Beyond MedQA: Towards Real-world Clinical Decision Making in the Era of LLMs

Anonymous ACL submission

Abstract

Large language models (LLMs) show promise for clinical use. They are often evaluated using datasets such as MedQA. However, Many medical datasets, such as MedQA, rely on simplified Question-Answering (Q&A) that underrepresents real-world clinical decision-making. Based on this, we propose a unifying paradigm that characterizes clinical decision-making tasks along two dimensions: Clinical backgrounds and Clinical questions. As the background and questions approach the real clinical environment, the difficulty increases. We summarize the settings of existing datasets and benchmarks along two dimensions. Then we review methods to address Clinical decision-making, including training-time and test-time techniques, and summarize when they help. Next, we extend evaluation beyond accuracy to include efficiency, explainability. Finally, we highlight open challenges. Our paradigm clarifies assumptions, standardizes comparisons, and guides the development of clinically meaningful LLMs.

1 Introduction

In the fast-evolving field of artificial intelligence, large language models (LLMs) have made remarkable progress, showcasing impressive capabilities across diverse applications and opening new possibilities for innovation (Hurst et al., 2024; Achiam et al., 2023; Roumeliotis and Tselikas, 2023; Liu et al., 2024a; Guo et al., 2025). The integration of large foundational models into real-world medical practice holds tremendous promise, offering the potential to transform clinical workflows and patient care (Singhal et al., 2023b, 2025; Wu et al., 2024c). Nowadays, more and more medical fine-tuned large language models are being developed, enriching the practice of professionals and researchers (Chen et al., 2023; Singhal et al., 2025; Saab et al., 2024; Wang et al., 2023; Wu et al., 2024a; Sellergren et al., 2025).

LLM as a tool needs to be tested before implementation. One aspect of interest to developers of LLM is the performance. Towards testing performance, the medical field has been focused on using datasets, most of which were sourced from standard tests. Among these datasets, the most common are question-answering datasets, including MedQA (Jin et al., 2021), MMLU (Hendrycks et al., 2020), MedmcQA (Pal et al., 2022), MedxpertQA (Zuo et al., 2025), Pubmed QA (Jin et al., 2019), and so on. Each case typically contains context (patient history, laboratory data, etc.), a question, and multiple answer options. The model needs to choose the most appropriate option. These datasets have many advantages, such as standardization and easy evaluation of indicators, and have therefore been widely used to evaluate the performance of large language models in the medical field (Sellergren et al., 2025; Saab et al., 2024; Xie et al., 2024a).

However, there are still many differences between the LLM-powered clinical use cases and the question-and-answer (Q&A). For example, during the patient's first clinical encounter, the doctor may not be able to obtain the patient's medical history, test results, and other information in advance; he or she must ask about the patient's condition and ask for necessary laboratory tests while avoiding meaningless questions and laboratory tests. But medical Q&A can often provide this information directly and precisely. The process of making choices in a clinical environment is called clinical decision making, which is a multistep process that requires gathering and synthesizing data from diverse sources and continuously evaluating the facts to reach an evidence-based decision on a patient's diagnosis and treatment (Hager et al., 2024). Therefore, medical Q&A can be understood as a clinical decision-making process that simplifies the data collection and analysis process and simplifies clinical solutions. Such simplification promotes the standardization of large language model evaluation in the

medical field, but it also raises questions about whether large language models can be applied in actual clinical scenarios.

Some existing research has begun to recognize the gap between medical Q&A and real-world clinical decision-making, and has begun to explore more realistic task settings. For example, MedQA can be transformed into a doctor-patient dialogue, where doctors must ask patients questions or request tests to gradually obtain information before ultimately making a decision. However, the settings for these tasks vary. For example, in addition to converting MedQA to a doctor-patient dialogue (Schmidgall et al., 2024; Li et al., 2024b, 2025b), recent work directly adds irrelevant content or modifies options in MedQA (Pan et al., 2025). Whether these settings truly reflect the clinical decision-making capabilities of large models, and to what extent, is not addressed by the unified paradigm.

