0
	CoT	10.5	12.3	22.7	18.3	8.0	4.3	12.8	25.0	30.7	15.3	13.7
Qwen3-32B non-reasoning	IR	11.2	13.7	21.3	19.7	6.0	4.3	10.0	24.0	29.0	18.7	12.3
	AP	16.9	13.0	19.7	19.7	8.7	5.3	18.3	26.0	21.0	12.7	11.7
	FS	10.9	11.3	17.7	15.3	6.7	8.3	14.2	26.7	21.7	18.3	14.3
	MP	13.9	10.3	23.0	18.7	8.7	6.0	13.2	27.7	20.7	18.0	12.3
	Base	11.2	12.7	24.3	21.0	11.0	12.7	20.1	21.7	19.0	52.7	12.0
cogito-v1-preview-llama-70B non-reasoning	CoT	15.0	12.7	19.0	27.7	12.3	10.3	21.5	29.7	18.0	34.3	23.7
	IR	12.7	13.7	22.3	24.3	10.0	14.3	23.3	27.0	19.7	46.3	15.3
	AP	18.0	12.3	20.3	23.3	14.3	9.7	21.0	33.7	19.7	34.7	21.3
	FS	14.2	10.3	21.0	24.0	11.0	13.3	22.8	31.3	16.0	37.7	22.0
	MP	16.9	11.3	20.0	25.0	10.7	9.0	21.9	33.0	5.0	43.0	23.3
DeepSeek-R1-Distill-Llama-70B	Base	15.4	16.3	21.0	21.3	11.0	9.0	35.6	34.0	12.7	33.0	18.0
	CoT	18.0	12.0	19.0	27.3	10.7	9.3	46.6	34.0	16.3	41.3	20.3
	IR	15.4	17.0	16.0	23.3	11.0	11.7	43.8	35.0	14.0	47.0	18.7
	AP	17.6	12.0	20.3	26.0	10.3	11.3	35.6	34.3	11.7	37.3	19.0
	FS	18.0	13.7	24.7	25.0	12.3	12.3	42.5	35.0	15.7	36.7	22.0
MiMo-7B-SFT	MP	17.6	15.3	19.3	28.0	10.7	10.7	43.4	37.0	13.0	48.0	19.7
	Base	24.0	13.7	24.7	33.3	14.7	12.7	28.8	26.3	13.7	36.0	19.0
	CoT	25.5	14.7	23.3	29.0	16.3	12.0	24.7	26.7	15.0	37.7	22.3
	IR	24.7	15.7	27.7	31.0	16.0	16.0	24.7	27.0	14.7	36.3	23.3
	AP	26.2	16.0	28.3	34.0	15.0	13.0	23.7	26.3	13.7	34.7	24.0
Phi-4-reasoning	FS	21.7	15.7	29.0	29.3	14.3	16.0	24.7	29.7	11.0	35.3	21.7
	MP	32.6	13.7	24.7	35.7	17.3	14.3	26.5	30.0	11.7	39.7	24.0
	Base	13.1	12.3	23.3	18.0	4.3	3.3	16.4	12.0	16.7	17.0	10.7
	CoT	7.1	11.3	21.3	16.0	3.7	3.3	17.4	9.3	14.3	15.3	15.0
	IR	8.6	10.7	21.0	14.3	5.3	2.7	16.9	12.3	18.7	13.7	13.7
Qwen3-32B reasoning	AP	8.6	12.0	18.7	19.3	4.0	6.0	16.9	13.3	15.3	14.7	11.7
	FS	9.4	10.7	21.0	15.7	5.3	3.0	14.6	12.3	14.0	16.7	9.3
	MP	9.4	8.7	20.3	16.0	5.7	2.0	16.0	12.0	18.3	23.0	12.0
	Base	28.8	12.3	23.3	33.3	12.0	14.0	31.1	31.7	26.0	36.7	26.3
	CoT	24.0	10.3	26.7	36.0	13.0	14.3	31.5	31.7	22.3	35.3	27.7
cogito-v1-preview-llama-70B reasoning	IR	24.7	14.0	25.0	30.3	10.0	13.7	29.7	30.7	25.7	39.3	28.7
	AP	25.1	12.3	22.3	30.0	12.3	15.0	24.2	28.0	24.7	31.3	23.7
	FS	23.2	12.7	23.3	23.3	12.3	12.7	30.1	29.7	22.3	37.3	19.0
	MP	28.8	14.7	23.0	35.3	13.7	17.0	33.3	31.3	26.3	37.7	27.7
	Base	23.2	12.7	24.3	30.3	19.0	9.3	26.0	31.3	18.0	45.0	25.0
Qwen2.5-Aloe-Beta-72B	CoT	27.7	13.7	22.0	35.7	18.0	12.0	26.0	30.0	17.0	39.3	27.3
	IR	29.2	13.0	21.3	36.3	16.3	12.3	26.0	29.7	17.7	39.7	31.0
	AP	25.5	13.0	17.0	33.7	16.3	13.3	25.1	29.7	17.7	39.0	26.0
	FS	24.0	13.3	23.0	33.0	16.7	15.3	28.8	34.7	15.0	35.7	24.3
	MP	32.6	15.7	24.0	43.7	18.0	12.7	30.6	33.3	8.3	41.3	34.0
Meditron3-70B	Base	21.3	14.0	23.0	32.3	11.0	10.0	38.4	34.7	16.0	24.0	24.0
	CoT	20.6	14.7	23.0	34.7	10.3	11.0	44.3	35.0	15.3	29.7	25.0
	IR	17.6	15.7	22.3	30.3	9.7	9.7	38.4	34.0	17.0	28.7	30.3
	AP	19.5	13.0	22.3	27.7	12.3	10.3	29.2	34.7	15.3	29.7	22.7
	FS	20.2	13.3	26.3	35.3	10.0	10.3	38.8	34.0	16.3	33.7	24.0
medgemma-27b-text-it	MP	16.9	13.7	19.3	30.3	10.3	11.3	39.3	33.0	15.7	33.7	24.3
	Base	12.4	20.3	26.3	16.3	8.3	11.0	25.6	29.3	14.0	55.7	13.3
	CoT	15.4	17.3	23.3	21.3	10.7	8.7	18.7	24.7	17.3	30.3	17.0
	IR	15.0	18.3	23.7	24.0	10.7	13.3	25.1	26.7	18.3	38.7	16.0
	AP	15.4	16.7	24.3	18.0	12.0	9.3	22.8	18.7	20.0	32.7	18.0
HuatioGPT-o1-72B	FS	25.1	19.0	29.7	27.0	8.3	13.7	22.4	31.3	21.0	41.7	22.7
	MP	18.0	14.7	24.3	23.3	9.7	8.7	20.1	26.7	14.7	37.3	17.0
	Base	13.9	12.3	25.0	21.7	10.3	8.7	19.6	24.7	6.3	44.0	13.7
	CoT	16.1	10.0	20.0	23.3	5.3	10.7	21.0	27.3	5.7	28.7	19.3
	IR	13.5	10.3	21.7	20.0	8.3	10.7	19.6	29.3	5.7	22.3	18.3
HuatuoGPT-o1-72B	AP	15.0	12.7	18.7	18.3	6.7	6.0	12.8	14.7	5.3	35.7	11.7
	FS	19.5	10.3	24.0	23.0	10.3	12.3	26.9	25.0	11.7	50.0	21.3
	MP	14.2	12.7	21.0	23.0	8.3	11.0	26.0	27.3	6.3	40.3	15.3
	Base	22.8	15.7	27.0	32.7	8.0	8.0	22.4	32.0	18.3	36.0	28.7
	CoT	5.6	11.7	22.3	11.7	6.7	5.7	20.1	29.0	16.7	20.0	14.0
HuatuoGPT-o1-72B	IR	16.1	12.0	24.7	25.3	6.0	8.3	21.5	30.7	13.7	34.7	25.0
	AP	27.7	13.0	28.7	34.3	14.0	10.3	22.4	31.3	18.7	34.0	27.3
	FS	24.3	11.7	23.3	39.7	11.3	11.3	23.7	33.0	14.3	32.0	32.3
	MP	1.1	11.7	25.3	7.7	4.0	4.7	11.9	22.3	9.3	23.7	6.0
	Base	33.0	15.7	36.0	46.0	21.0	13.3	31.1	36.0	17.3	37.0	39.3
HuatuoGPT-o1-72B	CoT	37.1	15.0	38.7	47.0	20.0	11.7	33.3	32.3	19.0	36.0	39.3
	IR	33.3	16.0	39.0	48.7	21.3	10.7	34.7	36.0	19.0	35.7	36.7
	AP	33.0	20.0	39.3	46.7	19.7	15.0	34.7	36.0	17.3	36.7	38.7
	FS	33.0	17.0	39.0	46.0	19.0	11.7	31.5	33.3	16.3	40.3	39.0
	MP	33.3	16.3	39.3	50.7	18.3	15.0	31.1	32.7	15.7	39.3	42.7

Table 4: Performance across benchmarks, showing the accuracy of one model per family under different prompting strategies, compared to a random baseline (values below and above the baseline). The model categories are general-purpose, reasoning, domain-specific, and domain-specific reasoning. Results are averaged over three runs with different random seeds to ensure robustness.

