

# Diplomarbeit

## Benchmarking Clinical Reasoning in Large Language Models: A Comparative Assessment Study

zur Erlangung des akademischen Grades

Doktor der gesamten Heilkunde

(Dr.med.univ.)

an der

Medizinischen Universität Wien

ausgeführt am

Zentrum für Medical Data Science - Institut für Artificial Intelligence

unter der Anleitung von

Priv.Doz. Mag. Dr. Matthias Samwald

eingereicht von

**Thomas Prade, BSc.**

Wien, 22.12.2025

# Eidesstattliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Abschlussarbeit selbstständig und ohne fremde Hilfe verfasst habe, andere als die angegebenen Quellen nicht verwendet habe und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 22.12.2025

\_\_\_\_\_  
Unterschrift, eh

# Danksagung

Meinem Betreuer, Dr. Matthias Samwald, danke ich aufrichtig für die Möglichkeit, am Institut für Artificial Intelligence meine Abschlussarbeit zu verfassen und meiner Neugier für Zukunftsfragen zum Spannungsfeld zwischen moderner Medizin und Künstlicher Intelligenz nachgehen zu können. Ebenfalls dankbar bin ich für den entscheidenden Impuls, der dieser Arbeit die notwendige Orientierung und den konkreten Rahmen gab.

Mein innigster Dank gilt meinen Eltern, Clara Inés und Norbert Prade, sowie meiner Schwester, Isabel Prade. Eure wohlwollende und liebevolle Unterstützung hat mich stets begleitet und war Licht in den Momenten größter Zweifel. Gracias por siempre estar, sin importar nada.

*Andar  
Creando  
Caminos*

# Contents

<b>1. Introduction</b>	<b>1</b>
1.1. AI Driven Clinical Decision Support Systems . . . . .	1
1.2. Clinical Reasoning . . . . .	3
1.3. Clinical Prompt Engineering . . . . .	8
1.4. Medical Benchmarking . . . . .	11
1.5. MedQA-USMLE . . . . .	12
1.6. Script Concordance Testing . . . . .	13
1.7. Evaluation Metrics . . . . .	14
1.8. Calibration and Uncertainty Quantification . . . . .	15
1.9. Research Focus and Goals . . . . .	16
1.9.1. Research Objectives . . . . .	17
1.9.2. Impact Goals . . . . .	17
<b>2. Methods and Resources</b>	<b>19</b>
2.1. Datasets . . . . .	19
2.1.1. MedQA-USMLE . . . . .	19
2.1.2. SCT-Bench . . . . .	19
2.2. Model and Data Generation . . . . .	19
2.3. Prompting and Uncertainty Quantification . . . . .	20
2.3.1. ROC Curve . . . . .	21
2.3.2. Expected Calibration Error . . . . .	21
2.4. Evaluation Metrics and Data Analysis . . . . .	22
<b>3. Results</b>	<b>24</b>
3.1. Benchmark Scores . . . . .	24
3.1.1. MedQA . . . . .	24
3.1.2. SCT-Bench . . . . .	25
3.2. Response Entropy . . . . .	26
3.3. Calibration Metrics . . . . .	28
3.3.1. Self-Elicited Confidence . . . . .	28
3.3.2. Sample Consistency . . . . .	29
3.4. ROC . . . . .	31
3.5. SCT Rating Distribution . . . . .	34
3.6. Clinical Reasoning Strategies . . . . .	35
3.7. Clinical Reasoning Strategy Distribution . . . . .	44
3.8. Rating Transitions . . . . .	45

3.9. Expert versus Model Rating . . . . .	50
<b>4. Discussion</b>	<b>52</b>
4.1. Benchmark Performance . . . . .	52
4.2. Rating Behavior . . . . .	53
4.3. Uncertainty Proxies . . . . .	56
4.3.1. Systematic Overconfidence . . . . .	56
4.3.2. Benchmark-Disparity in Uncertainty Assessment . . . . .	57
4.4. Self-reflecting Clinical Reasoning . . . . .	58
4.4.1. Strategy Implementation . . . . .	58
4.4.2. Strategy Distributions . . . . .	58
4.5. Limitations . . . . .	59
4.6. Conclusion and Future Outlook . . . . .	60
<b>List of Equations</b>	<b>71</b>
<b>List of Abbreviations</b>	<b>72</b>
<b>A. Clinical CoT Prompt Templates</b>	<b>I</b>
A.1. MedQA-USMLE . . . . .	I
A.2. SCT-Bench . . . . .	V
<b>B. Tool Table</b>	<b>X</b>

# Kurzfassung

Die Evaluierung von Large Language Models (LLMs) und deren klinisches Kompetenzniveau hat sich bisher vorwiegend auf konventionelle multiple-choice Fragenformate (MCQ) konzentriert, welche Benchmarks wie MedQA-USMLE hervorgebracht haben und wo Sprachmodelle Experten bereits übertreffen konnten. Allerdings wurden unlängst andere Bewertungsmethoden für Medical Benchmarking vorgeschlagen: SCT-Bench basiert auf Script Concordance Testing (SCT), wo klinisches Urteilsvermögen und probabilistisches Denken unter Unsicherheit evaluiert werden. Studien zeigten, dass sich Performance Merkmale auf SCTs wesentlich von denen auf traditionellen MCQ Tests unterscheiden. Inbesondere reasoning-optimierte Modelle schnitten auf SCT-Bench schlechter ab als non-reasoning Modelle, wobei sie auf alternativen Medical Benchmarks sonst bessere Leistungen erzielten.

Diese Abschlussarbeit führte anhand des Beispiels von OpenAI's GPT-4.1 eine komparative Analyse von Performance Metriken, Uncertainty Proxies und Clinical Reasoning Merkmalen zwischen MedQA-USMLE und dem öffentlich Subset von SCT-Bench durch. Baseline und Chain-of-Thought (CoT) prompting wurden über mehrere Sampling-Durchläufe verglichen, wobei CoT darauf abzielte, den Gebrauch vordefinierter Clinical Reasoning Strategien zu induzieren.

Die Ergebnisse zeigen, dass nur die Performance auf MedQA von klinischem CoT Prompting profitierte und sich von 86,4% auf 93,0% verbesserte, während sich der SCT-Bench Score von 77,7% auf 74,7% verschlechterte, wobei dieser Rückgang statistisch nicht signifikant war. Dies zeigte sich in der erhöhten Tendenz zu extremen Ratings auf SCT-Fragen. GPT-4.1 schien den Einfluss neuer Information bei expliziter Rationalisierung systematisch zu überschätzen, was zu unangemessener Overconfidence führte. Sample-basiertes Majority Voting hatte einen signifikanten Effekt auf die Verbesserung des MedQA-Scores mit CoT, jedoch nicht für Baseline Prompting. Auf SCT-Bench verbesserte Majority Voting die Scores nicht signifikant über den Durchschnittswert hinaus. Entropievergleiche offenbarten, dass CoT zu größerer Uneinigkeit unter Modellantworten führte, aber gleichzeitig Cluster korrekter Antworten auf MedQA erzeugte, was zu Vorteilen für Majority Voting führte. Dies war nicht der Fall für SCT-Bench. Qualitative Evaluierung bestätigte, dass GPT-4.1 konzeptionell in der Lage war, situationsadäquate Clinical Reasoning Strategien anzuwenden und über diese sinnvoll zu reflektieren. Das Experten-Rating-Verhalten in Abhängigkeit von im Output erwähnten Strategien legt nahe, dass GPT-4.1 teilweise expertenähnliche Logik anwendete.

Zusammenfassend haben sich Hinweise auf substantielle Einschränkungen im klinischen Urteilsvermögen von LLMs und Prompt Engineering in SCT Formaten weiter erhärtet. Die Reasoning Strategie Analyse legt nahe, dass zukünftige Reasoning-zentrierte Evaluierungsframeworks einen bedeutsamen Beitrag zur Medical Benchmark Landschaft für LLMs leisten könnten.

# Abstract

Evaluation of Large Language Models (LLMs) and their clinical competence has mainly focused on conventional multiple-choice (MCQ) formatted medical question answering exams, yielding benchmarks like MedQA-USMLE, where models have already exceeded expert-level performance. However, different assessment methods have recently been proposed as novel tools for medical benchmarking, such as SCT-Bench based on Script Concordance Testing (SCT), which evaluates clinical reasoning and probabilistic thinking under uncertainty. Previous research has shown that performance characteristics differ significantly on SCT compared to traditional MCQ tests. Unexpectedly, reasoning-optimized models scored particularly worse on SCT-Bench, even when outperforming non-reasoning models on other medical benchmarks.

This thesis conducted a comparative analysis of performance metrics, uncertainty proxies and clinical reasoning qualities between MedQA-USMLE and the public subset of SCT-Bench using OpenAI's GPT-4.1 model. Baseline and Chain-of-Thought (CoT) prompting were compared across sampled responses with CoT aiming to elicit the use of explicit predefined clinical reasoning strategies.

The results show that only the performance on MedQA benefited from clinical CoT prompting, improving from 86.4% to 93.0%, while the SCT-Bench score worsened from 77.7% to 74.7%, though this decline was not statistically significant. GPT-4.1 appeared to systematically overestimate the impact of new information when formulating explicit reasoning rationales, leading to inappropriate overconfidence in uncertain scenarios. This was evident in an increased tendency for extreme ratings on SCT questions. Sample-based majority voting had a small but significant effect on improving the MedQA score when CoT was applied, but not for baseline prompting. On SCT-Bench, majority voting did not significantly improve scores beyond the mean score. Response entropy analysis revealed that CoT led to greater overall disagreement among model answers, but at the same time created clusters of correct responses on MedQA leading to benefits for majority voting. This was not the case for SCT-Bench. Qualitative evaluation confirmed that GPT-4.1 was conceptually able to apply situation-appropriate clinical reasoning strategies and to self-reflect on them. Expert rating behaviour depending on reasoning strategies mentioned in the output suggests that GPT-4.1 partially applied expert-like logic.

In conclusion, substantial limitations in elicited clinical reasoning and clinical prompt engineering for SCT-based benchmarking were further corroborated. The reasoning strategy analysis suggests that future reasoning-aware evaluation frameworks could contribute meaningfully to the medical benchmark landscape for LLMs.

# 1. Introduction

## 1.1. AI Driven Clinical Decision Support Systems

With the rise of what is commonly known as "artificial intelligence" (AI) and the impact it is already having on our lives it is fair to say that humankind finds itself in the midst of a new technology revolution, which is already revealing its potential for reshaping the world we live in profoundly. The life sciences as well as clinical medicine are without any doubts one of the most significant protagonists in this revolution seeking to develop new technological tools in order to accelerate, improve and augment research and general healthcare. Implementation and constant development of AI-driven systems span across a wide spectrum and includes areas like modern imaging, drug-discovery and personalized precision medicine just to name a few.

Many of these rising technologies will be integrated in so called clinical decision support system (CDSS), designed to assist and enhance decision-making processes in a clinical context. CDSSs have evolved significantly from their early implementations in the 1970s, with pioneering systems doing foundational work in medical artificial intelligence [1]. These early systems established groundwork for today's more sophisticated AI systems.

Modern AI-based CDSSs demonstrate impressive effectiveness across multiple clinical disciplines as discussed quite thoroughly in a review article by Ouanes and Farhah [2]: Analysing studies published between 2018 and 2023, they identified important contributions in the following categories: Early detection and disease diagnosis, enhanced decision making, medication and error reduction. Examples for outcomes with high effectiveness include improved prediction of acute kidney injury or sepsis with high specificity and sensitivity. In radiology deep learning models have enhanced diagnostic accuracy and in oncology AI applications led to 5% change in treatment decisions affecting adherence to guidelines. New technologies have also shown promise in medication safety, demonstrating greater prescription accuracy compared to existing systems or being able to predict drug efficacy and guide accurate prescriptions [2].

Since a few years now already, the field of natural language processing (NLP) has been getting extensive world-wide attention in the context of large language models (LLM). A variety of models that have been released for mass use over the past 2 years have arguably even rekindled the general perception and discussion about AI technologies and their impact. The sudden introduction of LLMs like ChatGPT by OpenAI to the general public really caused a shock-wave of admiration and scepticism at the same time. Ever since, language models have progressed to be more powerful and accurate, able to facilitate or even take over a range of different tasks. This development can be seen as equally impressive as concerning.

A study by Liu et al. that addresses the problem of alert fatigue, where about 90% clinicians

override or ignore alerts, showed the practical potential of LLMs for clinical decision support optimization [3]. GPT-3.5 was able to generate suggestions that effectively enhanced the CDS alert logic, achieving high scores for relevance and understanding, but moderate for usefulness. The generated suggestions could compete with those written by human experts without showing redundancy between them. This highlights the possibilities the synergy of human and computational forces offers.

The potential applications of LLMs in medicine are diverse and span from direct patient care to medical education and research. Tasks in this field require domain-specific competence and this is why biomedical language models such as BioBERT, BioGPT, ClinicalBERT or MedAlpaca have been developed [4]. In previous studies, Models specialized in medical NLP have shown enhanced performance on clinical text mining, biomedical information extraction and medical question answering tasks compared to general-purpose models [4]. LLMs can also be used to enrich biomedical-related queries by adding relevant terms or concepts that improve the accuracy of search results [4]. In a study by Wang et al., ChatGPT was applied to formulate and refine boolean queries for systematic reviews, finding it showed higher precision compared to state-of-the-art methods [5]. Biomedical text summarization by LLMs showed to be useful for medical literature, radiology reports and clinical notes among others. However, challenges remain in maintaining clinical accuracy and avoiding omission of crucial medical information [4]. LLM powered information extraction can be helpful for presenting data from unstructured clinical notes in a structured narrative, helping capture valuable patient information and contributing to CDS [2]. The clinical workflow itself could be facilitated meaningfully by assisting doctors in healthcare delivery. This includes reducing their time needed for writing reports or performing other administrative tasks, making it possible to engage more in patient care or to collaborate more with colleagues [6]. The potential for medical education should not be overlooked either. LLMs could turn out to be effective interactive education tools able to support and enhance the learning process. By generating case scenarios, asking follow-up questions or with their conversational abilities they could significantly augment the student's learning experience [4]. Additionally, models have demonstrated counterfactual reasoning for hypothesizing why students might have made certain mistakes, helping to overcome misconception in an interactive manner [6].

However, implementing AI-powered CDSSs in healthcare also faces limitations and concerns as Ouanes and Farhah point out comprehensively [2]: There is real danger for systematic biases that origin from historical data that might be even discriminatory against certain groups. Biased treatment or diagnostic recommendations leading to disparities in healthcare could be the consequence. Technical challenges also need to be addressed properly, for example interoperability issues, as these new systems must integrate with existing healthcare IT and electronic health records. Physicians also rightfully express concerns about clinical accuracy of provided advice as well as legal liability in case of harm. Finally, further ethical considerations implying data security, document confidentiality and liability need to be made

too.

The highlighted potential use-cases offer great opportunities for optimizing healthcare workflow, particularly via human-AI hybrid systems. For developing even more sophisticated CDSSs, it is fundamental that LLMs are capable to correctly process medical data and use the information adequately when performing different kind of tasks. This can mean summarising patient records while preserving all crucial data, assisting triage tasks by identifying time sensitive conditions and even generating original content based on initial data avoiding hallucinations. The latter can imply engaging in clinical decision making by recommending next steps in diagnostic or therapeutic procedures and assisting doctors e.g. with generated clinical hypotheses, risk-stratification or guideline-adherence among others. Clinical reasoning is a vital ability needed to perform these type of tasks. This implies being able to reason about pathophysiological mechanisms and use encoded medical knowledge in an adequate way to make consistent and correct conclusions compatible with evidence-based medicine and the expert consensus. In addition it may also include the ability of conceptual clinical thinking strategies crucial to medical practice (section 3.6).

## 1.2. Clinical Reasoning

Clinical reasoning represents a complex and multi-layered cognitive process fundamental to medical practice. It involves a diverse set of thinking strategies and principles suited for clinical problem solving, operating through both conscious and unconscious cognitive mechanisms interacting with contextual factors [7]. The most basic components of clinical reasoning can be identified as information gathering, hypothesis generation, forming a problem representation, generating a differential diagnosis, selecting a leading diagnosis and developing a plan for further management [7].

In medical education, teaching this ability takes place in a rather implicit way and most of the times comes short in curricula of medical schools or is not explicitly taught at all. As Connor et al. stated, a theory-informed approach for learning and practicing clinical reasoning is commonly missing [8]. Despite inadequate emphasis in medical education, clinical reasoning is a core competency that should be transferred in validated ways. Connor et al. argue that such an approach could be based e.g. on dual process theory (described below), emphasizing the concept of illness scripts, problem representation, knowledge organisation, Bayesian reasoning, heuristics, context specificity and ultimately clinical reasoning strategies and the flexibility in their use [8].

What does this mean for the development and deployment of AI systems in clinical decision support? Understandably, we expect optimal clinical reasoning capacities and awareness of such technologies. This raises the question how we should train and educate these systems on this matter and what paradigms of clinical reasoning we choose to follow doing so. This is a particularly interesting question considering that clinical reasoning is not even

a skill that is being explicitly taught in most medical schools worldwide, where the pure transfer of knowledge across subjects such as anatomy, physiology etc. is still predominant. Knowledge alone is not enough though, not only for future doctors, but also for AI systems including LLMs. It is essential that they "internalize" the principles of clinical thinking and corresponding reasoning strategies. Subsequently we also need effective ways to evaluate whether they are capable of doing so. In humans, emotions most likely also play an important role in clinical decision making beside purely cognitive processes, as they have an impact on cognition, especially in high-stakes situations [9]. This is an element of decision making that almost certainly will be absent in AI systems.

The theoretical frameworks describing clinical reasoning include dual-process theory, situativity, cognitive load and deliberate practice as synthesized by Durning et al. [9]: Situativity theory essentially argues that clinical reasoning cannot be separated from the context in which it takes place, since cognition also entails interaction with the environment. Cognitive load theory emphasizes the cognitive limitations the human brain has, saying that it can only hold and process a limited amount of information in its working memory. This is where organized knowledge comes into play, an effective way of grouping information in order to free up working memory, turning slow thinking into highly effective fast thinking if done well. According to deliberate practice theory, becoming an expert requires deliberate practice of core skills under the guidance of a mentor [9]. Finally, dual-process theory (a well-established idea in cognitive psychology thoroughly reviewed by Evans and Stanovich [10]) describes two distinct mental processes: System 1 (fast, intuitive, pattern-recognition-based thinking) and System 2 (slower, deliberative, analytical reasoning). This framework has been extensively applied to clinical reasoning [11, 12].

System 1 is rather automatic and intuitive, resembling a reflex response, meaning that it is generated without much effort. Using information that is readily available its responses are based on pattern recognition and previous experience [12]. It processes information holistically, generating rapid impressions without conscious effort. This enables very experienced doctors to recognize familiar patterns and generate hypotheses within the first few minutes of patient encounters [12].

System 2 operates actively and slowly, engaging in systematic analysis and consciously applying rules previously acquired through learning [12]. This system requires active effort and structured problem-solving. Expert clinicians predominantly rely on System 1 thinking through extensive "illness scripts" and effective heuristics, while novices typically depend more heavily on System 2 [11].

The interaction between these systems creates the dynamic observed in clinical reasoning. It usually involves initial rapid hypothesis generation through System 1, followed by analytical evaluation through System 2 [12]. The positive impact of this synergy could be shown in a previous study that provided evidence that the combination of both strategies improved diagnostic accuracy in novices significantly compared to when only one of them was used

[13]. Studies have also shown the importance of intuition in diagnostics [14]: When a correct hypothesis is considered in the first five minutes, the likelihood of finding the correct diagnosis is about 95%. If this is not the case, there is the same probability of missing it.

Errors in reasoning play a significant role in diagnostic or therapeutic mistakes that may compromise patient safety. Framing clinical reasoning in the context of dual-process theory also led to evidence that both System 1 and System 2 are equally prone to errors for similar as well as different reasons [15]. A published book about the foundations of clinical reasoning by Cooper and Fraim categorized important error sources in the following manner [16]: (1) no-fault errors (e.g., unavoidable errors due to missing or withheld information), (2) system errors (organizational and contextual factors such as lack of resources, lack of supervision, deficient communication, i.a.), (3) knowledge gaps (ideal reasoning still cannot compensate substantial gaps in knowledge), (4) misinterpretation (of diagnostic tests and how they modify disease probabilities), and (5) cognitive biases. The latter account for the majority of causes for diagnostic errors and while they are usually associated with System 1, they can also compromise System 2 reasoning processes [16]. Prominent examples include anchoring, confirmation bias, and premature closure, among others. The question to what extent LLMs can fall into the pitfalls of such biases or tend to replicate them in the context of clinical decision making remains largely unexplored.

LLMs and other AI systems possess great potential to excel in each of the mentioned theoretical frameworks that describe clinical reasoning, and in some aspects like computational power, they have long surpassed human performance. Human experts do arguably still have performance advantages due to decades of clinical experience, including intensive and rigorous training. But does a computer have all the potential to become the ideal clinical reasoner?

Motivated by cognitive processes involved in clinical decision making, Table 1.1 gives an outlined overview of a set of reasoning strategies and their use cases. Many of these thinking schemes are interdependent and consist of elements from one another. Table 1.1 can be understood as a simplified toolkit proposition for clinical reasoning and was integrated in the prompts used for this study (chapter 3).

Table 1.1.: Overview of clinical reasoning strategies with brief descriptions and illustrative pathways.

<b>Reasoning Strategy</b>	<b>Description</b>	<b>Example Pathway</b>
<b>Deductive Reasoning</b>	Apply general medical rules, principles, or established facts to draw logical conclusions about the specific case. Analyze how certain circumstances can condition others and what might cause them or follow from them. This strategy applies broad knowledge and principles to infer targeted clinical decisions and is essential in almost any reasoning process—applicable in diagnosis, monitoring, or treatment planning, among others.	“All patients on warfarin require INR monitoring → This patient started warfarin → Order INR before next dose.”
<b>Hypothetico-Deductive Reasoning</b>	Generate multiple hypotheses about diagnosis, management options or expected outcomes, then systematically evaluate each through available evidence.	“Generate 3-5 differential diagnoses → Identify discriminating features → Test each hypothesis against the clinical data → Refine or reject hypotheses.” “Three possible treatments exist → Compare efficacy, contraindications, and side effects for this patient → Select treatment that best fits patient’s specific context.”

Table 1.1.: (continued)

<b>Reasoning Strategy</b>	<b>Description</b>	<b>Example Pathway</b>
<b>Inductive Reasoning</b>	Build general conclusions from individual observations. This approach synthesizes specific data points (findings, trends, responses) into a broader understanding, often guiding initial diagnoses, treatment direction, or research. This can also be useful in situations where an unidentified condition affects several patients simultaneously and certain findings in one or a few of them are transferred to all.	“Multiple cases with unknown condition → Identify pattern in first few patients → Generalize findings to all affected patients.”
<b>Abductive Reasoning</b>	Start with puzzling findings and work backwards to find the most plausible explanation. This is ideal for integrating new or unexpected clinical data into coherent decision-making.	“Patient shows these findings → Which causes could explain all of them? → Choose the explanation that accounts for the most features with the fewest assumptions.”
<b>Probabilistic Reasoning</b>	Make decisions based on the relative likelihood of possible conditions or outcomes based considering risk factors, prevalence and context. Frequently used for test selection, triage and risk-informed treatment decisions.	“Assess pretest probability → Factor in test characteristics → Calculate post-test probability → Decision based on likelihood thresholds.”
<b>Rule-Based / Categorical / Deterministic Reasoning</b>	Apply fixed clinical criteria or thresholds. Use validated scoring systems for diagnosis, treatment or risk stratification. This minimizes variability and is useful in high-stakes or guideline-driven settings.	“PERC rule negative + low clinical suspicion → D-dimer not needed → PE effectively ruled out.”

Table 1.1.: (continued)

<b>Reasoning Strategy</b>	<b>Description</b>	<b>Example Pathway</b>
<b>Causal Reasoning</b>	Trace pathophysiological sequences to understand findings, predict complications or anticipate treatment effects. Useful in explaining complex presentations or predicting downstream effects of disease or treatment.	“Intervention/condition A → Causes physiologic change B → Leads to clinical finding C → Expect downstream effect D/ explains current clinical picture D.”
<b>Heuristic / Pattern Recognition (Fast Thinking)</b>	Use intuitive, experience-based recognition of common clinical patterns. Fast, but potentially error-prone - best suited to familiar scenarios.	“Patient presents with classic triad X, Y, Z → Instantly fits known syndrome or condition → Immediate action without lengthy analysis.”
<b>Red Flag / Rule-Out Reasoning</b>	Prioritize identifying or excluding critical and time-sensitive conditions that would drastically change management, regardless of their probability. Especially important in emergency, triage, or risk-sensitive settings, whenever the cost of missing a diagnosis far exceeds the cost of over-investigating or over-treating.	“Chest pain + sudden onset, tearing quality, asymmetric arm blood pressures → Rule out aortic dissection before considering other causes.”
<b>Guideline-Based Reasoning</b>	Apply evidence-based protocols/ algorithms for diagnosis, treatment selection, monitoring, or follow-up based on established and standardized recommendations. Effective in test ordering, management decisions, or when new data emerge mid-course.	“Consult relevant clinical guideline → Map patient features to decision steps → Follow recommended pathway → Implement standardized approach.”

