REVIEW ARTICLE

Critique of impure reason: Unveiling the reasoning behaviour of medical large language models

Shamus Zi Yang Sim[1]*†, Tyrone Chen[2]*†

[1]QueueMed Healthtech, Kuala Lumpur, Malaysia; [2]Peter MacCallum Cancer Centre, Melbourne, Australia

Abstract Despite the current ubiquity of large language models (LLMs) across the medical domain, there is a surprising lack of studies which address their *reasoning behaviour*. We emphasise the importance of understanding *reasoning behaviour* as opposed to high-level prediction accuracies, since it is equivalent to explainable AI (XAI) in this context. In particular, achieving XAI in medical LLMs used in the clinical domain will have a significant impact across the healthcare sector. Therefore, in this work, we adapt the existing concept of *reasoning behaviour* and articulate its interpretation within the specific context of medical LLMs. We survey and categorise current state-of-the-art approaches for modelling and evaluating *reasoning* in medical LLMs. Additionally, we propose theoretical frameworks which can empower medical professionals or machine learning engineers to gain insight into the low-level reasoning operations of these previously obscure models. We also outline key open challenges facing the development of *large reasoning models*. The subsequent increased transparency and trust in medical machine learning models by clinicians as well as patients will accelerate the integration, application as well as further development of medical AI for the healthcare system as a whole.

***For correspondence:** shamus@qmed.asia (SZYS); tyrone.chen@petermac.org (TC)

†These authors contributed equally to this work

Competing interest: The authors declare that no competing interests exist.

Funding: See page 23

Preprinted: 20 December 2024
Received: 28 January 2025
Accepted: 07 October 2025
Published: 28 October 2025

Reviewing Editor: Yongliang Yang, Shanghai University of Medicine and Health Sciences, China

Introduction

Reasoning drives problem-solving activities and is ubiquitous in our daily lives. The rising adoption of the field of artificial intelligence and its proximity to the concept of reasoning then naturally provokes the question: what is the *reasoning behaviour* of machine learning models commonly used in artificial intelligence (***Figure 1***)? This is particularly pertinent with regard to the increasing use of large language models (LLMs).

In this review, we focus specifically on **transformer-based LLMs**, which are built on the transformer architecture, an attention-based mechanism capable of capturing complex dependencies in sequential data (***Vaswani et al., 2017***). These models are characterised by a large number of trainable parameters and are pre-trained on vast corpora of textual data.

We define medical LLMs in an application-centric manner, referring broadly to any LLMs that are employed as a core component in medical or clinical workflows. This broad definition accommodates a diverse range of use cases, including applications in diagnosis (***Yang et al., 2024a***; ***Nori et al., 2025***; ***Savage et al., 2024***), medical image analysis (***Pan et al., 2025***; ***Lai et al., 2025***), clinical summarisation (***Zi Yang et al., 2024***), and EHR question-answering tasks (***Shi et al., 2024***). This includes both domain-specific models pre-trained or fine-tuned on biomedical corpora, as well as general-purpose LLMs (e.g. GPT or LLAMA models) that are adapted or prompted for use in medical settings. While such general models may not have been originally developed for healthcare, they are increasingly leveraged in a wide range of medical applications and thus fall within the scope of our discussion. In this review, we observed the frequent use of various variants of GPT- and LLaMA-based models throughout the literature. Due to their sheer scale and complexity, LLMs are inherently less

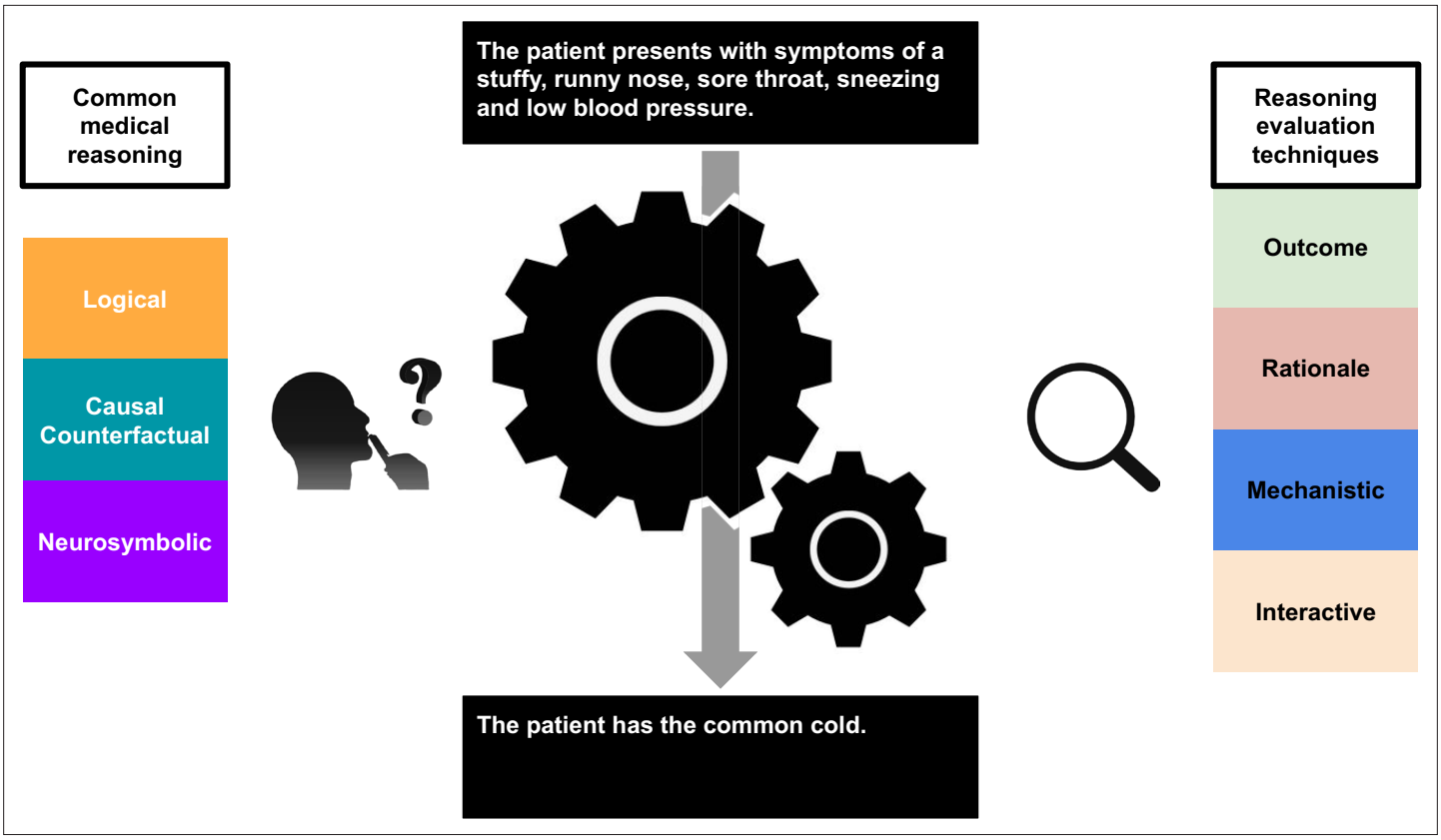

Figure 1. A graphical abstract illustrating the current state of medical large language models (LLMs) in the context of *reasoning behaviour*.

interpretable than traditional machine learning models, which are themselves often considered 'black boxes' under normal circumstances. We note, however, that although LLMs have also been applied directly to genomic and molecular sequence data (***Chen et al., 2023***) with potential medical use, the present review does not cover these applications.

The question of *how* LLMs arrive at their answers—particularly in high-stakes applications like medicine—is surprisingly underexplored. This gap is especially striking given the widespread deployment of these models across domains, often without a comprehensive understanding of their underlying reasoning mechanisms. Instead, evaluations tend to focus on performance metrics such as accuracy, F1 scores, precision, and recall. These metrics are typically benchmarked against curated subsets of state-of-the-art (SOTA) models as well as specialised datasets which consist of medical licence examination questions from the United States (USMLE), Mainland China (MCMLE), Taiwan (TWMLE) (***Jin et al., 2020***), India (AIIMS/NEET) (***Pal et al., 2022***), and other broader questions less directly related to clinical fields (***Jin et al., 2019***; ***Hendrycks et al., 2020***). While these may be effective in some cases, such metrics provide limited insight into the complex and obscure inferential processes that LLMs apply to generate answers.

This neglect in understanding *reasoning behaviour* leads to their unintentional misuse with direct real-world effects, including data fabrication (***Lorek, 2024***), false accusations (***Gegg-Harrison and Quarterman, 2024***), and suicide (***Dewitte, 2024***). Further obfuscating *reasoning behaviour* in LLMs, particularly generative models, is their ability to mimic the semantics of question and answer processes convincingly while being surprisingly accurate, to the point where individuals have mistakenly assumed their sentience (***Tiku, 2022***). Such issues are amplified in LLMs used for medical purposes, given their proximity to life-and-death decisions, for example, in the case of acute heart failure (***Kwon et al., 2019***).

We highlight the importance of gaining insight into the process-driven, *reasoning behaviour* by observing its functional similarity to *interpretability metrics* in machine learning (***Sundararajan et al., 2017***; ***Selvaraju et al., 2017***; ***Zhang et al., 2021***; ***Alammar, 2021***; ***Tenney et al., 2020***). In both

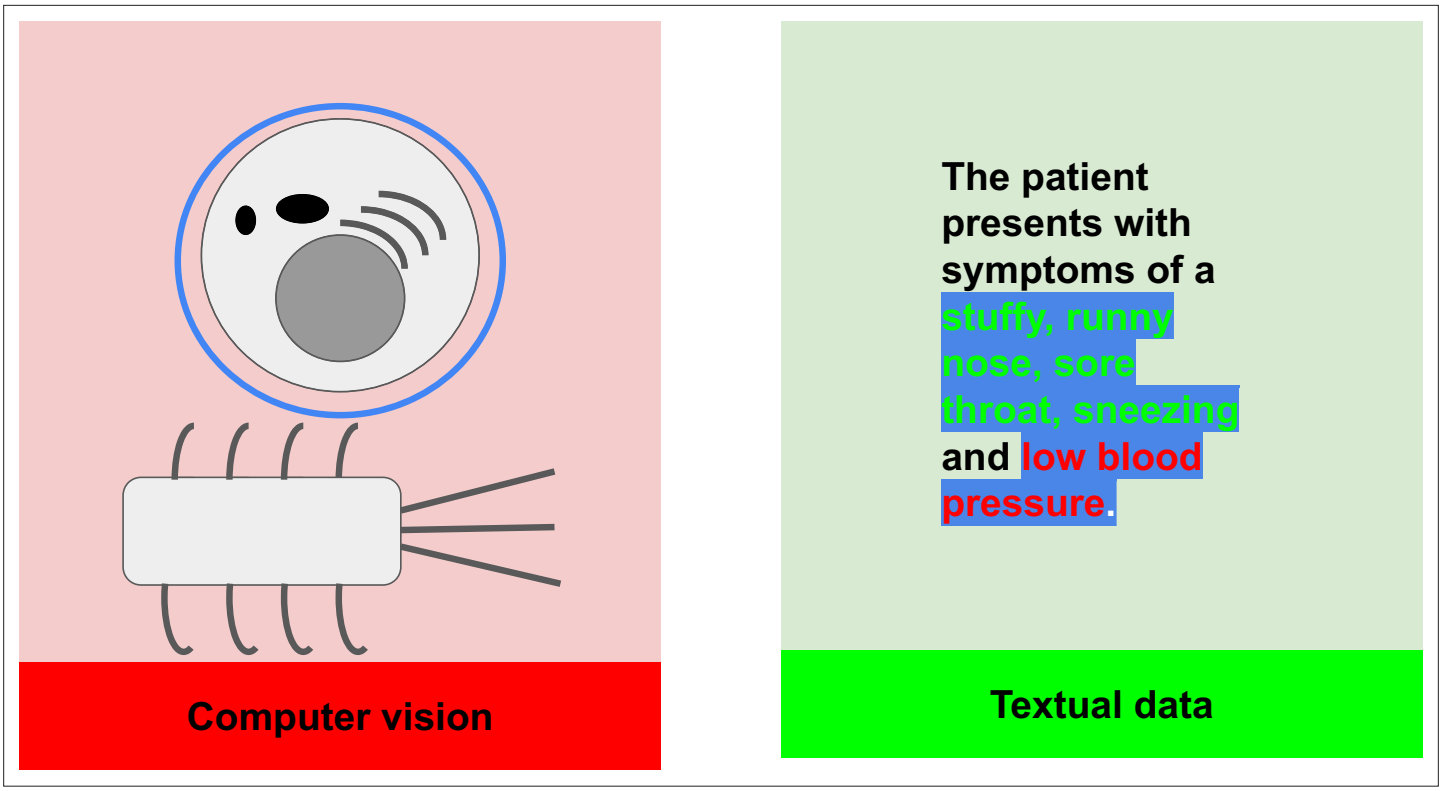

Figure 2. An illustration of the contrast in modalities between computer vision and natural language processing.

cases, the aim is to inspect the learning process of the model, with corresponding metrics used to gain information into how a model is correctly or incorrectly predisposed towards a certain outcome, in an attempt to address the common 'black-box' problem of machine learning models (***Figure 2***). We re-emphasise the lack of LLM studies tackling the question of general inference to focus on a noticeable gap in the field: **remarkably few studies investigate *reasoning behaviour* in the medical LLM space**.

Given the increased stakes of medical LLMs in clinical decision making, achieving a deeper understanding of medical LLMs carries a greater weight than with general-purpose LLMs. Thus, their intense scrutiny by both medical experts and the general public is unsurprising. Therefore, it is necessary to supplement clinicians with insight into the *reasoning behaviour* of medical LLMs to better understand how they arrive at their conclusions and expose potential logical fallacies. An ability for LLMs to provide reasoning for their outputs, for example, in a medical recommendation or diagnosis, allows clinicians to clarify discrepancies between machine and expert suggestions. This transparency fosters trust, encouraging integration of LLMs and other machine learning models into clinical decisions and subsequently improving patient outcomes.