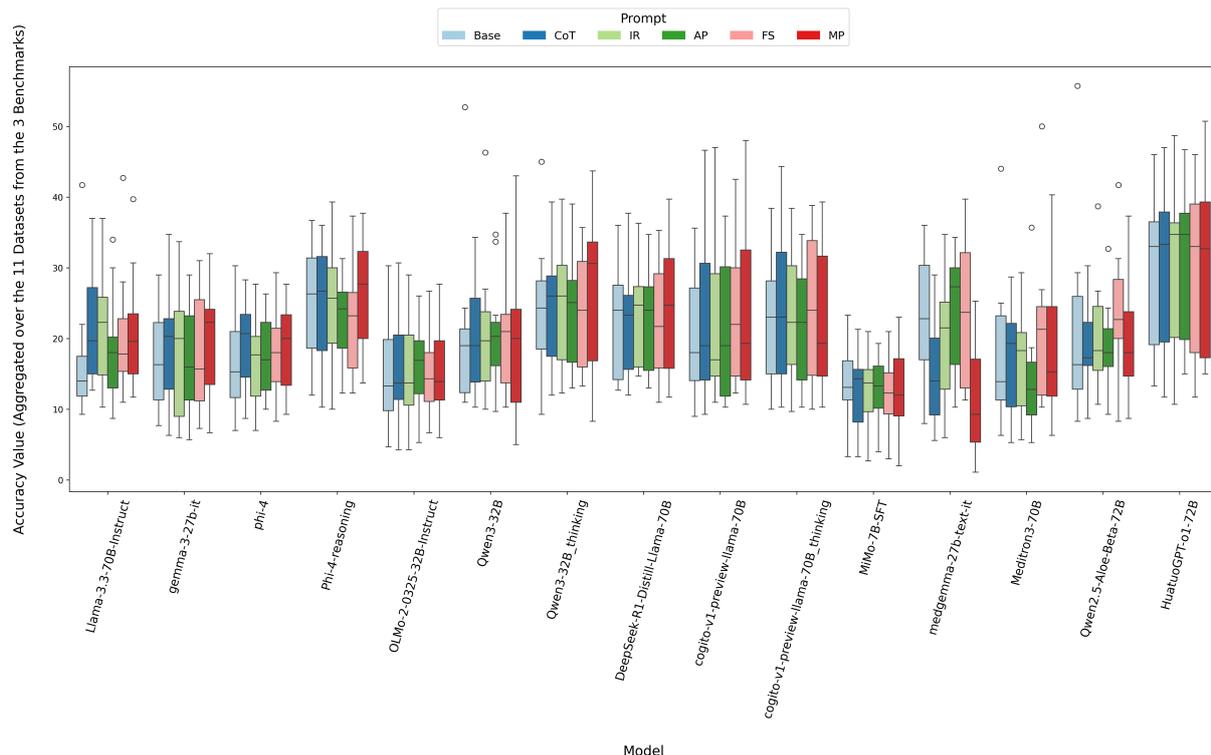


Figure 2: Model accuracy across the three benchmarks, averaged over three random seeds and grouped by prompt.

Family	Model	MedAgentsBench									MedARC-QA	MetaMedQA
		MedBullets	MedExQA	MedMCQA	MedQA	MedXpert-R	MedXpert-U	MMLU	MMLU-Pro	PubMedQA		
random	random-baseline	17.2	24.3	29.7	23.0	11.0	9.7	29.2	11.7	30.0	23.0	18.7
DeepCogito	cogito-v1-preview-llama-70B	15.4	16.3	21.0	21.3	11.0	9.0	35.6	34.0	12.7	33.0	18.0
Phi	Phi-4-reasoning	28.8	12.3	23.3	33.3	12.0	14.0	31.1	31.7	26.0	36.7	26.3
MedGemma	medgemma-27b-text-it	22.8	15.7	27.0	32.7	8.0	8.0	22.4	32.0	18.3	36.0	28.7
Huatuo	HuatuoGPT-o1-72B	30.3	17.0	36.0	41.0	22.0	13.0	30.1	33.0	22.0	36.0	40.0
Llama	llama-4-maverick	29.2	25.0	43.0	39.0	26.0	26.0	39.7	38.0	15.0	45.0	34.0
DeepSeek	deepseek-chat-v3-0324	22.5	12.0	22.0	31.0	11.0	12.0	21.9	34.0	17.0	45.0	26.0
	deepseek-r1-0528	44.9	22.0	28.0	54.0	25.0	22.0	34.2	34.0	19.0	56.0	48.0
Anthropic	claude-sonnet-4	38.2	26.0	40.0	53.0	20.0	25.0	45.2	44.0	17.0	53.0	40.0
Gemini	gemini-2.5-pro	48.3	23.0	40.0	59.0	34.0	26.0	38.4	34.0	16.0	78.0	52.0
OpenAI	gpt-4o-2024-08-06	25.8	16.0	24.0	45.0	8.0	13.0	30.1	34.0	18.0	34.0	32.0
	o3-2025-04-16	70.8	31.0	49.0	78.0	46.0	37.0	46.6	44.0	23.0	68.0	64.0

Table 5: Performance comparison across benchmarks, focusing on the accuracy of the best open-source medium-sized models for type against frontier models. Each model is compared to a random baseline (values below and above the baseline). The model categories are general-purpose, reasoning, domain-specific, and domain-specific reasoning.

generally remains low (see Section 5.1, Table 4 and Figure 2).

Fine-grained evaluation, conducted in this work for the first time, reveals an intrinsic limitation of LLMs. The results indicate that models encounter their greatest challenges with inductive reasoning (see Section 5.3 and Figure 3), an inference process that "uses specific examples to draw general conclusions" (Sooknanan and Seemungal, 2019), identifies patterns in observations and measurements to infer generalized rules, and can "increase human knowledge, leading to new theories" (Sooknanan

and Seemungal, 2019). In contrast, LLMs tend to perform better in abduction, the "faculties of making informed guesses and selections among those [...]. Correct guesses are often informed by experience" (Pietarinen and Stanley, 2025), and in deduction, which "refines hypotheses against established medical knowledge" (Pietarinen and Stanley, 2025). These findings are consistent with previous studies in medical reasoning, suggesting that LLMs rely heavily on parametric memory and often struggle to produce truly novel reasoning pathways (Soffer et al., 2024; Kim et al., 2025). Also with studies on

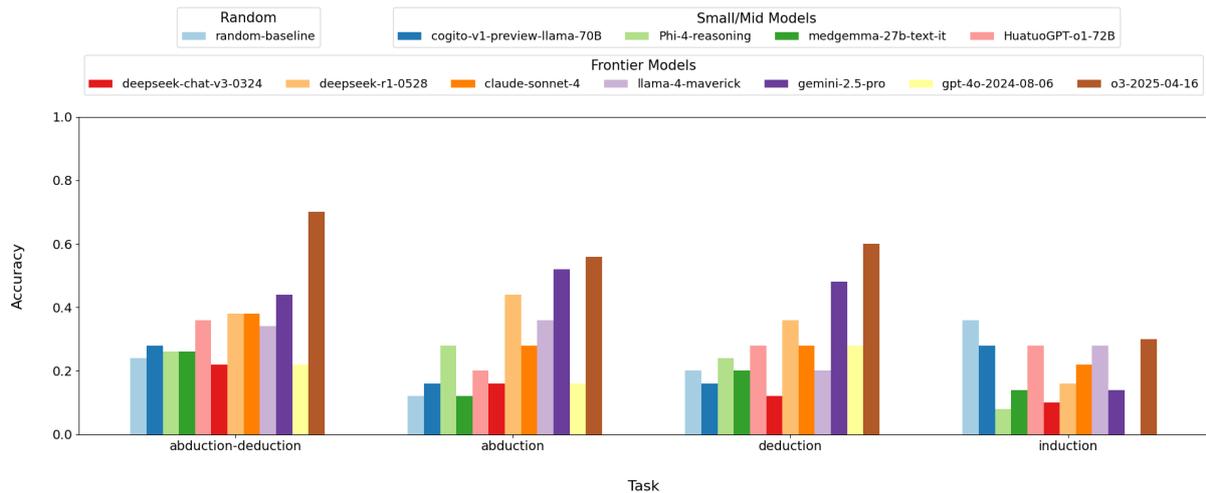


Figure 3: Model performance across reasoning types is shown, with the random baseline indicating subtask difficulty, which can vary according to the number of answer choices per question.

induction (Bowen et al., 2024), where LLM capabilities are assessed across multiple symbolic tasks in more general settings.

Although the evaluation of different reasoning types in LLMs have recently received attention (Mondorf and Plank, 2024; Sun et al., 2025), systematic assessments in complex medical reasoning remain limited. This study presents a structured methodology to identify potential limitations and highlight areas for improvement. Moreover, our experimental framework can function as a continuously updated resource for monitoring LLM progress in medical reasoning.

7 Conclusions

To our knowledge, this study is the first large-scale LLM evaluation on complex medical reasoning, spanning across 77 models of diverse size, tuning methods, and prompting strategies. Guided by the medical problem-solving literature, we selected three benchmarks addressing advanced knowledge and reasoning, robustness to cognitive biases, and metacognitive capabilities. Performance was generally low, often below random choice, with modest improvements only in frontier models fine-tuned for reasoning. We manually annotated a dataset capturing key reasoning types for medical diagnosis (abduction, deduction, induction), which have received limited prior attention, revealing persistent deficiencies across all types, particularly in inductive reasoning. This highlights the need for models that better approximate physician strategies.