### 1.3. Clinical Prompt Engineering

Several studies have examined strategies for enhancing clinical reasoning in LLMs with varying success using different prompt-engineering techniques [17, 18, 19, 20, 21, 22]. Chain-of-thought (CoT) prompting has demonstrated how it can enhance LLM performance on

complex reasoning tasks by making the processes explicit, generating a series of intermediate reasoning steps [23]. However, the effectiveness of CoT for medical problem solving varies depending on the tasks and their context. In a study by Singhal et al. from 2023, CoT did not perform better than standard few-shot prompting across several medical benchmarks like MedQA, MedMCQA and PubMedQA [17]. Instead they introduced an alternative approach named instruction prompt tuning, a way to efficiently train a model to follow instructions in a specific domain [17].

In a study by Savage et al. the effectiveness of several different clinical reasoning prompts on the MedQA-USMLE dataset was studied and compared to conventional CoT [18]. Five strategies were tested: Traditional CoT, Differential Diagnosis CoT (creating differential diagnoses then narrowing to final answer), Intuitive Reasoning CoT (using symptom-disease associations), Analytical Reasoning CoT (focusing on pathophysiology and biochemical mechanisms) and Bayesian Reasoning CoT (using probabilistic inference). GPT-3.5 struggled with advanced clinical reasoning strategies and its performance even decreased with analytical reasoning and differential diagnosis CoT. This was not the case with GPT-4, where performance remained consistent across all prompted strategies, but on the downside no significant improvement by any kind of diagnostic reasoning CoT could be demonstrated either. GPT-4 was able to successfully imitate the cognitive processes of physicians, but not to apply clinical reasoning like a human as Savage et al. state [18]. The authors argue this finding still as important step towards interpretability and explainability, mitigating the black-box limitations of LLMs [18].

Structured clinical reasoning approaches as introduced by Sonoda et al. showed a strong performance boost [19]. They developed a two-step approach that first systematically organizes clinical information into predefined categories based on common clinical practice (like patient information, medical history, symptoms, physical examination, laboratory results, imaging findings i.a.) before applying diagnostic reasoning. This structured approach significantly improved diagnostic accuracy for primary diagnosis and for top three diagnoses, demonstrating that the systematic organization of information and breaking it down into manageable problems enhances diagnostic performance. The authors suggested that the structured summarization mirrored human reasoning processes leading to more clinical alignment and potentially also reducing errors [19].

Lucas et al. proposed an iterative ensemble reasoning method where LLMs review multiple diverse reasoning pathways to refine their own answer [20]. After generating a set of different responses by zero-shot CoT with self-consistency, the responses were grouped into two different consistency groups ( $Q_{con}$  and  $Q_{inc}$ ). In case every response was consistently correct, they were all grouped in  $Q_{con}$  and no further processing was done. If there was a combination for correct and incorrect or only incorrect answers, then all of them would be grouped in the inconsistency group  $Q_{inc}$ . The ensemble method only took responses from  $Q_{inc}$  as input and the models were instructed to evaluate each reasoning and answer critically while not just

following the majority. After the models provided new responses with different reasoning paths the answers were organized in the same manner as already described. This process was repeated until convergence into  $Q_{con}$  was reached or a predefined exit condition was met. This approach outperformed standard CoT with self-consistency on GPT-3.5 turbo and Med42-70B, implying that exposure to different reasoning paths improved medical question answering and that convergence to correct answers could be reached iteratively [20].

In a study by Kwon et al. a clinical CoT based diagnosis framework divided in two modules proved effective for Alzheimer’s disease on ADNI dataset [21]: Before making predictions, real patient data, including the diagnosis, was augmented with diagnostic rationales by the model mirroring human reasoning processes. In a first module, LLMs were prompted with the clinical CoT augmented patient data and this approach outperformed standard few-shot prompting substantially in GPT-4. In a second module, models orders of magnitude smaller in parameter size were trained on the reasoning augmented patient data. Surprisingly, these “student models” outperformed the bigger “teacher models” after learning to generate clinical rationales based on the CoT of teacher models [21]. This shows how mimicking a clinical thought process alone can alter performance positively even though the model is otherwise in a disadvantageous position compared to the teacher model.

Nachane et al. used incremental reasoning for mimicking clinical thought processes for open-ended medical questions [22]. In their study a modified version of the MedQA-USMLE dataset (MEDQA-OPEN) was used, containing open-ended questions without multiple-choice options to better mimic clinical scenarios. The CLINICR prompt they developed uses a 5-shot CoT strategy that progressively integrates all patient data and medical findings present in the case vignette, mirroring prospective clinical reasoning for formulating a set of differential diagnosis. This incremental strategy clearly outperformed eliminative methods, as the latter’s backward reasoning from predefined options unsurprisingly proved ineffective in open-ended questions. A forward-backward method combining both approaches, first generating multiple differential diagnoses through CLINICR, then narrowing to a final diagnosis using elimination, achieved the highest expert agreement [22]. This framework replicated an aspect of real-world clinical decision-making where physicians develop multiple potential hypotheses without any constraints of limited answer options, before systematically narrowing them down.

Looking at the different approaches in the literature a common denominator usually is to imitate the human cognitive process in clinical decision making. However, the effectiveness of clinical reasoning enhancement techniques varies significantly across different assessment formats, particularly when evaluating clinical reasoning under uncertainty. McCoy et al. introduced Script Concordance Testing (SCT) as a novel benchmarking approach (SCT-Bench) for LLM clinical reasoning (section 1.6), comprising 750 questions from 10 internationally diverse medical datasets that assess how new information influences clinical decisions under ambiguity [24]. On many SCTs LLMs match or exceed medical student performance,

but consistently lag behind senior residents and expert physicians [24]. Most notably, the reasoning models among all studied models unexpectedly underperformed on their SCT-Bench compared to previous medical benchmark scores. The authors claimed their SCT-Bench to be the first medical benchmark where enhanced reasoning actually decreases performance rather than improving it [24]. This inversion suggests that current features to enhance LLM reasoning may not address the specific requirements of clinical decision making under uncertainty. SCTs evaluate fundamentally different cognitive capabilities than traditional multiple-choice questions and their findings highlight limitations in current LLM clinical reasoning that are not apparent in traditional medical benchmarks. This suggests that clinical judgment involves capabilities beyond encoded biomedical knowledge or explicit reasoning chains.

## 1.4. Medical Benchmarking

As clarified at the end of section 1.1, the ability of LLMs to accurately process and generate sound clinical information as well as reasoning is, among other qualities, crucial for integrating the technology in a clinical decision making pipeline. The assessment of these abilities is anything but trivial and requires a multidisciplinary approach. The evaluation of LLMs in medical contexts has been predominantly anchored by traditional medical licensing exams like the United States Medical Licensing Examination (USMLE). The USMLE has emerged as a cornerstone benchmark, representing a three-step process that tests clinical competency across different stages of medical training [6]: Step 1 covers core clinical knowledge like pathology and physiology. Step 2 tests clinical understanding by examining knowledge about diagnosis and patient management, while Step 3 assesses medical residents' ability to apply their knowledge in unsupervised practice. The prominence of USMLE-style questions in LLM evaluation has been reinforced by the development of benchmarks such as MultiMedQA, which brings together multiple medical question answering datasets like MedQA, PubMedQA, MedicationQA among others to enable evaluation beyond simple accuracy metrics [17]. Since the early days of mainstream LLM development it became impressively obvious how fast this field is evolving and performances improved rapidly within a short time period. GPT-4 showed impressive improvement over predecessors on official USMLE exam questions. It raised the score by over 30% compared to GPT-3.5 on the USMLE self-assessment and sample exam [6]. Medical LLMs have improved their performance within a short span of time too. Med-PaLM 2 achieved up to 86.5% accuracy on the MedQA dataset, outperforming Med-PaLM by over 19% [25]. The most recent general purpose models have reached superhuman levels, achieving unprecedented results. OpenAI's reasoning model o1 currently leads the MedQA leaderboard with an accuracy score of 96.5%, getting close to the practical ceiling of the benchmark [26].

The outstanding performances of LLMs have led to considerable optimism about future applications and their integration in clinical research as well as healthcare systems. However, it is important to emphasize that traditional medical benchmarks do have a range of strong limitations, which do not allow translating LLM accuracy to performance in real-life scenarios. It has been suggested that high scores on USMLE-style datasets likely reflect encoded knowledge and test-taking abilities rather than inherent (clinical) reasoning [24]. Furthermore, traditional multiple-choice tests do not adequately represent the complexity of real clinical decision making as Hager et al. pointed out in their publication about their innovative MIMIC IV dataset [27]. Clinical decision making is a multistep process that requires gathering and synthesizing data from diverse sources, continuously re-evaluating the situation. This is a very interactive process with a back and forth dynamic between the clinicians and all other involved parties, while commonly dealing with a lot of uncertainty. In contrast, LLMs' diagnostic capabilities have been tested primarily in structurally simple medical contexts like case vignettes with all required information provided upfront alongside answer options [27]. The limitations extend beyond assessment format and also include concerns about data contamination and memorization effects. Researchers have developed methods like MELD (memorization effects Levenshtein detector) to scan for training data memorization, finding no evidence in official USMLE datasets, but indeed in other datasets [6]. The paywall protection likely prevented USMLE content from being included in the training data [6]. Acknowledging the limitations certain assessment methods have has led to recognizing the necessity for alternative and more sophisticated evaluation frameworks. Hager et al. addressed this need by developing an evaluation framework around real patient cases from the MIMIC-IV database [27]: Their approach intends to simulate realistic clinical workflows, where LLMs are presented only a patient's history of present illness. Subsequently, they are tasked with autonomously requesting additional information like physical examinations, laboratory tests, imaging reports etc. in an iterative and interactive manner with the goal of arriving at a diagnosis and treatment plan. LLMs were evaluated on over 2000 patient cases, finding that state-of-the-art LLMs did not reach the diagnostic accuracy of physicians across all included pathologies and failed to follow diagnostic guidelines [27]. Their study also pointed out that LLMs struggle with interpreting laboratory results even when provided with reference ranges and are sensitive to both the quantity and order of information presentation [27].

## 1.5. MedQA-USMLE

When MedQA was introduced, it was the first large-scale open-domain MCQ dataset specifically for the medical domain [28]. Previously existing QA datasets had been largely solved by LLMs and they typically required only elementary or middle school knowledge. Real-world medical scenarios demand sophisticated professional knowledge and reasoning abilities that datasets usually did not challenge and MedQA addressed this gap. In total, MedQA comprises

just about over 60.000 questions across several languages, which were sourced from real medical board examinations [28]. Questions in English language are provided by the US medical licensing examination, forming MedQA-USMLE (hereafter referred to as MedQA). The dataset generally contains two primary question categories: i) questions expecting a single piece of knowledge and ii) clinical case vignettes asking for the most appropriate management in any aspect or about underlying mechanisms of the condition described [28].

Key challenges when taking the MedQA test include [28]:

- i. Professional Knowledge: Almost every question demands extensive medical knowledge, including understanding pathophysiological mechanisms.
- ii. Multi-hop Reasoning: Many questions require reasoning steps across multiple information snippets. For example, identifying patient symptoms, matching them to potential diagnoses and determining appropriate treatments or further steps of management.

## 1.6. Script Concordance Testing

SCTs have been designed to evaluate clinical reasoning capabilities under uncertainty, a fundamental skill hardly implied in common MCQ exams [24]: A clinical scenario is presented in a brief case vignette, including an intervention, diagnosis or treatment option. Then, new information to the case is introduced and SCT asks how it affects the likelihood of the initial hypothesis. The rating responses are scored against an expert panel [24]. SCT is grounded in script theory from cognitive psychology, which proposes that clinicians use organized knowledge structures (scripts) to make sense of clinical problems and guide decisions [29]. These scripts become more refined with clinical experience [29]. This approach addresses a critical gap in traditional assessment methods by focusing on how clinicians integrate new information to modify their clinical reasoning, rather than focusing on testing their ability to recall medical facts. Instead of providing scenarios with complete information that got definitive answers, SCT presents situations where new information must be weighed against existing hypotheses, mimicking the nature of clinical decision making. This requires the ability to reason with uncertainty, a crucial skill in clinical practice that is not measured by conventional multiple-choice examinations. This is done in the following way by providing:

- i. A clinical case vignette with e.g. diagnostic ambiguity
- ii. An initial hypothesis ("If you were thinking of...")
- iii. New clinical information ("And then you find...")

The test taker is then required to rate on a 5-point Likert scale (-2 to +2) how the new information changes the likelihood of the hypothesis, where -2 means it decreases strongly and +2 the opposite. Table 1.2 shows the rating system used SCTs.

Table 1.2.: 5-point Likert Scale used for rating SCT questions.

Rating	Likelihood of hypothesis
-2	strongly decreases
-1	slightly decreases
0	no effect
+1	slightly increases
+2	strongly increases

### SCT Question Example

#### SCT-STEM:

A 57-year-old female presents to the ED with a red right eye. She reports the eye is painful and has been red for the past 2-3 hours. The eye is painful all the time and moving the eye does not alter the pain. She has tried paracetamol and ibuprofen for the pain but this has not alleviated the pain significantly.

If you were thinking of: Using topical atropine to enable a more detailed ophthalmic examination”

And then you find: She reports seeing halos of lights around objects with the affected eye.

Example 1.1.: Example of an SCT question.

## 1.7. Evaluation Metrics

Beside evaluating abilities like medical knowledge and coherent clinical reasoning, there are more qualities that make a good clinician and that AI technology must adhere to for reliable and responsible use. It was shown that evaluating LLMs systematically and across many dimensions is far from trivial and there have been a few proposals on the lookout for a standardization of evaluation metrics. Abbasian et al. proposed a framework encompassing parameters like accuracy (factuality, robustness, up-to-dateness), trustworthiness (safety, privacy, bias, interpretability), empathy (emotional support, health literacy, fairness, personalization) and performance metrics (memory efficiency, latency) [30]. This multidimensional approach recognizes that clinical applications require evaluation beyond scientific accuracy considering the multifaceted nature of medical decision making, which apart from factual correctness, technological feasibility also implies ethical and humanistic aspects. For this reason it is crucial to evaluate and supervise future agentic AI technology across all implied dimensions

and develop standardized approaches to do so.

## 1.8. Calibration and Uncertainty Quantification

The ability to accurately quantify and communicate uncertainty represents a fundamental requirement for deploying large language models in clinical settings, where overconfident incorrect predictions could have life-threatening consequences. Calibration, meaning how well a model's confidence matches its actual accuracy, is especially important in medicine, where reliable probability estimates would help clinicians to relativize outputs accordingly [6]. However, the inability of LLMs to reliably communicate uncertainty remains a significant barrier to their clinical implementation [31]. Uncertainty quantification methods for LLMs can be broadly categorized into white-box approaches requiring internal model access and black-box methods using only model outputs. White-box methods leverage internal representations such as token logits and activations, but are often inaccessible for proprietary models [32]. Black box methods estimate confidence using only external access to model outputs, making them essential for closed-source commercial application programming interfaces (APIs). These include sample-based and self-evaluation approaches that can vary considerably in sophistication [32]. Discussing either white or black-box techniques would go beyond the scope of this thesis, but due to relevance two methods will be briefly mentioned:

- i. Prompt-based confidence elicitation directly asks models to express their uncertainty themselves. Verbalized confidence scores include numerical scales, verbal qualifiers and might even use CoT enhancement [33].
- ii. Sampling strategies use multiple generations to estimate how stable responses are. Assessing self-consistency within a sample of answers allows to measure the degree of internal agreement and this can be used as uncertainty parameter. There are advanced variants of this approach like confidence-informed self-consistency that achieve substantial reduction of computational power by weighted voting based on model confidence scores [34].

Studies suggest a performance gap between black- and white box approaches, where white box methods typically achieve higher AUROC scores (subsection 2.3.1) and lower ECE values (subsection 2.3.2) [33]. However, also if white box methods perform generally better, the difference is rather narrow.

Confidence elicitation, where models are prompted to express confidence verbally or numerically, consistently demonstrates systematic overconfidence across medical, but also other domains. Naderi et al. evaluated 48 different LLMs on gastroenterology questions, finding that average confidence exceeded average accuracy [35]. This overconfidence persists when models are explicitly asked to assess uncertainty, suggesting fundamental limitations in self-assessment capabilities and calibration plots show that confidence elicitation methods are

almost universally overconfident [31]. However, it is worth noting that these observations stay in contrast to the findings of other studies concluding that LLMs mostly "know what they know" [36]. Liévin et al. corroborated this notion in the medical context, showing that models such as Llama 2 were well calibrated on medical benchmarks (MedQA, MedMCQA, Pub-MedQA) when using 5-shot CoT prompting [37]. Recent work from Steyvers et al. proposes an interesting view on that matter, identifying the underlying problem in how confidence is communicated rather than poor "self-awareness" of the models [38]: While LLMs internally show good discrimination between correct and incorrect answers, there is more of a gap between what models know and users think they know. Provided with model generated explanations, humans' discrimination between correct and incorrect performed near chance level (AUROC  $\sim 0.59$ ; see subsection 2.3.1) despite fairly good internal model calibration [38]. These results suggest that the primary challenge lies not so much in the model's ability to assess their own knowledge, but in the accurate communication of uncertainty to a user [38].

Confidence scores or any kind of uncertainty parameter are necessary for estimating how well calibrated an AI system is and how good it is at discrimination between accurate and inaccurate predictions. Savage et al. conducted a comprehensive evaluation of uncertainty measurement methods specifically for medical diagnosis and treatment tasks, examining confidence elicitation, token level probability and sample consistency metrics across several models including GPT-3.5 and GPT-4 [31]: Among the tested approaches it was found, that sample consistency was generally the superior method. Additionally, it is also worth noting that uncertainty assessment showed to be task-specific. Almost all uncertainty metrics demonstrated better discriminative ability for diagnosis questions compared to ones about treatment [31].

## 1.9. Research Focus and Goals

The limitations of current medical benchmarking and the challenges for evaluating clinical competence in real-world scenarios underscore the need to further study how LLMs perform and behave across different assessment formats. While existing research has demonstrated impressive LLM performance on traditional medical examinations like MedQA-USMLE among others, the effectiveness of clinical reasoning enhancement across different benchmark types remains largely unexplored. There is a significant disparity in performance between knowledge-focused MCQ examinations, the most important representatives of medical benchmarking, and SCT as newly proposed method for benchmarking. Understanding the discrepancy better could further reveal current limitations in clinical reasoning in LLMs.

### 1.9.1. Research Objectives

**Medical Benchmark Comparison.** The primary focus of this study is analysing contrasting performance qualities as well as the impact of clinical chain-of-thought reasoning instruction prompting across two distinct medical benchmark types: MCQ-based MedQA and SCT-Bench using script concordance testing. This comparative analysis aims to reveal assessment-specific effects and limitations that may have implications for how we evaluate clinical competence in LLMs and generally language based CDSSs. Systematically comparing how clinical reasoning prompting affects LLM performance across MedQA and SCT evaluation formats includes analysing accuracy metrics, response consistency patterns and evaluation of reasoning strategies. Contrasting results between these benchmarks can provide insights into what qualities of clinical skills they actually challenge. Previous work has demonstrated the efficacy of clinical reasoning enhancement techniques on traditional medical licensing exams through domain-specific prompt engineering. However, it remains unclear whether such strategies are equally effective for SCT questions and this study addresses this gap by comparing reasoning enhancement effects:

- i. How does clinical reasoning instruction prompting affect MedQA and SCT performance by GPT-4.1 across several evaluation metrics?
- ii. How do GPT-4.1 rating distributions on the SCT benchmark compare to expert rating distributions and what is the effect of clinical reasoning instruction prompting in this regard?

**Clinical Reasoning Evaluation.** Using GPT-4.1, an instruction-following-optimized model, the ability of LLMs to authentically replicate and self-report clinical reasoning strategies when explicitly instructed to do so is investigated. This includes assessing whether reasoning pathways align with self-reported use of strategies as well as further statistical analysis of different reasoning approaches:

- i. Is GPT-4.1 able to effectively implement specific clinical thinking strategies and self-reflect on these reasoning pathways?

**Uncertainty Quantification.** Differences in confidence calibration and discrimination ability (AUROC) across MedQA and SCT-Bench are assessed using two different uncertainty proxies (self-elicited confidence, sample consistency):

- i. Is GPT-4.1 able to maintain calibration and discrimination ability across different types of medical benchmarking (MedQA vs SCT)?
- ii. Is either self-elicited confidence or sample-based consistency superior as a black-box method for quantifying uncertainty?

### 1.9.2. Impact Goals

This comparative analysis aims to identify limitations in current medical AI evaluation approaches and inform improvements to the assessment of clinical competence in LLMs. It may

corroborate already known as well as reveal further weaknesses of existing benchmarks and how LLMs handle uncertainty and incomplete information in a clinical context. Examining when clinical reasoning augmentation hurts performance may emphasize the limitations of CoT orientated prompt engineering and highlight aspects of clinical competence that remain challenging for language models. Findings from this thesis may contribute to the collective effort of understanding benchmark limitations and developing evaluation frameworks that address them.

## 2. Methods and Resources

### 2.1. Datasets

The primary sources for medical test questions and case vignettes were two different datasets: MedQA-USMLE, which has already been used extensively for benchmarking and evaluating LLMs in the context of biomedicine and SCT-Bench introduced by McCoy et al. [24]. These two benchmarks evaluate biomedical knowledge and reasoning following distinctive assessment paradigms and both measure clinical competence, yet different qualities and aspects of it (see section 1.6).

#### 2.1.1. MedQA-USMLE

As a well established benchmark, MedQA served as both the primary reference representing conventional MCQ assessments and the comparison point for SCT. A subset of 700 USMLE questions, ranging from step 1 to step 3, was used for evaluating OpenAI’s gpt4.1 model (section 2.2). Every test item is a classic MCQ question that provides 5 answer options, where one answer only is labelled as correct. The final MedQA score is the percentage of correctly answered questions.

#### 2.1.2. SCT-Bench

McCoy et al. published an SCT-based benchmark (SCT-Bench) comprising in total 750 questions across 10 different SCT datasets from a diverse range of medical specialties with origin in North America, Asia and Australia [24]. However, only a smaller subset including 174 questions is publicly available. The freely accessible dataset is a combination of Open Medical SCT (University of Florida, 72 paediatrics questions) and Adelaide SCT (University of Adelaide, 102 multi-specialty questions). For this comparative benchmark analysis the smaller subset accessible for the public was used.

The expert consensus is scored with a value of exactly 1.0 while all other ratings are weighted with a fraction of this maximum score based on the proportion of experts selecting each option. Responses that no expert opted for correspondingly receive 0 points.

## 2.2. Model and Data Generation

For this benchmark comparison OpenAI’s GPT-4.1 (gpt-4.1-2025-04-14) model was studied. Unlike more recent reasoning-optimized models (such as the o-series), GPT-4.1 does not have built-in CoT features and instead relies on explicit, instruction-heavy prompting. OpenAI notes that this model is particularly suitable for tasks requiring precise adherence to specified

formats and multi-step procedures [39]. This property was considered to be a good condition for studying the impact of CoT inducing instructional prompts and compare it to a zero instruction baseline. Further technical settings were rudimentary with temperature = 0.7 and top-p = 1.0 for all runs and queries.

Data generation was achieved by sending all prompts via OpenAI API keys to GPT-4.1 and writing generated responses to structured response files, which were used for further data extraction (2.4). All responses were stored in JSON format for subsequent analysis, including extracted answers, confidence scores and reasoning strategy data. The publicly available code for automated API queries from McCoy et al. [24] served as a template and was adapted for this study's analytical purposes. For certain text-extraction and further data analysis functions AI assistance for coding was used [40]. Information regarding code and data availability can be found in the supplementary materials (chapter A.2).

## 2.3. Prompting and Uncertainty Quantification

Two kind of data queries were run:

- i. Baseline prompts with no CoT: Zero-instruction prompts telling the model to answer straight without any kind of reasoning steps or CoT.
- ii. Clinical CoT prompts: Prompts augmented with instructions for clinical CoT, incremental information processing applying principles of clinical thinking and a list of clinical reasoning strategies described in the literature [11]. The prompt templates can be found in Appendix A.

In every prompt GPT-4.1 was instructed to quantify the level of confidence it has in its own response along to the answer itself. For confidence elicitation, a simple one-step approach was applied: After answering each question, the model was asked to rate its uncertainty on a scale from 0 to 100, where 0 represents complete uncertainty and 100 represents complete certainty. This kind of confidence elicitation has generally not proven to be very effective and accurate, but due to its simplicity and no real additional expenses it was employed to get this extra information.