For the reasons mentioned above, this survey aims to propose a paradigm for clinical decision-making tasks. We evaluate the complexity of a clinical medical decision-making task from two dimensions: Clinical Background and Clinical Questions.

We summarize our contributions as follows:

- We define the clinical decision-making from the perspective of task setting, which provides a reference for other researchers in the future.
- We propose a paradigm to describe the real-world clinical decision-making tasks and summarize the existing literature.
- We propose several open issues for future clinical decision-making tasks.

2 Definition of Clinical Decision-Making

Previous definitions of clinical decision-making were summarized in the clinical context, or healthcare (Banning, 2008; Thompson and Dowding, 2002; Hager et al., 2024). In the era of large language models, it is important to address the definition of clinical decision-making because it could facilitate the standardization of LLMs development, research, and evaluation in clinical decision-making. According to our observation and existing literature, different medical datasets and clinical decision-making tasks can be divided into two parts:

Clinical Backgrounds. Before making a decision, doctors need to collect necessary information, such as clinical conditions and patient history, which is collectively called the clinical background. Different background settings exist in different tasks. We divide them into four categories: (1) No background; (2) Precise Information Background; (3) Rich Information Background; (4) Incomplete Information Background.

Clinical Questions. In most cases, clinical questions represent decisions that need to be made in a clinical background. In many research works and surveys, “tasks” and “questions” are often used synonymously, and “questions” are categorized according to specific medical tasks, such as EHR summary questions, disease diagnosis questions, etc. In this paper, we classify questions based on the question types. We divide them into four categories: (1) True or false questions, (2) Multiple choice questions, (3) Short-answer questions, and (4) Open-ended questions.

We show different backgrounds and question settings in Figures 1 and 2. The combination of background and questions constitutes the case for clinical decision-making. In clinical practice, rich information background and incomplete information background, especially incomplete information background, and the short-answer questions and open-ended questions, especially open-ended questions, are common. However, most current medical question-answering datasets still use precise information backgrounds and multiple-choice questions. We will introduce each background and question in detail in the next two sections.

3 Clinical Backgrounds

Clinical backgrounds are part of clinical decision-making. In this section, we introduce various background settings for clinical decision-making. These settings can be the original datasets or benchmarks formed by adding various constraints to the original datasets.

3.1 No Clinical Backgrounds

This type of task often only includes a medical question without any context, such as "What are the symptoms of diabetes?" These simple knowledge questions appear in many medical question-answering datasets and benchmarks (Zhang et al., 2025a; Jin et al., 2019). They just test LLMs’ ability to master medical knowledge. However, they often appear together with cases with a precise information background in medical question an-