Given that reasoning-focused fine-tuning most

effectively enhances performance, future work should develop specialized datasets targeting the identified reasoning types, structured to mirror physician problem-solving strategies and incorporating design principles that must mitigate susceptibility to cognitive biases and metacognitive limitations. Evaluation against human experts can contextualize the gap between current models and expert-level reasoning, capturing both quantitative and qualitative aspects. Additional research directions include examining relationships between model parameters and reasoning types (Christ et al., 2024), as well as optimizing training strategies to guide models along correct reasoning paths (Zhang et al., 2024; Ferrazzi et al., 2025), particularly for reasoning chains.

Limitations

It is important to acknowledge several limitations that define the scope and interpretation of our findings. First, our conclusions are drawn from three specific medical benchmarks which, while designed to test complex reasoning, cognitive bias, and metacognitive ability, cannot fully capture the breadth or nuance of real-world clinical scenarios. Second, our evaluation is limited to a defined set of open-source and proprietary models, with frontier models accessed as black boxes via their APIs, restricting transparency into their architecture, training data, or internal reasoning processes. Third, reliance on static, exam-style multiple-choice questions oversimplifies the inherently dynamic, context-rich, and interactive nature

of actual clinical decision-making. Fourth, our study is confined to English-language datasets, and findings may not generalize to other languages or healthcare systems. Lastly, although we applied the MELD framework to detect potential data contamination, its recall remains unknown; therefore, the absence of detected overlap does not definitively rule out prior exposure to the benchmark data during model training.

Ethical considerations

This study evaluates publicly available language models using benchmark datasets tailored to assess clinical reasoning, such as MedAgentsBench, MedARC-QA, and MetaMedQA. All benchmarks are constructed from de-identified, publicly released data sources and do not contain personally identifiable health information. While our analysis relies on open benchmarks, the demographics of these datasets across dimensions such as age, gender, and ethnicity has not yet been systematically assessed. Given the relevance of these factors in medical contexts, future work should examine potential biases and disparities within these datasets to ensure equitable evaluation. Although the models we study are publicly accessible, particularly for proprietary frontier models, we cannot fully verify the nature of the data used during pre-training due to limited transparency regarding their training corpora. This lack of visibility presents challenges for assessing fairness, safety, and potential data contamination. We have followed responsible computing practices throughout this work, including rigorous checks for data leakage, the use of multiple random seeds for robustness, and the public release of our code to promote reproducibility. Finally, we acknowledge the broader societal implications of deploying large language models in high-stakes domains such as healthcare. This work does not advocate for the unsupervised use of LLMs in clinical practice. Rather, it aims to contribute to the responsible development of AI systems by identifying current limitations in clinical reasoning and underscoring the need for careful stewardship and continuous oversight in medical applications.

Acknowledgments

The authors gratefully acknowledge support from Villanova (financed by IPICEI-CIS, Prog. n. SA. 102519 - CUP B29J24000850005), the EU project LLMs4EU (DIGITAL-20234-AI-06-LANGUAGE-

01), IKUR HPC&AI project LOGOS, Ministerio de Ciencia, Innovación y Universidades, DeepThought project (PID2024-159202OB-C21) and the Ministerio para la Transformación Digital y de la Función Pública - Funded by EU - NextGenerationEU within the framework of the project *Desarrollo de Modelos ALIA*.

References

- Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Singh Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio Cesar Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, and 4 others. 2025. [Phi-4-reasoning technical report](#). *ArXiv*, abs/2504.21318.
- Marah Abdin, Jyoti Aneja, Harkirat Singh Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio Cesar Teodoro Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. [Phi-4 technical report](#). *ArXiv*, abs/2412.08905.
- Marcel Binz and Eric Schulz. 2021. [Reconstructing the einstellung effect](#). *Computational Brain & Behavior*, 6:526–542.
- Chen Bowen, Rune Sætre, and Yusuke Miyao. 2024. [A comprehensive evaluation of inductive reasoning capabilities and problem solving in large language models](#). In *Findings*.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. [Sparks of artificial general intelligence: Early experiments with gpt-4](#). *ArXiv*, abs/2303.12712.
- Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2024. [A survey on mixture of experts in large language models](#). *IEEE Transactions on Knowledge and Data Engineering*, 37:3896–3915.
- Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024a. [Benchmarking large language models on answering and explaining challenging medical questions](#). *ArXiv*, abs/2402.18060.
- Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025a. [Benchmarking large language models on answering and explaining challenging medical questions](#). In *Proceedings of the 2025 Conference*

- of the Nations of the Americas Chapter of the Association for Computational Linguistics: *Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.
- Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025b. [Benchmarking large language models on answering and explaining challenging medical questions](#). *Preprint*, arXiv:2402.18060.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024b. [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. 2024c. [Cod, towards an interpretable medical agent using chain of diagnosis](#). *ArXiv*, abs/2407.13301.
- Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. [Meditron-70b: Scaling medical pretraining for large language models](#). *Preprint*, arXiv:2311.16079.
- Bryan R Christ, Zack Gottesman, Jonathan Kropko, and Thomas Hartvigsen. 2024. [Math neurosurgery: Isolating language models’ math reasoning abilities using only forward passes](#). *arXiv preprint arXiv:2410.16930*.
- Salvatore Corrao and Christiano Argano. 2022. [Rethinking clinical decision-making to improve clinical reasoning](#). *Frontiers in Medicine*, 9:900543.
- DeepCogito Research. 2025. [Cogito v1 preview](https://www.deepcogito.com/research/cogito-v1-preview). <https://www.deepcogito.com/research/cogito-v1-preview>.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *ArXiv*, abs/2501.12948.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bing-Li Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 179 others. 2024. [Deepseek-v3 technical report](#). *ArXiv*, abs/2412.19437.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. [The llama 3 herd of models](#). *ArXiv*, abs/2407.21783.
- Alexander Viktor Eriksen, Sören Möller, and Jesper Ryg. 2023. [Use of gpt-4 to diagnose complex clinical cases](#). *NEJM AI*.
- Y. Esteitieh, S. Mandal, and G. Laliotis. 2025. [Towards metacognitive clinical reasoning: Benchmarking mdpie against state-of-the-art llms in medical decision-making](#). In *medRxiv*.
- Pietro Ferrazzi, Aitor Soroa, and Rodrigo Agerri. 2025. [Grounded multilingual medical reasoning for question answering with large language models](#). *arXiv preprint arXiv:2512.05658*.
- M Fieschi and J Gouvernet. 1999. [Reasoning foundations of medical diagnosis revisited](#). *Yearbook of Medical Informatics*, 8(01):78–82.
- Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrián Tormos, Daniel Hinjos, Pablo Bernabeu Perez, Anna Arias-Duart, Pablo A. Martin-Torres, Marta Gonzalez-Mallo, Sergio Álvarez-Napagao, Eduard Ayguad’e-Parra, and Ulises Cort’es. 2025. [The aloe family recipe for open and specialized healthcare llms](#).
- Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Maite Oronoz, and Rodrigo Agerri. 2023. [Explanatory argument extraction of correct answers in resident medical exams](#). *Preprint*, arXiv:2312.00567.
- Google DeepMind. 2025. [Medgemma: Advanced ai models for medical text and image analysis](https://medgemma.org/). <https://medgemma.org/>.
- Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025. [Large language models lack essential metacognition for reliable medical reasoning](#). *Nature Communications*, 16.
- Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel M Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer F Braren, Georgios Kaissis, and Daniel Rueckert. 2024. [Evaluation and mitigation of the limitations of large language models in clinical decision-making](#). *Nature Medicine*, 30:2613 – 2622.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. [Measuring massive multitask language understanding](#). *arXiv preprint arXiv:2009.03300*.
- Shengxin Hong, Liang Xiao, Xin Zhang, and Jian-Xing Chen. 2024. [Argmed-agents: Explainable clinical decision reasoning with llm discussion via argumentation schemes](#). *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5486–5493.

- OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, and 397 others. 2024. [Gpt-4o system card](#). *ArXiv*, abs/2410.21276.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.
- Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'è, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, and 191 others. 2025. [Gemma 3 technical report](#). *ArXiv*, abs/2503.19786.
- Zahir Kanjee, Byron Crowe, and Adam Rodman. 2023. [Accuracy of a generative artificial intelligence model in a complex diagnostic challenge](#). *JAMA*.
- Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. 2025. [Limitations of large language models in clinical problem-solving arising from inflexible reasoning](#). *ArXiv*, abs/2502.04381.
- Yu Han Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024a. [Mdagents: An adaptive collaboration of llms for medical decision-making](#). In *Neural Information Processing Systems*.
- Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024b. [Medexqa: Medical question answering benchmark with multiple explanations](#). *arXiv preprint arXiv:2406.06331*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Robert S Ledley and Lee B Lusted. 1959. Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science*, 130(3366):9–21.
- Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. [Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning](#). In *Neural Information Processing Systems*.
- Zheqing Li, Yiyang Yang, Jiping Lang, Wenhao Jiang, Yuhang Zhao, Shuang Li, Dingqian Wang, Zhu Lin, Xuanna Li, Yuze Tang, Jiexian Qiu, Xiaolin Lu, Hongji Yu, Shuang Chen, Yuhua Bi, Xiaofei Zeng, Yixian Chen, Junrong Chen, and Lin Yao. 2025a. [Gpbench: A comprehensive and fine-grained benchmark for evaluating large language models as general practitioners](#). *ArXiv*, abs/2503.17599.
- Zhongzhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025b. [From system 1 to system 2: A survey of reasoning large language models](#). *ArXiv*, abs/2502.17419.
- Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei A. Clifton, and David A. Clifton. 2024. [Large language models are poor clinical decision-makers: A comprehensive benchmark](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Abraham S Luchins. 1942. Mechanization in problem solving: The effect of einstellung. *Psychological monographs*, 54(6):i.
- Arjun R. Malghan. 2025. [Evaluating computational accuracy of large language models in numerical reasoning tasks for healthcare applications](#). *ArXiv*, abs/2501.13936.
- Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak N. Patel, Dale R. Webster, and 9 others. 2023. [Towards accurate differential diagnosis with large language models](#). *ArXiv*, abs/2312.00164.
- Nikita Mehandru, Brenda Y Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul Janardhan Butte, and Ahmed Alaa. 2024. [Evaluating large language models as agents in the clinic](#). *NPJ Digital Medicine*, 7.
- Philipp Mondorf and Barbara Plank. 2024. [Beyond accuracy: Evaluating the reasoning behavior of large language models – a survey](#). *Preprint*, arXiv:2404.01869.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. [Capabilities of gpt-4 on medical challenge problems](#). *ArXiv*, abs/2303.13375.

- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2024. [olmo 2 furious](#). *ArXiv*, abs/2501.00656.
- OpenAI. 2025. [Openai o3 and o4-mini system card](#).
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.
- Ahti-Veikko Pietarinen and Donald E Stanley. 2025. The logic of medical reasoning: toward an integrated inductive, deductive, and abductive approach to clinical practices. *Philosophy, Ethics, and Humanities in Medicine*, 20(1):16.
- Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Weike Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. [Quantifying the reasoning abilities of llms on real-world clinical cases](#). *ArXiv*, abs/2503.04691.
- Sumant R. Ranji. 2024. [Large language models-misdiagnosing diagnostic excellence?](#) *JAMA network open*, 7 10:e2440901.
- Justin T. Reese, Daniel Danis, J. Harry Caufield, Tudor Groza, Elena Casiraghi, Giorgio Valentini, Chris J. Mungall, and Peter N. Robinson. 2023. [On the limitations of large language models in clinical diagnosis](#). *medRxiv*.
- Thomas Savage, Ashwin Nayak, Roberta Gallo, Ekanath Srihari Rangan, and Jonathan H. Chen. 2023. [Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine](#). *NPJ Digital Medicine*, 7.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguerre y Arcas, Dale Webster, and 11 others. 2022. [Large language models encode clinical knowledge](#). *Preprint*, arXiv:2212.13138.
- S. Soffer, V. Sorin, G. N. Nadkarni, and E. Klang. 2024. [Chatgpt-01 and the pitfalls of familiar reasoning in medical ethics](#). In *medRxiv*.
- Joanna Sooknunan and Terence Seemungal. 2019. Not so elementary—the reasoning behind a medical diagnosis. *MedEdPublish*, 8:234.
- Gaurav Srivastava, Shuxiang Cao, and Xuan Wang. 2025. [Towards reasoning ability of small language models](#). *ArXiv*, abs/2502.11569.
- Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, and 1 others. 2025. [A survey of reasoning with foundation models: Concepts, methodologies, and outlook](#). *ACM Computing Surveys*, 57(11):1–43.
- Xiangru Tang, Yanjun 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](#). *ArXiv*, abs/2503.07459.
- Hao Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024a. [Beyond direct diagnosis: Llm-based multi-specialist agent consultation for automatic diagnosis](#). *ArXiv*, abs/2401.16107.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 others. 2024b. [Mmlu-pro: A more robust and challenging multi-task language understanding benchmark](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. [Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration](#). In *North American Chapter of the Association for Computational Linguistics*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. [Emergent abilities of large language models](#). *ArXiv*, abs/2206.07682.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022b. [Chain of thought prompting elicits reasoning in large language models](#). *ArXiv*, abs/2201.11903.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Xi Xia, Bowen Shen, Cici, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, Liang Zhao, Peidian Li, Peng Wang, Shi jie Yu, Shimao Chen, Weikun Wang, Wenhan Ma, Xia Deng, Yi Huang, and 44 others. 2025. [Mimo: Unlocking the reasoning potential of language model – from pretraining to posttraining](#).

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. [Towards large reasoning models: A survey of reinforced reasoning with large language models](#). *ArXiv*, abs/2501.09686.

Chenwei Yan, Xiangling Fu, Yuxuan Xiong, Tianyi Wang, Siu Cheung Hui, Ji Wu, and Xien Liu. 2025a. [Llm sensitivity evaluation framework for clinical diagnosis](#). In *International Conference on Computational Linguistics*.

Yang Yan, Yu Lu, Renjun Xu, and Zhenzhong Lan. 2025b. [Do phd-level llms truly grasp elementary addition? probing rule learning vs. memorization in large language models](#).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#).

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed Huai hsin Chi, and Denny Zhou. 2023. [Large language models as analogical reasoners](#). *ArXiv*, abs/2310.01714.

Mingze Yuan, Peng Bao, Jiajia Yuan, Yunhao Shen, Zi Chen, Yi Xie, Jie Zhao, Yang Chen, Li Zhang, Lin Shen, and Bin Dong. 2023. [Large language models illuminate a progressive pathway to artificial healthcare assistant: A review](#). *ArXiv*, abs/2311.01918.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024. [Chain of preference optimization: Improving chain-of-thought reasoning in llms](#). *Advances in Neural Information Processing Systems*, 37:333–356.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. [A survey of large language models](#). *ArXiv*, abs/2303.18223.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. [Large language models are not robust multiple choice selectors](#). *ArXiv*, abs/2309.03882.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, Zheng Li, and Fenglin Liu. 2023. [A survey of large language models in medicine: Progress, application, and challenge](#). *ArXiv*, abs/2311.05112.

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](#). *arXiv preprint arXiv:2501.18362*.

A Data Annotation

This appendix provides a detailed description of the process applied to annotate a subset of the medical benchmarks used in this work, evaluating models’ performance in the reasoning strategies followed by physicians.

Annotation criteria Medical problem-solving is a complex, multi-step process that typically involves the interplay of three reasoning types: abduction, deduction, and induction (Pietarinen and Stanley, 2025; Sooknanan and Seemungal, 2019). Abduction is used to generate hypotheses about potential causes of observed symptoms. These hypotheses then guide the reasoning process, requiring verification through deduction, for example, selecting the appropriate tests to confirm or refute the initial abductive hypotheses. Deduction applies general rules to relate causes and effects, reasoning downward to predict specific outcomes. Induction, in contrast, derives general rules from observations, identifying patterns to form broader conclusions. Inductive reasoning is particularly important in medicine, an empirical science that constantly deals with uncertainty and probabilistic outcomes. Guided by these definitions, we manually selected questions from the benchmarks adopted in this work that explicitly address these tasks. Therefore, we used a specific medical task as a proxy for the reasoning strategy required to solve it, allowing us to annotate these examples according to their reasoning category.

Annotation details Taking into account the types of medical tasks required to isolate reasoning strategies, we identified two suitable datasets: MedBullets and PubMedQA. MedBullets focuses on diagnosis and treatment decision-making, while PubMedQA contains questions designed to draw conclusions from biomedical studies. Accordingly, we expected tasks in MedBullets to require abduction and deduction, and tasks in PubMedQA to rely on induction. Additionally, both datasets include explanations or rationales manually provided by physicians, offering clear guidance on the reasoning strategies necessary to solve each task and serving as a reliable reference and validation for our annotations. In MedBullets, 42 of 89 questions referenced missing figures, which were poten-