Inspired by the Savage et al. study about uncertainty proxies in diagnostic and treatment related test questions, uncertainty was additionally quantified via sample consistency [31]. Every question item was sent 15 times ( $n = 15$ ) to the model in order to generate 15 different responses to the same problem. According to previous studies sample consistency performance has usually reached a plateau at  $n = 15$ , which was defined as the per question sample size [41]. The response or rating that was selected most frequently was considered as the model's reference answer for determining correctness. The sample consistency score for each question was calculated as the fraction of responses that matched this majority response ( $X/n$ ), providing a measure of response reliability that served as an uncertainty proxy.

Additionally, reasoning strategy extraction was implemented for CoT prompts. The model

was instructed to identify and report which clinical reasoning strategies it applied from a predefined list (see Table 1.1), enabling analysis of strategy distribution and their relationship to performance outcomes.

### 2.3.1. ROC Curve

The receiver operating characteristic (ROC) curve [42] is a tool for evaluating classification performance by plotting the true positive rate ( $\hat{=}$  sensitivity) against the false positive rate ( $\hat{=}$  1-specificity) [43] across all possible decision thresholds.

Formally, at each threshold  $\tau$ , predictions with confidence  $\geq \tau$  are classified as correct, yielding [43]:

$$\text{TPR}(\tau) = \frac{\text{TP}(\tau)}{\text{TP}(\tau) + \text{FN}(\tau)}, \quad \text{FPR}(\tau) = \frac{\text{FP}(\tau)}{\text{FP}(\tau) + \text{TN}(\tau)} \quad (2.1)$$

where TP, FN, FP, and TN denote true positives, false negatives, false positives and true negatives, respectively. The ROC curve is then the set of points  $\{(\text{FPR}(\tau), \text{TPR}(\tau)) : \tau \in [0, 1]\}$ .

In the context of LLM confidence estimation, the ROC curve visualizes how well confidence scores can distinguish between correct and incorrect model responses. The area under the ROC curve (AUROC) provides a single scalar metric summarizing the discriminative ability of confidence estimates across all thresholds. AUROC values range from 0 to 1, where 0.5 indicates random performance equivalent to chance, and 1.0 perfect classification (represented by a point at (0,1)) [43]. Furthermore, in this particular context it equals the probability that a randomly selected correct response receives higher confidence than a randomly selected incorrect response [44]. Mathematically, AUROC is literally what the name suggests it is: the area under the ROC, which can be approximated applying numeric integration, e.g. using the trapezoid rule [44, 43].

The AUROC parameter serves not only as a metric for measuring how well a specific method for confidence elicitation or assessment works, but can also be interpreted as reflecting a model's "self-awareness": While perfect AUROC can be achieved even with badly calibrated confidence scores, the estimates still have to correlate with correctness. In other words, higher confidence must correspond to greater likelihood of being correct, which given the case shows a sense for self-evaluation.

### 2.3.2. Expected Calibration Error

The expected calibration error (ECE) quantifies the alignment between predicted confidence levels and actual accuracy rates, serving as a metric for measuring confidence calibration. A model is well calibrated if an outcome predicted with probability  $p$  actually occurs approximately  $p$  fraction of the time for the whole predicted range of  $p$  [45]. The ECE is computed

by partitioning predictions into  $M$  fixed equal-width bins, before determining the weighted average of the absolute differences between mean confidence and accuracy within each bin [38]:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (2.2)$$

where  $M$  is the number of bins,  $B_m$  represents the  $m$ -th confidence bin,  $n$  is the total number of predictions,  $\text{acc}(B_m)$  is the true fraction of correct predictions in bin  $m$  (accuracy) and  $\text{conf}(B_m)$  is the mean of the predicted confidence scores in that bin. Lower ECE values indicate better calibration. In an ideal system, confidence and accuracy align perfectly.

## 2.4. Evaluation Metrics and Data Analysis

The primary evaluation metric for assessing and quantifying the model's performances on either benchmark is the correspondent test score. The score on the respective dataset was calculated in two ways: using the majority responses as reference answer, as well as by randomly sampling answers for determining an average score value and comparing it to the majority score. Further metrics include response entropy analysis, ROC and calibration parameters, quantification and qualitative analysis of self-reported use of reasoning strategies as well as LLM rating distributions for SCT-Bench.

The following list provides an overview of quantified parameters and characteristics for evaluating performance and other properties. Statistical testing for benchmark scores and response entropies was done using paired and unpaired t-tests, as well as McNemar's test. Results are discussed in chapter 4:

- i. Majority MedQA score: This is the MedQA test score determined by using the most frequently selected answer across  $n$  runs. It represents the score the model achieves when choosing the response it is most consistent with.
- ii. Mean MedQA score: Running  $n = 15$  queries per test item yielded a large sample of simulated test results by randomly selecting one arbitrary response per question. The mean of that sample is the mean MedQA score. For this purpose, the mean MedQA score was determined from a sample with size  $N = 100$ .
- iii. Majority SCT score: Equivalently to the MedQA score, the majority SCT score is calculated from the most frequently chosen ratings. The score itself is computed with a normed method like explained in subsection 2.1.2.
- iv. Mean SCT score: Mean SCT score of  $> 1$  test results with a sample size  $N = 100$ , analogous to the mean MedQA score.
- v. Response entropy: Mean response entropy was calculated to quantify the disagreement

among all responses to the same question using Shannon entropy for categorical data:

$$H = - \sum_{i=1}^k p_i \log_2(p_i) \quad (2.3)$$

This metric allows comparison of response variability between baseline and CoT prompting approaches across both datasets.

- vi. Calibration metrics: ECE (subsection 2.3.2) using both self-elicited confidence scores and sample consistency as uncertainty proxies. TPR and FPR were computed using the `calibration_curve()` function from the `sklearn` module for Python.
- vii. AUROC: The area under the ROC curve (subsection 2.3.1) was computed using the `roc_auc_score()` function from the `scikit-learn` module available for python, evaluated for both self-elicited confidence and sample consistency methods.
- viii. SCT rating distribution and disagreement analysis: A rating disagreement parameter  $\Delta r = r' - r_0$  was defined, where  $r'$  is any rating and  $r_0$  the expert consensus rating. LLM rating distribution around the expert consensus rating was compared to the rating distribution of the expert panel. Applying the Jensen-Shannon distance the similarity of both distributions was quantified for baseline as well as for clinical CoT prompting (section 2.3).
- ix. Rating transition analysis: Sankey diagrams were used to visualize how likelihood ratings transitioned between expert consensus and GPT-4.1 responses as well as between baseline and CoT ratings. Ratings were grouped into three categories (Strong:  $\pm 2$ , Moderate:  $\pm 1$ , Neutral: 0) to capture transition patterns and the effects of clinical reasoning on the model's rating behaviour.
- x. Reasoning strategies: Histograms of self-reported reasoning strategies show their distribution and predominant approaches in each benchmark. Concrete examples illustrate GPT-4.1's capability of mimicking and self-reflect on applied clinical thinking schemes within a decision-making process. Additional analysis examined the relationship between reported reasoning strategies and rating levels in SCT-Bench questions.

# 3. Results

## 3.1. Benchmark Scores

The primary benchmark scores (section 2.4) for both prompting strategies (section 2.3) and each dataset reveal the effect of clinical CoT prompting on the performance of GPT-4.1.

### 3.1.1. MedQA

The majority MedQA score at baseline prompting is 86.4%, while the clinical CoT approach achieved an improved score of 93% (Figure 3.1). This represents a quite impressive accuracy boost by almost 7 percentage points and a statistically significant performance difference ( $p < 0.05$ ).

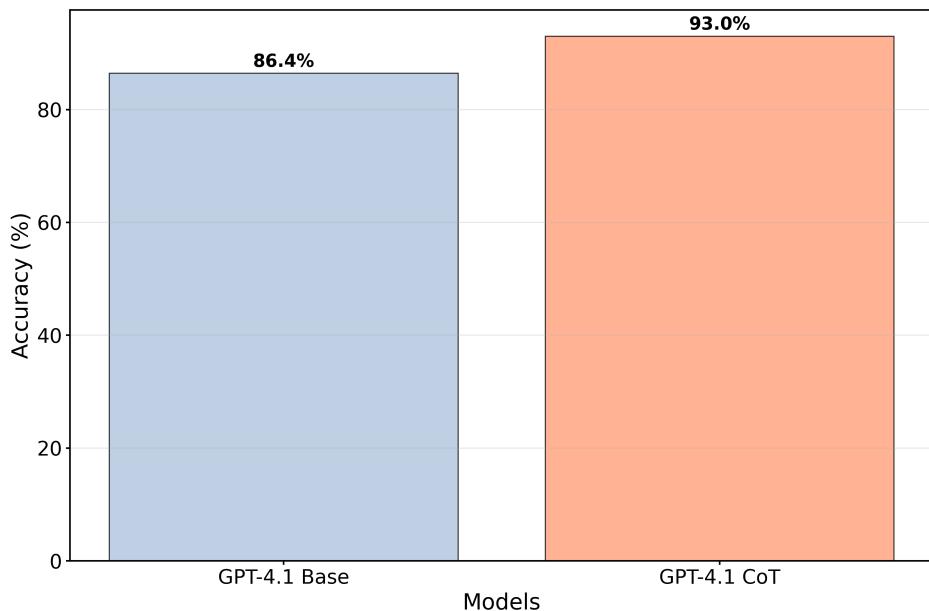


Figure 3.1.: Majority MedQA-USMLE score with vs. without clinical CoT prompting.

As expected, the mean MedQA score improved substantially too (Figure 3.2). We can see an increase in accuracy by 5 percentage points, lower than for the majority score, but significant ( $p < 0.05$ ). Additional observations can be made: At baseline, the majority and mean score are separated by just 0.1 percentage points, while with clinical CoT prompting this difference is 1.4 percentage points. Further analysis shows that the performance benefits significantly ( $p < 0.05$ ) from sample-based majority voting, but remarkably only in CoT mode. For the baseline score, sampling and majority voting was not significantly superior to just picking any response ( $p \approx 0.96$ ).

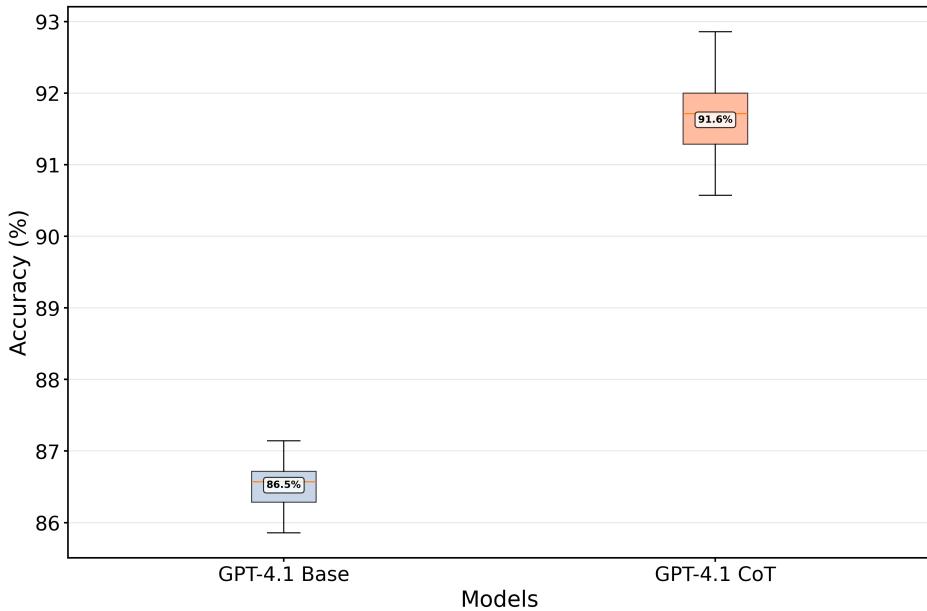


Figure 3.2.: Mean MedQA-USMLE score ( $N = 100$ ) with vs. without clinical CoT prompting

### 3.1.2. SCT-Bench

Results for the SCT benchmark are generally not consistent with those for MedQA (subsection 3.1.1). The SCT majority score at baseline reached 77.7%, but dropped to 74.7% with CoT prompting (Figure 3.3), a decrease of 3 percentage points. However, paired t-testing could not show statistical significance for this performance drop ( $p \approx 0.45$ ).

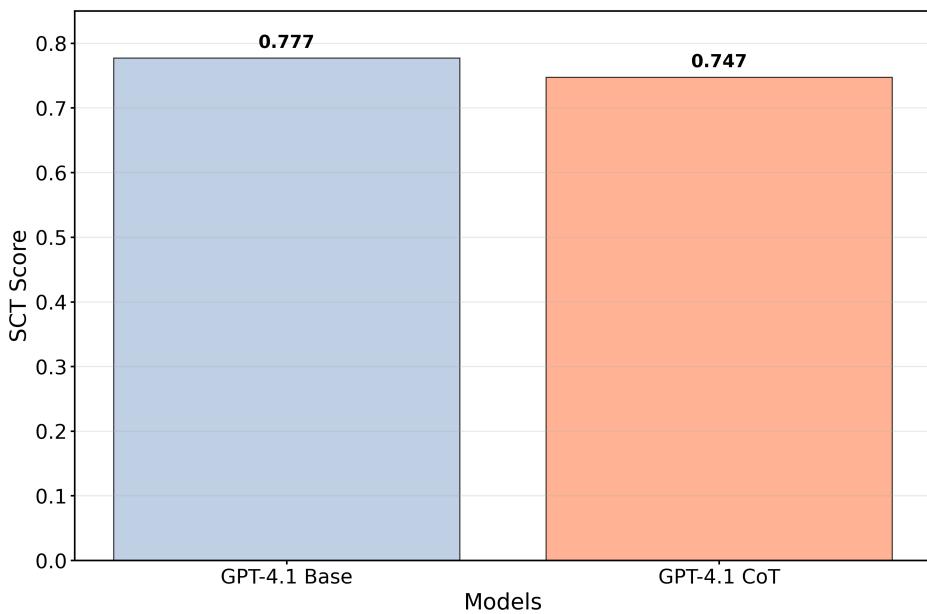


Figure 3.3.: Majority SCT-Bench score with and without clinical CoT prompting.

The mean SCT score dropped from 75.9% to 73.8% which was not significant either according to question-level paired t-testing ( $p \approx 0.54$ ). Majority voting was significantly ( $p < 0.05$ ) beneficial for the baseline as well as for the CoT score, beating the mean score

in both modes. However, baseline had greater advantage from majority voting, which is a symptom of CoT reducing the consistency of correct answers on SCT-Bench.

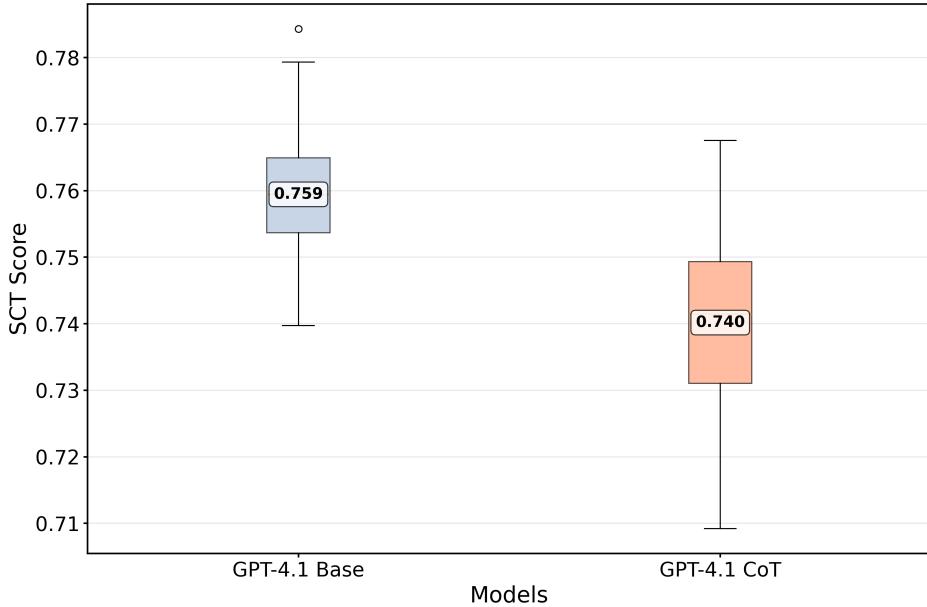


Figure 3.4.: Mean SCT-Bench score with vs. without clinical CoT prompting

Overall, clinical reasoning instruction prompting had a positive impact only on MedQA, by increasing benchmark scores as well as the frequency of correct answers leading to significant advantage of sampling and majority voting.

### 3.2. Response Entropy

Response entropy quantifies the disagreement among all LLM responses to the same question. Figure 3.5 shows the mean response entropy of GPT-4.1 on MedQA for baseline and clinical CoT prompting. With clinical CoT, mean entropy increases significantly ( $p < 0.05$ ) from  $\hat{H}_{base} = 0.034$  to  $\hat{H}_{CoT} = 0.089$ . This indicates that instruction-based clinical reasoning led to greater response variability, with answers becoming more dispersed. Notably, however, the correct answer became more likely to be selected by majority voting, indicating stronger clustering around the correct response despite the overall increase in variance.

As evident in Figure 3.6, clinical reasoning had the same effect on rating variability on the SCT dataset. However, unlike on MedQA, this increased variability did not increase the frequency of correct answers, but rather reduced it. Since SCT is based on multiple reference expert answers, there is also non-zero variance among expert ratings.

As shown in Figure 3.6, response entropy among experts is generally significantly higher than among model answers. While there is usually consensus among most experts, sometimes more or less pronounced, dissenting opinions are common. The response entropy of GPT-4.1 is considerably lower than the expert panel's. Clinical CoT significantly increased the model's

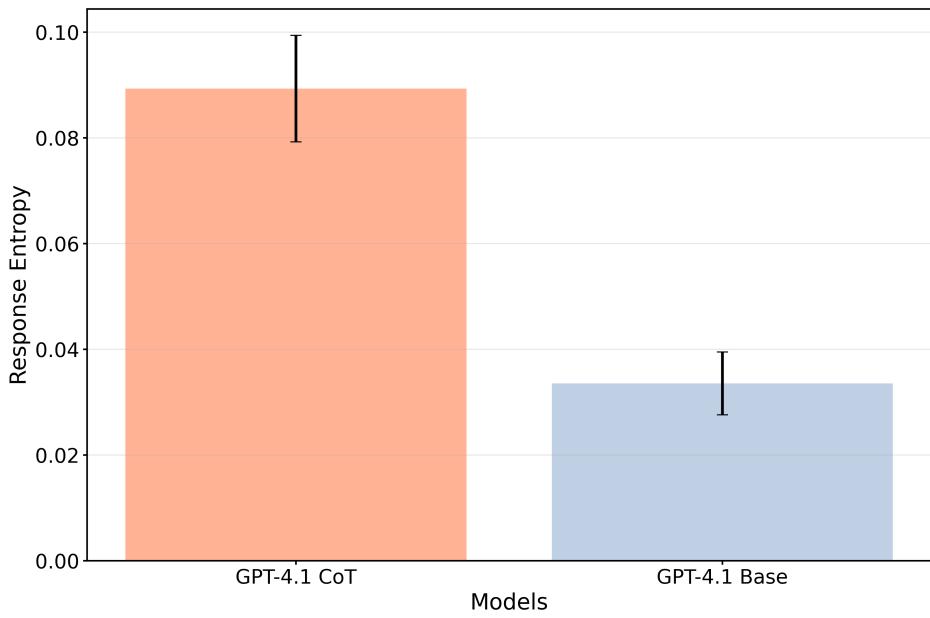


Figure 3.5.: GPT-4.1 response entropy on MedQA-USMLE with vs. without clinical CoT prompting.

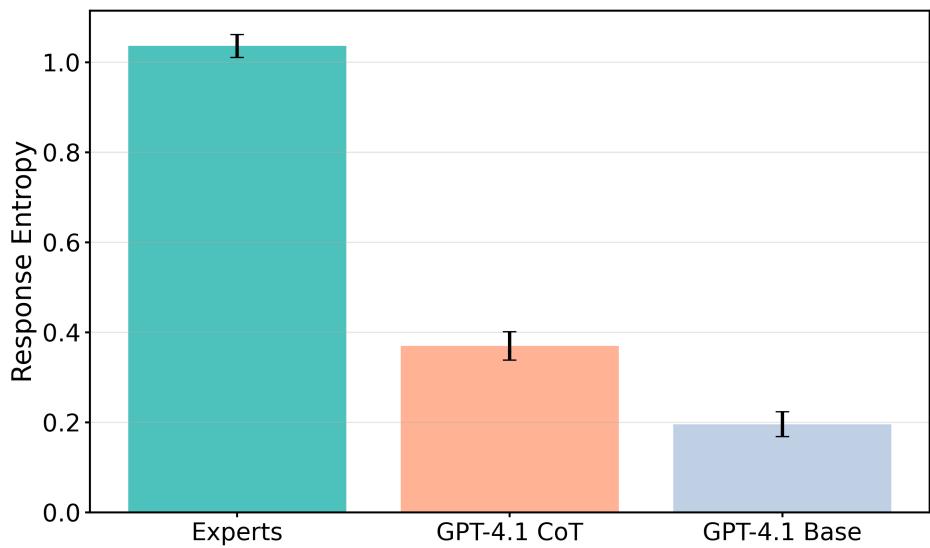


Figure 3.6.: GPT-4.1 rating entropy on SCT-Bench with vs. without clinical CoT prompting

response entropy compared to its baseline entropy ( $p < 0.05$ ), but without cluster formation (=greater concentration) of correct answers like on MedQA.

Table 3.1.: Mean entropy values depending on dataset and prompting mode.

Dataset	$\hat{H}_{\text{base}}$	$\hat{H}_{\text{CoT}}$	$\hat{H}_{\text{expert}}$
MedQA	0.034	0.089	–
SCT	0.196	0.370	1.036

### 3.3. Calibration Metrics

Calibration metrics were computed using both self-elicited confidence and sample consistency methods as described in section 2.3.

#### 3.3.1. Self-Elicited Confidence

Self-elicited confidence turned out to be ineffective as a calibration signal. GPT-4.1 stated uncertainty scores only in a very limited range, which made it impossible to plot complete calibration curves and interpret them accordingly. Figure 3.7 shows the baseline and CoT calibration curves on MedQA.

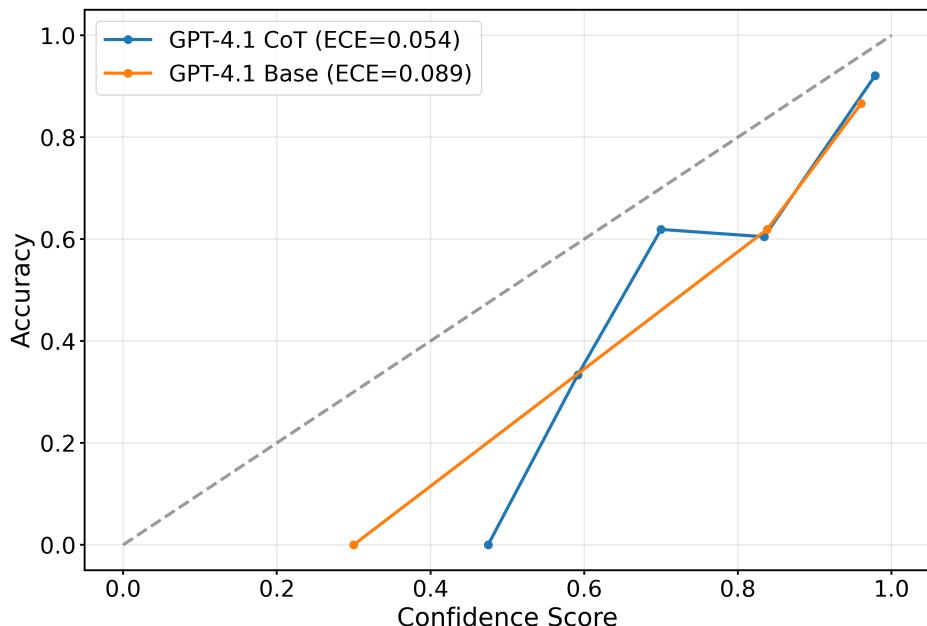


Figure 3.7.: MedQA-USMLE calibration curve for one-step self-elicited confidence with and without CoT prompting

In both cases, the model is overconfident in its responses within the provided data range. In fact, GPT-4.1 is so overconfident that it rarely reported confidence below 60%. The clinical

CoT prompt had minimal impact on confidence elicitation but reduced overconfidence within this narrow range, decreasing the ECE parameter by nearly half.

Systematic overconfidence is also evident in responses on SCT-Bench, where the range of self-elicited uncertainty scores is very limited (Figure 3.8). Clinical CoT had almost no impact on confidence elicitation quality, and since GPT-4.1’s accuracy on SCT (agreement with expert consensus) worsened with CoT, calibration also deteriorated.