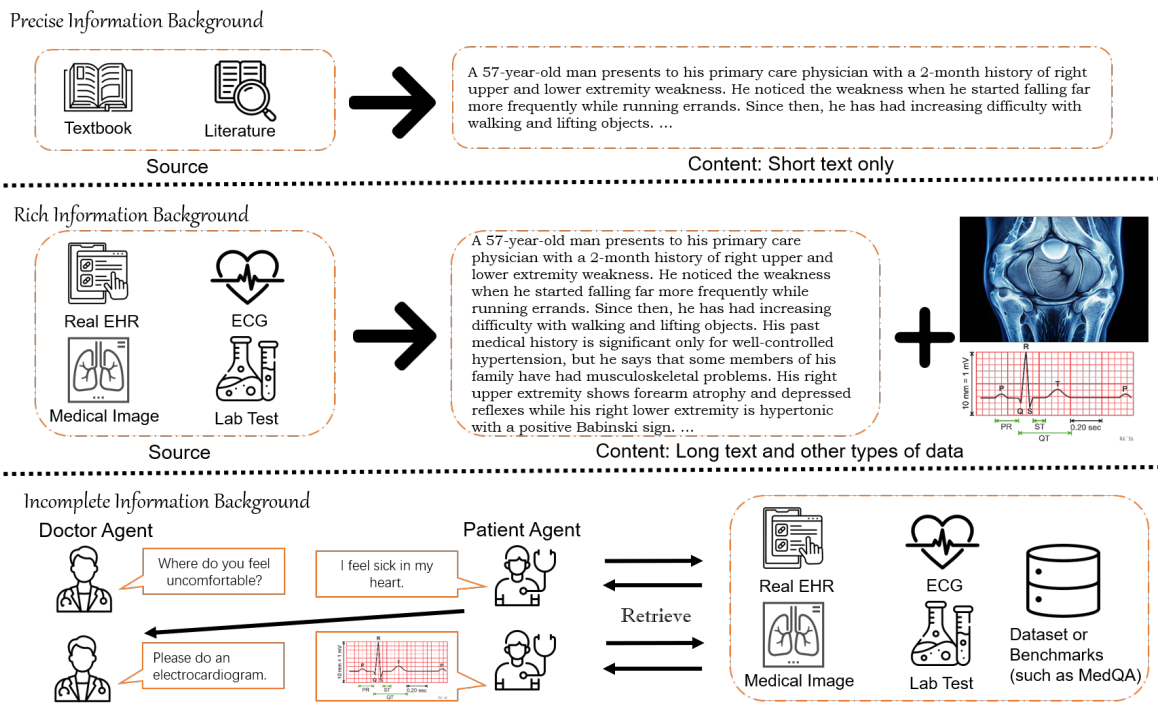

Figure 1: Three different clinical settings (No clinical background setting is not shown in the figure).

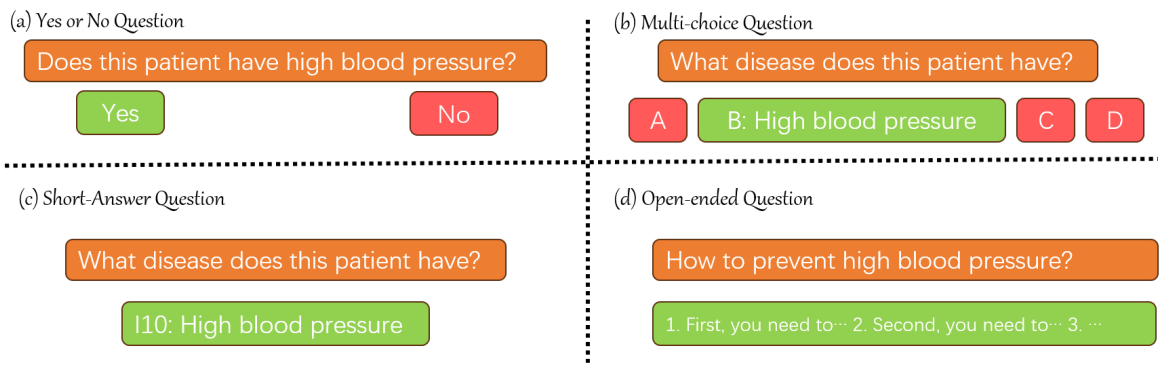

Figure 2: Four different clinical question settings.

swering datasets (Liu et al., 2023; Jin et al., 2021).

3.2 Precise Information Background

This is the most commonly used task setting in most datasets or benchmarks for clinical decision-making tasks. Some datasets use cases from textbooks (Jin et al., 2021; Hendrycks et al., 2020), while others use peer-reviewed literature (Jin et al., 2019; Lin et al., 2023). What they have in common is that, as teaching or published texts, most clinical environments have been simplified, and the resulting background information is dense, so the correct answers can be obtained by using existing medical knowledge and simple reasoning. This type of data set is often plain text data (Jin et al., 2025; Pal et al., 2022; Jin et al., 2021; Hendrycks et al., 2020; Liu et al., 2023). Some work has also begun to explore using LLM to generate specific backgrounds (Jin et al., 2025). Specifically, researchers use textbooks or guidelines as prompts and ask LLMs to generate a series of clinical backgrounds.

3.3 Rich Information Background

Compared to precise background information, this type of background information is richer; however, as the data increases, it also introduces redundant information and noise. At the same time, they put forward higher requirements on the retrieval and reasoning capabilities of LLMs; one example is that the texts in these tasks are usually longer, and long texts can lead to a decline in the reasoning ability of LLMs (Liu et al., 2025; Adams et al.,

2025).