Category	Dataset	ID	Question	Options	Answer	Reason/Rationale
abduction-deduction	MedBullets	20	A 4-week-old boy is brought to the pediatrician by his parents for an initial evaluation. His parents are concerned that he is not feeding well and has lost weight over the last 2 weeks. He was born at home via spontaneous vaginal delivery at 37 weeks of gestation to a G1P1 mother. His mother did not receive prenatal care. His blood pressure is 70/44 mmHg, pulse is 151/min, and respirations are 41/min. His weight is 3 kg (4th percentile) compared with 3.5 kg at birth (45th percentile). On physical exam, he is jaundiced with an enlarged liver and spleen. He does not appear to respond to visual stimuli, and further examination reveals bilateral clouding of the lens. Which of the following interventions could have avoided this patient's symptoms?	["Avoiding fruit juice and sweetened foods", "Changing to a soy based formula", "Providing imiglucerase enzyme replacement", "Removing phenylalanine from maternal diet during pregnancy", "Vitamin B6 supplementation"]	Changing to a soy based formula	<p>This patient who presents with failure to thrive, hepatosplenomegaly, and bilateral cataracts most likely has classic galactosemia.</p> <p>Patients with this disorder should avoid lactose-containing products by changing to a soy-based formula.</p> <p>Classic galactosemia is an autosomal recessive defect in galactose-1-phosphate uridylyltransferase. This enzyme is involved in the conversion of galactose to glucose, and a deficiency of this enzyme results in the accumulation of galactose 1-phosphate in the liver, kidney, and brain. This metabolite acts as a phosphate sink, meaning that it traps all free phosphate in the cytosol and inhibits the formation of other phosphate-dependent metabolites such as adenosine triphosphate. This results in failure to thrive, vomiting, lethargy, and hepatosplenomegaly after ingestion of galactose-containing products such as milk. Patients will also develop bilateral cataracts if the deficit is not detected in the first few weeks after birth.</p> <p>Avoiding galactose by changing to a soy-based formula will avoid the complications of this disorder.</p> <p>Turck discusses using soy protein for infant feeding. The author notes that there are safety concerns with regards to the high phyto-estrogenic isoflavone content of these formulas. The author recommends that despite these safety concerns, these formulas should still be used in patients with classic galactosemia.</p>
abduction	MedBullets	5	A 9-year-old girl presents to the emergency department with a fever and a change in her behavior. She presented with similar symptoms 6 weeks ago and was treated for an Escherichia coli infection. She also was treated for a urinary tract infection 10 weeks ago. Her mother says that last night her daughter felt ill, and her condition has been worsening. Her daughter experienced a severe headache and had a stiff neck. This morning she was minimally responsive, vomited several times, and produced a small amount of dark cloudy urine. The patient was born at 39 weeks and met all her developmental milestones. She is currently up to date on her vaccinations and did not have infections during early childhood. Her parents are divorced and her father has noted she does not seem to get sick when he takes care of her. Her temperature is 99.5°F (37.5°C), blood pressure is 60/35 mmHg, pulse is 190/min, respirations are 33/min, and oxygen saturation is 98% on room air. The patient is started on intravenous fluids, vasopressors, and broad-spectrum antibiotics. Which of the following is the most appropriate underlying explanation for this patient's presentation?*, "answer": "Intentional contamination"	["Gastroenteritis", "Immunodeficiency", "Intentional contamination", "Meningitis", "Urinary tract infection"]	Intentional contamination	<p>This patient is presenting with repeat sepsis, which is uncommon in a pediatric patient with no significant medical history. This abnormal concentration of infections when the patient is with her mother suggests a diagnosis of intentional contamination.</p> <p>Factitious disorder by proxy (formerly Munchausen syndrome) occurs when 1 individual creates symptoms in another for attention. Typically, this is a parent feigning or creating symptoms in a child in order to receive the psychiatric gain of attention from healthcare personnel. The patient should be treated and removed from the parents as this is considered child abuse. Factitious disorder by proxy should be suspected in any pediatric patient who has repeat infections that are atypical or not expected epidemiologically. Patients should also be evaluated for organic causes of symptoms such as immunodeficiency syndromes. Abeln and Love review the evidence regarding the diagnosis of factitious disorder by proxy. They discuss how these disorders have high morbidity and mortality when not accurately recognized. They recommend paying attention to red flags such as abnormal constellations of symptoms.</p>
deduction	MedBullets	9	A 34-year-old man is brought to a rural emergency department by ambulance after being involved in a motor vehicle accident. Paramedics report that the patient was driving the car and crashed into a tree at roughly 25 miles per hour. There were no passengers and he was awake but disoriented at the scene. His temperature is 97.9°F (36.6°C), blood pressure is 131/88 mmHg, pulse is 89/min, and respirations are 14/min. He is speaking but is confused, opens his eyes to voice commands, and follows simple commands. He has multiple lacerations on his face and arms and smells of alcohol and marijuana. His cardiac exam is normal and his lungs are clear to auscultation bilaterally. He has bruising over his abdomen without any tenderness to palpation, distension, or rigidity. Which of the following is the most appropriate next step in management?	["Abdominal and chest CT", "Chest radiograph anterior-posterior and lateral", "Diagnostic laparoscopy", "Diagnostic peritoneal lavage", "Focused abdominal sonography for trauma exam"]	Abdominal and chest CT	<p>This patient presents following a motor vehicle accident with a tender abdomen and stable vital signs. Given his stability, the most appropriate next step is an abdominal and chest CT. In blunt abdominal trauma, the next step in management depends upon whether the patient is hemodynamically stable.</p> <p>With a normal blood pressure and pulse within the normal range, patients can undergo imaging to evaluate for injury to the abdominal organs with the most appropriate study being an abdominal CT with contrast. This study can accurately localize bleeding and aid in operative planning. A bedside ultrasound known as the FAST exam is often performed in unstable patients. Unstable patients, patients with abdominal tenderness, and patients with a severe mechanism of trauma may have a bedside FAST exam performed. If the free fluid is localized, the patient can be transferred directly to the operating room. Note that in many hospitals, the FAST exam is done simultaneously with the primary and secondary survey; however, if asked to choose which exam to perform on a trauma patient who is stable with no signs of a surgical abdomen, a CT scan should be performed to more accurately assess the patient's injuries and assess for injuries that may be missed by a FAST exam. Feliciano reviewed the evidence surrounding the current diagnosis and management of abdominal trauma. He discusses how contrast-enhanced CT of the abdomen and pelvis is an important method for evaluating the etiology of intra-abdominal bleeding. He recommends considering damage control methods in the polytrauma patient.</p>
induction	PubMedQA	11146778	The APACHE II (Acute Physiology and Chronic Health Evaluation II) score used as an intensive care unit (ICU) admission score in emergency surgical patients is not independent of the effects of treatment and might lead to considerable bias in the comparability of defined groups of patients and in the evaluation of treatment policies. Postoperative monitoring with the APACHE II score is clinically irrelevant. Inception cohort study. Secondary referral center. Eighty-five consecutive emergency surgical patients admitted to the surgical ICU in 1999. The APACHE II score was calculated before surgery; after admission to the ICU; and on postoperative days 3, 7, and 10. APACHE II scores and predicted and observed mortality rates. The mean +/- SD APACHE II score of 24.2 +/- 8.3 at admission to the ICU was approximately 36% greater than the initial APACHE II score of 17.8 +/- 7.7, a difference that was highly statistically significant (P<.001). The overall mortality of 32% favorably corresponds with the predicted mortality of 34% according to the initial APACHE II score. However, the predicted mortality of 50% according to the APACHE II score at admission to the ICU was significantly different from the observed mortality rate (P=.02). In 40 long-term patients (>=10 days in the ICU), the difference between the APACHE II scores of survivors and patients who died was statistically significant on day 10 (P=.08). Risk stratification in emergency surgical patients: is the APACHE II score a reliable marker of physiological impairment?	["yes", "no", "maybe"]	yes	<p>For risk stratification in emergency surgical patients, it is essential to measure the APACHE II score before surgical treatment.</p> <p>Longitudinal APACHE II scoring reveals continuous improvement of the score in surviving patients but has no therapeutic relevance in the individual patient.</p>

Table 6: Example questions for each reasoning category, highlighting the abduction, deduction, and induction text snippets.

tially essential for answering them, so we excluded these samples. To compensate, we manually selected similar diagnostic questions from MedQA, MedXpertQA-R, MedXpertQA-U, MMLU, and MMLU-PRO that closely matched the MedBullets questions.

Examining data samples and their rationales manually showed that some questions involved implicit abductive reasoning before the subsequent deductive steps. For instance, in clinical case management, selecting the next step often relies on a

diagnostic hypothesis not explicitly stated in the text, which is then validated through deduction, such as choosing an appropriate diagnostic test. We annotated these samples as abduction-deduction, capturing the reasoning steps that physicians may sometimes perform implicitly. The process produced a dataset of 150 data samples, balanced as follows: 50 abduction-deduction, 25 abduction, 25 deduction, and 50 induction. This arrangement was designed to provide a balanced set of 50 samples for each reasoning type. Table 6 presents four

examples for each annotation category.

Annotation Quality Assessment To evaluate the quality of the generated subset, two expert.ai professional linguists were assigned to replicate the annotation process. The task was designed and deployed using a LabelStudio¹ instance. Annotators were provided with detailed guidelines describing both the data and the task, including definitions of the reasoning types adopted in this work, with a practical example provided for each.

Guidelines

Data Description

Each data sample contains a medical question along with options and the correct answer. When available, a physician’s rationale is included to help clarify the answer and explain the reasoning path.

Task Description

The goal of this task is to annotate the type of reasoning required to answer each question. You can choose from the following options: *abduction-deduction*, *abduction*, *deduction*, and *induction*. To select the correct reasoning type, consider the descriptions provided below.

Reasoning inferences:

- *abduction*: Reasoning to the most likely explanation for a set of observations.
- *deduction*: Reasoning from general rules to specific conclusions.
- *induction*: Reasoning from specific observations to general principles.

Practical examples:

- *abduction-deduction*: Sometimes a question asks for the next step in patient management (e.g., which test to perform or treatment to prescribe), but the underlying condition is not explicitly stated. First, *abduction* should be used to infer the most likely implicit diagnosis, and then *deduction* applied to determine the appropriate answer.
- *abduction*: Used when the question asks

for the most likely condition or diagnosis that explains a set of observations.

- *deduction*: Involves applying a known rule or principle to reach a specific conclusion, such as selecting a test to confirm a hypothesis or prescribing a treatment for a known condition.
- *induction*: Used when the question asks you to review observations, studies, or experiments and then derive general rules or principles from them.

An example of annotation task, as displayed in the LabelStudio user interface, is shown in Figure 4. Inter-annotator agreement between the authors and the linguists was assessed using Fleiss’ Kappa (κ), resulting in a score of 0.71. This value indicates substantial agreement, reflecting the quality of the annotations and the clarity of the provided guidelines.

B Data Contamination Analysis

We adopted the MELD (Memorization Affects Levenshtein Detector) framework (Nori et al., 2023), a strategy designed to detect data contamination between the pretraining or fine-tuning data and the test set in use. We selected one model per family for our evaluation in the case of open-source medium-sized models, whereas all frontier models were included individually. The approach involves splitting each test sample into two parts. The first part is provided to the model, which is then prompted to generate the missing portion, constrained to match the original length. Once the missing part is generated, it is compared using Levenshtein ratio. The percentage of generated parts that achieve a similarity of 95% or higher with the original is then calculated over a specific test set. Building on adaptations from prior work (Tang et al., 2025), and considering that all models in our study are instruction or reasoning tuned, we implemented a slightly modified MELD strategy. Instead of allowing the model to auto-complete the missing part, we provided explicit instructions to guide the generation, as illustrated in the boxes below:

System Prompt

You are a helpful assistant who execute instructions.