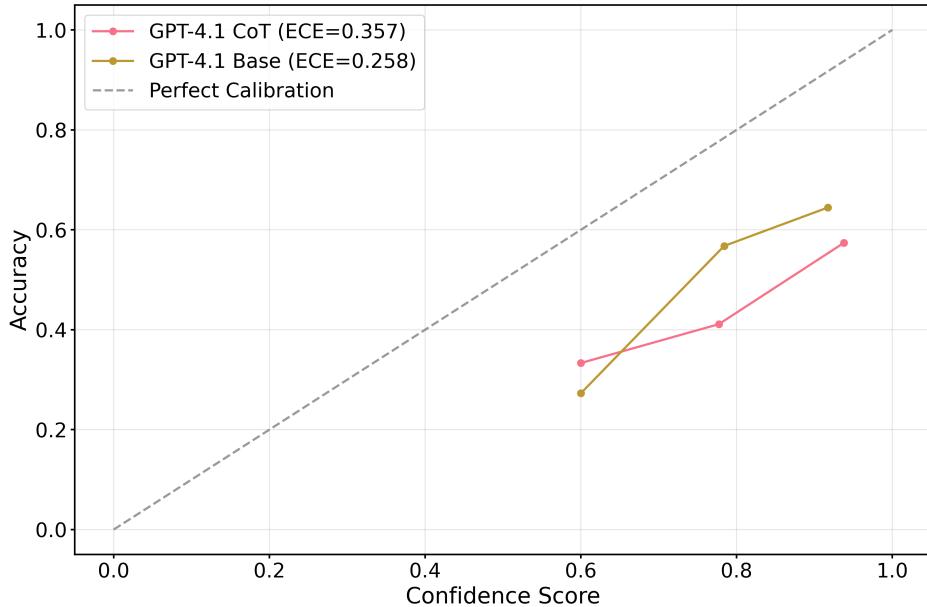


Figure 3.8.: SCT-Bench calibration curve for one-step self-elicited confidence with and without CoT prompting

### 3.3.2. Sample Consistency

Sample consistency analysis offers a broader range of uncertainty scores than GPT-4.1’s inadequately overconfident self-reported scores (subsection 3.3.1). Figure 3.9 shows the calibration data for on MedQA using sample-based uncertainty.

Calibration for MedQA improves considerably when using sample consistency as an uncertainty proxy. While clinical reasoning prompting had only a marginal effect on the ECE parameter, visual inspection of the calibration curves in Figure 3.9 reveals that GPT-4.1 is notably better calibrated when applying clinical CoT than without it.

This effect was not reproducible for the SCT dataset, where calibration is substantially worse overall. ECE values as well as the respective curves show generally poor calibration and unlike with MedQA, clinical reasoning prompting degraded it instead of improving it. An overview of all computed ECE values can be found in Table 3.2.

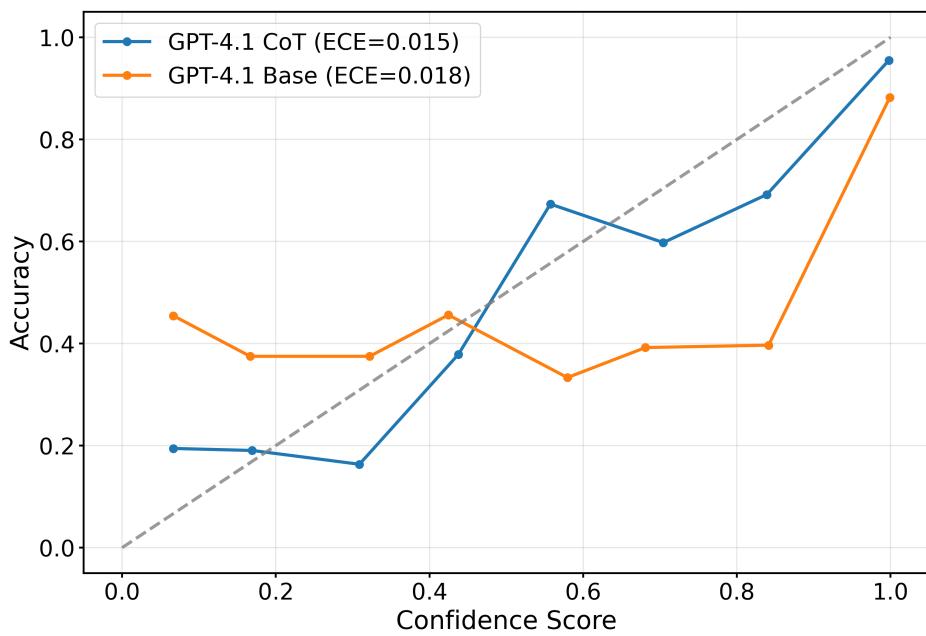


Figure 3.9.: MedQA-USMLE calibration curve for sample-based uncertainty estimation with and without clinical CoT prompting

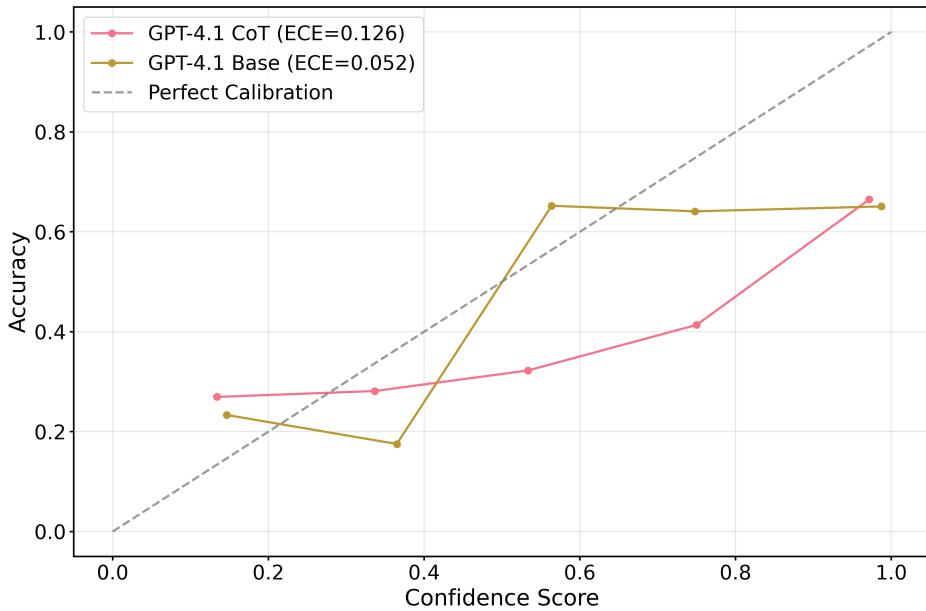


Figure 3.10.: SCT-Bench calibration curve for sample-based uncertainty estimation with and without clinical CoT prompting

### 3.4. ROC

ROC curves were computed for both self-elicted confidence and sample-based uncertainty scores. The clinical CoT approach consistently increased AUROC (subsection 2.3.1). Figure 3.11 shows both ROC curves on MedQA (baseline and CoT) with self-elicted confidence as uncertainty signal. AUROC improved by 14.7 percentage points and, as evident from the plot, discrimination between accurate and inaccurate predictions was enhanced particularly in the intermediate confidence range. This effect is also apparent when using sample consistency scores as uncertainty signal, as illustrated in Figure 3.12, where AUROC rose by 17.5 percentage points.

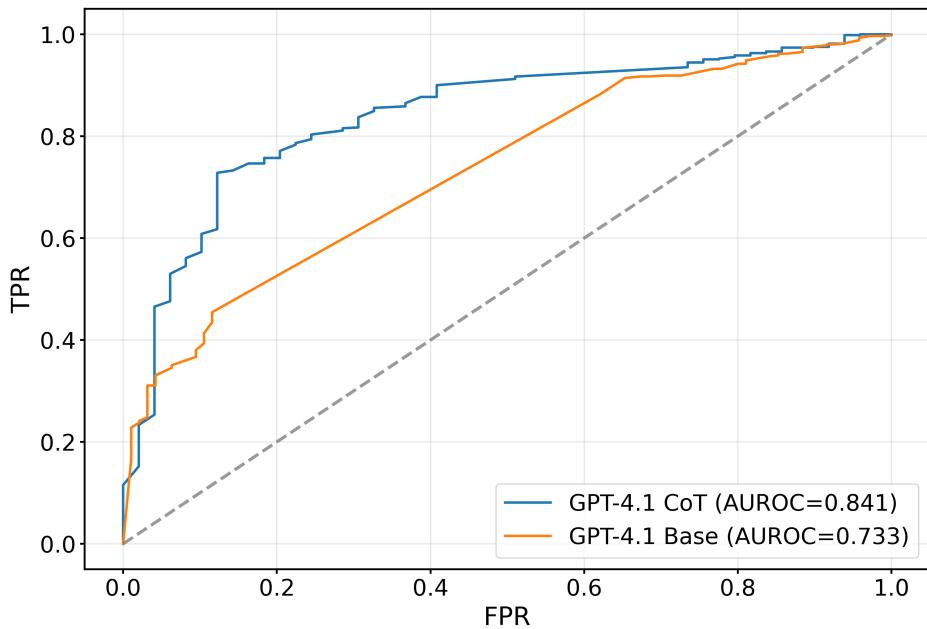


Figure 3.11.: ROC curves for one-step self-elicted confidence on MedQA-USMLE with and without clinical CoT prompting

For the SCT-Bench, clinical reasoning instruction prompting had the same effect as for MedQA (Figure 3.13 and Figure 3.14). It consistently improved the AUROC parameter, boosting the model’s discrimination capability. This impact is less pronounced when using self-elicted confidence as an uncertainty proxy, where AUROC gained only 10.7 percentage points, while it increased by 23.3 percentage points with sample-based uncertainty assessment.

Generally, AUROC is greater using self-reported confidence on both datasets, while clinical CoT had a larger relative impact on increasing AUROC using sample-based uncertainty assessment. The latter is particularly true for the SCT dataset, where the model at baseline performs barely better than random guessing. Overall, both calibration and ROC performance are substantially better for MedQA than for SCT (Table 3.2).

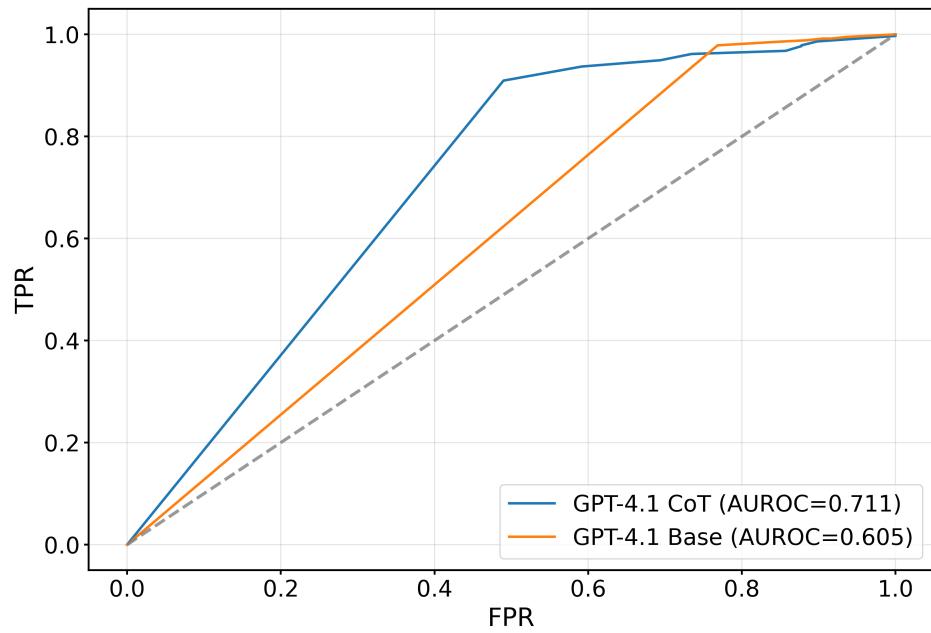


Figure 3.12.: ROC curves for sample-based confidence estimation on MedQA-USMLE with and without clinical CoT prompting

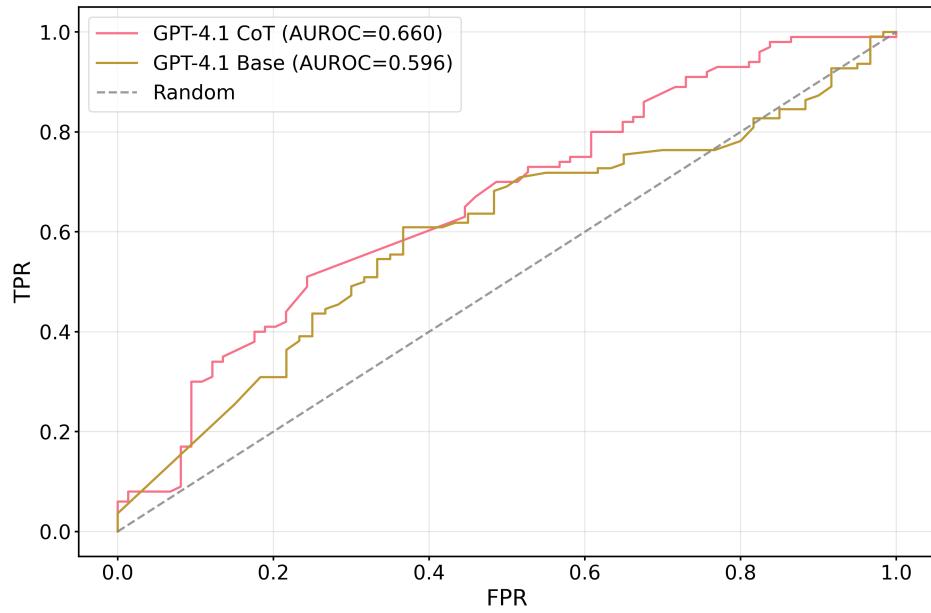


Figure 3.13.: ROC curves for one-step self-elicited confidence on SCT-Bench with and without clinical CoT prompting

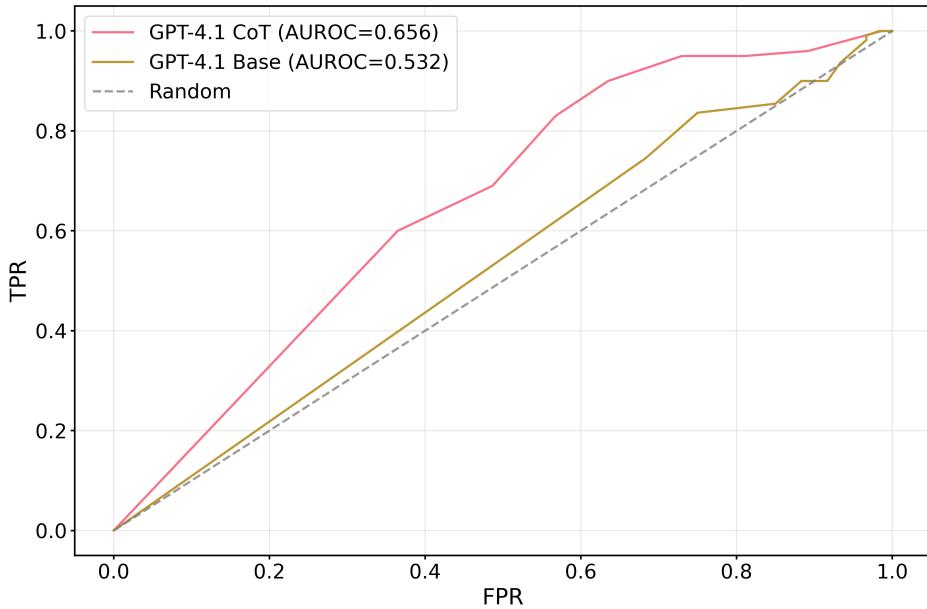


Figure 3.14.: ROC curves for sample-based confidence estimation on SCT-Bench with and without clinical CoT prompting

Table 3.2.: AUROC and ECE values on MedQA-USMLE and SCT-Bench depending on uncertainty estimation method and prompting mode

Benchmark	Uncertainty Proxy	Prompting	AUROC	ECE
MedQA	Self-elicted	Baseline	0.733	0.089
		CoT	0.841	0.054
	Sample-based	Baseline	0.605	0.018
		CoT	0.711	0.015
SCT-Bench	Self-elicted	Baseline	0.596	0.258
		CoT	0.660	0.357
	Sample-based	Baseline	0.532	0.052
		CoT	0.656	0.126

### 3.5. SCT Rating Distribution

The SCT score (section 2.4) measures how well a test taker, GPT-4.1 in this case, agrees with expert consensus. A perfect SCT score of 1.0 would mean the model always chooses the rating that the majority of experts agreed on, achieving 100% alignment with consensus. However, assessing a medical problem slightly differently than most experienced clinicians does not necessarily imply unreasonable judgment. The SCT format inherently involves uncertainty and ratings deviating from consensus do not necessarily indicate clinical incompetence. After all, not all experts agree on the same rating and diverging from consensus is still compatible with "thinking like a doctor."

It is interesting to look at the distribution of "rating distances" from the consensus rating  $r_0$ . It shows how frequently  $r_0$  or other ratings  $r'$  that do not align with  $r_0$  are chosen. Let's define a rating disagreement parameter  $\Delta r$  in the following manner:

$$\Delta r = r' - r_0 \quad (3.1)$$

In Equation 3.1  $r'$  is any chosen rating by experts or the LLM and  $r_0$  is, as already mentioned, the consensus expert rating. This means for example that when the model selects the rating +1 while most experts chose -1, the disagreement  $\Delta r$  is +2.

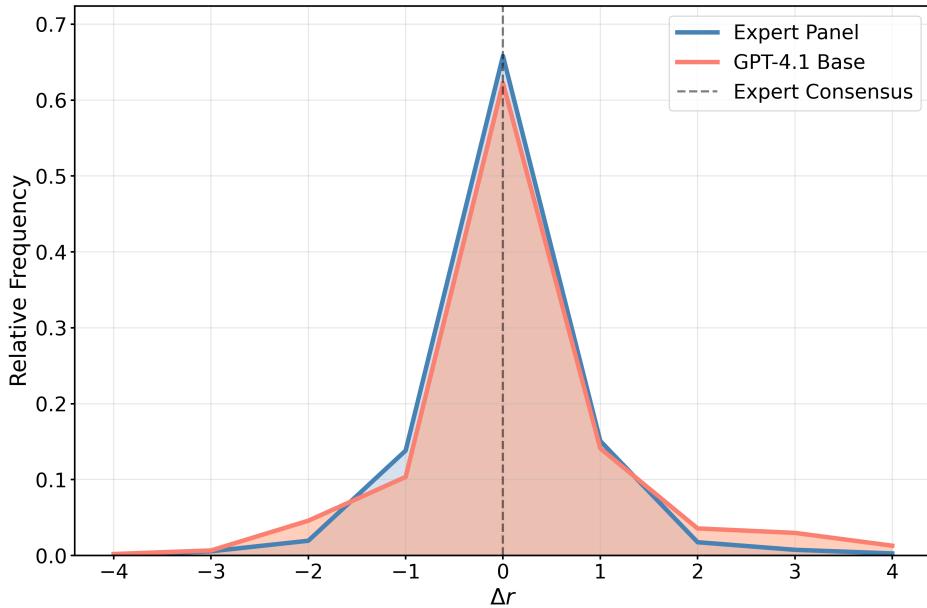


Figure 3.15.: Rating disagreement distributions around the centered expert consensus rating for experts (blue) and GPT-4.1 at baseline (red)

Figure 3.15 shows the distributions of rating disagreement among experts themselves and between baseline GPT-4.1 and experts. With a Jensen-Shannon distance of 0.109, the distributions are quite similar. This alignment can be interpreted as the model "rating like doctors" despite the imperfect SCT score. The model has a slightly greater likelihood of strong disagreement with consensus, giving ratings with  $|\Delta r| > 1$ , while physicians rarely disagreed

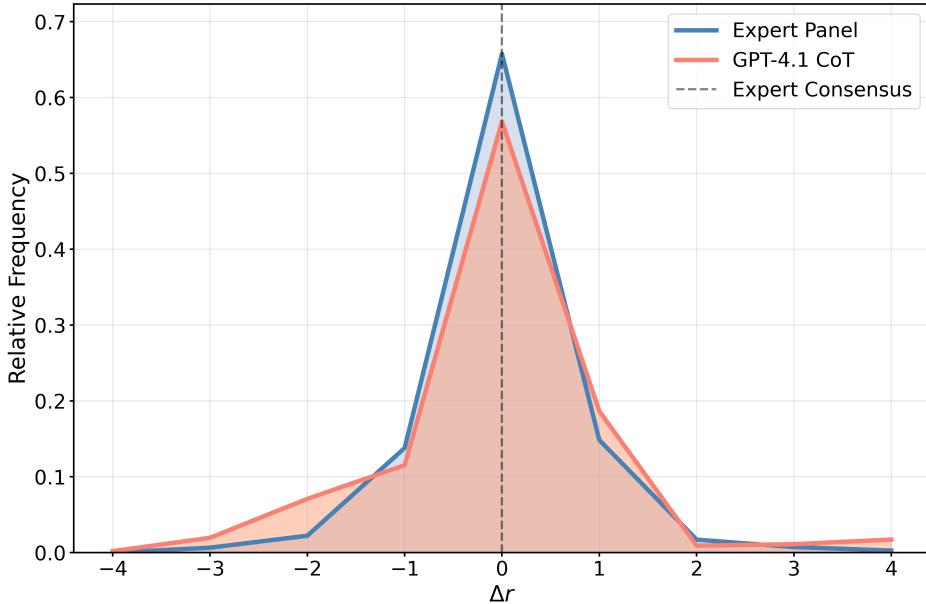


Figure 3.16.: Rating disagreement distributions around the centered expert consensus rating for experts (blue) and GPT-4.1 with CoT (red)

by more than  $|\pm 1|$ . The effect of clinical CoT prompting on this distribution is illustrated in Figure 3.16. Visual inspection reveals clear impact: the distributions align somewhat less compared to baseline prompting. Most notably, agreement with expert consensus ratings decreased as the model favored extreme ratings with  $|\Delta r| > 1$ . Nevertheless, with a Jensen-Shannon distance of 0.127, the distributions remain moderately similar.

### 3.6. Clinical Reasoning Strategies

For each dataset, the model was instructed to apply clinical reasoning strategies as presented in Table 1.1. Along with CoT prompting, a comprehensive list of clinical reasoning pathways with brief instructions was provided. The goal was to further stimulate GPT-4.1 to emulate “clinical thinking,” explicitly enriching its reasoning process with these tactics and to assess whether this would measurably affect benchmark scores. Furthermore, this raised the question of whether the model would effectively implement these strategies in its chain-of-thought. Is GPT-4.1 able to identify specific combinations of clinical data and settings in order to apply appropriate reasoning approaches?

A priori, it cannot be assumed that the model’s self-reported clinical reasoning strategies are actually aligned with the output CoT steps. Given the nature of LLMs, there are obviously no real cognitive processes taking place when they “reason” about something. Nonetheless, LLMs can mirror and mimic human thought processes. Ideally, the CoT would genuinely mimic the reasoning pathways that the model reports using for approaching a clinical problem. The only way to evaluate this is by reading GPT-4.1’s outputs and manually assessing whether the illustrated reasoning paths truly depict the mentioned strategies and emulate clinical

thinking processes.

Generally, the self-reported strategies are quite consistent with the output CoT steps and presented rationale. In the following sections, several CoT question-answer pairs illustrate how the model systematically applied the clinical thinking schemes it was instructed to use. Example 3.1 - Example 3.5 show how GPT-4.1 applies concrete clinical reasoning pathways and correctly identifies them as such (most relevant text passages are highlighted). They lend validity to the model's self-reported strategies and demonstrate its ability to follow reasoning instructions for clinical problem solving and to self-reflect on applied strategies. Example 3.1 shows a good example of causal reasoning, combining pathomechanisms with anatomical knowledge that leads to the correct assessment of the patient's situation.

### Example 3.1 - Causal Reasoning

#### MedQA:

A 68-year-old male comes to the physician for evaluation of right flank pain. He has a history of diabetes and peripheral artery disease. His blood pressure is 160/90 mm Hg. Physical examination shows abdominal tenderness and right flank tenderness. An ultrasound shows **dilation of the right ureter and renal pelvis**. Which of the following is the most likely underlying cause of this patient's condition?