In our review, we will address a few specific points:

1. We provide a primer introducing fundamental AI concepts covered in this review.
2. We adopt the existing concept of *reasoning behaviour* and articulate its interpretation within the specific context of medical LLMs.
3. We discuss the importance of evaluating *reasoning behaviour* in addition to performance metrics
4. We compare and contrast the current SOTA in *reasoning behaviour* for the medical field, and note a surprising lack of such studies.
5. We propose strategies to improve and evaluate the *reasoning behaviour* of medical LLMs, which will grant greater transparency

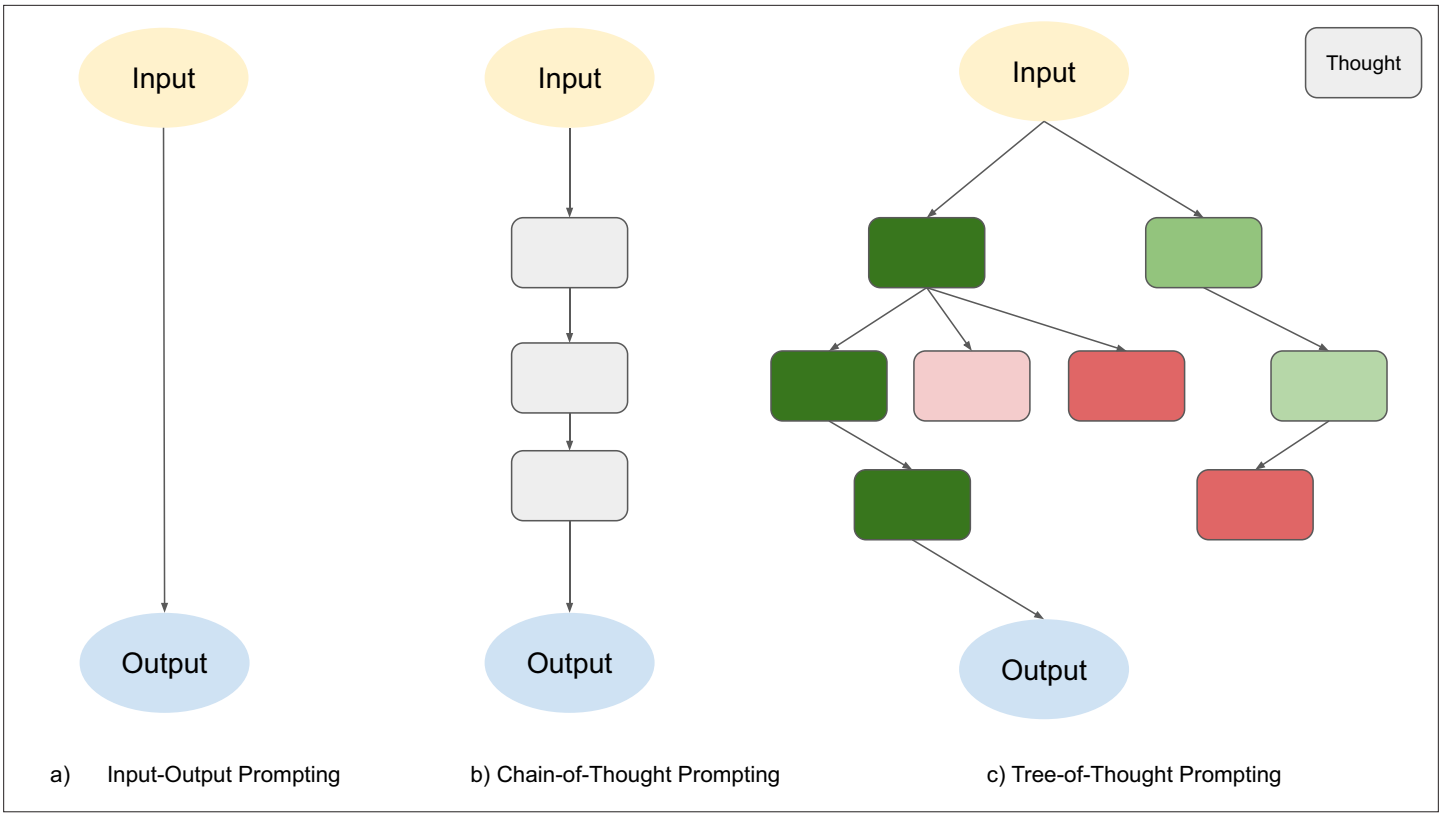

Figure 3. A schematic diagram illustrating different strategies for solving problems using large language models (LLMs). Each rectangular box represents a 'thought'—a meaningful segment of language that functions as an intermediate step in the reasoning or problem-solving process.

Primer of foundational concepts in reasoning

Prompting methods for reasoning

Prompting methods are lightweight techniques that guide LLMs to perform more structured reasoning without additional training. In this section, we focus on two representative prompting-based strategies: chain-of-thought (CoT) prompting, which encourages sequential reasoning through intermediate steps, and tree-of-thought (ToT) prompting, which allows the model to explore multiple reasoning paths through structured search and self-evaluation.

Chain of thought reasoning

Chain-of-Thought (CoT)CoT (***Figure 3***) reasoning is a prompting technique used to improve an LLM's ability to solve reasoning-related problems. Rather than producing an answer directly, the model decomposes the problem into a sequence of intermediate steps, thereby facilitating more structured and transparent reasoning (***Wei et al., 2022***). However, in some cases, LLMs may still generate incorrect intermediate steps (***Wang et al., 2022***).

Tree of thought reasoning

ToT (***Figure 3***) is an advanced reasoning framework that extends the problem-solving capabilities of LLMs. (***Yao et al., 2023***). Instead of token-by-token generating responses in a linear fashion, ToT prompts the model to generate coherent reasoning steps or 'thoughts', which are organised into a tree structure (***Yao et al., 2023***). Each node in this tree represents a partial solution, and the model can explore multiple possible continuations (branching), evaluate their promise using heuristic reasoning (state evaluation), and make decisions through structured search strategies such as breadth-first or depth-first search (***Yao et al., 2023***). This modular approach includes four core components: decomposing problems into thought steps, generating multiple candidate thoughts, heuristically evaluating them, and using search algorithms to navigate the tree (***Yao et al., 2023***). By combining these

elements, ToT enables LLMs to reason with greater depth, backtrack to previous nodes when necessary, and make more globally informed decisions, features that are particularly relevant for complex biomedical tasks such as differential diagnosis or clinical planning.

Agentic-based methods

Agent-based methods for LLM reasoning are an emerging paradigm designed to enhance the problem-solving capabilities of LLMs by structuring their operations as autonomous or semi-autonomous agents. These approaches address the limitations of standard LLM prompting by introducing explicit planning, memory, iterative decision-making, and tool usage.

In the medical domain, agent-based methods leverage LLMs as modular agents, each assigned a specific function. Examples include clinical triage, medical literature retrieval, summarisation of patient data (***Zi Yang et al., 2024***), decision support, or guideline compliance checking.

Common features of agent-based medical reasoning with LLMs include

- **Iterative planning:** The model decomposes complex clinical problems into sub-questions, enabling step-by-step analysis and dynamic adjustments as new information becomes available (***Wang et al., 2025***).
- **Memory integration:** Some agents maintain both short- and long-term memory, allowing them to track patient history, previous actions, and evolving diagnostic hypotheses over time (***Wang et al., 2025***).
- **Tool augmentation:** LLM agents can interact with external tools—such as medical databases like EHRs (***Shi et al., 2024***), drug databases (***Yue et al., 2024***), or medical knowledge graphs (***Yue et al., 2024***), medical calculators (***Zhu et al., 2024***), or literature search engines—to retrieve up-to-date information and perform specialised computations.
- **Reflection:** Agents incorporate feedback mechanisms (***Hong et al., 2024***) and reflective decision-making (***Yue et al., 2024***) to revise and improve reasoning dynamically, reducing hallucinations and adapting to new inputs.
- **Multi-agent collaborative group reasoning:** A multiagent reasoning framework involves deploying multiple specialised LLM-based agents (***Hong et al., 2024***; ***Yue et al., 2024***)—such as efficacy, safety, or diagnostic agents (***Yue et al., 2024***)—that collaboratively analyse clinical information, challenge each other's conclusions, and synthesise decisions through structured dialogue or argumentation (***Hong et al., 2024***). This mirrors multidisciplinary clinical teams and enhances reasoning transparency, robustness, and safety.

Learning-based approaches for reasoning: Supervised to reinforcement learning

LLMs can be prompted into producing CoT- sequence of tokens representing intermediate steps in the reasoning process. However, LLMs lack explicit training objectives that encourage deep, stepwise CoT before arriving at a final answer. Large reasoning models (LRMs), a subclass of LLMs, close this gap by being trained to perform extended reasoning in CoT, before taking actions or producing a final answer. The release of OpenAI's o1 series exemplifies this emerging trajectory.

In the emerging landscape of *learning-to-reason* in LRMs, the training process often begins with **supervised fine-tuning (SFT)**, where models are trained on labelled reasoning datasets to capture task-specific logic and patterns. However, as reasoning tasks grow more nuanced and open-ended, SFT alone becomes insufficient. To overcome these limitations, **reinforcement learning from human feedback (RLHF)** has been introduced. RLHF refines model output by training a reward model based on human preferences, enabling the LLM to generate responses that are more aligned with human-like reasoning: coherent, responsible and contextually appropriate.

Beyond training, recent research also demonstrates the benefits of *inference-time scaling*, sometimes also known as *test-time scaling* (***Snell et al., 2024***; ***Zhao et al., 2024***), where prompting LLMs to generate longer or multiple reasoning paths (e.g. via tree-of-thought prompting) can significantly improve inference-time performance. These advances in both training and inference signal a shift towards what has been described as LRMs—LLMs specifically optimised for robust, interpretable, and multi-step reasoning in CoT.

In the sections that follow, we describe two foundational learning strategies—SFT and RLHF—both of which have been shown to align LLM outputs more closely with expert medical reasoning and improve clinical utility.

Supervised fine-tuning

SFT is a process used to improve a pre-trained AI model, such as an LLM, so it performs better on specific tasks or domains (***Singhal et al., 2023***). It involves training the model on a carefully prepared dataset that includes input examples paired with the correct outputs (labels). This helps the model learn the desired behaviour more precisely by adjusting its parameters based on these examples.

The key steps in SFT are (***Singhal et al., 2023***)

1. Starting with a pre-trained model that already understands general language or knowledge.
2. Preparing a labelled dataset relevant to the specific task, where each input has a correct output.
3. Training the model on this dataset to fine-tune its parameters, so it responds accurately in the targeted context.
4. Evaluating and iterating to ensure the model improves without overfitting.

This method transforms a general-purpose model into a specialised model that delivers better performance on domain-specific tasks (***Singhal et al., 2023***).

Reinforcement learning with human feedback

RLHF (***Christiano et al., 2017***) in the context of LLMs is a technique used to teach LLMs to behave in ways that align better with human preferences (***Singhal et al., 2023***; ***Achiam et al., 2023***). Instead of relying solely on fixed rules or labelled data, RLHF uses human feedback to guide the learning process.

The standard RLHF pipeline has three stages: (***Ouyang et al., 2022***; ***Achiam et al., 2023***):

1. **SFT (optional):** Often one starts with an LLM that has been fine-tuned on high-quality responses with reasoning traces to some sample of prompts (datasets).
2. **Reward model training:**
   - Humans see multiple candidate outputs for the same prompt.
   - They rank or compare outputs by which one they prefer.
   - A small neural network (*reward model*) is trained to predict those human preference scores as a single scalar reward.
3. **Policy optimisation:**
   - The LLM (now called the *policy*) is fine-tuned with a reinforcement-learning algorithm (e.g. PPO).
   - At each update, the policy generates outputs, the reward model scores them, and the RL algorithm nudges the policy towards higher-reward actions.

Over successive iterations, the model learns to produce answers that are more helpful, more accurate, and more aligned with human values (***Ouyang et al., 2022***; ***Achiam et al., 2023***). As an example of its applicability, RLHF underpins the effectiveness of systems like GPT-4 (***Achiam et al., 2023***).

There are several variants of the RL step (PPO [***Schulman et al., 2017***], DPO [***Rafailov et al., 2023***], GRPO [***Shao et al., 2024***], etc.), each with its own trade-offs in stability, efficiency, and memory usage. We will not dive into their details here; instead, our focus will be on how this human-in-the-loop process is key to building LLMs that can carry out complex medical reasoning tasks.

Directed acyclic graphs

A directed acyclic graph (DAG) is a graphical structure made up of nodes (also called vertices) connected by arrows (edges) that indicate a specific direction of influence from one variable to another (***Foraita et al., 2014***). The term 'acyclic' means that the graph contains no closed loops—following the arrows from node to node will never bring you back to the starting point (***Foraita et al., 2014***). This enforces a clear, unidirectional flow of information, which is essential for modelling causal relationships (***Figure 4***).

DAGs are commonly used to represent processes or systems where order matters and repetition is not allowed (***Foraita et al., 2014***; ***Figure 4***). For example, DAGs can clarify whether the observed link between paracetamol use and childhood wheezing is due to a true causal effect or confounded by factors like viral infections (***Williams et al., 2018***).

This structure is particularly useful in biomedical AI for modeling causal relationships (***Williams et al., 2018***), clinical reasoning pathways (***Kiciman et al., 2023***), or decision-making logic in a transparent and interpretable way (***Foraita et al., 2014***; ***Naik et al., 2023***; ***Wang et al., 2021***).