¹<https://labelstud.io/>

UI Preview

Question:
Distance to provider might be an important barrier to timely diagnosis and treatment for cancer patients who qualify for Medicaid coverage. Whether driving time or driving distance is a better indicator of travel burden is also of interest.
Driving distances and times from patient residence to primary care provider were calculated for 3,917 breast, colorectal (CRC) and lung cancer Medicaid patients in Washington State from 1997 to 2003 using MapQuest.com. We fitted regression models of stage at diagnosis and time-to-treatment (number of days between diagnosis and surgery) to test the hypothesis that travel burden is associated with timely diagnosis and treatment of cancer.
Later stage at diagnosis for breast cancer Medicaid patients is associated with travel burden (OR = 1.488 per 100 driving miles, P= .037 and OR = 1.270 per driving hour, P= .016). Time-to-treatment after diagnosis of CRC is also associated with travel burden (14.57 days per 100 driving miles, P= .002 and 5.86 days per driving hour, P= .018).
Is distance to provider a barrier to care for medicaid patients with breast, colorectal, or lung cancer?

Options:
- yes
- no
- maybe

Answer:
no

Rationale:
Although travel burden is associated with timely diagnosis and treatment for some types of cancer, we did not find evidence that driving time was, in general, better at predicting timeliness of cancer diagnosis and treatment than driving distance. More intensive efforts at early detection of breast cancer and early treatment of CRC for Medicaid patients who live in remote areas may be needed.

Select the required reasoning type

abduction-deduction^[1] abduction^[2] deduction^[3] induction^[4]

Figure 4: Example of LabelStudio annotation task.

User Prompt

Instructions:

You will receive a passage enclosed within <text_to_continue>...</text_to_continue> tags. This passage contains the initial segment of a medical question.

Carefully analyze the passage to determine the most probable medical question being asked. Then, generate a clear, concise, and clinically appropriate continuation that completes the second half of the question.

Enclose your completion within <continue>...</continue> tags, maintaining a style consistent with the original text.

Task:

<text_to_continue>{TEXT}</text_to_continue>

For the benchmark adopted in our experiments, none of the models tested returned a significant MELD score. As previously noted by prior work (Tang et al., 2025), datasets such as MMLU-Pro and MedMCQA exhibit some outlier points with unusually high similarity scores, suggesting possible contamination in those specific data instances. In contrast, datasets like MedARC-QA and MetaMedQA display values that are more consistent and indicative of low contamination. It is important to remember that MELD achieves high precision, meaning that detected matches are strong indicators of memorization; however, its recall is unknown, so the absence of a match does not necessarily imply that the model has not encountered the data during training.

C Computational Resources

Experiments with small and medium models (up to 72B parameters) were conducted on a cluster of four NVIDIA L40S GPUs, using the HuggingFace Transformers library (Wolf et al., 2020) for model loading and vLLM (Kwon et al., 2023) for efficient inference. All experiments used consistent generation hyperparameters: temperature 0.6, top-p 0.9, and a maximum output length of 4096 tokens. Each was repeated three times with seeds 0, 32, and 64 to ensure robustness. Frontier model evaluations were conducted via their respective APIs in a single run, using the same hyperparameters.

D Prompt Template Details

In this appendix, we present the prompt templates used in the experiments on prompting strategies. When the model supports both a system and a user template, we use them accordingly. Otherwise, we embed the system template within the user message. Below, we list the templates adopted.

D.1 System prompts

We used several system prompts, which are presented below.

Base This prompt is used for the base setting, where the role of the model is simply to act as an assistant that executes instructions.

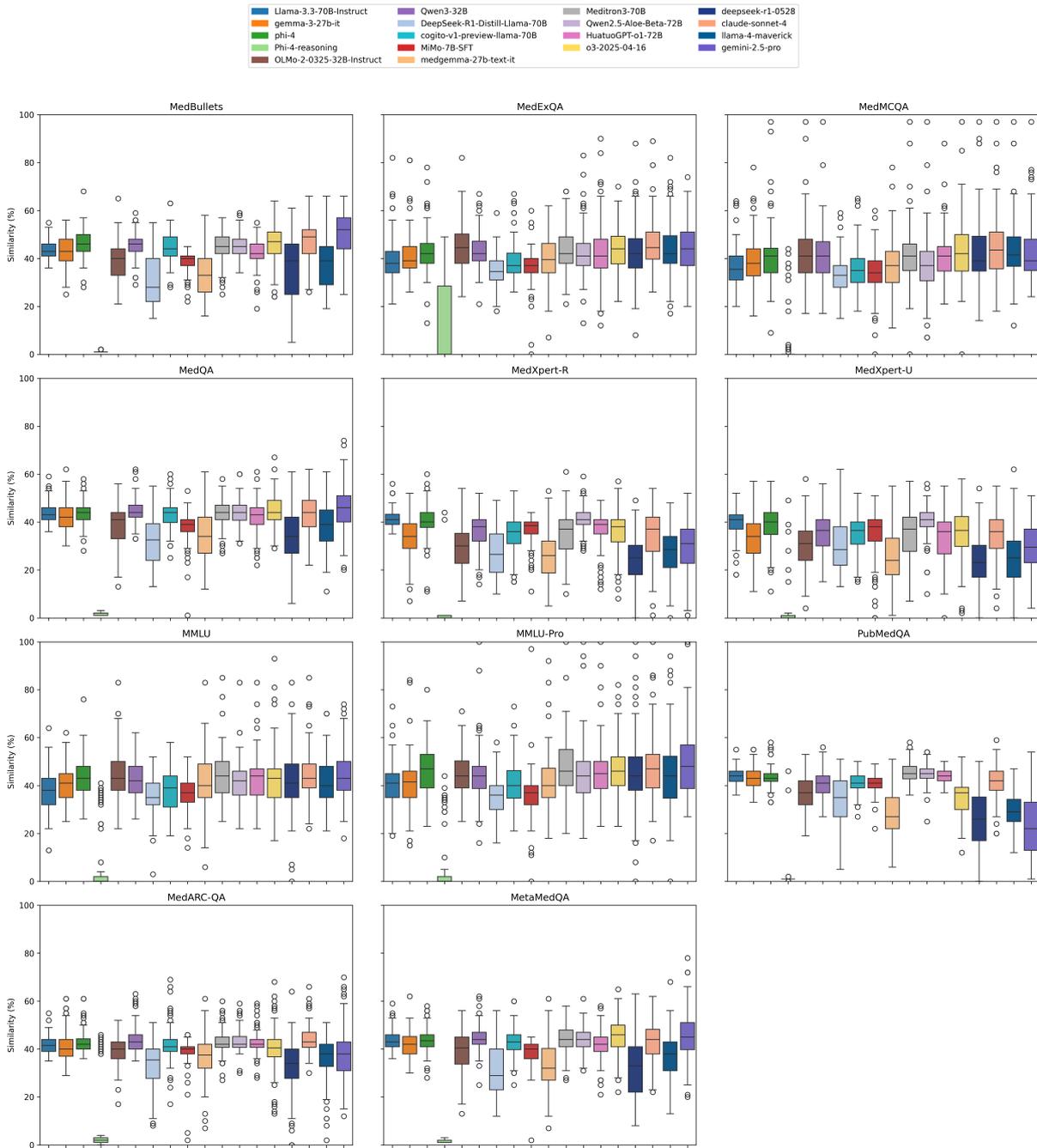


Figure 5: Distribution of Levenshtein similarity scores per model across the eleven datasets from the three benchmarks.

Base System Prompt
 You are a helpful assistant who execute instructions.

CoT/IR/AP/FS System Prompt
 You are a helpful assistant who execute instructions to solve clinical problems.

CoT/IR/AP/FS Similar to the base prompt, this prompt includes additional information about the medical domain. It is shared across the Chain-of-Thought (CoT), Intuitive Reasoning (IR), Analogical Prompting (AP), and Few-Shot (FS) strategies.

MP Multi-Persona Prompting is a strategy designed to simulate diverse expert perspectives by assigning distinct professional roles to the model. Each persona is characterized by a unique clinical viewpoint, reasoning style, or level of expertise. In our setup, we used multiple personas such as

Innovative Medical Thinker, MD, Critical Medical Analyst, Medical Professor, and Clinical Decision Specialist, Medical Researcher. Additionally, we included a persona representing a Senior Medical Expert. The goal of this strategy is to enrich the model's reasoning by framing responses through varied medical lenses, reflecting the diversity of real-world clinical expertise.

#1 MP System Prompt
You are a {ROLE}.

#2 MP System Prompt
You are a senior medical expert.

D.2 User and assistant prompts

Depending on the prompting strategy, we used a variety of user and assistant prompts.

Base The base prompt is the simplest form, in which the model is asked to provide a solution to a given clinical problem.

Base User Prompt
Clinical problem:
{QUESTION}
Options:
{OPTIONS}
Instructions:
Select the most appropriate solution from the provided 'Options' list and report it within square brackets []

CoT The Chain-of-Thought (CoT) prompt guides the model to explicitly articulate its reasoning process step by step prior to providing a solution, aiming to improve the quality and interpretability of the response.