- A. Renal artery stenosis
- B. Benign prostatic hyperplasia
- C. Diabetic nephropathy
- D. Common iliac artery aneurysm
- E. Urethral stricture

#### GPT-4.1:

Step-by-step clinical reasoning: 1. **Key clinical findings**: - 68-year-old male - Right flank pain - History: diabetes, peripheral artery disease - Hypertension (BP 160/90 mm Hg) - Abdominal and right flank tenderness - Ultrasound: dilation of right ureter and renal pelvis 2. **Interpretation of imaging**: - **Dilation of the right ureter and renal pelvis indicates hydronephrosis, which is most commonly due to obstruction of urine flow downstream of the kidney.** 3. **Evaluate possible causes (Hypothetico-Deductive Reasoning)**: - **A. Renal artery stenosis**: Would cause hypertension and possibly renal dysfunction, but does NOT cause hydronephrosis; it does not block urine flow. - **B. Benign prostatic hyperplasia (BPH)**: Very common in older men, can cause lower urinary tract obstruction, typically presents with bilateral hydronephrosis because BPH affects the prostatic urethra (distal to both ureters), unless there is a unilateral anatomical anomaly. - **C. Diabetic nephropathy**: Causes proteinuria and chronic kidney disease, but not hydronephrosis. - **D. Common iliac**

artery aneurysm\*\*: Can compress the ipsilateral (right) ureter as it crosses over the iliac vessels, leading to unilateral hydronephrosis (dilation of right ureter and renal pelvis). -

\*\*E. Urethral stricture\*\*: Like BPH, typically causes bilateral hydronephrosis unless the stricture is off-center (rare), and usually presents with lower urinary tract symptoms (difficulty urinating, weak stream). 4. \*\*Risk factors and context\*\*: - The patient has vascular disease (peripheral artery disease), which increases the risk of aneurysms (including iliac artery aneurysms). - He has unilateral hydronephrosis, which suggests a focal, unilateral obstruction. - His diabetes increases risk for atherosclerosis and vascular complications. 5. \*\*Most likely underlying cause\*\*: - The only answer that explains \*unilateral\* hydronephrosis in the context of a man with vascular disease is \*\*D. Common iliac artery aneurysm\*\* compressing the right ureter. 6. \*\*Rule out red flags\*\*: - No evidence of infection or acute surgical abdomen, so focus remains on obstructive uropathy due to external compression. answer: D uncertainty: 95 strategy: Hypothetico-Deductive Reasoning, Causal Reasoning, Probabilistic Reasoning justification: Hypothetico-deductive reasoning allowed systematic evaluation of each possible answer in light of the clinical findings. Causal reasoning was used to connect an iliac artery aneurysm to extrinsic ureteral compression and hydronephrosis. Probabilistic reasoning factored in the patient's age, vascular history, and the frequency of unilateral versus bilateral causes of hydronephrosis.

Example 3.1.: Clinical CoT example for applied causal reasoning. The highlighted text passages show reasoning relevant to the strategy.

After briefly summarising and integrating all given clinical data, GPT-4.1 proceeds to interpret the ultrasound finding in deductive manner to correctly determine that the patient has hydronephrosis. It continues to apply mainly 3 reasoning strategies which lead to the final conclusion about this case: (i) hypothetico-deductive, (ii) causal/ anatomical and (iii) probabilistic reasoning.

The rationale confirms these are indeed the key strategies used and the reasoning pathways GPT-4.1 reported. Hypothetico-deductive reasoning leads to considering all answer options as hypotheses and evaluating each one of them. Causal reasoning applies anatomical knowledge and links it to a possible pathomechanism (the dilated common iliac artery compressing the right ureter). Finally, in addition to these conclusions, the model also considers the patient's context and risk factors for estimating which hypothesis is most likely, essentially applying probabilistic thinking. It also demonstrates 'red flag/rule-out reasoning' by considering if there is evidence pointing to any other serious and time-sensitive cause, though this is not stated among the applied strategies. Perhaps the absence of acutely threatening red flags and the lack of clinical consequences led to it not being acknowledged as an applied reasoning strategy, even though the mindset was present.

It is reasonable to assume that the identification and self-reporting of these strategies are suboptimal and susceptible to errors. Some reasoning schemes may not be explicitly indicated, even when certain CoT steps correspond to them. Conversely, others may be mismatched or mentioned among the strategies despite not being actually implemented. Although these limitations may exist to an unknown extent (reading and manually evaluating the required volume of material would have gone well beyond the scope of this thesis) it is valid to state that the model was able to emulate clinical reasoning processes and appeared to “know” that it did so. Keeping potential limitations in mind, the data provided by this form of self-reflection remain insightful.

Example 3.2 illustrates how GPT-4.1 adequately points out the heuristic reasoning approach in a typical case for this kind of intuitive and fast pattern recognition, which plays an important role in clinical medicine and even more so in experienced physicians.

### Example 3.2 - Heuristic Pattern Recognition

#### SCT-STEM:

A five-year child presents to his primary care provider with a rash on his upper and lower extremities for 2 weeks.

If you were thinking of a diagnosis of: Impetigo

And then you find: Honey colored crusting on erythematous plaques in the popliteal creases.

#### GPT-4.1:

First, let's break down the clinical scenario and hypothesis: - A five-year-old child presents with a rash on the upper and lower extremities, duration two weeks. - The diagnostic hypothesis is impetigo. - New information: Honey-colored crusting on erythematous plaques in the popliteal creases. Step-by-step reasoning: 1. \*\*Integration of Clinical Information:\*\* - Age: Five years—impetigo is more common in young children. - Distribution: Rash on upper and lower extremities, especially in the popliteal (behind the knee) creases. - Duration: Two weeks—fits with impetigo, which can persist if untreated. - Morphology: Honey-colored crusting on erythematous plaques—this is the classic description for non-bullous impetigo. 2. \*\*Deductive Reasoning:\*\* - Medical fact: Impetigo typically presents as erythematous macules that rapidly evolve into vesicles or pustules and then rupture, leaving honey-colored crusted plaques. Most commonly occurs in children and often involves exposed areas or areas of minor trauma. - The finding of honey-colored crusting directly matches the textbook description for impetigo. 3. \*\*Heuristic / Pattern Recognition:\*\* - "Honey-colored crust" is essentially pathognomonic for non-bullous impetigo. Most clinicians

would recognize this instantly. 4. \*\*Hypothetico-Deductive Reasoning:\*\* - Differential diagnoses for a two-week rash in a child include eczema (atopic dermatitis), contact dermatitis, psoriasis, fungal infection, and impetigo. - The honey-colored crusting is not typical of eczema (which is usually weepy or scaly), nor of psoriasis (silvery scale), nor of tinea (annular with central clearing). - The description is most consistent with impetigo. 5. \*\*Probabilistic Reasoning:\*\* - Impetigo is common in this age group and setting; combined with the classic morphological finding, the probability becomes very high. 6. \*\*Guideline-Based Reasoning:\*\* - CDC, AAP, and dermatology guidelines universally cite "honey-colored crusting" as a key diagnostic feature of impetigo. \*\*Conclusion:\*\* The new information (honey-colored crusting on erythematous plaques) strongly increases the likelihood of impetigo as the diagnosis. rating: +2 uncertainty: 95 strategy: Deductive Reasoning, Heuristic/Pattern Recognition, Hypothetico-Deductive Reasoning, Probabilistic Reasoning, Guideline-Based Reasoning justification: I used deductive reasoning and pattern recognition to directly match the classic description of impetigo, considered and ruled out alternate diagnoses, factored in prevalence and guidelines, and found the new information to be strongly supportive of the hypothesis. The only reason for not assigning 100% certainty is that rare mimickers exist, but the finding is nearly pathognomonic.

Example 3.2.: Clinical CoT example for heuristic pattern recognition. The highlighted text passages show reasoning relevant to the strategy.

The diagnosis of impetigo contagiosa is typically determined clinically with no need for further microbiologic assays. The clinical features are highly characteristic and easily recognized by doctors in a manner where diagnosis is established almost instantly. This is exactly what heuristics and fast pattern matching consist of in medicine. In (Example 3.2), GPT-4.1 emphasizes the representative diagnostic feature of the disease, describing the "honey-colored crusting on erythematous plaques" as "essentially pathognomonic." The model clearly recognizes this description as a typical scenario in which heuristics play a significant role and applies this approach in combination with other reasonable reasoning strategies (hypothetico-deductive after formulating possible differential diagnoses, probabilistic when taking factors like clinical presentation, age and context into account and finally guideline-based for revisiting guideline instructions/descriptions).

### Example 3.3 - Probabilistic Reasoning

#### SCT-STEM:

A five-year child presents to his primary care provider with a rash on his upper and lower extremities for 2 weeks.

If you were thinking of a diagnosis of: Atopic Dermatitis

And then you find: A past medical history of allergic rhinitis and asthma

**GPT-4.1:** Step-by-step clinical reasoning: 1. **Clinical Information Integration:** - The child is 5 years old and has had a rash on both upper and lower extremities for 2 weeks. - Atopic dermatitis (AD) is a common chronic, pruritic, inflammatory skin disease in children, often appearing as eczematous lesions (usually on the flexural surfaces) and typically associated with chronicity and pruritus. - The distribution and chronicity (2 weeks) are compatible, though not specific, for AD. 2. **Hypothesis:** - The working hypothesis is atopic dermatitis. 3. **New Information:** - The child has a past medical history of allergic rhinitis and asthma. 4. **Application of Clinical Reasoning Strategies:** - **Deductive Reasoning:** Atopic dermatitis is frequently associated with other atopic conditions (the "atopic triad": atopic dermatitis, asthma, allergic rhinitis). **The presence of allergic rhinitis and asthma makes AD more likely.** - **Probabilistic Reasoning:** Children with one atopic condition have an increased prevalence of others due to shared genetic and environmental risk factors. - **Pattern Recognition:** The combination of skin findings in a young child with a history of other atopic disorders fits the classic presentation of AD. 5. **Evidence-Based Medicine:** Guidelines (e.g., American Academy of Dermatology) highlight that the diagnosis of AD is supported by a history of atopy (personal or family), including asthma and allergic rhinitis. 6. **Differential Diagnosis Consideration:** - While other rashes are possible, the combination of atopic history with chronic eczematous rash is most suggestive of AD. rating: +2 uncertainty: 95 strategy: Deductive reasoning, probabilistic reasoning, pattern recognition, guideline-based reasoning justification: **The presence of other atopic conditions (asthma, allergic rhinitis) strongly supports the diagnosis of atopic dermatitis per clinical guidelines and pathophysiologic understanding. This information makes the hypothesis of AD much more likely.**

Example 3.3.: Clinical CoT example for probabilistic reasoning. The highlighted text passages show reasoning relevant to the strategy.

Example 3.3 illustrates a combination of pattern recognition ("atopic triad"), probabilistic, evidence-based guideline and general deductive approaches. The clinical features are only rudimentarily described and the focus is on a characteristic constellation of risk factors that together increase the likelihood of the hypothesis. It is a good example of integrating risk factors in probabilistic thinking. Stating that "atopic dermatitis is frequently associated with other atopic conditions" is arguably very compatible with rule-based/categorical thinking too (Table 1.1), yet this strategy is not reported here.

### Example 3.4 - Red Flags & Rule-Out Reasoning

#### SCT-STEM:

A 64-year-old male presents to ED with sharp central chest pain for the past 45 minutes. The pain started suddenly while he was at rest. The pain also radiates to his back. He has also been feeling nauseous. He has long-term poorly controlled hypertension. An ECG demonstrates ST elevation in leads II, III and aVF.

If you were thinking of: CT angiogram

And then you find: An absent right brachial and radial pulse. Left radial pulse present.

#### GPT-4.1:

Let's break down and integrate all clinical information, focusing on the hypothesis "CT angiogram" in the context of this new finding:

- \*\*Step 1: Initial clinical scenario and hypothesis evaluation\*\* - 64-year-old man with sudden-onset, severe, sharp central chest pain radiating to the back. - Symptoms began at rest, with associated nausea.
- Long-standing, poorly controlled hypertension (a significant risk factor for aortic dissection).
- ECG: ST elevation in II, III, and aVF (suggestive of inferior STEMI).
- \*\*Step 2: Initial differential diagnosis\*\* - Inferior STEMI (supported by ECG). - Aortic dissection (supported by pain characteristics, hypertension, radiation to the back). - Other causes of chest pain (less likely given context).
- \*\*Step 3: Impact of new finding (absent right brachial and radial pulse; left radial pulse present)\*\* - Pulse deficit is a classic, high-yield physical finding in aortic dissection.
- Asymmetric pulses, especially in the context of chest pain radiating to the back and hypertension, are highly suggestive of aortic dissection involving the aortic arch and affecting the right brachiocephalic trunk.
- This finding makes aortic dissection much more likely, even in the presence of STEMI findings on ECG (as aortic dissection can involve the coronary arteries, especially the right coronary, causing STEMI patterns).
- \*\*Step 4: How does this affect the hypothesis (CT angiogram)?\*\* - The gold standard for diagnosing aortic dissection is CT angiography.
- The new finding of a pulse deficit strongly increases the suspicion for aortic dissection, making CT angiography urgently indicated.
- Early diagnosis is crucial because management (e.g., anticoagulation or thrombolysis for MI) could be fatal in dissection.
- \*\*Step 5: Synthesis, guideline, and risk assessment\*\* - ACC/AHA guidelines recommend immediate imaging (CT angiogram, TEE, or MRI) in suspected aortic dissection when clinical features (pain, pulse deficit, hypertension) are present.
- The presence of "red flag" findings (like pulse deficit) in chest pain should always prompt consideration and urgent rule-out of aortic dissection.

rating: +2 uncertainty: 98 strategy: Red Flag/Rule-Out Reasoning, Guideline-Based Reasoning, Deductive Reasoning, Causal Reasoning justification: Red

Flag/Rule-Out Reasoning was prioritized because aortic dissection is a life-threatening emergency where missing the diagnosis could be fatal. Guideline-Based Reasoning was used because current recommendations call for urgent advanced imaging in this scenario. Deductive Reasoning applied general knowledge about the classic features and investigation pathway for aortic dissection. Causal Reasoning was used to link the specific pathophysiology (aortic dissection causing pulse deficit) to the need for immediate CT angiogram. The combination of these strategies ensures the safest and most evidence-based approach.

Example 3.4.: Clinical CoT example for red flag and rule-out reasoning. The highlighted text passages show reasoning relevant to the strategy.

Example 3.4 shows a clinical setting in which the presence of typical "red flags" guides the whole approach and influences decision-making. The combination of pain characteristics and ECG findings suggests a myocardial infarction. However, the additional information about asymmetry of radial pulses represents an extra red flag that calls for immediate action to rule out or confirm an aortic dissection. Via causal reasoning, it is also acknowledged that an aortic dissection can involve the coronary arteries and therefore cause STEMI-like ECG patterns. Both diagnoses are life-threatening and time-sensitive, with an aortic dissection requiring urgent surgery and a STEMI requiring immediate revascularization, making it crucial to identify which emergency is actually taking place. Missing either diagnosis would likely be fatal in this case and prioritizing a characteristic red flag is a reasonable approach.

### Example 3.5 - Guideline-Orientated

#### MedQA:

A 65-year-old man with a history of hypertension, obesity, and alcoholic cirrhosis is seen in clinic for follow-up. He feels well and currently drinks 5 glasses of wine each night. Medications include atenolol and lisinopril. On physical exam, temperature is 98.1 deg F (36.7 deg C), blood pressure is 151/82 mmHg, pulse is 71/min, and respirations are 14/min. He has spider angiomata on his chest; no asterixis, jaundice, ascites, or peripheral edema is noted. Screening ultrasound reveals a new liver nodule, and follow up CT demonstrates a 2 cm right hepatic lobe lesion with enhancement in the arterial phase. No hypodense attenuation is seen on the venous or delayed phase. What is the next step in management?

- A. Proceed with liver biopsy
- B. Refer for surgical resection
- C. Refer for radiofrequency ablation
- D. Treat with sorafenib

E. Observe and get follow-up imaging in 3 months

**GPT-4.1:**

Step-by-step clinical reasoning: 1. \*\*Identify the clinical context and risk factors:\*\* - 65-year-old man with alcoholic cirrhosis (well-established risk factor for hepatocellular carcinoma, HCC). - Ongoing alcohol use, hypertension, obesity (further increasing HCC risk). 2. \*\*Interpretation of clinical findings:\*\* - Asymptomatic; no encephalopathy, jaundice, ascites, or edema (indicates compensated cirrhosis; likely Child-Pugh A). - Physical exam: spider angioma, but no other decompensation. 3. \*\*Key diagnostic finding:\*\* - Liver nodule detected on screening ultrasound. - CT: 2 cm right hepatic lobe lesion with arterial phase enhancement and \*no\* washout (no hypodense attenuation) on venous/delayed phase. 4. \*\*Integration with diagnostic criteria for HCC:\*\* - In patients with cirrhosis, the diagnosis of HCC can be made \*\*non-invasively\*\* if a lesion > 1 cm shows classic imaging findings: arterial phase enhancement \*\*and\*\* washout on portal venous or delayed phase. - In this case, the lesion is 2 cm, but \*does not\* show washout—so it \*\*does not meet classic radiological criteria for HCC\*\*. 5. \*\*What is the next step?\*\* - According to guidelines (AASLD, EASL), if a new lesion > 1 cm in a cirrhotic liver is detected and does \*not\* meet classic imaging criteria for HCC on dynamic imaging (arterial enhancement + venous/delayed washout), \*\*the next step is to obtain a tissue diagnosis\*\* before proceeding with treatment. - Surgical resection, ablation, or systemic therapy (sorafenib) are only considered \*after\* a definitive diagnosis of HCC. - Observation is not appropriate given the risk and abnormal imaging. - Biopsy is indicated to establish diagnosis. 6. \*\*Rule out other options:\*\* - B (resection): Not without tissue diagnosis. - C (ablation): Not without tissue diagnosis. - D (sorafenib): Reserved for advanced/unresectable HCC, and not without tissue diagnosis. - E (observation): Not appropriate for a new nodule in cirrhosis with suspicious but non-diagnostic imaging. answer: A uncertainty: 95 strategy: Guideline-Based Reasoning, Deductive Reasoning, Red Flag / Rule-Out Reasoning justification: Guideline-based reasoning was selected to apply AASLD/EASL recommendations for evaluation of liver nodules in cirrhosis. Deductive reasoning was used to apply the principle that non-invasive diagnosis of HCC requires both arterial enhancement and washout, which is lacking here. Red-flag reasoning was used to avoid missing an early treatable malignancy.

Example 3.5.: Clinical CoT example for explicit guideline-orientated reasoning. The highlighted text passages show reasoning relevant to the strategy.

Example 3.5 is a case vignette where the evaluation of the next best step in management is particularly guideline-oriented. It is reasonable to look for specific guideline recommendations

in this kind of scenario, which GPT-4.1 does in order to apply them in its decision-making. In Example 3.5 explicit clinical CoT flipped the majority answer from C ("Refer for radiofrequency ablation") to the correct answer A ("Proceed with liver biopsy"). The model cited the AASLD and EASL guidelines, using them to argue for its choice. When reviewing all 15 responses to this question, it was interesting to see that even though these guidelines were referred to each time, the answer was not always the same. In very few responses, GPT-4.1 stuck to answer C despite citing the same source. With all available information, the patient's imaging from Example 3.5 would classify as LI-RADS 4 according to current AASLD guidelines [46]. This classification implies an HCC diagnosis with high likelihood (60-70%), yet does not confirm it, and a biopsy should be considered in the diagnostic scheme following AASLD guidelines. Although the guidelines are not cited correctly or only incompletely in several responses, implementing them explicitly in the CoT still modified the answer positively and made the model choose the correct option in most cases.

### 3.7. Clinical Reasoning Strategy Distribution

Looking into the data of self-reported reasoning strategies allows quantifying them and examining how they are distributed. Certain reasoning schemes were reported significantly more frequently than others.

Figure 3.17 shows a histogram of reasoning strategies that GPT-4.1 claims to have applied on MedQA. Three of them dominate: deductive, guideline-based and hypothetico-deductive reasoning. It is not surprising that these reasoning pathways are so omnipresent. Deductive reasoning is a core element of almost any kind of problem-solving and is always present when inferring specific conclusions from given premises. Hypothetico-deductive reasoning is not fundamentally different from this but has specific clinical hypotheses (e.g., a differential diagnosis) as its starting point. Generally, tackling medical problems commonly involves formulating a diverse range of hypotheses about their causes or other aspects. In the context of a medical question-answering exam like MedQA, the multiple-choice answer options already represent hypotheses the model iterates through. It is fairly obvious that this strategy plays an important role for these reasons. Guideline-oriented management is indeed pivotal in clinical real-life scenarios and was also referred to very frequently by GPT-4.1.

Two reasoning strategies that had almost no relevance were inductive and abductive reasoning. Considering abductive reasoning, one of the most obvious use cases in clinical medicine is as part of diagnostic reasoning when generating a hypothesis, e.g. a differential diagnosis that best explains a patient's symptoms. A test like the USMLE exam, however, already provides a set of different hypotheses (= answer options) and the LLM always systematically went through all of them applying a hypothetico-deductive approach. With the "supply" of available hypotheses, there is no need to generate them on its own and it would likely not be the most "intuitive" approach either, considering that one of the options has to be chosen.

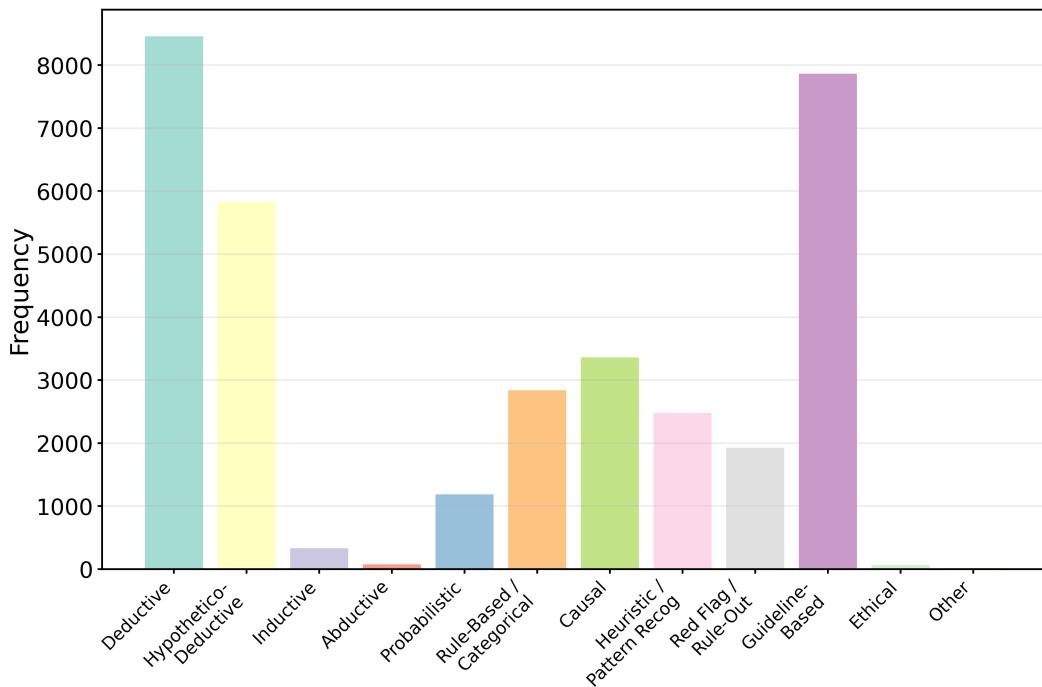


Figure 3.17.: Distribution of self-reported clinical reasoning strategies on MedQA-USMLE

Arguably these kinds of exams hardly test abductive reasoning and it is not unreasonable to assume that human candidates approach multiple-choice tests in a similar manner. Since inductive reasoning has more of its place in the context of clinical research and public health, it is not surprising that it had little relevance.

It is interesting to notice that in comparison to the distribution on SCT-Bench (Figure 3.18), rule-based and heuristic reasoning are relatively more represented on MedQA question answering than on SCT-Bench. This observation could be compatible with the nature of both test formats. MedQA questions tend to be brief yet quite elaborate case vignettes providing a lot of information and drawing a fairly complete clinical picture. It may be that fast pattern recognition and categorical thinking are generally more applicable on MedQA. Since SCT aims to assess clinical reasoning under uncertainty, it may also be that questions aiming at such strategies are underrepresented.

The distribution of self-reported reasoning strategies is shown in Figure 3.18 and reveals a major difference compared to the MedQA benchmark: Probabilistic reasoning was referred to significantly more frequently and is within the top three reasoning strategies. Given the inherent uncertainty of SCT questions and their nature of assessing likelihood shifts this is consistent with the benchmark format.