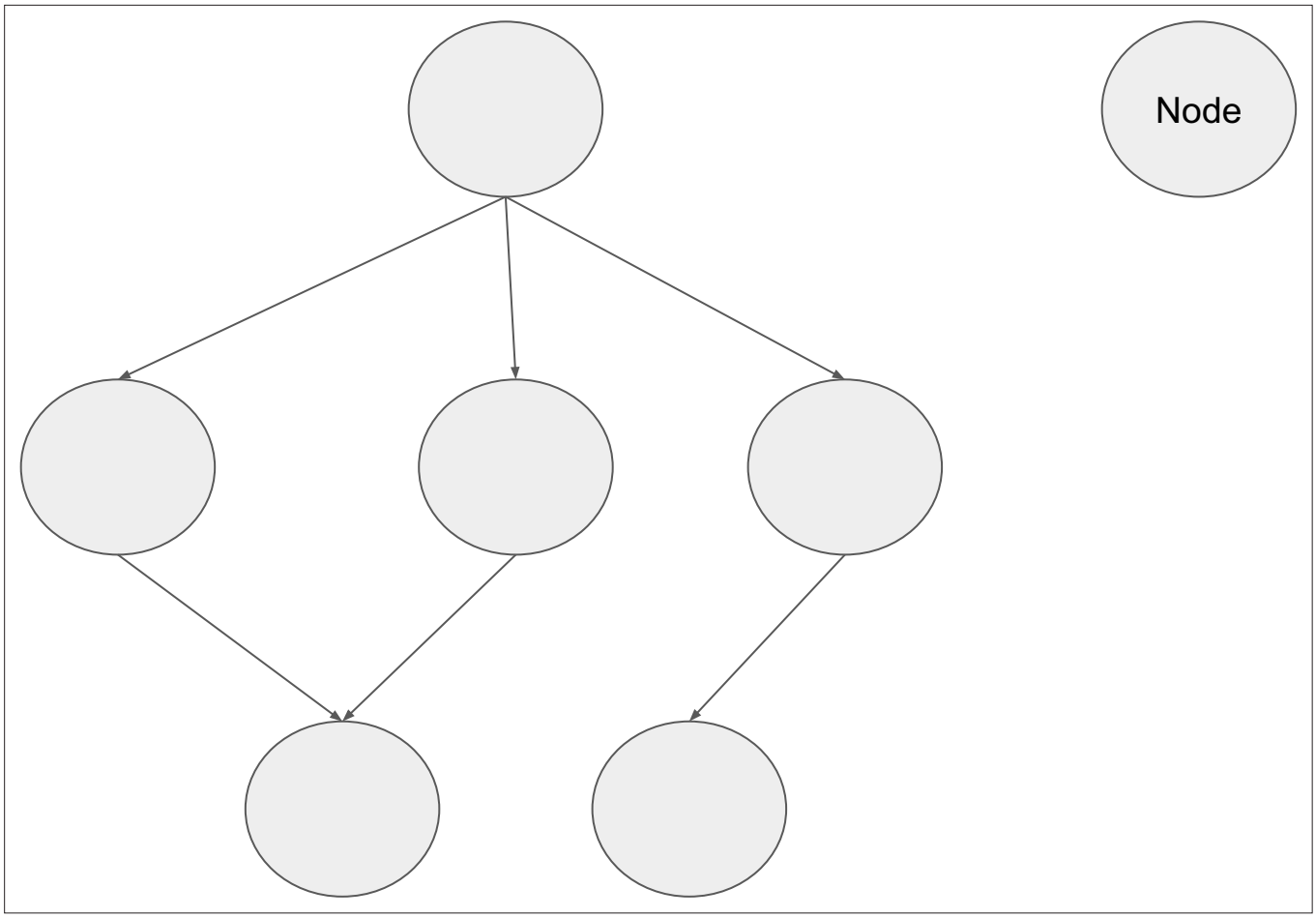

Figure 4. A sample directed acyclic graph.

What is reasoning behaviour in the context of medical LLMs?

First, we specifically define *reasoning behaviour* in the context of our review. It is important to note that the general term *reasoning* is used loosely across LLM-related literature, and often *reasoning behaviour* is not the focus of the experiment but high-level performance metrics are. Here, we slightly adapt the specific definitions of *reasoning* and *reasoning behaviour* respectively from ***Mondorf and Plank, 2024***, who define these concepts in the context of general LLMs. We also add a third definition: *reasoning outcome*.

> *Reasoning: 'The process of drawing conclusions based on available information'.*
> *Reasoning outcome: 'An event where reasoning reaches a conclusion.'*
> *Reasoning behaviour: 'The specific flow of logic within the scope of available information in a system that leads to a reasoning outcome.'*

Specifically, while the *outcome* of *reasoning* is an event where a conclusion is obtained, *reasoning behaviour* describes the *process of how* the conclusion is obtained (***Figure 5***). The vast majority of generic and medical LLMs focus on the former while disregarding the latter.

We apply the same definition to this review for medical LLMs.

Types of reasoning applicable to medical LLMs

This section reviews studies that extend beyond a high-level focus on task accuracy, focusing instead on evaluating the *reasoning behaviour* of LLMs. *Reasoning behaviour* can be subdivided into multiple categories (***Table 1***). However, for the purposes of this study, we focus mainly on subtypes of logical reasoning (***Holyoak and Morrison, 1999***) and causal reasoning (***Sloman, 2009***) that are common in

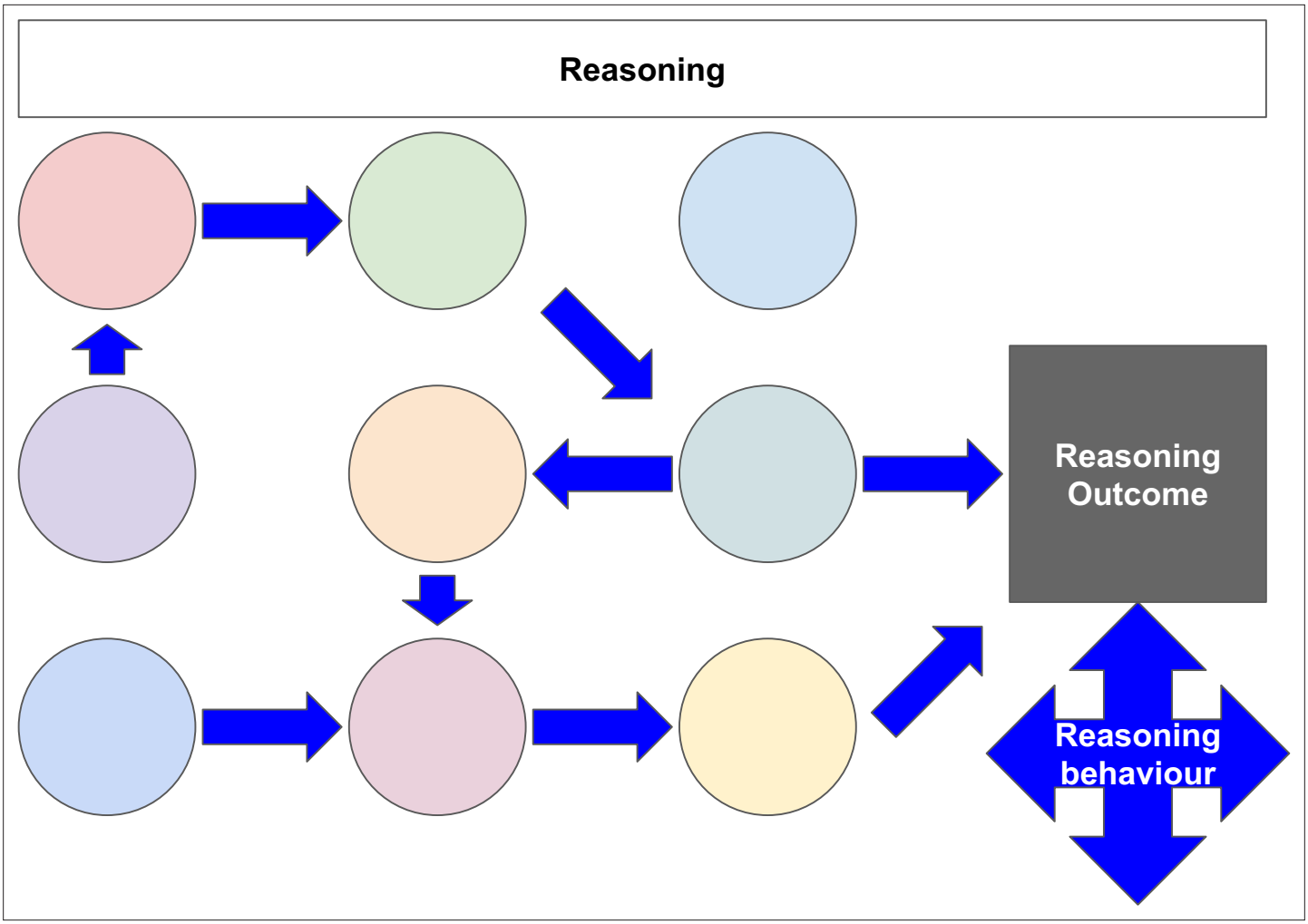

Figure 5. A graphical representation of *reasoning*, *reasoning outcome*, and *reasoning behaviour*. *Reasoning* encapsulates the process of drawing conclusions, arriving at a *reasoning outcome*. At a more fundamental level, *reasoning behaviour* describes the logical flow through the system that occurs during *reasoning*.

medical LLMs. In addition, we explore the less visible field of neurosymbolic reasoning. We note that other reasoning types such as mathematical reasoning (***Horsten, 2007***) may be more applicable to other categories of LLMs, which are not the focus of this review.

Logical reasoning

The study of logical reasoning addresses the question of how individuals infer valid conclusions from a set of given premises within a structured framework of logical rules and principles (***Mondorf and Plank, 2024***). ***Mondorf and Plank, 2024*** classify logical reasoning into deductive, inductive, and abductive reasoning. Deductive and inductive reasoning both work towards a general conclusion, with the key distinction being that deductive reasoning begins with a premise while inductive reasoning begins with observations. On a broader scale, abductive reasoning involves formulating plausible hypotheses that explain incomplete observations. The key distinction between deductive reasoning and both inductive/abductive reasoning is that deductive reasoning results in clear conclusions, while inductive/abductive reasoning may not necessarily achieve this. For more nuanced and low-level details on the distinction between the three types of logical reasoning in the context of LLMs, we refer the reader to other publications (***Mondorf and Plank, 2024***; ***Sun et al., 2023***; ***Yu et al., 2023b***).

Table 1. A table showing types of reasoning, their definition and examples.
Reasoning types are colour-coded for clarity. Logical reasoning encompasses abductive, deductive, and inductive subtypes.

Type of reasoning	Definition	General example	Medical example
Abductive	Inferring the most likely explanation for observed data or evidence.	Ali, Muthu, and Ah Hock breathe oxygen. Therefore, Ali, Muthu, and Ah Hock are likely human.	A patient has increased intercranial pressure, blurred vision and nausea. Therefore, the patient may have a brain aneurysm or ischaemic stroke.
Deductive	Reasoning from a set of premises to reach a certain conclusion.	All humans breathe oxygen. Rentap is human. Therefore, Rentap breathes oxygen.	A patient has increased intercranial pressure, blurred vision and nausea. A CT scan shows no bleeding or swelling. Therefore, the patient does not have a brain aneurysm.
Inductive	Inferring general principles based on specific observations.	All humans that I have seen breathe oxygen. Therefore, Rentap probably breathes oxygen.	A patient has increased intercranial pressure, blurred vision and nausea. A CT scan shows no bleeding or swelling. Therefore, the patient probably has an ischaemic stroke.
Symbolic*	The abstraction of a system into its component parts, which enables a more direct application of mathematics.	Rule: If an organism breathes oxygen and nitrogen, and exhales carbon dioxide → likely human. Observation: Ali, Muthu, Ah Hock, and Rentap exhibit this respiratory pattern. Conclusion: Therefore, they are probably human.	Rule 1: If a patient presents with increased intracranial pressure (ICP), blurred vision, and nausea → infer high intracranial pathology. Observation 1: The patient shows increased ICP, blurred vision, and nausea. Rule 2: If CT scan shows no bleeding or swelling → rule out haemorrhagic causes. Observation 2: CT scan reveals no evidence of bleeding or swelling. Rule 3: If high ICP and haemorrhage is ruled out → suspect ischaemic stroke. Conclusion: Therefore, the patient most likely has an ischaemic stroke.
Causal/ Counterfactual	Establishing a cause-and-effect relationship between events.	Ali, Muthu, Ah Hock, and Rentap exhibit this respiratory pattern.	A blood clot probably caused blockage in the brain leading to the stroke.

*We note that the term *symbolic reasoning* may be misleading as it is fundamentally an abstract *data representation* which simplifies the process of translating a scenario into a *reasoning* framework.

Causal/counterfactual reasoning

Causal reasoning refers to the ability to connect cause and effect in scenarios. In the context of medical and general LLMs, their capabilities are a matter of debate. Intuitively, providing cause and effect relationships improves one's understanding of a system. Correspondingly, providing this information to medical LLMs would improve a model's 'understanding' and has unsurprisingly emerged as an area of interest. This capability is essential in applications like medical diagnosis, where identifying causal links—such as between symptoms and potential conditions—can inform accurate and actionable insights. Causal reasoning involves not just recognising associations but distinguishing *directional* influences. In theory, knowing directionality grants the model the ability to infer, for instance, whether 'A causes B' or 'B causes A'. In real-world medical applications, larger LLMs like GPT-4.0 are capable of inferring causal direction between variables, allowing accurate diagnosis of neuropathic pain (***Kiciman et al., 2023***).

Symbolic reasoning

Symbolic reasoning—also known as *symbolic AI* or 'good old-fashioned artificial intelligence'—is a process that involves the use of mathematical symbols to represent concepts, objects, or relationships in order to facilitate reasoning, problem-solving, and decision-making. Unlike machine learning which relies on learning from vast datasets, this form of reasoning is characterised by its reliance on formal logic and structured representations, allowing for the manipulation of abstract symbols according to defined rules without requiring vast datasets. Symbolic systems execute explicit inference chains—for example, 'if symptom A and test result B, then diagnosis C'—mirroring the logic of clinical guidelines and expert systems such as MYCIN (***van Melle, 1978***) or INTERNIST (***Miller et al., 1985***). This approach enhances interpretability and trustworthiness in biomedical applications, since the reasoning steps are transparent and auditable.

Neurosymbolic reasoning

Neuro-symbolic AI (N-SAI) is an interdisciplinary field that aims to harmoniously integrate neural networks with symbolic reasoning techniques. Its overarching objective is to establish a synergistic connection between symbolic reasoning and statistical learning, harnessing the strengths of each approach. In the context of N-SAI, the symbolic system is used to represent predefined rulesets and knowledge bases, which then streamlines the process of making inferences and highlighting relationships between entities. Crucially, it is more transparent and more interpretable to humans. Meanwhile, the *neuro* component refers to artificial neural networks in the context of large-scale statistical learning. Artificial neural networks are adept at scale in classification, prediction, and pattern recognition as well as processing unstructured data. Therefore, unifying and leveraging the strengths of both would hypothetically lead to the best of both worlds (***Sheth et al., 2023***).

Trends in existing medical LLMs

While there is no shortage of LLMs applied to medical problems, there is a striking lack of methods which leverage *reasoning behaviour* in their operation (***Table 2***). Comparing and contrasting this small subset of methods reveals some interesting trends.

First, inspecting their foundational or base models shows that unsurprisingly, most of these methods are built on generic LLMs, commonly GPT (***Wu et al., 2023***) or LLaMA (***Touvron et al., 2023***) variants. This is likely due to their demonstrated effectiveness in day-to-day tasks, with more modern variants being shown to be surprisingly effective in clinical applications as-is (***Homolak, 2023***). However, it is notable that many approaches utilise multiple base models, with no single method relying on one model type. Relying on multiple models is unsurprising, as combining the strengths of multiple models is likely to boost overall effectiveness.