CoT User Prompt
Clinical problem:
{QUESTION}
Options:
{OPTIONS}
Instructions:
Reason step-by-step to determine the correct solution. Select the most appropriate solution from the provided 'Options' list and report it within square brackets []

IR The Intuitive Reasoning prompt is based on a clinical problem-solving strategy that encourages the model to rely on experience-driven reasoning, mimicking the cognitive processes leveraged by physicians.

IR User Prompt
Clinical problem:
{QUESTION}
Options:
{OPTIONS}
Instructions:
Use the available medical information to deduce and determine the correct solution. Select the most appropriate solution from the provided 'Options' list and report it within square brackets []

AP Analogical Prompting is a strategy that guides a language model to solve a problem by drawing analogies to previously seen or understood examples. It relies on the notion that humans often reason by analogy, mapping knowledge from familiar situations to new, similar ones. The prompt encourages the model to generate three analogous examples before providing its final answer.

AP User Prompt
Clinical problem:
{QUESTION}
Options:
{OPTIONS}
Instructions:
Retrieve three distinct medical problems, each different from one another and from the provided problem, followed by your step-by-step reasoning to determine the correct solution. Select the most appropriate solution from the provided 'Options' list and report it within square brackets []

FS Few-Shot Prompting is a strategy in which the model is guided by a small number of input-output examples before being asked to solve a new task. These examples help the model infer the expected format, reasoning style, and domain-specific knowledge required for the task. In our case, we used examples from the Casimedicos dataset, where each example is structured as a triplet: question-answer-explanation. The user presents a clinical case along with multiple-choice options, and the assistant provides an explanation followed by the correct answer.

FS User Prompt (Example)

Clinical problem:
{QUESTION}
Options:
{OPTIONS}

FS Assistant Prompt (Example)

Explanation:
{EXPLANATION}
Answer:
{ANSWER}

FS User Prompt

Clinical problem:
{QUESTION}
Options:
{OPTIONS}
Instructions:
Select the most appropriate solution from the provided 'Options' list, furnish an explanation for your choice, and write the selected solution inside square brackets [].

MP Multi-Persona Prompting is an agent-based workflow composed of three distinct steps. First, each persona introduced in the system prompts independently analyzes the case and provides an initial solution. Next, a debate phase occurs where each persona receives the initial thoughts of the other two roles and is asked to revise and reconsider its original position. Finally, an aggregator collects all the debated responses and generates a consolidated synthesis.

MP User Prompt (Init)

Clinical problem:
{QUESTION}
Options:
{OPTIONS}
Instructions:
Based on your medical expertise, please analyze the provided problem and options. Think step-by-step and provide your reasoning and answer clearly. Please include distinct sections for your thinking and your answer in your response. Report your reasoning steps within angle brackets < >. Select the most appropriate solution from the provided 'Options' list and report it within square brackets []

MP User Prompt (Debate)

Considering the following insights from your peers:
{ROLE}'s previous opinion:
{OPINION}
{ROLE}'s previous opinion:
{OPINION}
Clinical problem:
{QUESTION}
Options:
{OPTIONS}
Instructions:
Please update your analysis for the question. Think carefully step-by-step and revise your answer accordingly. Provide your response with clear sections for your updated thinking and updated answer. Report your reasoning steps within angle brackets < >. Select the most appropriate solution from the provided 'Options' list and report it within square brackets []

MP User Prompt (Final)

Considering all the following debate insights and answers:
{ROLE}'s final opinion:
{OPINION}
{ROLE}'s final opinion:
{OPINION}
Clinical problem:
{QUESTION}
Options:
{OPTIONS}
Instructions:
Please carefully review all the information and provide the final decision. Offer a detailed rationale and clearly state your final answer, indicating your final reasoning and the chosen option. Report your reasoning steps within angle brackets < >. Select the most appropriate solution from the provided 'Options' list and report it within square brackets []

E Failure Examples

To better understand why several models, including frontier ones, sometimes fail to outperform a simple random baseline, we present a representative failure example for each reasoning strategy. We selected OpenAI's o3 model because it achieved the best overall performance. Table 7 presents these examples, including the complete question data along with the physician's rationale, the corresponding model outputs, and, where available, physician-provided refutations of incorrect answers from the original dataset.