### 3.8. Rating Transitions

As shown in section 1.4, clinical CoT prompting had a significant effect on GPT-4.1's MedQA performance, improving the score by over 6 percentage points, while it had a negative impact

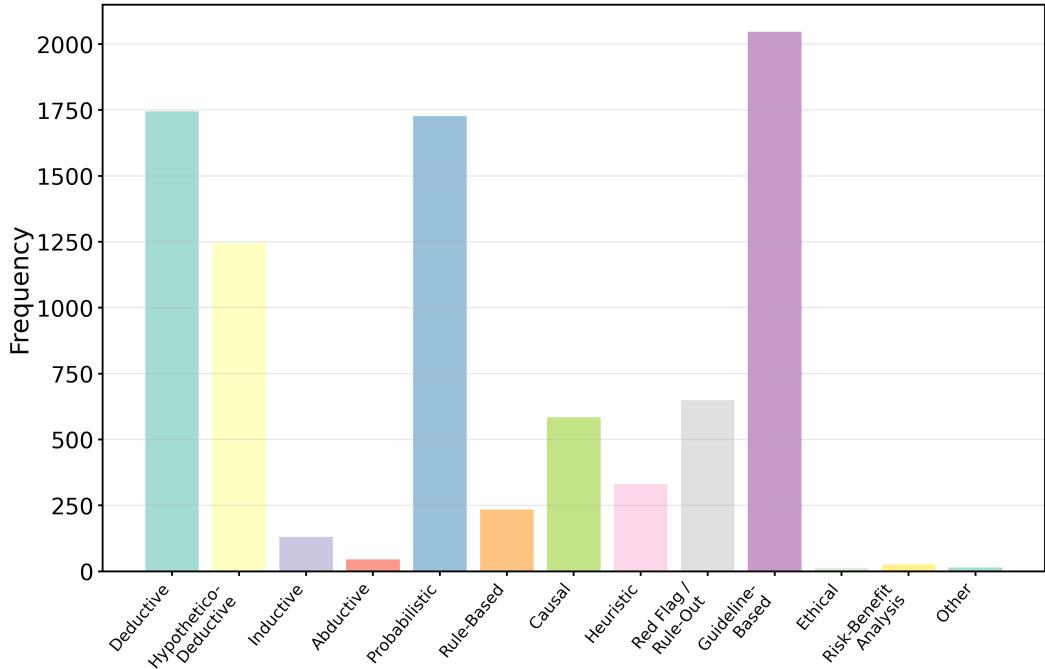


Figure 3.18.: Distribution of self-reported clinical reasoning strategies on SCT-Bench

on the SCT dataset. Manual review of a sample of SCT outputs provided qualitative insight into the effect of CoT prompting. The model seemed to be more prone to choose strong ratings of  $\pm 2$  after generating a clinical CoT sequence than without it. This appeared to happen especially more frequently in situations where it had chosen more moderate likelihood ratings of  $\pm 1$  with baseline prompting, which seemed to coincide more often with expert consensus. This suggests that GPT-4.1 adopted a general inclination of assessing the positive or negative impact of new information with greater likelihood. In other words, one could say it started seeing things with more certainty when there was the slightest evidence in its rationale.

Figure 3.19 shows how ratings are distributed among all experts and all GPT-4.1 answers at baseline. When compared to Figure 3.20, the difference in rating behavior by GPT-4.1 with CoT becomes evident. With CoT, strong ratings ( $+2, -2$ ) are more frequent, while moderate and neutral ratings ( $+1, 0, -1$ ) become rarer, although this is not entirely true for  $r = -1$ , which increases slightly due to unexpected flipping from  $r = +1$  to  $r = -1$  ratings (see below).

The histograms seem to be consistent with the empirical observation of clinical CoT shifting the model's ratings towards stronger ratings, but they cannot show how these were affected individually for a particular SCT item. Quantifying and moreover visualizing how likelihood ratings transitioned within the same test questions depending on rater (experts vs. GPT-4.1) and/or CoT can be done with Sankey diagrams. For better capturing certain trends in how rating is affected by this and also by clinical CoT prompting, the ratings were grouped into 3 categories: i) Strong, ii) Moderate and iii) Neutral.

Figure 3.21 gives insight into how GPT-4.1 rated SCT-Bench questions with baseline

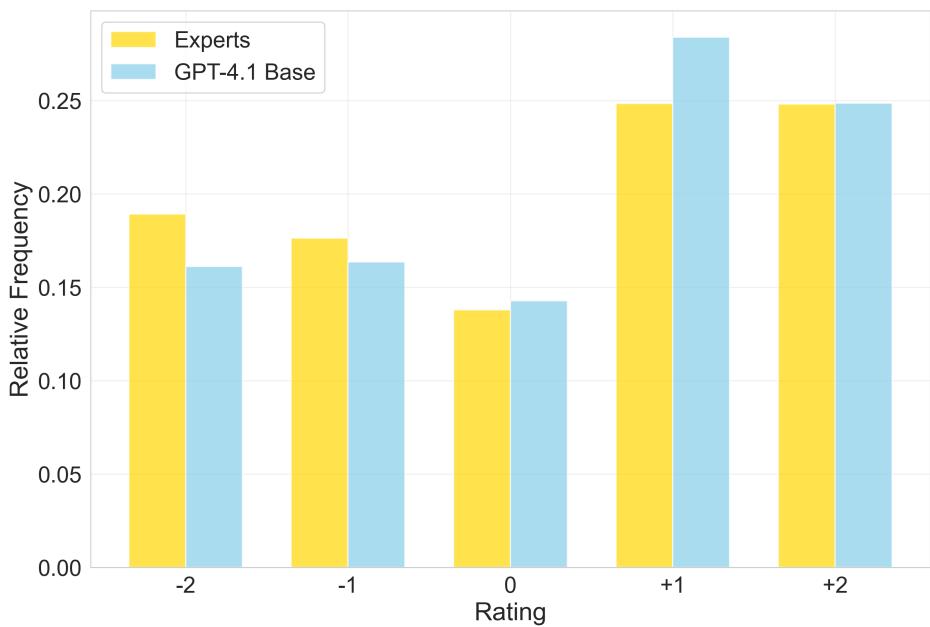


Figure 3.19.: Distribution of expert and GPT-4.1 baseline ratings.

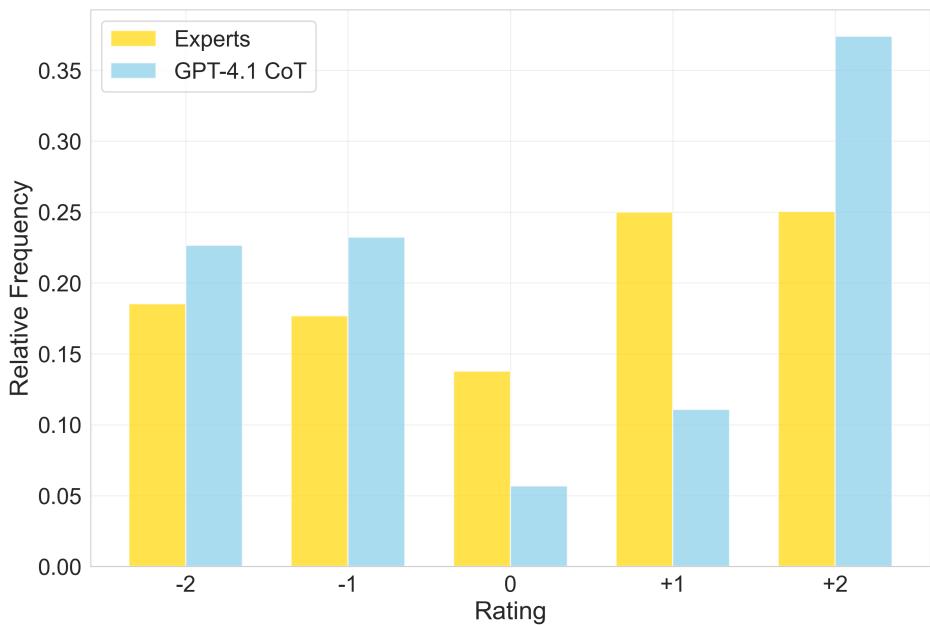


Figure 3.20.: Distribution of expert and GPT-4.1 CoT ratings.

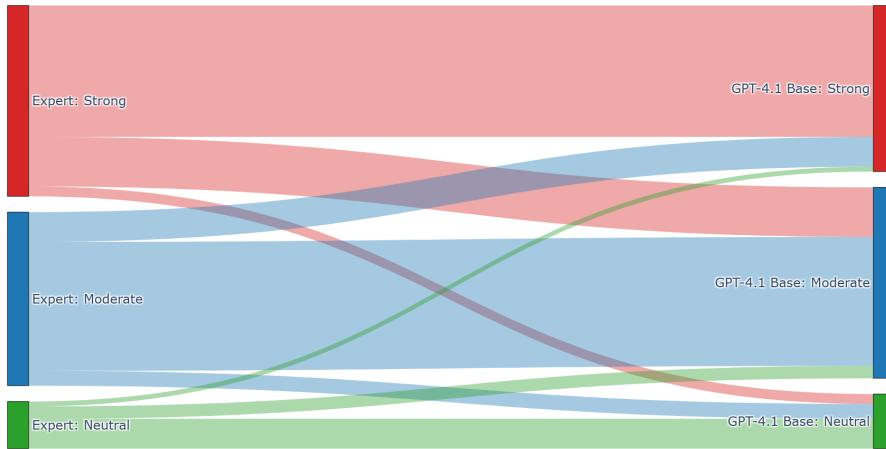


Figure 3.21.: Rating transitions between expert consensus and GPT-4.1 baseline ratings.

prompting compared to expert consensus ratings. Cases that were rated differently by the model than by the experts are shown to flow from one rating category to a different one, essentially making all transitions visible.

In most cases GPT-4.1 agrees with the expert consensus. The total distributions of rating levels are not too different, with GPT-4.1 at baseline rating slightly more moderate. However, there are flows that almost compensate each other and this is information which remains masked in the histograms: Partially, the model gives moderate ratings where experts assign strong ones and vice versa. The same situation exists with neutral and moderate ratings.

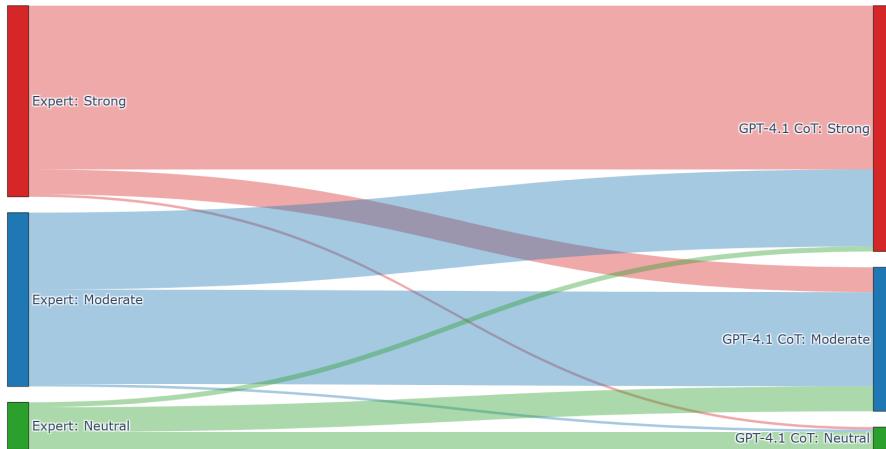


Figure 3.22.: Rating transitions between expert consensus and GPT-4.1 CoT ratings.

Figure 3.22 shows how ratings transitioned between expert consensus and GPT-4.1 CoT ratings. One can see how there is now a clear tendency of the model to assign strong ratings for questions where experts assigned moderate ratings and this effect is much greater than the opposite direction (strong → moderate). Finally, it also becomes evident that neutral ratings transition mostly to moderate ones. Generally, the ratings shift from "weaker" to "stronger" ones.

This effect becomes even more obvious when visualizing the impact of clinical reasoning on the model itself, hence the rating transitions between GPT-4.1 at baseline and GPT-4.1 with CoT prompting. In Figure 3.23 it is evident how the flows go "upstream" illustrating the effect clinical reasoning has on the model's rating behaviour:



In conclusion, clinical CoT prompting made the model overestimate neutral and moderate impacts of new information on the original hypothesis when compared to the expert consensus. Despite this evident trend, there is one disclaimer to be made: Transition analysis between all ratings (not grouped) using Sankey diagrams revealed that a non-negligible fraction of +1 ratings transitioned to -1. This CoT-induced phenomenon, where GPT-4.1 would opt for the opposite direction of its previous rating and therefore change its opinion completely, was not analysed in more detail. If it also contributes significantly to the CoT-triggered deviation from the expert consensus or not cannot be said at this point.

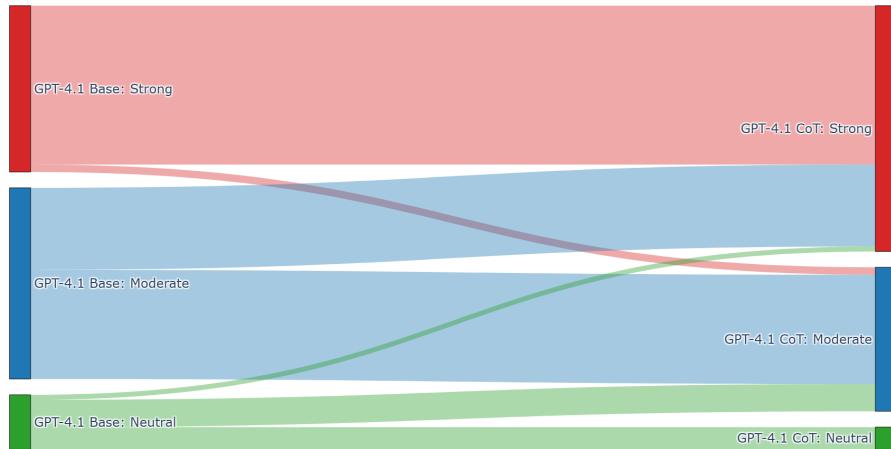


Figure 3.23.: Rating transitions between GPT-4.1 baseline and CoT ratings. CoT leads to systematic overestimation of neutral and moderate impacts of new information.

### 3.9. Expert versus Model Rating

The histograms (Figure 3.17, Figure 3.18) in section 3.7 show the overall distribution of self-reported reasoning strategies for both MedQA and SCT-Bench. An interesting observation can be made when examining the SCT rating distributions across different reasoning strategies. Figure 3.24 and Figure 3.25 show the rating preferences of GPT-4.1 and experts when a specific reasoning strategy was mentioned by GPT-4.1. The results in section 3.8 already showed the tendency of GPT-4.1 to overestimate likelihoods in comparison to the experts. This trend is also visible in Figure 3.24: Strong ratings clearly dominate across almost all reasoning strategies. This shows the general inclination for extreme ratings with clinical CoT prompting.

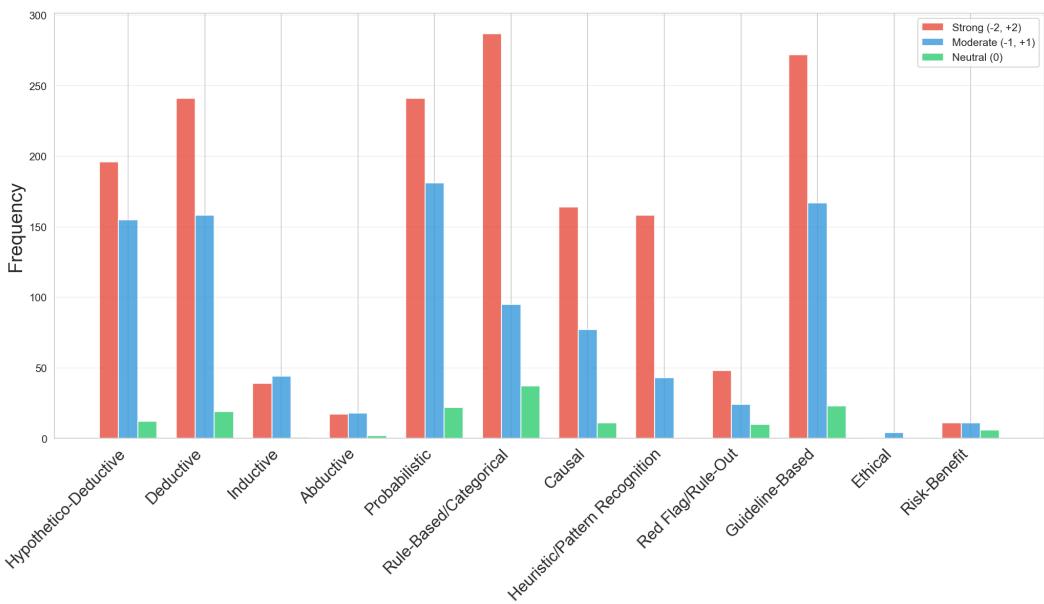


Figure 3.24.: Distribution of GPT-4.1 CoT rating levels across model-reported clinical reasoning strategies.

An obvious contrast appears when looking at the distributions of expert ratings (Figure 3.25). It is important to point out that there is no way to know what kind of cognitive approach expert clinicians had when completing the SCT-Bench questionnaire. The strategies referenced in Figure 3.25 are those reported back by GPT-4.1. The data shows that experts generally opted more frequently for moderate or neutral ratings rather than extreme ones, which is compatible with all observations so far. However, there are two reasoning strategies, where experts seem to have a similar distribution of rating preferences on respective questions, also favoring extreme ratings over moderate and neutral ones: **i) Rule-based/ categorical reasoning and ii) red-flag reasoning.**

Considering the key characteristics of both rule-based and red-flag reasoning, this can be seen as fairly consistent and compatible with "genuine" clinical reasoning. It aligns with the inherent deterministic nature of these reasoning strategies. Rule-based reasoning has a rather rigid algorithmic approach with clear "if-then" logic, leading to more definitive judgments

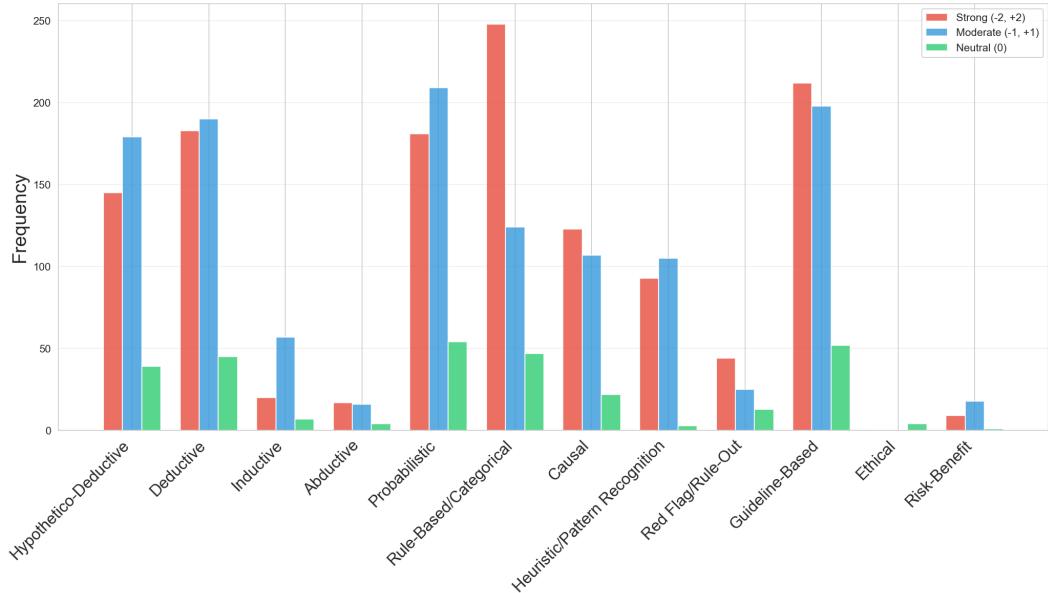


Figure 3.25.: Distribution of expert consensus rating levels across model-reported clinical reasoning strategies.

about the impact of new clinical information. Similarly, red-flag reasoning functions as effective detection of potentially life-threatening conditions, naturally resulting in stronger convictions about whether new information significantly increases or decreases the likelihood of appropriate next actions. This observed compatibility between model strategies and expert rating levels may be evidence that GPT-4.1 was effectively able to apply case-appropriate strategies and self-reflect on them accordingly. It could further suggest that the experts themselves, consciously or not, had similar reasoning approaches that align well with the model's self-reported logic. This kind of indirect evidence for alignment gives GPT-4.1's self-reflection more validity and also lends further credibility to its elicited clinical reasoning capabilities. Whenever guideline-based reasoning was reportedly involved, strong likelihood ratings by experts were only a slim majority over moderate and neutral ones. This circumstance could reflect the more nuanced approach of clinical guidelines, which often provide graded recommendations rather than absolute consensus-driven directives or deterministic algorithmic schemes.

# 4. Discussion

## 4.1. Benchmark Performance

Comparing the benchmark performances by GPT-4.1 on MedQA and SCT-Bench shows a clear picture of the substantially different impact clinical reasoning prompting had on both formats. By eliciting a clinical reasoning chain, the MedQA majority score increased significantly by 6.6 percentage points. In times where new LLMs already perform impressively well straight "out of the box" with rich knowledge across domains including medicine, it is remarkable how significant induced clinical thinking still is for enhancing accuracy on medical licensing exams. The question remains whether we might encounter some kind of "ceiling effect" where domain-specific CoT prompting does not significantly enhance performance by newer-generation state-of-the-art models. Not only the majority score but also the mean MedQA score improved significantly by about 5 percentage points.

These results stand in great contrast to the effect of clinical CoT on SCT-Bench. The majority SCT score declined by exactly 3 percentage points with clinical reasoning instruction prompting, while the mean SCT score decreased by 2.1 percentage points. Performance on SCT-Bench degraded significantly when clinical reasoning was elicited in GPT-4.1. This phenomenon was observed by McCoy et al. when they noticed that reasoning-inherent models underperformed on the SCT dataset compared to other medical benchmarks [24]. OpenAI's o1-preview model e.g. performed worse than GPT-4o, even though it usually outperformed that model on any other evaluation dataset, including medical ones [24]. The o1 model was introduced as one of the newer reasoning models that inherently performs internal chain-of-thought as part of its problem-solving approach. This feature seemed to be disadvantageous when being evaluated on the SCT dataset. To the knowledge of McCoy et al., it was the first medical benchmark test to invert the pattern where reasoning enhancement would usually boost a model's performance [24]. This study managed to reproduce this phenomenon within the non-reasoning model GPT-4.1 itself, by comparing baseline scores with explicitly induced chain-of-thought reasoning. It is further evidence that neither traditional nor domain-tailored CoT currently improves performance on clinical SCTs. It is striking, though, that this strategy not only fails to boost the scores but rather seems to hurt them somewhat. In this particular context it means clinical CoT led to less alignment with expert consensus overall.

With clinical CoT, sample-based majority voting turned out to be effective for enhancing the MedQA score, hence making the model choose its most frequent answer led to better outcomes. Interestingly, this was not the case for baseline prompting. Even though baseline scores were consistently inferior to CoT scores, the assumption that majority voting might also improve the baseline score does not seem far-fetched. Yet, majority voting did not have this impact at baseline, which further shows the strengths of CoT prompting for benchmarks

like MedQA. However, this pattern is once again broken by SCT-Bench where majority voting could not significantly improve the benchmark score and if anything had a bigger advantage at baseline than in CoT mode. This circumstance further corroborates how explicit reasoning seems to even deteriorate performance on SCT type questions.

When it comes to response entropy (section 3.2), clinical reasoning elicitation had a similar effect on both medical benchmark tests. On both formats, response entropy was significantly increased, implying that CoT led to more variation and disagreement among the model's answers. The absolute entropy values are much lower for MedQA than for SCT, which is expected given the different nature of both datasets. While MedQA provides for each question an answer option that is widely accepted as the only factually correct choice, every SCT question contains considerably more uncertainty among all rating options, ultimately leading to disagreement among clinician experts themselves. This is why an SCT score measures how much the model agrees with what the majority of experts think instead of how "objectively accurate" it is.

Having said this, CoT prompting had a similar relative effect on response entropy, which more than doubled on both benchmarks. Interestingly, while it generally augmented the variety of different responses (implying decreased sample consistency), it still increased the number of correct answers in the case of MedQA. Basically, responses are spread wider but at the same time clusters of correct answers are created. On SCT-Bench, however, it reduced the number of ratings that aligned with expert consensus. This means that while clinical CoT prompting yielded greater rating variability, it did not lead to clusters aligning with expert consensus ratings.

The rating entropy among experts (Figure 3.6) reflects genuine clinical uncertainty despite there being a consensus most experts gravitate toward. It should be the goal to develop AI systems that have greater likelihood of agreeing with expert consensus while keeping entropy as low as possible to reproduce reliable and stable outputs. Clinical CoT managed to increase the probability of correctness/consensus alignment on MedQA but also increased answer entropy, meaning it included more diversely incorrect responses as possible outputs. While the latter is also true for SCT-Bench, it also decreased the likelihood of agreeing with the expert consensus. CoT turned out to be more of a disadvantage for performing well on SCT-Bench.

## 4.2. Rating Behavior

In order to study further the effects of clinical reasoning elicitation in performance at script concordance testing, it is insightful to evaluate how GPT-4.1 adapted its rating behavior on the level of individual questions instead of only comparing macroscopic benchmark scores. In section 3.5, the distributions of rating disagreement with expert consensus for both experts and model answers were presented. At baseline, the distribution of all model ratings aligned quite well with the distribution of expert ratings around the consensus, as evident in Figure 3.16.