Second, most of their *reasoning behaviour* is derived from variants of CoT (***Wei et al., 2022***) processes or reinforcement learning, likely because both techniques closely mimic cognitive processes fundamental to reasoning. CoT enables models to break down complex medical cases into a series of logical steps, mirroring the structured, stepwise reasoning that healthcare professionals apply. Additionally, few-shot learning complements CoT, allowing LLMs to 'learn' from the input prompt, generalise from minimal clinical examples, and adapt quickly to nuanced cases—a useful capability in medicine where data can be sparse or highly specialised. Meanwhile, reinforcement learning allows models to refine their reasoning capabilities through practice in a virtual simulation, improving accuracy through iterative feedback.

We note with interest that SFT and RLHF have been widely adopted to train medical LLMs for reasoning tasks, giving rise to several domain-specific models such as Huatuo GPT-o1 (***Chen et al., 2024***), MedR (***Lai et al., 2025***), MedVLM-R1 (***Pan et al., 2025***), and MedFound (***Liu et al., 2025***). We refer to this class of systems as LTMs, a subclass of LLM, trained to perform extensive CoT reasoning. These LRMs employ a variety of RLHF strategies, including Proximal Policy Optimization (PPO) (***Schulman et al., 2017***) for Huatuo GPT-o1 (***Chen et al., 2024***), Direct Preference Optimization (DPO) (***Rafailov et al., 2023***) for MedFound (***Liu et al., 2025***), and Group Relative Policy Optimization (GRPO) (***Shao et al., 2024***) for both MedR1 (***Lai et al., 2025***) and MedVLM-R1 (***Pan et al., 2025***). Interestingly, findings from the MedR1 study revealed that models trained with GRPO not to output intermediate reasoning traces performed better in terms of final accuracy compared to those trained to explicitly generate intermediate steps (***Lai et al., 2025***). This challenges the prevailing assumption that 'more reasoning always leads to better outcomes', suggesting that, in some cases, compressed or implicit reasoning may yield higher task performance.

Third, the *reasoning behaviour* of most methods can be categorised as deductive reasoning, although there are a few cases of abductive and causal/counterfactual reasoning. Here, it is also worth noting that while LLMs excel in abductive reasoning tasks in multiple-choice scenarios, they are considerably less effective in generating hypotheses from scratch which may be of value in clinical use (***Gouveia and Malík, 2024***). Since the overall goal of clinical diagnosis is to determine the disease affecting a patient from causative agents, the prevalence of deductive and, to a lesser extent, causal/counterfactual reasoning makes sense.

Finally, training datasets used vary widely in both scope and size, ranging across many different medical conditions, source material and between hundreds to thousands of samples. We observed no single standardised training dataset used by each approach, and as with architecture types many

Table 2. A table showing medical reasoning methods, their defining characteristics, and approach to reasoning.

Method name	Base architecture/ method	Reasoning improvement strategy	Type of Reasoning	Advantages	Disadvantages	Dataset	Github
Savage et al., 2024	GPT-3.5; GPT-4.0	Chain-of-thought (diagnostic reasoning)	Deductive	Easy to implement	Scope is limited to GPT models, focusing exclusively on English medical questions	Modified MedQA USMLE; NEJM (*New England Journal of Medicine*) case series	
Kwon et al., 2024	GPT-4.0; OPT; LLaMA-2; 3D ResNet	Chain-of-thought; knowledge distillation (via SFT)	Deductive	Lightweight and practical to use	Tight scope to limited disease conditions	Alzheimer's Disease Neuroimaging Initiative (ADNI); Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL)	https://github.com/ktio89/ClinicalCoT
MEDDM Binbin et al ***Li et al., 2023***	GPT	Chain-of-thought; clinical decision trees	Deductive	Adaptable to different systems	Heavy data collection to generate clinical guidance trees	Medical books, treatment guidelines, and other medical literature	
DRHOUSE ***Yang et al., 2024a***	GPT-3.5; GPT-4.0; LLaMA-3 70b; HuatuoGPT-II; MEDDM	Chain-of-thought; clinical decision trees	Deductive	Objective sensor measurement	Available datasets are currently limited	MedDG; KaMed; DialMed	
DR. KNOWS ***Gao et al., 2023b***	Vanilla T5; Flan T5; ClinicalT5; GPT	Chain-of-thought; extracted explainable diagnostic pathway	Deductive; neurosymbolic	Hybrid method improves accuracy; provides explainable diagnostic pathways	Particularly fragile to missing data	MIMIC-III; In-house EHR	
TEMED-LLM ***Bisercic et al., 2023***	text-davinchi-003; GPT-3.5; logistic regression; decision tree; XGBoost	Few-shot learning, tabular ML modelling; Neurosymbolic	Deductive	End-to-end interpretability, from data extraction to ML analysis	Requires human experts	EHR dataset (kaggle; see referenced publication for details)	
EHRAgent ***Shi et al., 2024***	GPT-4	Autonomous code generation and execution for multi-tabular reasoning in EHRs	Deductive	Facilitates automated solutions in complex medical scenarios	Non-deterministic; limited generalisability	MIMIC-III; eICU; TREQS	https://github.com/wshi83/EhrAgent; https://wshi83.github.io/EHR-Agent-page
AMIE ***Tu et al., 2024***	PaLM 2	Reinforcement learning	Deductive	Effectively handles noisy and ambiguous real-world medical dialogues	Computationally expensive and resource-intensive; simulated data may not fully capture real-world clinical nuances	MedQA; HealthSearchQA; LiveQA; Medication QA in MultiMedBench, MIMIC-III	
ArgMed-Agents ***Hong et al., 2024***	GPT-3.5-turbo; GPT-4	Chain-of-Thought; symbolic reasoning: neurosymbolic	Deductive	Training-free enhancement; explainability matches fully transparent, knowledge-based systems	Artificially restricted responses that do not match real-world cases	MedQA; PubMedQA	

Table 2 continued on next page

Table 2 continued

Method name	Base architecture/ method	Reasoning improvement strategy	Type of Reasoning	Advantages	Disadvantages	Dataset	Github
Fansi ***Tchango et al., 2022a***	BASD (baseline ASD): multi-layer perceptron (MLP) diaformer	Reinforcement learning	Deductive	Closely align with clinical reasoning protocols	Limited testing on real patient data	DDxPlus	https://github.com/mila-iqia/Casande-RL
MEDIQ ***Li et al., 2024***	LLaMA-3-Instruct (8B, 70B); GPT-3.5; GPT-4	Chain-of-thought; information-seeking dialogues	Abductive	Robust to missing information	Available datasets are limited Proprietary; artificially restricted responses that do not match real-world cases	iMEDQA; iCRAFT-MD	https://github.com/stellalisy/mediQ
Naik et al., 2023	GPT-4	Causal network generation	Causal/ counterfactual	Uses general LLMs	Lacks a specialised medical knowledge base	Providence St.Joseph Health (PSJH's) clinical data warehouse	
Gopalakrishnan et al., 2024	BioBERT; DistilBERT; BERT; GPT-4; LLaMA	Causality extraction	Causal/ counterfactual	Easy to implement	Tight scope to limited disease conditions	American Diabetes Association (ADA); US Preventive Services Task Force (USPSTF); American College of Obstetrics Gynecology (ACOG); American Academy of Family Physician (AAFP); Endocrine Society	https://github.com/gseetha04/LLMs-Medicaldata
InferBERT ***Wang et al., 2021***	ALBERT; Judea Pearl's Do-calculus	Causal inference using do-calcus	Causal/ counterfactual; mathematical	Establishes causal inference	Tight scope to limited disease conditions; highly restrictive input format	FAERS case reports from the PharmaPendium database	https://github.com/XingqiaoWang/DeepCausalPV-master
Emre Kıcıman	text-davinci-003, GPT-3.5-turbo, and GPT-4	Determine direction of causality between pairs of variables	Causal/ counterfactual	Highly accurate for large models	Limited reproducibility due to dependency on tailored prompts	Tübingen cause-effect pairs dataset.	https://github.com/py-why/pywhy-llm
Huatuo GPT-o1 ***Chen et al., 2024***	LLaMA-3.1-8B-Instruct and LLaMA-3.1-70B-Instruct	Supervised fine-tuning and PPO	Deductive reasoning	Instils multi-step reasoning in medical LLMs; built-in interpretability as LLM can output reasoning traces along with answer	Limited evaluations as evaluations cover accuracy scores on medical MCQ benchmarks	Adapted from MedQA-USMLE and MedMcQA	https://github.com/FreedomIntelligence/HuatuoGPT-o1
Med-R1 ***Lai et al., 2025***	Qwen2-VL-2B	Supervised fine-tuning and GRPO	Deductive reasoning	Joint image-text and multi-task reasoning; built-in interpretability as LLM can output reasoning traces along with answer	Rethinking the 'More Thinking is Better' Assumption	OmniMedVQA	https://github.com/Yuxiang-Lai117/Med-R1

Table 2 continued on next page

Table 2 continued

Method name	Base architecture/ method	Reasoning improvement strategy	Type of Reasoning	Advantages	Disadvantages	Dataset	Github
MedVLM-R1 ***Pan et al., 2025***	Qwen2-VL-2B	Supervised Fine-tuning and GRPO	Deductive Reasoning	Joint image-text reasoning; Built-in interpretability as LLM can output reasoning traces along with answer	Limited evaluations as evaluations cover accuracy scores on medical MCQ benchmarks	VQA-RAD, SLAKE, PathVQA, OmniMedVQA, and PMC-VQA	https://huggingface.co/JZPeterPan/MedVLM-R1
MedFound ***Liu et al., 2025***	176 billion parameter LLM pretrained from scratch	Supervised fine-tuning and DPO	Deductive reasoning	- Self-bootstrapped Chain-of-Thought fine-tuning; Rigorous human evaluation of reasoning traces with rubric; built-in interpretability as LLM can output reasoning traces along with answer	Proprietary EHR datasets aren't fully open, hindering exact reproduction	MedCorpus, MedDX-FT and MedDX-Bench	https://github.com/medfound/medfound
DeepSeekR1 ***Guo et al., 2025***	DeepSeek-V3-Base	Supervised fine-tuning and GRPO	Deductive reasoning	Built-in interpretability as LLM can output reasoning traces along with answer	Pre-training and reasoning datasets aren't open-sourced	-	https://github.com/deepseek-ai/DeepSeek-R1

approaches used multiple training datasets. Most datasets were of the same modality (text data only), though some medical imaging datasets (MRI scans) were present. MIMIC-III was the most commonly used text dataset, with a combination of medical literature and other publicly available datasets (***Johnson et al., 2016***). Therefore, due to the differences in scope, strategy and data used by each approach, directly comparing *reasoning behaviour* across all models simultaneously is not presently feasible.

Aside from deductive reasoning, causal reasoning and neurosymbolic reasoning have also been demonstrably effective (***Table 2***). However, example use cases are considerably less common. Current causal inference tests have a limited scope, such as determining the direction of causality between variable pairs, and their performance in more open-ended or nuanced causal inference as well as counterfactual reasoning remains unexplored. Meanwhile, neurosymbolic reasoning strategies exploit their inherently grounding properties to address the more fundamental issue of hallucinations in LLMs (***Huang et al., 2023***). The diversity of strategies is striking—some methods exploit agent-based approaches to tailor argumentation schema and symbolic solvers for clinical reasoning (***Hong et al., 2024***), while others integrate dynamic medical ontologies in an attempt to more closely align *reasoning behaviour* with medical knowledge (***Gao et al., 2023a***).

As each approach varies widely in scope and implementation, the advantages and disadvantages of each approach are broad. Generally, approaches using graph and decision tree-based strategies are easier to interpret due to their more deterministic nature, but may be less effective in ambiguous or complex cases (which are common in clinical practice). Meanwhile, methods which are more robust to noise or complex use cases are limited by a highly restricted scope, availability of training resources, and a large computational footprint. Among these methods, DR HOUSE (***Yang et al., 2024a***) is unique due to its EHR-free approach, only relying on objective sensor data to circumvent variance in clinical note interpretation. Unfortunately, the code associated with many of these methods is not publicly available under an open-source licence, which limits our ability to inspect them in close detail. It is worth mentioning that medical LLMs are equally affected by some of the deeply rooted issues that plague general purpose LLMs as well, for instance, memorisation (***Hartmann et al., 2023***) and hallucination (***Huang et al., 2023***).

Evaluating reasoning behaviour in medical LLMs

To date, a standardised methodology for assessing the reasoning capabilities of LLMs is absent. We review the current state-of-the-art in evaluation frameworks for analysing the *reasoning behaviour* of LLMs in medical tasks and we categorise evaluation methodologies into four distinct groups: (i) conclusion-based, (ii) rationale-based, (iii) interactive, and (iv) mechanistic evaluations (***Table 3***).

Conclusion-based evaluation

In conclusion-based evaluation schemes, the focus is on the model's final answer rather than the reasoning process that led to it. Although this outcome-focused approach may overlook the model's underlying rationales, it can still offer valuable but limited insights into the model's reasoning patterns,

Table 3. Comparison of reasoning evaluation paradigms in medical LLMs.

Evaluation paradigm	Conceptual focus	Typical implementation and metrics
Conclusion-based	Assesses correctness of the **final answer** only, without inspecting the reasoning path.	Automated scoring on Q-&-A benchmarks (e.g. MedQA, MedMCQA). Metrics: Accuracy, exact-match, F1. Fast, reproducible, but offers only high-level insight.
Rationale-based	Evaluates the **logic chain** or narrative explanation produced by the model. Focuses on **coherence, validity, and completeness** of reasoning traces.	Manual expert review or rubric-based grading of CoT. Automated graph checks (e.g. DAG similarity, causal-direction tests). Metrics: Bayesian Dirichlet score, Normalised Hamming Distance.
Mechanistic	Probes **low-level internal signals** to answer "*why did the model arrive here?*". Targets feature attribution and internal attention contributions.	Explainable-AI toolkits (Integrated Gradients, SHAP, attention rollout). Outputs saliency maps or keyword heat-maps for clinician inspection.
Interactive	Treats evaluation as a **dialogue or game**; dynamically stresses the model in real time. Explores the *response space* by challenging, re-prompting, or role-playing.	Game-theoretic tasks (e.g. debate, self-play). Rich insights but lower reproducibility; requires human-in-the-loop or scripted agents.

especially if there is a clear cause and effect between different premises and conclusions in which the *reasoning behaviour* may be more self-evident.