Category	Dataset	ID	Question	Options	Answer	Reason/Rationale	Model Output	Refutation
abduction-deduction	MedBullets	77	A 57-year-old man is brought to the emergency department by his wife with confusion. He was seen in the emergency department 2 days ago for profuse epistaxis. His primary care provider was concerned and had advised him to go to the hospital, as the patient is on warfarin for atrial fibrillation. However, his bleeding resolved with pressure and nasal packing. Upon presentation to the ED today, his temperature is 97.6°F (36.5°C), blood pressure is 164/53 mmHg, pulse is 110/min, respirations are 22/min, and oxygen saturation is 100% on room air. Physical exam reveals an obtunded man who cannot answer questions. A dark, purpuric rash is noted on his torso and extremities. The patient's nasal packing is removed with no further bleeding. The patient is given IV fluids and blood cultures are drawn. Which of the following should be administered first?	<p>["Clindamycin", "Meningococci", "Penicillin", "Piperacillin-tazobactam", "Vancomycin"]</p>	Clindamycin	<p>This patient is presenting with nasal packing, hypotension, tachycardia, a dark, purpuric rash, and confusion, which is concerning for toxic shock syndrome. The most important initial step in management is to administer clindamycin.</p> <p>Toxic shock syndrome is a life-threatening diagnosis caused by Streptococcus pyogenes or Staphylococcus aureus and occurs commonly secondary to retained tampons or nasal packing. Patients are critically ill and require immediate management with fluids, vasopressors, and blood cultures. Source control is imperative, and the foreign body causing the infection must be removed. Next, broad-spectrum antibiotics should be started immediately.</p> <p>Clindamycin is often preferred as the first agent as this lincomamide antibiotic not only covers Staphylococcus aureus and Streptococcus pyogenes, but also has the theoretical benefit of suppressing toxin production.</p> <p>Further management is centered on fluid administration, vasopressors, and monitoring and care in an ICU setting as these patients are typically critically ill. Abruzzo and Buchsack review the treatments for toxic shock syndrome. They note that clindamycin is particularly important in toxic shock syndrome that is complicated by toxin production. They recommend that administration may reduce toxin production and could improve outcomes.</p>	<p>Correct answer: Penicillin. Explanation: The patient is in fulminant septic shock (BP 64/43 mmHg, tachycardia 160/min, tachypnea 32/min, obtundation).</p> <ul style="list-style-type: none"> The dark, non-blanching purpuric rash that rapidly involves the trunk and extremities is classic for meningococci (disseminated Neisseria meningitidis infection) and is frequently accompanied by disseminated intravascular coagulation and Waterhouse-Friderichsen syndrome (adrenal hemorrhage) that produces profuse, refractory hypotension. Meningococci progress extremely quickly; outcome depends on giving an effective antibiotic as soon as the diagnosis is suspected—ideally within minutes of drawing blood cultures. N. meningitidis remains uniformly susceptible to high-dose intravenous penicillin G (or a third-generation cephalosporin such as ceftazidime). Of the drugs listed, penicillin is the agent that most directly and reliably covers meningococci. Although patients in septic shock are ultimately treated with broader coverage (eg, vancomycin plus a beta-lactam with Pseudomonas activity) until cultures return, when a characteristic presentation of meningococci is recognized, targeted therapy with intravenous penicillin G is the first, lifesaving step. Therefore, penicillin should be administered immediately after blood cultures are obtained. 	<p>Penicillin will be used for infections such as streptococcal pharyngitis, which presents with a sore throat and is not used in a critically ill patient with toxic shock syndrome.</p>
abduction	MedBullets	244	A 57-year-old man presents to the emergency department with several days of malaise, weakness, and night sweats. Today, he experienced a headache with blurry vision, thus prompting his presentation. The patient has a history of diabetes and is followed closely by an endocrinologist. Otherwise, he has lost 10 pounds over the past month. His temperature is 100°F (37.8°C), blood pressure is 122/90 mmHg, pulse is 86/min, respirations are 16/min, and oxygen saturation is 99% on room air. Physical exam is notable for a thin man but is otherwise unremarkable. His visual acuity is 20/100 in both eyes, though he says that he normally has 20/20 vision. His gait is mildly ataxic as well, but his neurological exam is otherwise normal. Laboratory studies are ordered as seen below. Hemoglobin: 10 g/dL, Hematocrit: 31%, Leukocyte count: 57,500/mm ³ with 35% blasts. Platelet count: 109,000/mm ³ . Serum: Na ⁺ : 139 mEq/L, Cl ⁻ : 100 mEq/L, K ⁺ : 4.1 mEq/L, HCO ₃ ⁻ : 22 mEq/L, BUN: 30 mg/dL, Glucose: 99 mg/dL, Creatinine: 1.1 mg/dL, Calcium: 10.0 mg/dL. Which of the following is the most likely diagnosis?	<p>["Acute lymphoblastic leukemia", "Acute myelogenous leukemia", "Chronic lymphocytic leukemia", "Chronic myelogenous leukemia", "Multiple myeloma"]</p>	Chronic myelogenous leukemia	<p>This patient is presenting with symptoms of malignancy (malaise, weakness, and night sweats) and has a white blood cell count > 50,000/mm³ in the setting of > 20% blasts, which is concerning for chronic myelogenous leukemia (CML). Given the blurry vision and ataxia in the setting of increased blasts, this patient may be suffering from a hyperviscosity syndrome secondary to a blast crisis, which can be seen in CML.</p> <p>CML is a neoplastic proliferation of myeloid stem cells. Patients are often asymptomatic and may present with a white blood cell count > 50,000/mm³. There are many different presentations.</p> <p>The chronic phase presents with < 10% blasts and is often asymptomatic. The accelerated phase has increased blasts, and patients present with fatigue, weight loss, malaise, night sweats, abdominal pain, and hepatosplenomegaly. Finally, the blast phase presents similarly to the accelerated phase but with more severe symptoms. The blast phase may evolve into a blast crisis (> 20% blasts), which could cause a hyperviscosity syndrome.</p> <p>Hyperviscosity syndrome may require IV fluids and leukapheresis. The most common translocation in CML is the 9:22 translocation, which forms the BCR-ABL1 gene product. Jain et al. review CML. They note the outcomes in patients in the blast phase in CML and how there are poorer outcomes associated with this phase. They recommend paying close attention and obtaining a CBC with a differential to determine the percent blasts.</p>	<p>Correct answer: Acute myelogenous leukemia. Explanation: Acute versus chronic disease. The presence of ≥ 20% blasts in the peripheral blood or bone marrow is diagnostic of an acute leukemia. This patient's differential shows 35% blasts → acute process.</p> <p>Myeloid versus lymphoid lineage • ALL occurs predominantly in children; AML is the most common acute leukemia in adults (median age ≥ 65 years). • CNS findings in this patient (headache, blurred vision, ataxia) are best explained by leukostasis/hyperviscosity produced by large myeloblasts at a WBC > 50,000/mm³—a classic presentation of AML. • Although ALL can affect adults and may infiltrate the meninges, hyperleukocytosis with symptomatic leukostasis is far more characteristic of AML. Chronic leukemias and plasma-cell dyscrasias ruled out. • Chronic myelogenous leukemia (CML) usually shows very high WBC counts with a full spectrum of myeloid precursors, but < 10% blasts in chronic phase.</p> <ul style="list-style-type: none"> Chronic lymphocytic leukemia (CLL): marked lymphocytosis without blasts and is often discovered incidentally. Multiple myeloma produces anemia, bone pain, hypercalcemia, renal failure, but not peripheral blasts. Supporting laboratory findings • Anemia and thrombocytopenia result from marrow replacement by blasts—typical in AML. • No hypercalcemia, lytic bone lesions, or renal failure to suggest multiple myeloma. Taken together, an older adult with leukostasis containing > 20% blasts, associated cytopenias, and symptoms caused by leukostasis most strongly indicates acute myelogenous leukemia. 	<p>Acute myelogenous leukemia is a malignancy of myeloblast and presents in the elderly. Patients will present with fatigue, malaise, weight loss, anorexia, fever, leukocytosis, anemia (causing fatigue, thrombocytopenia (causing bleeding)), and neutropenia (leading to life-threatening infections). Auer rods can be seen on peripheral smear.</p>
deduction	MedBullets	22	A 25-year-old woman is brought to the emergency department by her family. They found her somnolent next to her computer with an empty pill bottle around 3 hours ago. The family cannot recall the types of pill bottles that they found and did not bring them to the ED. The patient's medical history is notable for anxiety, obesity, hypertension, and polycystic ovarian syndrome. She is currently not on any medications. She was given an exercise program but she has not been compliant with these exercises. Her temperature is 99.5°F (37.5°C), pulse is 82/min, blood pressure is 128/85 mmHg, respirations are 11/min, and oxygen saturation is 97% on room air. Physical exam is within normal limits. The patient begins communicating with the physician and states that she did take acetaminophen but it was only a few pills. Her parents are certain the bottle was new. Which of the following is the most appropriate next step in management?	<p>["Blood acetaminophen levels", "Charcoal", "Lactated ringers bolus", "N-acetylcysteine", "Syrup of ipecac"]</p>	N-acetylcysteine	<p>This patient is presenting with acute acetaminophen intoxication. The most appropriate next step in management is to administer the antidote N-acetylcysteine (NAC). In a toxicology case where the ingestion is unknown, the most appropriate initial step in management is to begin treatment empirically with IV fluids, charcoal, NAC, naloxone, and glucose if needed. These agents are effective against the most common causes of intoxication. In cases where there is known acetaminophen ingestion, the most important treatment is prompt administration of NAC. This medication works by blocking the hepatic production of glutathione, which detoxifies the toxic metabolite of acetaminophen (N-acetyl-p-benzoquinone imine [NAPQI]). NAC is a safe, cheap, and effective treatment that could be life-saving if administered in a timely fashion.</p> <p>In massive overdoses, it is appropriate to start treatment empirically as it can be discontinued later. Otherwise, a 4-hour acetaminophen level should be collected and the level should be plotted on the Rumack-Matthew nomogram to determine treatment. In chronic acetaminophen toxicity, the acute serum level play no role in guiding treatment. Chen et al. performed a meta-analysis regarding the treatment of acetaminophen overdose. They found that the administration of charcoal immediately after ingestion can decrease the rate of absorption of acetaminophen. They recommend early administration of N-acetylcysteine in order to decrease the hepatotoxicity of this substance.</p>	<p>Correct choice: Blood acetaminophen levels. Explanation: The only specific information available is that the patient may have taken acetaminophen, and the time of ingestion is known (about 3 hours ago).</p> <ul style="list-style-type: none"> Management of a possible acetaminophen overdose is guided by the Rumack-Matthew nomogram, which requires a serum acetaminophen concentration obtained 2-4 hours after ingestion. Because she is still within that 4-hour window, the next step is to obtain (or order now so it will be drawn at the 4-hour mark) a serum acetaminophen level. N-acetylcysteine (the antidote) is started if the 4-hour level plots at or above the treatment line on the nomogram, or immediately if the time of ingestion is unknown or > 6 hours. Activated charcoal is most useful if given within 1 hour of ingestion; at 3 hours it offers little benefit. Syrup of ipecac is obsolete, and a fluid bolus is not indicated. Therefore, measuring the blood acetaminophen level is the appropriate next step. 	<p>Blood acetaminophen levels should be ordered, but it is much less due to confirm the diagnosis than it is to begin treatment given the high fatality rate with untreated acetaminophen overdose. Testing for all common intoxication should proceed while treatment is being administered. In this case, the concern for consumption of an entire bottle of acetaminophen would make it reasonable to treat first.</p>
induction	PubMedQA	20538207	It is generally considered that kidney grafts should be preserved at 4 degrees C during cold storage. However, actual temperature conditions are not known. We decided to study the temperature levels during preservation with the BioStar storage can and Vitapak transport pack. Temperature was monitored using the ThermoScan probe during preservation of pig kidneys, in the same conditions used with human grafts. The probe recorded the temperature level every 10 minutes during four days. We compared the results found with the new storage can with results obtained in the same conditions with the storage can formerly used by our team. We also studied the influence of the amount of ice within the transport pack on the temperature level. We then monitored temperature during the conservation of actual human kidney grafts harvested at our institution from August 2007 to May 2008. The temperature levels were the same regardless of the position of the probe within the transport pack. The lowest temperature was maintained during 15 hours, and the temperature level stayed below 5 degrees C for 57 hours with the new storage can. The former storage can maintained the lowest temperature level for 80 minutes, and temperature reached 5 degrees C after 10 hours 40 minutes. Temperature levels were similar when 2 or 4 kg of crushed ice were used. We observed similar results when monitoring the conservation of human grafts. Similar results were monitored during kidney allograft preservation?	<p>["yes", "no", "maybe"]</p>	no	<p>The new storage can affords more stable temperature levels when compared to the formerly used can. Since temperature is stable during conservation, continuous monitoring in every day practice does not seem warranted.</p>	<p>Correct option: yes. Explanation: The study shows that the temperature inside a kidney-transport pack does not automatically stay at the recommended 4 °C for the whole journey and that different containers keep the graft cold for very different lengths of time (e.g., the former can reached 5 °C after only 10 h 40 min, whereas the new can stayed below 5 °C for 57 h). Because graft viability depends on staying in the narrow 0-4 °C range and because neither the operator nor the recipient center can know how long a particular container will maintain that range without feedback, continuous temperature monitoring is needed to detect inadvertent warming and to decide whether corrective action (adding ice, shortening ischemic time, etc.) is required. Therefore, temperature should indeed be monitored during kidney allograft preservation.</p>	

Table 7: Representative failures for each reasoning type, highlighting in green the correct reasoning passages from physician explanations and in red the incorrect ones generated by the OpenAI's o3 model. Note that in the abduction-deduction example, the model incorrectly identifies a shock syndrome rather than a toxic syndrome (abduction), resulting in inappropriate antibiotic administration (deduction).

As illustrated by the **abduction-deduction** example, the model is able to use abduction to hypothesize the correct diagnosis, but it fails in the subsequent deductive step. It misclassifies the microorganism based on the interpretation of a symptom and proposes a treatment that is inappropriate, as more effective options are available for the case at hand. In the **abduction** example, the model demonstrates deep knowledge of the specific domain but hypothesizes an acute syndrome instead of a chronic one. From the perspective of symptoms and lab results, the differences between the two are minimal, but abduction involves making the best guess while also considering missing information, such as a leukemia test. Here, the model underestimates the possibility of a chronic illness presenting with parameters of an acute condition while being in the blast phase of the disease. In the

deduction example, the model, although knowledgeable, does not accurately determine the best course of action given the available information. In this case, the most appropriate step would be to administer the life-saving medication immediately and skip measuring the level of intoxication. Lastly, in the **induction** example, the model is unable to generalize the results of a study to formulate a rule capable of answering the question.