The model had greater tendency to disagree with the expert consensus and to disagree more strongly than the general body of experts. This disagreement increased with clinical CoT prompting, which also resulted in greater misalignment between the expert and model rating distributions. Just by looking at the distributions, it is evident that apart from agreeing less frequently with the consensus, GPT-4.1’s responses were generally more aligned with the full range of expert opinions at baseline than with CoT. Perfect alignment with the whole spectrum of expert opinions would yield an optimal majority SCT score of 100%. However, this would not be the only way to achieve a perfect score, since it is all about agreeing with expert consensus. A model that invariably aligns with expert consensus will achieve perfect performance on a SCT test, while its answer distribution will bear no resemblance to that of the experts, manifesting instead as a sharp line in the center. This raises the question of which would be preferable: an AI that achieves a perfect score by aligning perfectly with the full spectrum of expert opinions, or one that does so by consistently agreeing with the expert consensus while barely considering different options? In the context of trustworthy and consistent AI, it is arguably desirable to train systems that reliably align with expert consensus, substantially reducing outliers, and are able to communicate uncertainty accordingly.

The analysis in section 3.9 breaks down further how clinical reasoning elicitation affected rating behavior on a detailed level. Comparing both histograms (Figure 3.15, Figure 3.16) shows that CoT creates a shift toward stronger ratings, especially thinning out neutral ones. While at baseline the general distribution of ratings is not so far off from the experts, with CoT there is more obvious discrepancy. Since the histograms in section 3.9 only show the total distribution of ratings, we need to look at the Sankey diagrams (Figure 3.21 to Figure 3.22) to evaluate how clinical CoT prompting affected each individual rating. Grouping all ratings into categories by estimated impact strength without directional preference (neutral, moderate, strong) reveals a clear trend in rating transitions: Clinical reasoning elicitation leads to systematic overestimation of the impact of new information in SCT-Bench questions. When applying CoT, GPT-4.1 showed a significantly increased likelihood of changing its previous rating to a stronger one for the very same question. CoT leads to assessing likelihoods in SCT questions with less uncertainty than most experts think is appropriate. This appears to be the mechanism for why CoT reasoning turns out to be disadvantageous for script concordance testing. Keeping in mind the increased rating entropy, one could say that CoT increased “behavioral uncertainty,” meaning that the model offered a greater variety of different responses but in contrast decreased uncertainty in individual assessments of likelihoods. However, this conclusion cannot be generalized to other LLMs and effect of inherent or induced reasoning on SCT-Bench scores would need further evaluation. These results do not suggest that reasoning is inherently bad for SCT formats, as this would obviously not make any sense. What they do show, though, is that clinical reasoning and decision-making in LLMs is not appropriate enough for dealing with a high degree of uncertainty in

clinical scenarios. Unexpected CoT triggered transitions from  $+1$  to  $-1$  and their possible contribution for misalignment with the expert consensus was not analysed further as already mentioned in section 3.9.

Clinical reasoning elicitation has generally proven, also in this very same study, to enhance scores on other benchmark tests like MedQA (section 3.1). So why is it able to significantly boost the MedQA score while it hurts performance on SCT? This question reveals a fundamental gap in how we assess clinical competence and decision-making in LLMs and what we actually measure by doing so. The clinical CoT rationales seem to be more appropriate and applicable for MedQA questions, leading to more accurate answers. A possible explanation could be that CoT is particularly good at bringing deeply encoded knowledge to the surface in a mostly coherent and logical form. This might enhance the "awareness" of knowledge and how it relates to other bits of information (the reasoning part). Subsequently, this may lead to higher quality and accuracy when assessing medical questions that are heavily knowledge-based and provide a high degree of complete information regarding the question topic. However, this ability is evidently not good enough when the model is tasked with clinical decision-making under uncertainty: it does not contribute positively to the capacity of assessing and acknowledging uncertainty. The latter is a crucial element in real-life clinical scenarios, though, and we need to reflect on the implications this has for AI-driven clinical decision-making and evaluation.

In their article, McCoy et al. emphasize the evident gap in current benchmark performance between traditional MCQ and their SCT-Bench dataset [24]. While LLMs have been able to outperform clinicians in USMLE-like assessments, they consistently fall short of expert clinician levels in SCTs. As mentioned in their paper and shown in different studies, there is an additional fundamental observation: A certain pattern of discordance between SCT and MCQ performance is not unique to LLMs but can also be seen in humans. While medical students' early performance shows correlation between MCQs and SCTs, this correlation disappears in later years of their training [47]. It has also been found that emergency medicine residents did not show any significant correlation between MCQ and SCT scores [48]. This divergence becomes even more pronounced with increasing expertise [24, 47]. These findings imply that SCTs target different areas of clinical competence than MedQA-like exams. SCTs seem to reflect physicians' acquisition of clinical capabilities over many years that go beyond biomedical knowledge. While traditional MCQ formats likely test predominantly the latter, SCT scores appear to correlate strongly with years of experience as a doctor dealing with clinical decision-making on a daily basis. Without theorizing what exact nuance of clinical reasoning and decision-making is being tested in SCTs that conventional MCQs fail to capture, it seems safe to say that it is something that books and an academic education alone do not transmit.

At this point we have to raise the question of what all of this implies for LLMs. The

performance gap between MCQs and SCTs definitely reveals important limitations in clinical competence in large language models. Together with the evidence provided that chain-of-thought prompting is not able to enhance model performance on SCT unlike it does on MedQA, this might reveal fundamental properties and also limitations of (clinical) reasoning in LLMs. Primarily, it exposes that LLMs' way of reasoning does not handle uncertainty adequately and struggles with contextualizing it within elicited encoded knowledge.

The collected and generated data gave evidence for a substantial SCT-MCQ disparity in LLMs. CoT prompting, even when domain-tailored and enhanced for clinical reasoning, is not able to augment model performance on SCT-based questions and rather hurts the score. Rating behavior analysis showed how clinical CoT reasoning in GPT-4.1 causes the model to systematically overestimate the impact of new information in clinically uncertain settings. This might be directly linked to a kind of implicit "overconfidence" problem, where reasoning and active phrasing of knowledge-rich arguments provokes a false sense of certainty or relevance.

### 4.3. Uncertainty Proxies

Reliable uncertainty quantification could help physicians understand how confident AI is regarding its own decisions and assess their relevance more adequately. Methods that are able to determine the real uncertainty of systems are fundamental for the concept of trustworthy AI, and high-stake fields like medicine have a particular need for this. This study evaluated two black-box methods: self-elicited confidence, where GPT-4.1 explicitly reports uncertainty and sample consistency, measuring response agreement across n=15 queries. Results reveal systematic overconfidence and limited utility depending on method and benchmark.

#### 4.3.1. Systematic Overconfidence

Self-elicited confidence demonstrated clear overconfidence across both benchmarks, rarely dropping below 60% even when performance did. This reduced confidence range limited calibration assessment, making comprehensive evaluation challenging. The finding aligns with Naderi et al.'s observation that average confidence consistently exceeded accuracy across a range of different LLMs [35]. The "inverted benchmark pattern" observed for performance also manifested in calibration quality: Clinical CoT very modestly improved MedQA calibration while considerably worsening SCT calibration. This suggests that the negative impact of explicit reasoning on uncertainty-based assessment extends beyond accuracy implying confidence estimation too.

### 4.3.2. Benchmark-Disparity in Uncertainty Assessment

MedQA-USMLE showed relatively effective uncertainty quantification. Clinical CoT improved calibration modestly using sample consistency (ECE reduced by approximately half) and substantially enhanced discrimination for both methods: 14.7 percentage points AUROC increase for self-elicted confidence and 17.5 percentage points for sample consistency. Notably, self-elicted confidence yielded higher absolute AUROC values than sample-based methods despite poorer calibration. SCT-Bench proved more problematic. Both calibration and discrimination were markedly worse than on MedQA across all conditions. Baseline AUROC using sample consistency was barely above chance ( $\approx 0.5$ ). Clinical CoT worsened calibration considerably for both uncertainty methods, but improved discrimination, particularly with sampling (23.3 percentage points AUROC increase vs. 10.7 percentage points for self-elicted confidence). However, this enhancement coincided with worse calibration: the model became better at identifying relatively uncertain predictions, but worse at accurately quantifying uncertainty. Self-elicted confidence maintained slightly higher absolute AUROC than sample consistency, consistent with the pattern in MedQA-USMLE.

The results reveal clear functional discrepancy: sampling provides superior calibration through broader score distributions and better confidence-accuracy alignment, while self-elicted confidence provides superior discrimination despite systematic overconfidence. Both methods are black-box approaches, relatively simple for implementation, but limited compared to white-box methods where transparent access to internal states is key. They typically achieve higher AUROC and lower ECE values than black-box methods do [33, 49]. The results show that even within black-box constraints, meaningful distinctions exist. However, for trustworthy medical AI, black-box methods have obvious limitations for assessing true uncertainty. The computational cost of sample-based approaches is a factor to be considered, though recent confidence-informed approaches achieve 40% reduction of computational power [34]. Yet, the findings of how self-elicted confidence in LLMs is being consistently overestimated also reveal how vulnerable this approach could eventually be. The disparity in calibration and discrimination between MedQA-USMLE and SCT-Bench reinforces the fundamental differences between the two benchmarks. On MedQA, both metrics proved relatively reliable and clinical CoT improved discrimination ability substantially. On SCT-Bench, uncertainty quantification degraded: baseline discrimination barely exceeded chance, calibration was poor and clinical CoT worsened rather than improved it. This degradation of both performance and uncertainty assessment reveals a critical limitation: GPT-4.1 cannot adequately perform Bayesian reasoning about clinical situations involving uncertainty and demonstrates better uncertainty awareness on MedQA than on SCT-Bench.

## 4.4. Self-reflecting Clinical Reasoning

Beyond evaluating whether clinical reasoning prompting improves benchmark scores, this study investigated further the question: can GPT-4.1 authentically implement and self-reflect on specific clinical reasoning strategies? The results demonstrate that the model has the ability for strategy-appropriate reasoning and something resembling meta-cognitive awareness, though quality may vary and is not always optimal.

### 4.4.1. Strategy Implementation

The qualitative analysis of reasoning chains shows that GPT-4.1 can effectively organize its clinical reasoning according to prompted strategies instead of merely citing them superficially. Example 3.1 to Example 3.5 illustrate the mostly coherent application of diverse reasoning approaches matching the correspondent pathways.

Example 3.5 demonstrates authentic guideline-based reasoning where the model explicitly references AASLD and EASL guidelines to justify proceeding with liver biopsy rather than immediate treatment for a suspected HCC lesion. While the guideline citations were occasionally incomplete or imprecise across all responses, the fundamental reasoning principles, including consulting specific clinical guidelines and applying their recommendations to the case, remained consistent. Example 4 (Example 3.4) showcases red-flag and rule-out reasoning applied to a cardiac emergency scenario. The model appropriately prioritizes ruling out aortic dissection based on the critical "red flag" of pulse asymmetry, despite ECG findings suggesting myocardial infarction. The reasoning explicitly acknowledges that missing aortic dissection would be fatal and that dissection can cause STEMI-like patterns through coronary involvement. Considering that both potential diagnoses can be equally fatal if ignored, the priorities were set adequately and were clinically sound within the explicit reasoning framework. However, even though GPT-4.1 demonstrated adherence to such frameworks as illustrated in Example 3.1 to Example 3.5, the execution remains imperfect.

### 4.4.2. Strategy Distributions

The analysis of reasoning strategy distributions provides further evidence for at least functional deployment of thinking schemes and something resembling awareness about problem characteristics and task-appropriate selection of strategies. The most important distinction is that probabilistic reasoning emerged as significantly more prominent for SCT-Bench compared to MedQA-USMLE. The increased representation of probabilistic approaches indicates that GPT-4.1 recognizes SCT questions as inherently uncertain, requiring the evaluation of probabilities rather than reaching definitive conclusions. This interpretation is limited, though, and it is important to emphasize that SCT question prompts explicitly ask the model to assess likelihood shifts based on new information, so this could merely be instruction following

behavior. The probably most important evidence for "meta-cognitive" awareness comes from associating self-reported reasoning strategies with expert rating certainty levels. For SCT questions where GPT-4.1 applied rule-based and red-flag reasoning, experts provided strong likelihood ratings ( $\pm 2$ ) significantly more than any other, indicating greater clinical certainty. These are among the most deterministic and rigid strategies, yielding more definitive judgments. Considering these characteristics, they will naturally lead more frequently to decisions with greater certainty while reducing uncertainty. For this reason this link has a very fitting context and is probably no coincidence, possibly even a clue that experts applied the same logic. It can be seen as moderate evidence that GPT-4.1 is effectively capable of scenario-appropriate reasoning and self-reflection on it. For any other mentioned strategy, expert were way more likely to give moderate or neutral ratings.

## 4.5. Limitations

This study's findings must be interpreted within some methodological constraints that shape the scope and generalizability of the results.

**Limited model scope.** The exclusive focus on GPT-4.1 represents a primary limitation. Results and respective conclusions cannot be generalized to other LLMs without empirical verification. Different architectures, training backgrounds and inherent reasoning features among others may behave substantially differently. However, acknowledging this limitation does not eliminate validity in the conceptual findings. The fundamental differences between MedQA and SCT-Bench as benchmark formats remain model-independent and the observed mechanisms in GPT-4.1, particularly systematic overestimation under uncertainty, could potentially represent inherent properties shared among contemporary LLMs. This thesis work can also be seen as proof of concept for reasoning-aware evaluation approaches applicable beyond GPT-4.1.

**Dataset constraints.** This study utilized only the publicly accessible portion of the SCT benchmark provided by McCoy et al., comprising 174 questions compared to approximately 700 in the complete dataset [24]. This public subset covers fewer specialized medical areas and thus a more limited range of clinical scenarios. Consequently, the SCT scores reported here are not directly comparable to results published by McCoy et al. or at their official SCT-Bench website [24, 50]. Apart from this, the limited size of the public subset of SCT-Bench implies a statistical disadvantage compared to the MedQA dataset.

**SCT limitations.** The SCT concept faces certain constraints itself that need to be addressed: The task might not always be interpreted appropriately by human test takers, where some experts interpret strong ratings as confidence markers instead of shifts in likelihoods [51]. The latter, as well as a general expert tendency of avoiding extreme responses [52], may lead to bias and favour models replicating this kind of conservatism [24]. Furthermore, SCT by definition treats every expert rating as equally valid. This can be problematic considering

circumstances where there is disparity in experience, knowledge and ultimately competence among the rating expert panel. Finally, even though the expert consensus is expected to align well with decisions that hold the most evidence, a hypothetical superior model that disagrees with it would be penalized due to the inherent scoring paradigm. For measuring how well a model aligns with expert panels, this scoring system makes perfect sense and arguably it might be one of the most reasonable ways to evaluate clinical AI systems. However, this methodology makes it impossible to reflect hypothetical superiority of a model over expert consensus responses.

**Limited prompting method variety.** Only one single clinical CoT prompting approach was employed that provided general clinical reasoning instruction along with specific reasoning strategies applicable across diverse clinical scenarios. Alternative prompting techniques discussed in recent literature, including confidence-informed sample consistency, ensemble methods or highly task-specific prompt engineering were not explored. Notably, conventional CoT prompting was not evaluated as a secondary baseline, missing the opportunity to analyse how it compares to clinical CoT. However, the thesis' primary focus was not prompt optimization, but rather investigating differential impact on MedQA versus SCT-Bench.

**Reasoning assessment constraints.** The evidence for authentic self-reflection on reasoning pathways and associated cognitive properties faces possible limitations. SCT prompting explicitly instructs the model to assess likelihood shifts in every question and the model's awareness of taking a script concordance test may bias it toward self-reporting probabilistic reasoning strategies through associative priming and instruction following rather than genuine strategy selection. This suggestive element in the prompt design could artificially inflate the observed frequency of probabilistic reasoning in SCT responses. However, the qualitative analysis of reasoning chains demonstrates essential alignment between self-reported strategies and actual reasoning patterns exhibited in Example 3.1 - Example 3.5, suggesting the model does implement strategy-consistent reasoning beyond solely listing expected approaches. The link between self-reported reasoning strategies and expert ratings provides additional, though indirect, evidence for meaningful use rather than purely superficial strategy attribution. In future studies, it could be more insightful to evaluate correlations between model and expert strategies directly, e.g. in the context of a reasoning-centered evaluation framework.

## 4.6. Conclusion and Future Outlook

As AI technology advances at a fast pace while regulatory surveillance and safeguards remain limited, there is an urgent need to keep developing robust evaluation frameworks that can transparently and accurately assess clinical AI competence with minimal interpretive ambiguity. Furthermore, it will be crucial to create standardized "reasoning-aware" evaluation approaches able to assess the nuanced, context-dependent and often uncertain nature of clinical decision-making.

Current evaluation methods rely predominantly on knowledge-focused examination formats that provide only incomplete assessment of clinical reasoning abilities. SCTs contribute to closing this gap, testing probabilistic, context-sensitive reasoning under uncertainty that conventional benchmarks fail to capture. At the same time, SCT-Bench by McCoy et al. reveals significant limitations of state-of-the-art models that struggle with this new benchmark approach while achieving superhuman performance on MCQ formats. As discussed in section 4.2, the challenges that models face in SCTs are tightly linked to adequate, expert-compatible uncertainty handling. Inappropriate overconfidence as a result of active reasoning led to poorer results than expected. This circumstance also informs on evident gaps in clinical decision-making by LLMs and should raise awareness about this issue in the development of medical AI.

SCT as a benchmarking concept is without any doubt a valuable contribution that not only reveals limitations in LLMs, but also emphasizes the need for evaluation frameworks able to measure nuanced clinical thinking under uncertainty. The SCT approach itself already addresses this necessity partially. However, this concept should receive more attention and motivate the research community to build on it in order to enhance evaluation methodology. While scoring against an expert panel is one of the strengths of SCTs, as this implies human validation, it could also turn out to be one of the fundamental limitations this benchmark has. SCTs recognize that appropriate expert disagreement exists and award models for agreeing with the consensus. Points are generally awarded following the distribution of expert disagreement. If a model exhibited superhuman performance, how could this ever be reflected in SCTs, where disagreement with expert consensus is systematically penalized? Consensus-based scoring creates a fundamental evaluation paradox: a model with superior clinical reasoning that appropriately deviates from the expert majority opinion would, by definition, receive lower scores. Superiority and human validation become incompatible.

Regardless of these limitations, the SCT benchmark proposal is an interesting addition to the evaluation landscape of medical AI systems. Future benchmarks could potentially attempt to merge more explicitly the strengths of both worlds: testing basic competence in medical sciences like traditional MedQA-like assessments and nuanced clinical decision-making with a focus on probabilistic reasoning as SCT does, thereby providing a more complete a more complete assessment of clinical competence. How this could exactly look like has yet to be investigated in future research work.

The inherently interactive nature of clinical decision-making, characterized by iterative information gathering and continuous re-evaluation of strategies has motivated recent efforts towards more sophisticated evaluation frameworks. Hager et al. developed a MIMIC-IV-based benchmark simulating realistic clinical settings by asking LLMs to autonomously gather patient information and refine diagnostic reasoning within a back-and-forth dynamic [27]. Their experiments revealed that models that had achieved excellent scores on traditional exams failed to match this performance when confronted with iterative decision-making

tasks [27]. Similarly, Schmidgall et al. introduced AgentClinic, transforming static medical questions into dynamic simulations where agents must engage in patient dialogue, order tests and interpret multimodal data [53]. They demonstrated that diagnostic accuracies dropped to below 10% when problems were converted from static to interactive formats [53]. However, such interactive frameworks face significant challenges: computational demands, difficulty establishing ground truth in environments with unlimited possible interactions and isolating specific cognitive competencies [53].

Future evaluation frameworks may also become more reasoning-centered, explicitly assessing whether models apply contextually appropriate cognitive strategies or logical frameworks as proposed and analysed in the scope of this thesis work. Different clinical scenarios under distinct conditions demand different reasoning approaches. Analysing the compatibility between actual reasoning pathways and appropriate clinical reasoning strategies could potentially enhance evaluation of medical AI in a meaningful way. Inspired by how GPT-4.1 was instructed to apply and self-reflect on explicit clinical reasoning strategies, a similar approach could be implemented with an expert panel, creating an expert-validated clinical reasoning evaluation framework.

Which benchmarks maintain long-term value is a key question, given that many benchmarks quickly reach performance saturation (near-perfect scores) and fail to achieve widespread adoption [54]. Benchmarks stand a better chance at remaining relevant by standing out with broad applicability and real-world utility rather than narrow focus on isolated tasks [54]. It is preferable to create benchmarks that address a range of different performance aspects, which, if feasible, should be represented in a collection of metrics instead of one single score [54]. Advancing medical AI requires standardized and robust evaluation across diverse benchmark types capable of measuring multiple dimensions clinical competence for ensuring systems deployed in healthcare are reliable and trustworthy.

# Bibliography

- [1] B. Middleton, D. F. Sittig, and A. Wright. “Clinical Decision Support: a 25 Year Retrospective and a 25 Year Vision”. In: *Yearbook of Medical Informatics Suppl 1*.Suppl 1 (Aug. 2, 2016), S103–S116. DOI: 10.15265/IYS-2016-s034.
- [2] K. Ouanes and N. Farhah. “Effectiveness of Artificial Intelligence (AI) in Clinical Decision Support Systems and Care Delivery”. In: *Journal of Medical Systems* 48.1 (Aug. 2024), p. 74. ISSN: 0148-5598. DOI: 10.1007/s10916-024-02098-4.
- [3] S. Liu et al. “Using AI-generated suggestions from ChatGPT to optimize clinical decision support”. In: *Journal of the American Medical Informatics Association* 30.7 (June 20, 2023), pp. 1237–1245. DOI: 10.1093/jamia/ocad072.
- [4] Shuai Tian et al. “Opportunities and challenges for ChatGPT and large language models in biomedicine and health”. In: *Briefings in Bioinformatics* 25.1 (Nov. 2023), bbad493. DOI: 10.1093/bib/bbad493.
- [5] Shuai Wang et al. *Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?* 2023. arXiv: 2302.03495 [cs.IR]. URL: <https://arxiv.org/abs/2302.03495>.
- [6] Harsha Nori et al. *Capabilities of GPT-4 on Medical Challenge Problems*. 2023. arXiv: 2303.13375 [cs.CL]. URL: <https://arxiv.org/abs/2303.13375>.
- [7] M. Daniel et al. “Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance”. In: *Academic Medicine* 94.6 (June 2019), pp. 902–912. DOI: 10.1097/ACM.0000000000002618.
- [8] D. M. Connor, S. J. Durning, and J. J. Rencic. “Clinical Reasoning as a Core Competency”. In: *Academic Medicine* 95.8 (Aug. 2020), pp. 1166–1171. DOI: 10.1097/ACM.0000000000003027.
- [9] S. J. Durning et al. “Teaching clinical reasoning: principles from the literature to help improve instruction from the classroom to the bedside”. In: *Korean Journal of Medical Education* 36.2 (June 2024), pp. 145–155. DOI: 10.3946/kjme.2024.292.
- [10] Jonathan St. B. T. Evans and Keith E. Stanovich. “Dual-Process Theories of Higher Cognition: Advancing the Debate”. In: *Perspectives on Psychological Science* 8.3 (May 2013), pp. 223–241. DOI: 10.1177/1745691612460685.
- [11] Salvatore Corrao and Chiara Argano. “Rethinking clinical decision-making to improve clinical reasoning”. In: *Frontiers in Medicine* 9 (Sept. 2022), p. 900543. DOI: 10.3389/fmed.2022.900543.