A wealth of benchmark data exists for the purpose of straightforward score-wise conclusion-based evaluation. A subset of simple benchmarks is available on the open medical LLM leaderboard, which covers question-and-answer tasks (***Jin et al., 2020***; ***Pal et al., 2022***; ***Jin et al., 2019***; ***Hendrycks et al., 2020***). Notably, some datasets consist of multiple-choice questions and some consist of short-answer questions. More sophisticated examples incorporating detailed data in multiple formats (***Tchango et al., 2022b***) or with more refined metrics also exist (***Liao et al., 2024***).

Inroads into more nuanced frameworks to gain deeper insight in *reasoning behaviour* for conclusion-based evaluation have been made. These take the form of frameworks evaluating paradigms at various levels, with high-level theoretical frameworks available (***Bragazzi and Garbarino, 2024***). A lower-level and detailed framework more suited for direct implementation is DR BENCH, which specifically assesses medical knowledge representation, clinical evidence integration and diagnosis accuracy (***Gao et al., 2023a***). In the process, an expanded suite of specific tasks is carried out to evaluate these three interdependent elements, and an accuracy score is reported for each sub-task to evaluate reasoning at a high level. However, we emphasise that conclusion-based evaluation is only capable of yielding high-level information due to its intrinsic nature.

Rationale-based evaluation

In contrast to high-level conclusion-based evaluation schemes, rationale-based evaluation methods are process-driven instead of being outcome-driven. Their focus is on examining the reasoning process or *reasoning traces* generated by the model, typically assessing their logical validity and coherence. As rationale-based evaluation methods targeted at medical language models are relatively scarce and operate under distinct paradigms, we will discuss them individually on a case-by-case basis.

The most straightforward but resource-heavy approach was to manually evaluate answers using the skills of domain experts. These domain experts were blinded to the questions and identified logical fallacies as well as inaccuracies directly in provided rationale (***Savage et al., 2024***). More structured variants of this method have emerged, where expert clinicians employ clinically validated rubrics to score LLM-generated reasoning traces. Notable examples include the CLEVER (CLinical EValuation for Effective Reasoning in Diagnosis) rubric (***Liu et al., 2025***) and the Revised-IDEA (R-IDEA) rubric (***Esteitieh et al., 2025***; ***Cabral et al., 2024***).

Furthermore, clinicians have conducted deeper analyses of models such as DeepSeekR1 using USMLE-style multiple-choice questions, uncovering several notable failure modes. These included anchoring on initial symptoms while disregarding contradictory findings, omission of standard-of-care treatments, misuse of laboratory values or drug mechanisms, and confusion between clinically similar entities (e.g. troponin vs. CK-MB) (***Moëll et al., 2025***). Of particular interest was a failure mode termed CoT mismatch (***Moëll et al., 2025***), where the reasoning path supported one answer, but the model selected a different final option (***Moëll et al., 2025***)—highlighting that reasoning traces, while useful, are not a perfect proxy for model interpretability or correctness. We note the value and effectiveness of clinician-led evaluations of LLM reasoning traces as they offer nuanced insights into clinical reasoning quality. However, these methods are inherently time-consuming and reliant on the availability of expert clinicians.

To reduce the reliance on time-intensive clinician reviews, recent work has turned to automated reasoning-trace evaluation (***Zhou et al., 2025***; ***Wu et al., 2025***). When a high-quality 'gold-standard' CoT exists, generated traces can be compared step-by-step using text-similarity metrics such as BLEU, METEOR, or BERTScore (***Zhou et al., 2025***). Additionally, a secondary 'judge' LLM can be prompted to assess another model's chain of thought—either by comparing it against a provided standard (***Wu et al., 2025***; ***Zhou et al., 2025***; ***Qiu et al., 2025***) or in the case where no references exist the 'judge' LLM can be prompted to evaluate for logical coherence and clinical relevance based on its own internal reasoning (***Zhou et al., 2025***). Another LLM evaluation workflow integrated online search to fact check reasoning traces (***Qiu et al., 2025***).

Conversely, an automated approach applied DAG to represent underlying relationships in complex medical datasets, including cancer (***Naik et al., 2023***). In implementation, a DAG was constructed by predicting which factors might influence others, and accuracy was scored with a Bayesian Dirichlet metric measuring the similarity of the resultant graph with the ground truth of real patient data. In

addition, a separate method also applied DAG, but exploited it to infer the direction of causality between variable pairs (***Kiciman et al., 2023***). Accuracy was then measured using normalised Hamming distance (NHD) as a similarity metric between the resultant and ground truth patient outcome or diagnostic graph.

Mechanistic evaluation

Similarly, mechanistic evaluation of *reasoning behaviour* is process driven with the aim of examining low-level reasoning traces. In contrast to rationale-based evaluation, mechanistic evaluation delves deeper into the underlying processes that drive a model's response, aiming to uncover the fundamental questions of 'how' and 'why' associated with an outcome.

In practice, feature attribution methods can be exploited to study *reasoning behaviour* by highlighting keywords which are conceptually identical to features of interest in medical LLMs. These explainable AI (XAI) methods compute an attribution score for each input feature to represent its contribution to a model's prediction, which can be calculated and represented with a variety of metrics (***Sundararajan et al., 2017***; ***Lundberg, 2017***). For example, a hypothetical medical scenario may show that the key words "blocked nose" are strongly weighted in a positive influenza prediction. In this context, the key word is equivalent to the reasoning trace, and is shown to impact the model's *reasoning behaviour*. A conceptually similar strategy has been applied to explain predicted diseases from patient-doctor dialogues (***Ngai and Rudzicz, 2022***).

Interactive evaluation

Finally, a more open-ended approach to reviewing *reasoning behaviour* is interactive evaluation. Unique to other strategies, it is reactive and engages with the LLM directly during evaluation, adjusting questions to fit the model's response. This deeper exploration of the 'response space' tests and further exposes the model's reasoning capabilities as well as limitations (***Zhuang et al., 2024***). Variants of interactive evaluation may challenge the model's conclusions directly (***Wang et al., 2023***) or use game-theoretical scenarios to probe reasoning depth (***Bertolazzi et al., 2024***).

One notable implementation of this paradigm is Sequential Diagnosis Benchmark (SDBench) (***Nori et al., 2025***). Most existing medical LLM benchmarks present models with all clinical facts at once and assess multiple-choice accuracy—conditions far removed from real-world diagnostic workflows. To address this, SDBench (***Nori et al., 2025***) introduces a simulation-based framework that converts 304 NEJM Clinicopathological Conference (CPC) cases into interactive diagnostic scenarios. Here, the LLM or physician must sequentially ask patient history questions, order diagnostic tests (each with an associated monetary cost), and commit to a final diagnosis (***Nori et al., 2025***)—closely mimicking real clinical cases under uncertainty and cost constraints. This setup not only enables evaluation of diagnostic accuracy but also the diagnostic yield of tests as measured by total cost per case. By transforming passive vignette-based evaluation into a structured, decision-based simulation, SDBench enhances realism while maintaining standardisation and reproducibility within interactive evaluation. However, it is not without limitations: the CPC dataset is biased towards complex cases where there are no healthy patients in the medical cases, test costs are drawn only from US price lists, and benchmarking physicians were restricted from using web search to prevent them from finding the original CPC cases (***Nori et al., 2025***).

Unfortunately, a critical flaw of this evaluation method is its lack of reproducibility and standardisation due to its reactive nature. Currently, one exception exists, circumventing irreproducibility by sidestepping the requirement for a prompt (***Wang and Zhou, 2024***). Nevertheless, we note the strong advantages of interactive evaluation and note that it remains relatively unexplored in the current medical LLM literature. Refinement of the core method and further investigation of strategies (such as the aforementioned prompt skipping; ***Wang and Zhou, 2024***) to counteract its limitations have the potential to raise its reproducibility to reasonable levels.

Summary of evaluation strategies

To our surprise, we found few existing studies of *reasoning behaviour* evaluation in a medical LLM context. Nevertheless, we note several broad insights from the few existing studies matching our scope: (a) graph-theoretic approaches are intuitively applicable to evaluating causal or counterfactual *reasoning behaviour* due to their representation of cause-and-effect, (b) feature attribution methods

provide a low-level glimpse into medical reasoning, (c) interactive evaluations like SDBench (***Nori et al., 2025***) transform static multiple-choice medical benchmarks into interactive simulations that more realistically reflect clinical reasoning under uncertainty and resource constraints, (d) while manual clinician evaluations of reasoning traces offer nuanced insights that traditional metrics miss, but they are difficult to scale due to time and expertise requirements, and (e) *reasoning behaviour* evaluation methods are complementary, with the potential of being applied simultaneously to obtain a better understanding in cases where the model configuration allows.

Towards transparency in medical LLMs

Given our findings, we pose the central question: **How can we develop methods that expose the underlying *reasoning behaviour* in medical LLMs?**

To answer this, we introduce two conceptual and theoretical frameworks intended to expose reasoning behaviour in medical LLMs with orthogonal design principles. These frameworks are designed to meet two important criteria: (a) low-level *reasoning behaviour* must be visible and the framework should be (b) task-agnostic. Each framework would consist of three broad stages: (a) data preprocessing, (b) model training, and (c) interpretability via extraction of *reasoning behaviour*. These designs—one straightforward and one relatively more complex—are detailed in 'Theoretical frameworks for transparent reasoning behaviour' and visualised in ***Figure 6***.

Additionally, we observe with interest the growing prominence and impressive capabilities of LRMs such as Anthropic's Claude 3.7 Sonnet, Google's Gemini 2.5, and OpenAI's o1. Similarly, these models demonstrate key attributes aligned with our framework criteria: (a) visibility into low-level reasoning behaviour that supports interpretability, and (b) task-agnostic flexibility across domains. Although the precise details of their training pipelines remain undisclosed, these models likely rely on a combination of extensive supervised fine-tuning on curated reasoning datasets, large-scale reinforcement learning for reasoning alignment, and inference-time scaling techniques. These approaches have enabled such models to excel on existing benchmarks from math and logic benchmarks to coding benchmarks (***Guo et al., 2025***; ***Yang et al., 2024b***; ***Bai et al., 2025***). Open-source alternatives such as DeepSeek R1 (***Guo et al., 2025***) and Alibaba's Qwen 2.0 (***Yang et al., 2024b***), Qwen 2.5 (***Bai et al., 2025***), as well as domain-specialised models like MedVLM-R1 (***Pan et al., 2025***), MedR1 (***Lai et al., 2025***), and MedFound (***Liu et al., 2025***), have adopted broadly similar strategies tailored to medical contexts. In **'**Open challenges in large reasoning models', we outline several open challenges that constrain further progress in this space.

Theoretical frameworks with more transparent *reasoning behaviour*

We begin the simplistic framework by restricting input data scope to standardised data formats. To this end, TEMED-LLM (***Bisercic et al., 2023***) can be used to parse textual data into tables in the preprocessing stage with a predetermined format. Structured input has multiple advantages, being consistent and more easily ingested into software. In addition, a side effect is further simplification of data. An advantage of this which may benefit machine learning algorithms is a noise reduction while increasing variance in the data. However, we note that a degree of low-level feature loss is possible. Next, we consider that while deep learning is a powerful tool, more conventional machine learning approaches are often sufficient in many cases. We exploit the tabular nature of the data and leverage tree-based methods, which include examples such as XGBoost (***Chen and Guestrin, 2016***) or Random Forest (***Breiman, 2001***). While straightforward, these are effective and particularly suited to $P>>n$ problems common across the biological sciences, where there are far more features per sample than there are samples, that is, the 'curse of dimensionality' (***Xu and Jackson, 2019***). In addition, tree-like approaches have additionally benefited from properties that make them more interpretable. Exploiting this property allows us to generate decision sets which are interpretable during model training (***Yu et al., 2023a***), hence exposing *reasoning behaviour*.

Our second proposed framework requires specific context. We refer to the categorisation of reasoning into two systems of thinking (***Evans and Stanovich, 2013***; ***Figure 7***). 'System 1' thinking refers to more instinctive decision-making based on learned patterns and experiences. Although conventional LLMs often fall into this category, they perform effectively given their vast input corpora, similar to a human memorising vast quantities of data. Conversely, 'System 2' thinking refers to more

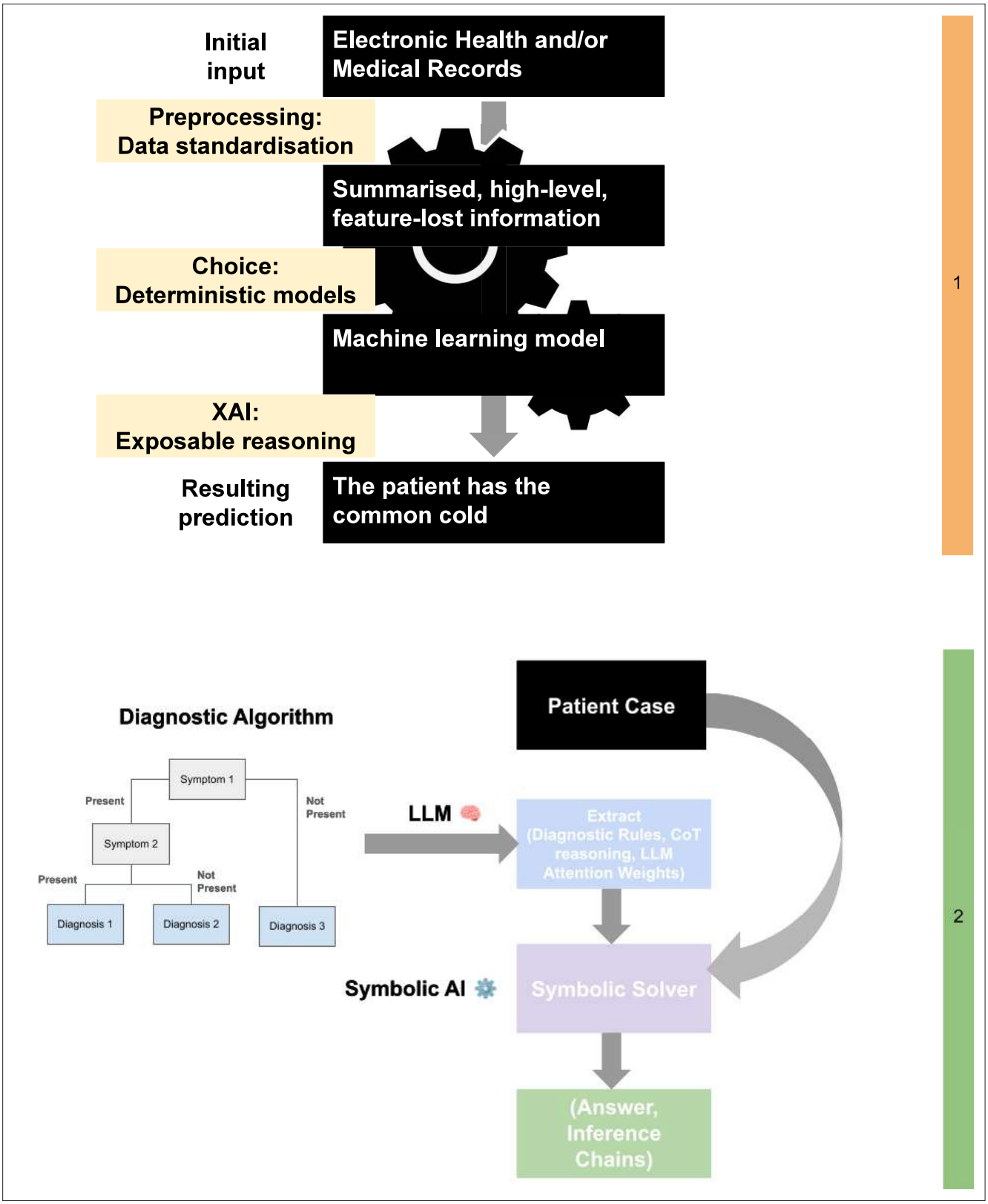

Figure 6. Two frameworks with a focus on exposing *reasoning behaviour*. Note that the two frameworks are independent but shown together to facilitate easier comparison. Top: input data is standardised and fed to tree-based models. The deterministic nature of trees is exploited for achieving transparency for *reasoning behaviour*. Bottom: an integrative framework of combining the complementary strengths of LLM and Symbolic Reasoning. The medical LLM extracts diagnostic rules from clinical algorithms, along with its chain-of-thought (CoT) reasoning and attention weights. These

Figure 6 continued on next page

Figure 6 continued

diagnostic rules, together with patient case inputs, are provided to the symbolic solver, which determines the final diagnosis and generates inference chains as its reasoning trace.

thorough and conscious thinking used in problem-solving and requires relatively more effort to achieve, both in humans and machine learning. Neither system operates fully independently, where the rapid assessments in 'System 1' thinking form the foundation for more methodical 'System 2' thinking. We note that we neither support nor disregard this overall viewpoint, but find value in using this angle to frame our proposed strategy, which addresses 'System 1' limitations inherent to conventional LLMs.

Next, we observe with interest that Symbolic AI (SAI) incorporating symbolic reasoning excels in 'System 2' thinking. In contrast, SAI has limited performance in 'System 1' thinking, lacking the capacity for rapid, intuitive pattern recognition and memorisation, hindering its performance in tasks where large volumes of unstructured data are processed. In order to utilise SAI more effectively, a more defined knowledge representation representing the search space with an optimal solution is often required.

Therefore, our final proposed framework synthesises LLMs and SAI, leveraging their complementary traits to supplement each other. The methodology consists of three core stages: (a) data preprocessing, (b) parallel model integration, and (c) reasoning extraction. One would first aim to generate a structured dataset which can be leveraged by LLM and SAI in data preprocessing, considering that

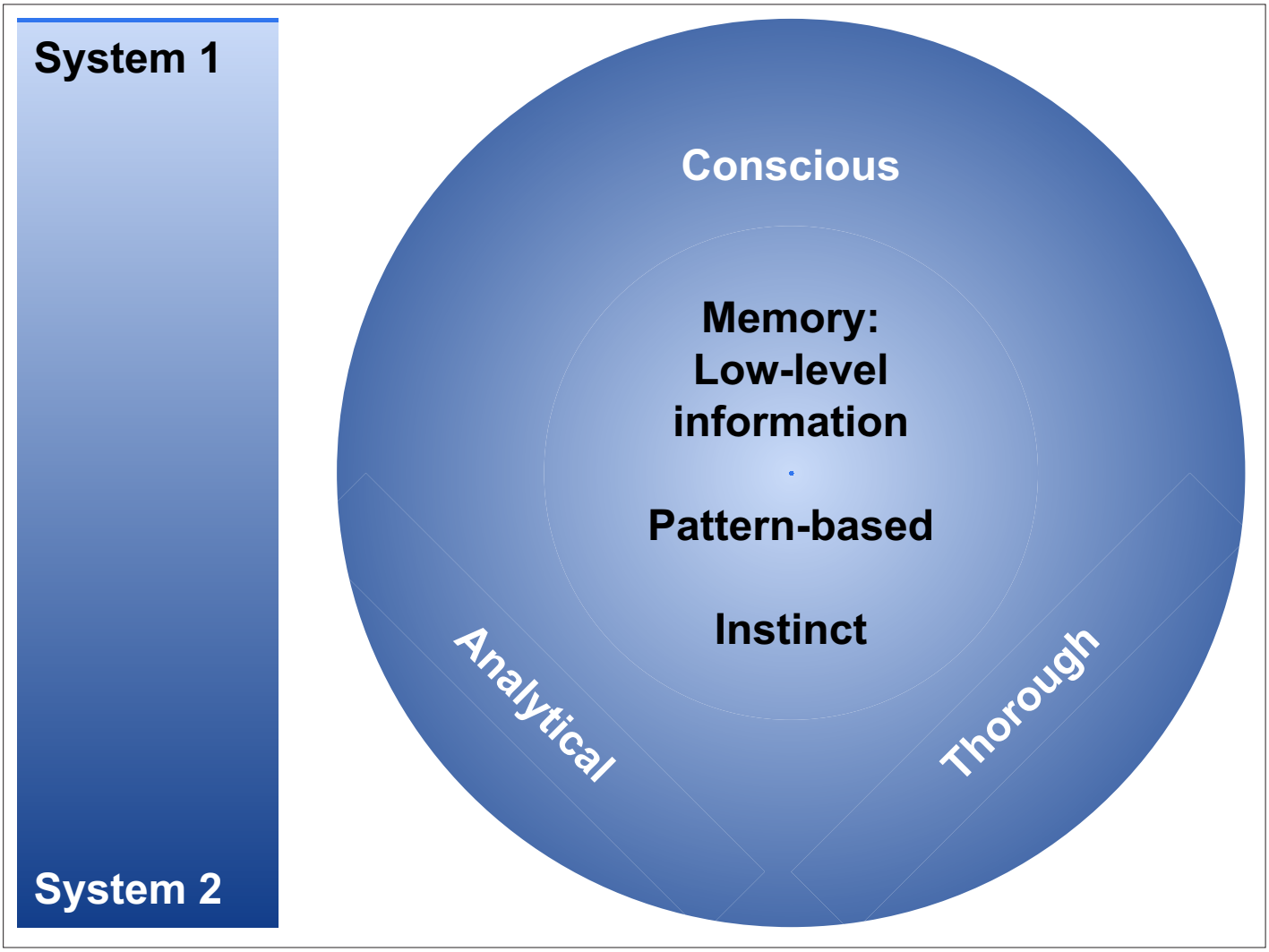

Figure 7. An illustration of the spectrum of 'System 1' fundamental thought processes to 'System 2' analytical thought processes.

many medical datasets contain highly unstructured data. Next, both LLM and SAI would be implemented. The purpose of the LLM would be to form a knowledge base capable of generating initial hypotheses, suggesting pertinent medical literature and proposing logical diagnostic rules inferred from clinical practice guidelines. Meanwhile, one would develop a formal SAI reasoning system encoding medical knowledge, clinical guidelines and diagnostic rules. By applying logical reasoning to patient data, conclusions would be consistent with established medical protocols. Finally, to trace *reasoning behaviour* it is possible to extract both attention weights and CoT-generated explanations from LLMs, as well as output logical inference chains from SAI. Subsequently combining the reasoning traces provides a more comprehensive glimpse into how the final conclusions were reached for medical professionals. Similar to its use in our second proposal, the PRM can be trained on reasoning traces—CoT explanations from the LLM, and inference chains from SAI—to evaluate reasoning behaviour. More specifically, the complementary nature of (a) assessing the logical coherence of these traces and (b) identifying potential logical fallacies lends additional credibility to the combined reasoning outputs.

Open challenges in large reasoning models

Reasoning data and evaluation

Reasoning-knowledge entanglement

Current benchmarks conflate reasoning ability with domain knowledge, reflecting a tension between intuitive, memory retrieval-based (System 1) and analytic (System 2) reasoning processes. It remains challenging to determine whether the performance of a model is the result of genuine logical reasoning or simply retrieval of memorised facts with some generalisation from what was memorised. Evaluating genuine reasoning performance requires benchmarks intentionally designed to minimise prior core knowledge dependency, to isolate reasoning capabilities from knowledge recall. The **ARC-AGI** ('Abstraction and Reasoning Corpus' for Artificial General Intelligence) benchmark exemplifies this approach, assessing how efficiently an AI can abstract and generalise from limited input on visual tasks with minimal knowledge priors, relying on *reasoning* rather than memorised content (***Chollet, 2019***; ***Chollet et al., 2025***). To the best of our knowledge, no equivalent benchmark currently exists for the medical domain.

Limited availability of high-quality reasoning traces

CoT data are critical for training reasoning models, yet richly annotated, diverse medical CoT corpora remain relatively scarce and costly to produce with human annotators (i.e. clinicians). One promising remedy is to harness advanced LLMs themselves to generate synthetic CoT examples: their long context windows and fine-grained instruction-following abilities enable them to produce detailed, multi-step rationales on demand.

Evaluation metrics beyond accuracy

Evaluating reasoning behaviour in medical LRMs requires moving beyond simple accuracy metrics, which often fail to capture the quality and structure of clinical reasoning processes. Instead, the focus should shift towards metrics that assess reasoning trace fidelity—such as logical soundness, coherence, and completeness. This approach is supported by recent work on CoT monitorability (***Korbak et al., 2025***), which argues that CoTs offer a fragile but valuable window into model reasoning. Integrating CoT monitoring into model oversight enables early detection of incorrect or unsafe reasoning patterns, an important consideration in high-stakes domains like medicine.

Rubrics like CLEVER (***Liu et al., 2025***) and R-IDEA (***Esteitieh et al., 2025***; ***Cabral et al., 2024***), which have been developed to evaluate clinical reasoning in humans, offer a promising template for benchmarking LRMs. By scoring LRM-generated CoTs against these rubrics and comparing them directly with evaluations of human clinicians on the same tasks, we can derive more nuanced and clinically relevant assessments of LRM reasoning performance.

While the use of 'judge' LRMs offers a scalable path for automated reasoning trace evaluation—discussed in detail in the rationale-based evaluation section—it is important to acknowledge their limitations, particularly concerning bias (***Li et al., 2025***). These models inherit patterns from their training data, which may embed societal stereotypes associated with race, gender, religion, culture,

and ideology (***Li et al., 2025***). Evaluating the robustness and fairness of such systems in medical contexts remains an essential direction for future work.

Training and inference optimisations

Scalability of RL-based approaches

Reinforcement learning (RL) methods such as PPO ***Schulman et al., 2017***, DPO ***Rafailov et al., 2023***, and GRPO ***Shao et al., 2024*** have shown strong potential in improving LRM reasoning capabilities. However, these approaches are computationally intensive and require significant resources, which can hinder widespread adoption. The key challenge is to develop RL-based reasoning techniques that are less computationally intensive without compromising performance.

Medicine/healthcare specific reward models or reward functions

The generalisability of reward models and functions developed for broad-domain LRMs to medical reasoning tasks remains uncertain, given the domain-specific nuances of clinical decision-making. One promising direction involves incorporating high-quality evaluations of reasoning traces directly into the reward modelling process, thereby enhancing reasoning alignment. This approach aligns with the principles of process supervision (***Lightman et al., 2023***), which has been shown to outperform outcome supervision—particularly in tackling complex problems in mathematics (***Lightman et al., 2023***; ***Uesato et al., 2022***). While process supervision traditionally depends on labour-intensive human annotation of step-by-step reasoning, this limitation can be addressed through process reward models (PRMs) (***Lightman et al., 2023***): customised reward systems trained on annotated CoT traces to automate reasoning evaluation during RLHF. These models assess logical coherence and detect fallacies, preserving transparency without requiring constant human input. An intriguing extension is to integrate LRM-based reasoning trace evaluators as those covered in 'Rationale-based evaluations' directly into the RLHF pipeline, enabling scalable process supervision as explored in the rationale-based evaluation paradigm.

Scaling laws for LRM inference

Studies have demonstrated that increasing compute during inference-time—rather than solely during training—can significantly enhance the reasoning performance of LRMs (***Snell et al., 2024***; ***Zhao et al., 2024***). Techniques like CoT and ToT (***Yao et al., 2023***) allow models to generate multiple intermediate reasoning paths, which can then be evaluated or expanded upon. This scaling behaviour is further supported by methods such as self-reflection (***Zhao et al., 2024***), Monte Carlo Tree Search (MCTS) (***Zhao et al., 2024***), and PRMs (***Snell et al., 2024***), which help verify and select coherent reasoning sequences during inference. Additionally, integrating fact-checking mechanisms via web-enabled LLMs adds a layer of external validation, enabling more accurate and trustworthy output. Together, these approaches illustrate how smarter use of inference-time compute can unlock deeper reasoning capabilities without retraining.

Alternative reasoning architectures

LLMs and subsequently LRMs rely on left-to-right, next-token prediction, limiting their capacity to plan, revise, or backtrack during inference. This property occurs as a result of the autoregressive nature of generative LLMs, which produce an output token by selecting a token with the highest probability score based on previously generated tokens. Should a model choose one 'nonsense' token in context, subsequent tokens are similarly affected. A compounding effect can occur, quickly worsening the error unless the model has the ability to backtrack. To achieve more structured and accurate reasoning, alternative architectures that emphasise planning-based non-autoregressive strategies, such as JEPA (***Assran et al., 2023***), warrant further exploration. Another promising alternative is latent reasoning (***Zhu et al., 2025***; ***Hao et al., 2024***), where models perform internal, iterative computations on hidden states before outputting tokens, enabling internal planning, error correction, and revision prior to generating any output.

Multimodal clinical reasoning integration

Clinical data is inherently multimodal, involving the integration of information across text, medical imaging, electronic health records (EHRs), sensor data (e.g. from smartwatches), and physiological time-series such as ECG or EEG. The ability of LRMs to unify understanding and reasoning across these diverse medical modalities within a single cohesive system indicates an exciting direction for future research. To advance LRM's multimodal reasoning capabilities, we require complex reasoning data synthesis pipelines that go beyond the current paradigm, which largely focuses on single modalities. The path forward involves aligning multiple modalities simultaneously and building rich, interleaved reasoning chains that reflect how clinicians synthesise diverse data sources to arrive at diagnostic and treatment decisions.