- [12] T. Pelaccia et al. “An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory”. In: *Medical Education Online* 16 (Mar. 2011), p. 5890. DOI: 10.3402/meo.v16i0.5890.
- [13] T. K. Ark, L. R. Brooks, and K. W. Eva. “Giving learners the best of both worlds: do clinical teachers need to guard against teaching pattern recognition to novices?” In: *Academic Medicine* 81.4 (Apr. 2006), pp. 405–409. DOI: 10.1097/00001888-200604000-00017.
- [14] Vernon R. Neufeld et al. “Clinical problem-solving by medical students: a cross-sectional and longitudinal analysis”. In: *Medical Education* 15.5 (Sept. 1981), pp. 315–322. DOI: 10.1111/j.1365-2923.1981.tb02495.x.
- [15] Geoff R. Norman et al. “The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking”. In: *Academic Medicine* 92.1 (Jan. 2017), pp. 23–30. DOI: 10.1097/ACM.0000000000001421.
- [16] Nicola Cooper and John Frain. *ABC of clinical reasoning*. John Wiley & Sons, 2022.
- [17] Karan Singhal et al. “Large language models encode clinical knowledge”. In: *Nature* 620.7972 (Aug. 2023). Epub 2023 Jul 12. Erratum in: Nature. 2023 Aug;620(7973):E19. doi: 10.1038/s41586-023-06455-0, pp. 172–180. DOI: 10.1038/s41586-023-06291-2.
- [18] Tyler Savage et al. “Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine”. In: *NPJ Digital Medicine* 7.1 (Jan. 24, 2024), p. 20. DOI: 10.1038/s41746-024-01010-1.
- [19] Yuki Sonoda et al. “Structured clinical reasoning prompt enhances LLM’s diagnostic capabilities in diagnosis please quiz cases”. In: *Japanese Journal of Radiology* 43.4 (Apr. 2025). Epub 2024 Dec 3, pp. 586–592. DOI: 10.1007/s11604-024-01712-2.
- [20] Matthew M. Lucas et al. “Reasoning with large language models for medical question answering”. In: *Journal of the American Medical Informatics Association* 31.9 (Sept. 2024), pp. 1964–1975. DOI: 10.1093/jamia/ocae131.
- [21] Taeksoo Kwon et al. “Large Language Models Are Clinical Reasoners: Reasoning-Aware Diagnosis Framework with Prompt-Generated Rationales”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16. 2024, pp. 18417–18425. DOI: 10.1609/aaai.v38i16.29802. URL: <https://doi.org/10.1609/aaai.v38i16.29802>.
- [22] Saeel Sandeep Nachane et al. *Few shot chain-of-thought driven reasoning to prompt LLMs for open ended medical question answering*. 2024. arXiv: 2403.04890 [cs.CL]. URL: <https://arxiv.org/abs/2403.04890>.

- [23] Jason Wei et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: 2201.11903 [cs.CL]. URL: <https://arxiv.org/abs/2201.11903>.
- [24] Liam G. McCoy et al. “Assessment of Large Language Models in Clinical Reasoning: A Novel Benchmarking Study”. In: *NEJM AI* 2.10 (2025), AIdbp2500120. DOI: 10.1056/AIdbp2500120. eprint: <https://ai.nejm.org/doi/pdf/10.1056/AIdbp2500120>. URL: <https://ai.nejm.org/doi/full/10.1056/AIdbp2500120>.
- [25] Karan Singhal, T. Tu, J. Gottweis, et al. “Toward expert-level medical question answering with large language models”. In: *Nature Medicine* 31 (2025), pp. 943–950. DOI: 10.1038/s41591-024-03423-7. URL: <https://doi.org/10.1038/s41591-024-03423-7>.
- [26] Vals AI. *MedQA*. Vals AI. 2025. URL: <https://www.vals.ai/benchmarks/medqa> (visited on 10/21/2025).
- [27] Paul Hager et al. “Evaluation and mitigation of the limitations of large language models in clinical decision-making”. In: *Nature Medicine* 30.9 (Sept. 2024). Epub 2024 Jul 4, pp. 2613–2622. DOI: 10.1038/s41591-024-03097-1.
- [28] Di Jin et al. “What disease does this patient have? a large-scale open domain question answering dataset from medical exams”. In: *Applied Sciences* 11.14 (2021), p. 6421.
- [29] Tarek Nouh et al. “The script concordance test as a measure of clinical reasoning: a national validation study”. In: *American Journal of Surgery* 203.4 (Apr. 2012), pp. 530–534. DOI: 10.1016/j.amjsurg.2011.11.006.
- [30] Mohammad Abbasian et al. “Foundation metrics for evaluating effectiveness of health-care conversations powered by generative AI”. In: *NPJ Digital Medicine* 7.1 (Mar. 2024), p. 82. DOI: 10.1038/s41746-024-01074-z.
- [31] T. Savage et al. “Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment”. In: *Journal of the American Medical Informatics Association* 32.1 (Jan. 2025), pp. 139–149. DOI: 10.1093/jamia/ocae254.
- [32] Evan Becker and Stefano Soatto. *Cycles of Thought: Measuring LLM Confidence through Stable Explanations*. 2024. arXiv: 2406.03441 [cs.CL]. URL: <https://arxiv.org/abs/2406.03441>.
- [33] Miao Xiong et al. *Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs*. 2024. arXiv: 2306.13063 [cs.CL]. URL: <https://arxiv.org/abs/2306.13063>.
- [34] Amir Taubenfeld et al. *Confidence Improves Self-Consistency in LLMs*. 2025. arXiv: 2502.06233 [cs.CL]. URL: <https://arxiv.org/abs/2502.06233>.

- [35] Nariman Naderi et al. *Self-Reported Confidence of Large Language Models in Gastroenterology: Analysis of Commercial, Open-Source, and Quantized Models*. 2025. arXiv: 2503.18562 [cs.CL]. URL: <https://arxiv.org/abs/2503.18562>.
- [36] Saurav Kadavath et al. *Language Models (Mostly) Know What They Know*. 2022. arXiv: 2207.05221 [cs.CL]. URL: <https://arxiv.org/abs/2207.05221>.
- [37] Valentin Liévin et al. “Can large language models reason about medical questions?” In: *Patterns* 5.3 (Mar. 2024), p. 100943. DOI: 10.1016/j.patter.2024.100943.
- [38] Mark Steyvers, Hector Tejeda, Anil Kumar, et al. “What large language models know and what people think they know”. In: *Nature Machine Intelligence* 7 (2025), pp. 221–231. DOI: 10.1038/s42256-024-00976-7. URL: <https://doi.org/10.1038/s42256-024-00976-7>.
- [39] OpenAI. *Introducing GPT-4.1 in the API*. Model version: gpt-4.1-2025-04-14. 2025. URL: <https://openai.com/index/gpt-4-1/> (visited on 10/27/2025).
- [40] Anthropic. *Claude*. Models used: Sonnet 4.0, Opus 4.0 and Sonnet 4.5 for AI coding assistance. 2025. URL: <https://www.claude.ai> (visited on 10/27/2025).
- [41] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. *SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models*. 2023. arXiv: 2303.08896 [cs.CL]. URL: <https://arxiv.org/abs/2303.08896>.
- [42] W. W. Peterson, T. G. Birdsall, and W. C. Fox. “The theory of signal detectability”. In: *Transactions of the IRE Professional Group on Information Theory* 4.4 (1954), pp. 171–212.
- [43] Tom Fawcett. “Introduction to ROC analysis”. In: *Pattern Recognition Letters* 27 (June 2006), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.
- [44] J. A. Hanley and B. J. McNeil. “The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve”. In: *Radiology* 143.1 (Apr. 1982), pp. 29–36. DOI: 10.1148/radiology.143.1.7063747.
- [45] M. P. Naeini, G. F. Cooper, and M. Hauskrecht. “Obtaining Well Calibrated Probabilities Using Bayesian Binning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Jan. 2015, pp. 2901–2907.
- [46] Amit G. Singal et al. “AASLD Practice Guidance on prevention, diagnosis, and treatment of hepatocellular carcinoma”. In: *Hepatology* 78.6 (Dec. 2023), pp. 1922–1965. DOI: 10.1097/HEP.000000000000466.
- [47] A. Collard et al. “Reasoning versus knowledge retention and ascertainment throughout a problem-based learning curriculum”. In: *Medical Education* 43.9 (Sept. 2009), pp. 854–865. DOI: 10.1111/j.1365-2923.2009.03410.x.

- [48] C. Boulouffe, B. Charlin, and D. Vanpee. “Evaluation of clinical reasoning in basic emergencies using a script concordance test”. In: *American Journal of Pharmaceutical Education* 74.10 (Dec. 2010), p. 194. DOI: 10.5688/aj7410194.
- [49] Tejaswini Pedapati et al. *Large Language Model Confidence Estimation via Black-Box Access*. 2025. arXiv: 2406.04370 [cs.CL]. URL: <https://arxiv.org/abs/2406.04370>.
- [50] *SCT-Bench: Script Concordance Test Benchmark*. Official leaderboard for the Script Concordance Test benchmark. URL: <https://www.concor.dance> (visited on 11/04/2025).
- [51] S. Lurie. “Towards greater clarity in the role of ambiguity in clinical reasoning”. In: *Medical Education* 45.4 (Apr. 2011), pp. 326–328. DOI: 10.1111/j.1365-2923.2011.03938.x.
- [52] Matthew Lineberry, Clarence D. Kreiter, and Georges Bordage. “Threats to validity in the use and interpretation of script concordance test scores”. In: *Medical Education* 47.12 (Dec. 2013), pp. 1175–1183. DOI: 10.1111/medu.12283.
- [53] Samuel Schmidgall et al. *AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments*. 2025. arXiv: 2405.07960 [cs.HC]. URL: <https://arxiv.org/abs/2405.07960>.
- [54] S. Ott, A. Barbosa-Silva, K. Blagec, et al. “Mapping global dynamics of benchmark creation and saturation in artificial intelligence”. In: *Nature Communications* 13.1 (Nov. 2022), p. 6793. DOI: 10.1038/s41467-022-34591-0.

# List of Figures

3.1. Majority MedQA score . . . . .	24
3.2. Mean MedQA score . . . . .	25
3.3. Majority SCT-Bench score . . . . .	25
3.4. Mean SCT score . . . . .	26
3.5. GPT-4.1 response entropy on MedQA . . . . .	27
3.6. GPT-4.1 rating entropy on SCT-Bench . . . . .	27
3.7. MedQA calibration curve for self-elicited confidence . . . . .	28
3.8. SCT-Bench calibration curve for self-elicited confidence . . . . .	29
3.9. MedQA calibration curve for sample-based uncertainty estimation . . . . .	30
3.10. SCT-Bench calibration curve for sample-based uncertainty estimation . . . . .	30
3.11. ROC curves for one-step self-elicited confidence on MedQA . . . . .	31
3.12. ROC curves for sample-based confidence estimation on MedQA . . . . .	32
3.13. ROC curves for one-step self-elicited confidence on SCT-Bench . . . . .	32
3.14. ROC curves for sample-based confidence estimation on SCT-Bench . . . . .	33
3.15. SCT-Bench rating disagreement distributions (baseline) . . . . .	34
3.16. SCT-Bench rating disagreement distributions (CoT) . . . . .	35
3.17. Distribution of self-reported clinical reasoning strategies on MedQA-USMLE	45
3.18. Distribution of self-reported clinical reasoning strategies on SCT-Bench . .	46
3.19. Distribution of expert and GPT-4.1 baseline ratings. . . . .	47
3.20. Distribution of expert and GPT-4.1 CoT ratings. . . . .	47
3.21. Rating transitions between expert consensus and GPT-4.1 baseline ratings. .	48
3.22. Rating transitions between expert consensus and GPT-4.1 CoT ratings. . . .	48
3.23. Rating transitions between GPT-4.1 baseline and CoT ratings. . . . .	49
3.24. Distribution of GPT-4.1 CoT rating levels . . . . .	50
3.25. Distribution of expert consensus rating levels . . . . .	51

# List of Tables

1.1.	Overview of clinical reasoning strategies . . . . .	6
1.2.	5-point Likert Scale used for rating SCT questions. . . . .	14
3.1.	Mean entropy values . . . . .	28
3.2.	AUROC and ECE values on MedQA and SCT-Bench . . . . .	33
B.1.	Tool Table . . . . .	X

# List of Examples

1.1. Example of an SCT question. . . . .	14
3.1. Clinical CoT example for applied causal reasoning. The highlighted text passages show reasoning relevant to the strategy. . . . .	37
3.2. Clinical CoT example for heuristic pattern recognition. The highlighted text passages show reasoning relevant to the strategy. . . . .	39
3.3. Clinical CoT example for probabilistic reasoning. The highlighted text passages show reasoning relevant to the strategy. . . . .	40
3.4. Clinical CoT example for red flag and rule-out reasoning. The highlighted text passages show reasoning relevant to the strategy. . . . .	42
3.5. Clinical CoT example for explicit guideline-orientated reasoning. The highlighted text passages show reasoning relevant to the strategy. . . . .	43

# List of Equations

2.1	TPR and FPR . . . . .	21
2.2	ECE . . . . .	22
2.3	Shannon entropy . . . . .	23
3.1	Rating disagreement . . . . .	34

# List of Abbreviations

<b>AI</b>	Artificial Intelligence
<b>API</b>	Application Programming Interface
<b>AUC</b>	Area Under the Curve
<b>AUROC</b>	Area Under the Receiver Operating Characteristic Curve
<b>CDSS</b>	Clinical Decision Support System
<b>CoT</b>	Chain-of-Thought
<b>ECE</b>	Expected Calibration Error
<b>FPR</b>	False Positive Rate
<b>GPT</b>	Generative Pre-trained Transformer
<b>LLM</b>	Large Language Model
<b>MCQ</b>	Multiple-Choice Question
<b>MedQA</b>	Medical Question Answering
<b>NLP</b>	Natural Language Processing
<b>ROC</b>	Receiver Operating Characteristic
<b>SCT</b>	Script Concordance Testing
<b>TPR</b>	True Positive Rate
<b>USMLE</b>	United States Medical Licensing Examination

# A. Clinical CoT Prompt Templates

## A.1. MedQA-USMLE

### Clinical CoT Prompting - MedQA

## Medical question answering

You are taking a medical question answering exam, which evaluates your medical knowledge and your ability for clinical decision making. The following question provides different answer options labelled from A to E. Choose the correct answer and rate the uncertainty of your proposed answer from 0 to 100. 0 represents definitely uncertain. 100 represents definitely certain. Apply one or several reasoning strategies of your choice from the list below or an alternative with one if you consider it more appropriate.

## Response format

Respond with a clear, step-by-step clinical reasoning process that leads to your chosen answer.

Start by identifying and integrating all available clinical information such as symptoms, signs, medical examination results (physical examination, laboratory tests, diagnostic imaging tests) and patient history to determine step-by-step the correct answer. Apply all your clinical, anatomical, physiological, pharmacological and general scientific knowledge in every reasoning step. Consider key aspects like risk factors, predispositions, possible biases, evidence-based medicine, gold standards, indications/ contraindications, differential diagnoses, patient- centered benefit- risk assessment as well as the principles of medical ethics. You may also find it helpful to consider official medical guideline recommendations for diagnostic or treatment procedures (from WHO or other subfield-specific professional medical societies).

Following the step- by- step clinical reasoning process respond at the end only within the following format:

answer: your letter choice only

uncertainty: your uncertainty score only

strategy: all your selected clinical reasoning strategies only

justification: a brief justification for why you chose the applied clinical reasoning strategies

## ## Clinical Reasoning Strategies

### # Deductive Reasoning

Apply general medical rules, principles, or established facts to draw logical conclusions about the specific case. Analyze how certain circumstances can condition others and what might cause them or follow from them. This strategy applies broad knowledge and principles to infer targeted clinical decisions and is essential in almost any reasoning process-applicable in diagnosis, monitoring, or treatment planning, among others.

Example pathway:

"All patients on warfarin require INR monitoring -> This patient started warfarin -> Order INR before next dose"

### # Hypothetico-Deductive Reasoning

Generate multiple hypotheses about diagnosis, management options or expected outcomes, then systematically evaluate each through available evidence.

Example pathway:

"Generate 3-5 differential diagnoses -> Identify discriminating features -> Test each hypothesis against the clinical data -> Refine or reject hypotheses"

"Three possible treatments exist -> Compare efficacy, contraindications, and side effects for this patient -> Select treatment that best fits patient's specific context"

### # Inductive Reasoning

Build general conclusions from individual observations. This approach synthesizes specific data points (findings, trends, responses) into a broader understanding, often guiding initial diagnoses, treatment direction, or research. This can also be useful in situations where an unidentified condition affects several patients simultaneously and certain findings in one or a few of them are transferred to all.

Example pathway:

"Multiple cases with unknown condition -> Identify pattern in first few patients -> Generalize findings to all affected patients."

### # Abductive Reasoning

Start with puzzling findings and work backwards to find the most plausible explanation. This is ideal for integrating new or unexpected clinical data into coherent decision-making.

Example pathway:

"Patient shows these findings -> Which causes could explain all of them? -> Choose the explanation that accounts for the most features with the fewest assumptions."

### # Probabilistic Reasoning

Make decisions based on the relative likelihood of possible conditions or outcomes based considering risk factors, prevalence and context.

Frequently used for test selection, triage and risk-informed treatment decisions.

Example pathway:

"Assess pretest probability -> Factor in test characteristics ->  
Calculate post-test probability -> Decision based on likelihood thresholds"

### # Rule-Based / Categorical / Deterministic Reasoning

Apply fixed clinical criteria or thresholds. Use validated scoring systems for diagnosis, treatment or risk stratification. This minimizes variability and is useful in high-stakes or guideline-driven settings.

Example pathway:

"PERC rule negative + low clinical suspicion -> D-dimer not needed ->  
PE effectively ruled out"

### # Causal Reasoning

Trace pathophysiological sequences to understand findings, predict complications or anticipate treatment effects. Useful in explaining complex presentations or predicting downstream effects of disease or treatment.

Example pathway:

"Intervention/condition A -> Causes physiologic change B -> Leads to clinical finding C -> Expect downstream effect D/ explains current clinical picture D"

### # Heuristic / Pattern Recognition (Fast Thinking)

Use intuitive, experience-based recognition of common clinical

patterns. Fast, but potentially error-prone - best suited to familiar scenarios.

Example pathway:

"Patient presents with classic triad X, Y, Z -> Instantly fits known syndrome or condition -> Immediate action without lengthy analysis"

#### # Red Flag / Rule-Out Reasoning

Prioritize identifying or excluding critical and time-sensitive conditions that would drastically change management, regardless of their probability. Especially important in emergency, triage, or risk-sensitive settings, whenever the cost of missing a diagnosis far exceeds the cost of over-investigating or over-treating.

Example pathway:

"Chest pain + sudden onset, tearing quality, asymmetric arm blood pressures -> Rule out aortic dissection before considering other causes"

#### # Guideline-Based Reasoning

Apply evidence-based protocols/ algorithms for diagnosis, treatment selection, monitoring, or follow-up based on established and standardized recommendations. Effective in test ordering, management decisions, or when new data emerge mid-course.

Example pathway:

"Consult relevant clinical guideline -> Map patient features to decision steps -> Follow recommended pathway -> Implement standardized approach"

#### # Final remark

Many clinical scenarios require combining several reasoning strategies. An example for combining fast familiar pattern recognition (heuristics) with a red flag or worst-case strategy:

"Fever + headache + neck stiffness -> consider bacterial meningitis -> Immediate empiric antibiotics before diagnostic confirmation"

## A.2. SCT-Bench

### Clinical CoT Prompting - SCT Bench

```
# Script Concordance Testing
```

You are taking a script concordance test, which evaluates your medical knowledge and your ability for clinical decision making. In this test, you will evaluate how new information impacts a specific hypothesis.

Use the following scale to rate the impact:

- 2: Strongly decreases the likelihood of the hypothesis
- 1: Slightly decreases the likelihood of the hypothesis
- 0: No effect on the likelihood of the hypothesis
- +1: Slightly increases the likelihood of the hypothesis
- +2: Strongly increases the likelihood of the hypothesis

Provide the most appropriate rating and rate the uncertainty of your proposed answer from 0 to 100. 0 represents definitely uncertain. 100 represents definitely certain. Apply one or several reasoning strategies of your choice from the list below or an alternative one if you consider it more appropriate.

```
## Response format
```

Respond with a clear, step-by-step clinical reasoning process that leads to your chosen answer.

Start by identifying and integrating all available clinical information such as symptoms, signs, medical examination results (physical examination, laboratory tests, diagnostic imaging tests) and patient history to determine step-by-step the correct answer. Apply all your clinical, anatomical, physiological, pharmacological and general scientific knowledge in every reasoning step. Consider key aspects like risk factors, predispositions, possible biases, evidence-based medicine, gold standards, indications/ contraindications, differential diagnoses, patient- centered benefit- risk assessment as well as the principles of medical ethics. You may also find it helpful to consider official medical guideline recommendations for diagnostic or treatment procedures (from WHO or other subfield-specific professional medical societies).

Following the step- by- step clinical reasoning process respond at the

end only within the following format:

rating: your rating only

uncertainty: your uncertainty score only

strategy: all your selected clinical reasoning strategies only

justification: a brief justification for why you chose the applied clinical reasoning strategies

## ## Clinical Reasoning Strategies

### # Deductive Reasoning

Apply general medical rules, principles, or established facts to draw logical conclusions about the specific case. Analyze how certain circumstances can condition others and what might cause them or follow from them. This strategy applies broad knowledge and principles to infer targeted clinical decisions and is essential in almost any reasoning process-applicable in diagnosis, monitoring, or treatment planning, among others.

Example pathway:

"All patients on warfarin require INR monitoring -> This patient started warfarin -> Order INR before next dose"

### # Hypothetico-Deductive Reasoning

Generate multiple hypotheses about diagnosis, management options or expected outcomes, then systematically evaluate each through available evidence.

Example pathway:

"Generate 3-5 differential diagnoses -> Identify discriminating features -> Test each hypothesis against the clinical data -> Refine or reject hypotheses"

"Three possible treatments exist -> Compare efficacy, contraindications, and side effects for this patient -> Select treatment that best fits patient's specific context"

### # Inductive Reasoning

Build general conclusions from individual observations. This approach synthesizes specific data points (findings, trends, responses) into a broader understanding, often guiding initial diagnoses, treatment direction, or research. This can also be useful in situations where an unidentified condition affects several patients simultaneously and certain findings in one or a few of them are transferred to all.

Example pathway:

"Multiple cases with unknown condition -> Identify pattern in first few patients -> Generalize findings to all affected patients."

#### # Abductive Reasoning

Start with puzzling findings and work backwards to find the most plausible explanation. This is ideal for integrating new or unexpected clinical data into coherent decision-making.

Example pathway:

"Patient shows these findings -> Which causes could explain all of them? -> Choose the explanation that accounts for the most features with the fewest assumptions."

#### # Probabilistic Reasoning

Make decisions based on the relative likelihood of possible conditions or outcomes based considering risk factors, prevalence and context.

Frequently used for test selection, triage and risk-informed treatment decisions.

Example pathway:

"Assess pretest probability -> Factor in test characteristics -> Calculate post-test probability -> Decision based on likelihood thresholds"

#### # Rule-Based / Categorical / Deterministic Reasoning

Apply fixed clinical criteria or thresholds. Use validated scoring systems for diagnosis, treatment or risk stratification. This minimizes variability and is useful in high-stakes or guideline-driven settings.

Example pathway:

"PERC rule negative + low clinical suspicion -> D-dimer not needed -> PE effectively ruled out"

#### # Causal Reasoning

Trace pathophysiological sequences to understand findings, predict complications or anticipate treatment effects. Useful in explaining complex presentations or predicting downstream effects of disease or treatment.

Example pathway:

"Intervention/condition A -> Causes physiologic change B -> Leads to clinical finding C -> Expect downstream effect D/ explains current

"clinical picture D"

# Heuristic / Pattern Recognition (Fast Thinking)

Use intuitive, experience-based recognition of common clinical patterns. Fast, but potentially error-prone - best suited to familiar scenarios.

Example pathway:

"Patient presents with classic triad X, Y, Z -> Instantly fits known syndrome or condition -> Immediate action without lengthy analysis"

# Red Flag / Rule-Out Reasoning

Prioritize identifying or excluding critical and time-sensitive conditions that would drastically change management, regardless of their probability. Especially important in emergency, triage, or risk-sensitive settings, whenever the cost of missing a diagnosis far exceeds the cost of over-investigating or over-treating.

Example pathway:

"Chest pain + sudden onset, tearing quality, asymmetric arm blood pressures -> Rule out aortic dissection before considering other causes"

# Guideline-Based Reasoning

Apply evidence-based protocols/ algorithms for diagnosis, treatment selection, monitoring, or follow-up based on established and standardized recommendations. Effective in test ordering, management decisions, or when new data emerge mid-course.

Example pathway:

"Consult relevant clinical guideline -> Map patient features to decision steps -> Follow recommended pathway -> Implement standardized approach"

# Final remark

Many clinical scenarios require combining several reasoning strategies. An example for combining fast familiar pattern recognition (heuristics) with a red flag or worst-case strategy:

"Fever + headache + neck stiffness -> consider bacterial meningitis -> Immediate empiric antibiotics before diagnostic confirmation"

# Supplementary Material

All supplementary materials for this thesis, in particular code and data, are available at:

<https://github.com/thomasprade1/thesis-meduniwien-codebase>

This includes code for automated API queries and data analysis as well as dataset references.

## B. Tool Table

Table B.1.: Tool Table

Tool	Purpose	Section
Claude Sonnet 4.0	Coding assistance and debugging for data analysis scripts and LaTeX syntax; literature search and BibLaTeX citation formatting; linguistic refinement	Throughout the thesis
Claude Sonnet 4.5		
Claude Opus 4.0		

**Note:** AI models were used only as supporting tools. All scientific ideas, content, analyses and conclusions were developed independently. Linguistic refinement (for ensuring linguistic accuracy) was carefully reviewed and adjusted as necessary.