Discussion

A striking lack of evaluation methods for reasoning behaviour

As part of our study, we intended to investigate the current SOTA of medical LLM performance in the context of *reasoning behaviour*. However, we found that only a few existing evaluation frameworks adequately capture the nuances involved in assessing *reasoning behaviour* in medical LLMs. This is perhaps unsurprising, given the inherent complexity of such evaluations—particularly those requiring human clinicians. To manually assess reasoning traces of LLMs is time-consuming and difficult to scale.

While conventional conclusion-based evaluations are not without value, they offer only a limited view of *reasoning behaviour* as they focus solely on final answers without analysing the reasoning process. Evaluating reasoning traces in LLMs is crucial, especially in medical contexts, as it is possible for models to reach the correct conclusion through erroneous reasoning. Therefore, an AI that not only gives a diagnosis but also provides a corresponding rationale behind it can significantly aid adoption among clinicians and other healthcare professionals by virtue of its transparency which fosters trust in its recommendations. Broadly, we consider that rationale-based, interactive and mechanistic evaluation are more naturally predisposed to decrypting the *reasoning behaviour* of medical LLMs. Interactive evaluations such as the SDBench offer a promising alternative to static MCQ benchmarks by simulating interactive diagnostic scenarios that better mirror real clinical workflows.

Memorisation vs. planning: The 'stochastic parrot' problem

Another potential issue is the lack of memorisation tests and benchmarks in LLM. This is pertinent as (like humans) medical LLMs have the ability to memorise the dataset they are given but on a much grander scale, giving the illusion of reasoning, although in reality regurgitating related or unrelated information from a vast knowledge base (like humans) (***Hartmann et al., 2023***). Hence, in many cases it was not possible to answer the question: "**to what extent is this model a stochastic parrot and to what extent is this model performing logical reasoning?**" To answer this question, a structured approach would involve testing its ability to work from foundational first principles, or 'base facts', embedded within its training corpus. These base facts encompass simple yet essential principles across areas such as medicine, physiology, and pharmacology, often representing core medical knowledge. For example, a base fact in medicine could be: 'The heart pumps blood throughout the body'. Reasoning tests can then be designed to see if the model can apply such base facts to answer more complex questions.

In practice, however, access to high-volume training corpora for closed-source enterprise models like GPT-4 (***Wu et al., 2023***), Gemini (***Anil et al., 2023***), or Anthropic (***Enis and Hopkins, 2024***) is restricted. This limitation calls for designing medical tests that embed low-level fundamental knowledge, and relying primarily on the model's ability to reason from these base facts. Nevertheless, we do not intend to diminish the usefulness of 'System 1' rote-memorisation type LLMs (which are a prerequisite for 'System 2' advanced-reasoning type systems), but instead wish to highlight the lack of insight into a model's *reasoning behaviour* without this layer of validation.

Hallucination and security implications of opaque reasoning

A natural side effect of obtaining transparency into medical *reasoning behaviour* is its neutralising effect on hallucination events common across LLMs in all domains (***Huang et al., 2023***). Such events occur due to the autoregressive nature of generative LLMs, which produce an output by selecting

a token with a highest probability score. Should a model choose one 'nonsense' token in context, subsequent tokens are similarly affected. A compounding effect can occur, quickly worsening the error unless the model has the ability to backtrack. However, exploring *reasoning behaviour* allows greater insight into hallucination events by exposing the involved logic chain, complementing the current state of the art of retrieval-augmented generation (RAG) (***Xia et al., 2024b***).

More ominously, it is an unfortunate reality that cybercrime is increasingly common, and it is not impossible that healthcare infrastructure, including associated medical LLMs, may be targeted. While this commonly takes the form of ransomware, it takes disturbingly little effort to 'poison' medical LLMs with misinformation (***Han et al., 2024***), with myriad implications for those used for clinical diagnosis or hospital operations. This aspect of medical LLMs is a relatively unexplored field, with most studies focusing on generic use cases (***Peng et al., 2024***), though a more comprehensive framework incorporating the paradigms discussed in our review as well as this security aspect exists for medical vision language models (***Xia et al., 2024a***). As with exposing hallucination events, greater transparency into medical models will assist in identifying such events should the situation arise.

Reasoning behaviour as a form of explainable AI

Enhanced medical reasoning transparency may contribute to solving ongoing problems such as addressing differential diagnosis or providing clinical management plans. Differential diagnosis is an ongoing problem in medical AI development, where similar conditions may confound a prediction with potentially severe consequences. By viewing reasoning traces, both method developers and clinicians will be able to better discern why an accurate or inaccurate choice is made and adjust the model accordingly as well as gaining potentially unknown clinical insights.

Finally, we highlight the point that understanding *reasoning behaviour* of medical LLMs is functionally equivalent to achieving XAI and is not mutually exclusive with other XAI techniques or evaluation paradigms discussed. Given the understandably high level of scrutiny placed on medical methods, achieving this deeper level of understanding is necessary to demonstrate the effectiveness of medical LLMs. At the same time, we may be able to answer an interesting question: 'Is improved reasoning correlated to improved performance?' Ultimately, understanding the *reasoning behaviour* of medical LLMs and LLMs in general has applications across all domains, especially since LLMs are also effective in computer vision tasks.

Conclusion

In summary, it is intuitive that a greater understanding of the *reasoning behaviour* of medical LLMs empowers clinicians, improves patient trust in them, and allows machine learning engineers to troubleshoot underperforming models. However, the lack of studies focusing on understanding *reasoning behaviour* is striking, where the majority of studies are focused on achieving high-level performance metrics with a conspicuous lack of focus on XAI. Most *reasoning behaviour* evaluation strategies are in their infancy, though there is notable potential for growth and further studies. Our theoretical proposed frameworks, while limited, can contribute to XAI in clinical LLMs, with the ultimate goal of improving transparency in medical AI and subsequently patient outcomes.

Acknowledgements

We thank Eric Lee Kuan Hui for facilitating the initial meeting between the authors and subsequent collaboration. We thank Jim Liew Jun Fei, Pui Wee Yang, and Winston Lee Meng Yih for critically reviewing the manuscript.

Additional information

Funding

No external funding was received for this work.

Author contributions

Shamus Zi Yang Sim, Conceptualization, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing – original draft, Project administration, Writing – review and editing; Tyrone Chen, Conceptualization, Formal analysis, Supervision, Validation, Investigation, Visualization, Methodology, Writing – original draft, Project administration, Writing – review and editing

Author ORCIDs

Shamus Zi Yang Sim https://orcid.org/0009-0000-1701-7747
Tyrone Chen https://orcid.org/0000-0002-9207-0385

References

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S. 2023. Gpt-4 technical report. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2303.08774

Alammar J. 2021. Ecco: an open source library for the explainability of transformer language models. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 249–257. DOI: https://doi.org/10.18653/v1/2021.acl-demo.30

Anil R, Borgeaud S, Alayrac J, Yu J, Soricut R, Schalkwyk J, Dai AM, Hauth A, Millican K. 2023. Gemini: A family of highly capable multimodal models. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2312.11805

Assran M, Duval Q, Misra I, Bojanowski P, Vincent P, Rabbat M, LeCun Y, Ballas N. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Vancouver, BC, Canada, . 15619–15629. DOI: https://doi.org/10.1109/CVPR52729.2023.01499

Bai S, Chen K, Liu X, Wang J, Ge W, Song S, Dang K, Wang P, Wang S, Tang J. 2025. Qwen2. 5-vl technical report. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2502.13923

Bertolazzi L, Gatt A, Bernardi R. 2024. A systematic analysis of large language models as soft reasoners: the case of syllogistic inferences. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. DOI: https://doi.org/10.18653/v1/2024.emnlp-main.769

Bisercic A, Nikolic M, Delibasic B, Lio P, Petrovic A. 2023. Interpretable medical diagnostics with structured data extraction by large language models. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2306.05052

Bragazzi NL, Garbarino S. 2024. Toward clinical generative ai: conceptual framework. *JMIR AI* **3**:e55957. DOI: https://doi.org/10.2196/55957, PMID: 38875592

Breiman L. 2001. Random Forests. *Machine Learning* **45**:5–32. DOI: https://doi.org/10.1023/A:1010933404324

Cabral S, Restrepo D, Kanjee Z, Wilson P, Crowe B, Abdulnour R-E, Rodman A. 2024. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Internal Medicine* **184**:581–583. DOI: https://doi.org/10.1001/jamainternmed.2024.0295, PMID: 38557971

Chen T, Guestrin C. 2016. Xgboost: a scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 785–794. DOI: https://doi.org/10.1145/2939672.2939785

Chen T, Vahab N, Tyagi N, Cummins E, Peleg AY, Tyagi S. 2023. *genomicBERT*: A light-weight foundation model for genome analysis using unigram tokenization and specialized DNA vocabulary. *bioRxiv*. DOI: https://doi.org/10.1101/2023.05.31.542682

Chen J, Cai Z, Ji K, Wang X, Liu W, Wang R, Hou J, Wang B. 2024. Huatuogpt-O1, towards medical complex reasoning with Llms. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2412.18925

Chollet F. 2019. On the measure of intelligence. *arXiv*. DOI: https://doi.org/10.48550/arXiv.1911.01547

Chollet F, Knoop M, Kamradt G, Landers B, Pinkard H. 2025. Arc-agi-2: a new challenge for frontier ai reasoning systems. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2505.11831

Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. 2017. Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems 30.

Dewitte P. 2024. Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. *Computer Law & Security Review* **54**:106019. DOI: https://doi.org/10.1016/j.clsr.2024.106019

Enis M, Hopkins M. 2024. From Llm to Nmt: advancing low-resource machine translation with claude. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2404.13813, PMID: 39472359

Esteitieh Y, Mandal S, Laliotis G. 2025. Towards metacognitive clinical reasoning: benchmarking MD-PIE against state-of-the-Art LLMs in medical decision-making. *medRxiv*. DOI: https://doi.org/10.1101/2025.01.28.25321282

Evans JSBT, Stanovich KE. 2013. Dual-process theories of higher cognition: advancing the debate. *Perspectives on Psychological Science* **8**:223–241. DOI: https://doi.org/10.1177/1745691612460685, PMID: 26172965

Foraita R, Spallek J, Zeeb H. 2014. Directed acyclic graphs. In: Foraita R, Spallek J (Eds). *In Handbook of Epidemiology*. Springer Nature. p. 1481–1517. DOI: https://doi.org/10.1007/978-0-387-09834-0_65

Gao Y, Dligach D, Miller T, Caskey J, Sharma B, Churpek MM, Afshar M. 2023a. DR.BENCH: diagnostic reasoning benchmark for clinical natural language processing. *Journal of Biomedical Informatics* **138**:104286. DOI: https://doi.org/10.1016/j.jbi.2023.104286, PMID: 36706848

Gao Y, Li R, Croxford E, Tesch S, To D, Caskey J, W. Patterson B, M. Churpek M, Miller T, Dligach D, Afshar M. 2023b. Large language models and medical knowledge grounding for diagnosis prediction. *medRxiv*. DOI: https://doi.org/10.1101/2023.11.24.23298641

Gegg-Harrison W, Quarterman C. 2024. AI detection's high false positive rates and the psychological and material impacts on students. In: Gegg-Harrison W, Quarterman C (Eds). *In Academic Integrity in the Age of Artificial Intelligence*. IGI Global. p. 199–219. DOI: https://doi.org/10.4018/979-8-3693-0240-8.ch011

Gopalakrishnan S, Garbayo L, Zadrozny W. 2024. Causality extraction from medical text using Large Language Models (LLMs). *Information* **16**:10013. DOI: https://doi.org/10.3390/info16010013

Gouveia SS, Malík J. 2024. Crossing the trust gap in medical AI: building an abductive bridge for xAI. *Philosophy & Technology* **37**:105. DOI: https://doi.org/10.1007/s13347-024-00790-4

Guo D, Yang D, Zhang H, Song J, Zhang R, Xu R, Zhu Q, Ma S, Wang P, Bi X. 2025. Deepseek-R1: incentivizing reasoning capability in Llms via reinforcement learning. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2501.12948

Han T, Nebelung S, Khader F, Wang T, Müller-Franzes G, Kuhl C, Försch S, Kleesiek J, Haarburger C, Bressem KK, Kather JN, Truhn D. 2024. Medical large language models are susceptible to targeted misinformation attacks. *NPJ Digital Medicine* **7**:288. DOI: https://doi.org/10.1038/s41746-024-01282-7, PMID: 39443664

Hao S, Sukhbaatar S, Su D, Li X, Hu Z, Weston J, Tian Y. 2024. Training large language models to reason in a continuous latent space. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2412.06769, PMID: 39398205

Hartmann V, Suri A, Bindschaedler V, Evans D, Tople S, West R. 2023. SoK: memorization in general-purpose large language models. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2310.18362

Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J. 2020. Measuring massive multitask language understanding. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2009.03300

Holyoak KJ, Morrison RG. 1999. *Cambridge Handbook of Thinking and Reasoning*. Cambridge University Press.

Homolak J. 2023. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. *Croatian Medical Journal* **64**:1–3. DOI: https://doi.org/10.3325/cmj.2023.64.1, PMID: 36864812

Hong S, Xiao L, Zhang X, Chen J. 2024. ArgMed-agents: explainable clinical decision reasoning with LLM disscusion via argumentation schemes. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2403.06294

Horsten L. 2007. *Philosophy of Mathematics*. Stanford Encyclopedia of Philosophy.

Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B. 2023. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2311.05232

Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. 2019. PubMedQA: A dataset for biomedical research question answering. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) Hong Kong, China. DOI: https://doi.org/10.18653/v1/D19-1259

Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2009.13081

Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* **3**:160035. DOI: https://doi.org/10.1038/sdata.2016.35, PMID: 27219127

Kiciman E, Ness R, Sharma A, Tan C. 2023. Causal reasoning and large language models: opening a new frontier for causality. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2305.00050

Korbak T, Balesni M, Barnes E, Bengio Y, Benton J, Bloom J, Chen M, Cooney A, Dafoe A, Dragan A. 2025. Chain of thought monitorability: a new and fragile opportunity for Ai safety. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2507.11473

Kwon JM, Kim KH, Jeon KH, Lee SE, Lee HY, Cho HJ, Choi JO, Jeon ES, Kim MS, Kim JJ, Hwang KK, Chae SC, Baek SH, Kang SM, Choi DJ, Yoo BS, Kim KH, Park HY, Cho MC, Oh BH. 2019. Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. *PLOS ONE* **14**:e0219302. DOI: https://doi.org/10.1371/journal.pone.0219302, PMID: 31283783

Kwon T, Ong KT, Kang D, Moon S, Lee JR, Hwang D, Sohn B, Sim Y, Lee D, Yeo J. 2024. Large language models are clinical reasoners: reasoning-aware diagnosis framework with prompt-generated rationales. Proceedings of the AAAI Conference on Artificial Intelligence. 18417–18425. DOI: https://doi.org/10.1609/aaai.v38i16.29802

Lai Y, Zhong J, Li M, Zhao S, Yang X. 2025. Med-R1: reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2503.13939

Li B, Meng T, Shi X, Zhai J, Ruan T. 2023. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2312.02441

Li SS, Balachandran V, Feng S, Ilgen J, Pierson E, Koh PW, Tsvetkov Y. 2024. MEDIQ: question-asking llms for adaptive and reliable medical reasoning. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2406.00922

Li D, Jiang B, Huang L, Beigi A, Zhao C, Tan Z, Bhattacharjee A, Jiang Y, Chen C, Wu T. 2025. From generation to judgment: opportunities and challenges of Llm-as-a-Judge. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2411.16594

Liao Y, Meng Y, Wang Y, Liu H, Liu H, Wang Y, Wang Y. 2024. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2403.08495

Lightman H, Kosaraju V, Burda Y, Edwards H, Baker B, Lee T, Leike J, Schulman J, Sutskever I, Cobbe K. 2023. Let's verify step by step. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2305.20050

Liu X, Liu H, Yang G, Jiang Z, Cui S, Zhang Z, Wang H, Tao L, Sun Y, Song Z, Hong T, Yang J, Gao T, Zhang J, Li X, Zhang J, Sang Y, Yang Z, Xue K, Wu S, et al. 2025. A generalist medical language model for disease diagnosis assistance. *Nature Medicine* **31**:932–942. DOI: https://doi.org/10.1038/s41591-024-03416-6, PMID: 39779927

Lorek LA. 2024. *AI Legal Innovations: The Benefits and Drawbacks of Chat-Gpt and Generative AI in the Legal Industry*. Ohio Northern University Law Review.

Lundberg S. 2017. A unified approach to interpreting model predictions. *arXiv*. DOI: https://doi.org/10.48550/arXiv.1705.07874

Miller RA, Pople HE, Myers JD. 1985. Internist-i, an experimental computer-based diagnostic consultant for general internal medicine. In: Miller RA, Pople HE (Eds). *In Computer-Assisted Medical Decision Making*. Springer. p. 139–158. DOI: https://doi.org/10.1007/978-1-4612-5108-8_8

Moëll B, Sand Aronsson F, Akbar S. 2025. Medical reasoning in LLMs: an in-depth analysis of DeepSeek R1. *Frontiers in Artificial Intelligence* **8**:1616145. DOI: https://doi.org/10.3389/frai.2025.1616145, PMID: 40607450

Mondorf P, Plank B. 2024. Beyond accuracy: evaluating the reasoning behavior of large language models–a survey. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2404.01869

Naik N, Khandelwal A, Joshi M, Atre M, Wright H, Kannan K, Hill S, Mamidipudi G, Srinivasa G, Bifulco CB, Piening B, Matlock K. 2023. Applying large language models for causal structure learning in non small cell lung cancer. 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI) Orlando, FL, USA, . 688–693. DOI: https://doi.org/10.1109/ICHI61247.2024.00110

Ngai H, Rudzicz F. 2022. Doctor XAvIer: explainable diagnosis on physician-patient dialogues and XAI evaluation. Proceedings of the 21st Workshop on Biomedical Language Processing Dublin, Ireland. DOI: https://doi.org/10.18653/v1/2022.bionlp-1.33

Nori H, Daswani M, Kelly C, Lundberg S, Ribeiro MT, Wilson M, Liu X, Sounderajah V, Carlson J, Lungren MP. 2025. Sequential diagnosis with language models. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2506.22405

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A. 2022. Training language models to follow instructions with human feedback. NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems. 27730–27744.

Pal A, Umapathi LK, Sankarasubbu M. 2022. Medmcqa: a large-scale multi-subject multi-choice dataset for medical domain question answering. In Conference on health, inference, and learning. 248–260.

Pan J, Liu C, Wu J, Liu F, Zhu J, Li HB, Chen C, Ouyang C, Rueckert D. 2025. Medvlm-R1: incentivizing medical reasoning capability of Vision-Language Models (Vlms) via reinforcement learning. *arXiv*. DOI: https://doi.org/10.1007/978-3-032-04981-0_32

Peng B, Chen K, Li M, Feng P, Bi Z, Liu J, Niu Q. 2024. Securing large language models: addressing bias, misinformation, and prompt attacks. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2409.08087

Qiu P, Wu C, Liu S, Zhao W, Chen Z, Gu H, Peng C, Zhang Y, Wang Y, Xie W. 2025. Quantifying the reasoning abilities of llms on real-world clinical cases. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2503.04691

Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. 2023. Direct preference optimization: your language model is secretly a reward model. Advances in Neural Information Processing Systems. 53728–53741.

Savage T, Nayak A, Gallo R, Rangan E, Chen JH. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine* **7**:20. DOI: https://doi.org/10.1038/s41746-024-01010-1, PMID: 38267608

Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. 2017. Proximal policy optimization algorithms. *arXiv*. DOI: https://doi.org/10.48550/arXiv.1707.06347

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV) Venice, . 618–626. DOI: https://doi.org/10.1109/ICCV.2017.74

Shao Z, Wang P, Zhu Q, Xu R, Song J, Bi X, Zhang H, Zhang M, Li Y, Wu Y. 2024. Deepseekmath: pushing the limits of mathematical reasoning in open language models. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2402.03300

Sheth A, Roy K, Gaur M. 2023. Neurosymbolic artificial intelligence (why, what, and how). *IEEE Intelligent Systems* **38**:56–62. DOI: https://doi.org/10.1109/MIS.2023.3268724

Shi W, Xu R, Zhuang Y, Yu Y, Zhang J, Wu H, Zhu Y, Ho JC, Yang C, Wang MD. 2024. EHRAgent: code empowers large language models for few-shot complex tabular reasoning on electronic health records. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing Miami, Florida, USA. DOI: https://doi.org/10.18653/v1/2024.emnlp-main.1245

Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera Y Arcas B, et al. 2023. Large language models encode clinical knowledge. *Nature* **620**:172–180. DOI: https://doi.org/10.1038/s41586-023-06291-2, PMID: 37438534

Sloman S. 2009. *Causal Models: How People Think about the World and Its Alternatives*. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780195183115.001.0001

Snell C, Lee J, Xu K, Kumar A. 2024. Scaling Llm test-time compute optimally can be more effective than scaling model parameters. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2408.03314

Sun J, Zheng C, Xie E, Liu Z, Chu R, Liu J, Xu J, Ding M, Li H, Geng M, Wu Y, Wang W, Chen J, Yin Z, Ren X, Fu J, He J, Yuan W, Liu Q, Liu X, et al. 2023. A survey of reasoning with foundation models. *arXiv*. https://doi.org/10.48550/arXiv.2312.11562

Sundararajan M, Taly A, Yan Q. 2017. Axiomatic attribution for deep networks. *arXiv*. https://doi.org/10.48550/arXiv.1703.01365

Tchango AF, Goel R, Martel J, Wen Z, Caron GM, Ghosn J. 2022a. Towards trustworthy automatic diagnosis systems by emulating doctors' reasoning with deep reinforcement learning. *arXiv*. https://doi.org/10.48550/arXiv.2210.07198
Tchango AF, Goel R, Wen Z, Martel J, Ghosn J. 2022b. Ddxplus: A new dataset for automatic medical diagnosis. *arXiv*. https://doi.org/10.48550/arXiv.2205.09148
Tenney I, Wexler J, Bastings J, Bolukbasi T, Coenen A, Gehrmann S, Jiang E, Pushkarna M, Radebaugh C, Reif E, Yuan A. 2020. The language interpretability tool: extensible, interactive visualizations and analysis for NLP Models. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. DOI: https://doi.org/10.18653/v1/2020.emnlp-demos.15
Tiku N. 2022. The google engineer who thinks the company's ai has come to life. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/ [Accessed June 11, 2022].
Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F. 2023. Llama: open and efficient foundation language models. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2302.13971
Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, Wang A, Li B, Amin M, Tomasev N. 2024. Towards conversational diagnostic Ai. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2401.05654
Uesato J, Kushman N, Kumar R, Song F, Siegel N, Wang L, Creswell A, Irving G, Higgins I. 2022. Solving math word problems with process- and outcome-based feedback. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2211.14275
van Melle W. 1978. MYCIN: a knowledge-based consultation program for infectious disease diagnosis. *International Journal of Man-Machine Studies* **10**:313–322. DOI: https://doi.org/10.1016/S0020-7373(78)80049-2
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. *arXiv*. DOI: https://doi.org/10.48550/arXiv.1706.03762
Wang X, Xu X, Tong W, Roberts R, Liu Z. 2021. InferBERT: A transformer-based causal inference framework for enhancing pharmacovigilance. *Frontiers in Artificial Intelligence* **4**:659622. DOI: https://doi.org/10.3389/frai.2021.659622, PMID: 34136800
Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, Chowdhery A, Zhou D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2203.11171
Wang Z, Zhang G, Yang K, Shi N, Zhou W, Hao S, Xiong G, Li Y, Sim MY, Chen X. 2023. Interactive natural language processing. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2305.13246
Wang X, Zhou D. 2024. Chain-of-thought reasoning without prompting. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2402.10200
Wang W, Ma Z, Wang Z, Wu C, Ji J, Chen W, Li X, Yuan Y. 2025. A Survey of LLM-based Agents in Medicine: How far are we from Baymax?. Findings of the Association for Computational Linguistics Vienna, Austria. DOI: https://doi.org/10.18653/v1/2025.findings-acl.539
Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2201.11903
Williams TC, Bach CC, Matthiesen NB, Henriksen TB, Gagliardi L. 2018. Directed acyclic graphs: a tool for causal studies in paediatrics. *Pediatric Research* **84**:487–493. DOI: https://doi.org/10.1038/s41390-018-0071-3, PMID: 29967527
Wu T, He S, Liu J, Sun S, Liu K, Han QL, Tang Y. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* **10**:1122–1136. DOI: https://doi.org/10.1109/JAS.2023.123618
Wu K, Wu E, Thapa R, Wei K, Zhang A, Suresh A, Tao JJ, Sun MW, Lozano A, Zou J. 2025. MedCaseReasoning: evaluating and learning diagnostic reasoning from clinical case reports. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2505.11733
Xia P, Chen Z, Tian J, Gong Y, Hou R, Xu Y, Wu Z, Fan Z, Zhou Y, Zhu K, Zheng W, Wang Z, Wang X, Zhang X, Bansal C, Niethammer M, Huang J, Zhu H, Li Y, Sun J, et al. 2024a. CARES: a comprehensive benchmark of trustworthiness in medical vision language models. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2406.06007
Xia P, Zhu K, Li H, Wang T, Shi W, Wang S, Zhang L, Zou J, Yao H. 2024b. MMed-Rag: versatile multimodal rag system for medical vision language models. *arXiv*. DOI: https://doi.org/10.18653/v1/2024.emnlp-main.62
Xu C, Jackson SA. 2019. Machine learning and complex biological data. *Genome Biology* **20**:1–4. DOI: https://doi.org/10.1186/s13059-019-1689-0
Yang B, Jiang S, Xu L, Liu K, Li H, Xing G, Chen H, Jiang X, Yan Z. 2024a. DrHouse: An LLM-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **8**:1–29. DOI: https://doi.org/10.1145/3699765, PMID: 39639863
Yang AN, Yang B, Hui B. 2024b. Qwen2 technical report. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2407.10671
Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, Narasimhan K. 2023. Tree of thoughts: deliberate problem solving with large language models. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2305.10601
Yu J, Ignatiev A, Stuckey PJ. 2023a. From formal boosted tree explanations to interpretable rule sets. In 29th International Conference on Principles and Practice of Constraint Programming.
Yu W, Jiang M, Clark P, Sabharwal A. 2023b. IfQA: A dataset for open-domain question answering under counterfactual presuppositions. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing Singapore. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.515

Yue L, Xing S, Chen J, Fu T. 2024. ClinicalAgent: clinical trial multi-agent system with large language model-based reasoning. Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics Shenzhen, China, . 1–10. DOI: https://doi.org/10.1145/3698587.3701359

Zhang Q, Rao L, Yang Y. 2021. A novel visual interpretability for deep neural networks by optimizing activation maps with perturbation. Proceedings of the AAAI Conference on Artificial Intelligence. 3377–3384. DOI: https://doi.org/10.1609/aaai.v35i4.16450

Zhao Y, Yin H, Zeng B, Wang H, Shi T, Lyu C, Wang L, Luo W, Zhang K. 2024. Marco-O1: towards open reasoning models for open-ended solutions. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2411.14405

Zhou S, Xie W, Li J, Zhan Z, Song M, Yang H, Espinoza C, Welton L, Mai X, Jin Y. 2025. Automating expert-level medical reasoning evaluation of large language models. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2507.07988

Zhu Y, Wei S, Wang X, Xue K, Zhang S, Zhang X. 2024. MeNTi: bridging medical calculator and LLM agent with nested tool calling. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics Albuquerque, New Mexico. DOI: https://doi.org/10.18653/v1/2025.naacl-long.263

Zhu R, Peng T, Cheng T, Qu X, Huang J, Zhu D, Wang H, Xue K, Zhang X, Shan Y. 2025. A survey on latent reasoning. *arXiv*. DOI: https://doi.org/10.48550/arXiv.2507.06203

Zhuang Y, Lin YH, Liyanawatta M, Saputro AH, Utami YD, Wang JH. 2024. An interactive programming learning environment supporting paper computing and immediate evaluation for making thinking visible and traceable. *Interactive Learning Environments* **32**:5253–5266. DOI: https://doi.org/10.1080/10494820.2023.2212709

Zi Yang SS, Fye GM, Yap WC, Yu DZ. 2024. Enhancing medical summarization with parameter efficient fine tuning on local CPUs. 2024 International Conference on Electrical, Communication and Computer Engineering (ICECCE) Kuala Lumpur, Malaysia. DOI: https://doi.org/10.1109/ICECCE63537.2024.10823619