Some datasets and benchmarks are natural extensions of the precise information background. Their sources are still textbooks or medical literature, but compared to the precise information background, the texts are often longer (Adams et al., 2025) and include data from other modalities, such as medical images and electrocardiograms, forming multimodal data backgrounds (Lin et al., 2023; Zuo et al., 2025; Ben Abacha et al., 2019; Singhal et al., 2025; Lau et al., 2018; Xie et al., 2024b; Zhang et al., 2025c). In addition to using real images, some work has also begun to explore synthetic medical images (Zhao et al.).

Compared to datasets generated directly from textbooks and literature, datasets constructed directly from clinical data, such as the MIMIC dataset (Johnson et al., 2016, 2023), contain a lot of information about patients. For a specific clinical question, only some of the information may be useful, while other information may be redundant. For example, in the task of predicting the length of hospital stay and mortality rate of COVID-19 patients, researchers need to collect a large number of laboratory indicators, only a few of which are correlated with mortality or hospitalization rate (Yan et al., 2020). For the MIMIC dataset (Johnson et al., 2016, 2023), researchers need to ignore some of the patient data (e.g., Doctors record patient conditions using natural language) during the data input stage (Jiang et al., 2024; Ma et al., 2024; Chen et al., 2024a; Yu et al., 2025b; Fang et al., 2025; Yan et al., 2024). However, there may still be a lot of redundant information. Because the patient's EHR still contains a lot of laboratory indicators, disease diagnosis records, and medication records, etc.

In addition to using real clinical data, redundant information can also be added to the clinical context manually. Some work directly inserts irrelevant information into the clinical context as part of red team attacks (Pan et al., 2025; Balazadeh et al., 2025). There are also some works showing that redundant information has an impact on LLM reasoning performance (Rajeev et al., 2025).

3.4 Incomplete Information Background

The clinical background information can be incomplete for doctors or clinical decision makers. That is, doctors can only obtain part of the clinical information initially. Some earlier works studied the ability of LLMs to diagnose multi-turn dialogues, and they often used online doctor-patient dialogue for evaluation (Xu et al., 2024; Zhang et al., 2023; Bao et al., 2023; Qiu et al., 2023). Online consultation is a new medical method, but it may deviate from the real clinical environment. In order to simulate the real clinical environment, the existing task set an incomplete information background as a multi-agent model. These tasks at least include two agents: the doctor agent and the patient agent. The doctor agent needs to communicate with the patient agent, ask the patient agent to perform the corresponding laboratory tests, gradually obtain clinical information, and finally make clinical decisions. The doctor agent is responsible for asking questions, and the patient agent retrieves the corresponding information from the dataset to answer the doctor agent's questions. Researchers often use existing datasets to create benchmarks. The datasets include MedQA (Schmidgall et al., 2024; Johri et al., 2025; Li et al., 2024b), PMC-OA (Qiu et al., 2025), MIMIC (Schmidgall et al., 2024; Hager et al., 2024) and so on.

3.5 Summary

We categorize the clinical background settings into four types. To our knowledge, the first two settings are often found in medical qualification exams, while the latter two settings are closer to clinical settings. Some datasets and benchmarks may have more than one clinical background setting. For example, a benchmark that uses both the MedQA and MIMIC datasets may have two clinical background settings (Liu et al., 2024b). Existing papers all show that the performance of the same dataset in an incomplete information background setting is significantly lower than that in a complete information background setting (Schmidgall et al., 2024; Li et al., 2024b).

4 Clinical Questions

Clinical questions are part of clinical decision-making. In this section, we introduce various question settings for clinical decision-making.

4.1 True/False and Multiple Choice Questions.

True/False and multiple-choice questions are collectively called closed-end questions; they are also the question formats used in most medical datasets and benchmarks (Pal et al., 2022; Jin et al., 2021; Liu et al., 2023; Hendrycks et al., 2020). Some questions have infinite solution spaces, such as

the problem of hospitalization days prediction, but some existing work stipulates it to a specific range and turns it into a multiple-choice question (Chen et al., 2024a; Yu et al., 2025b; Fang et al., 2025). The advantage of multiple-choice questions is that the correctness of the answers is relatively easy to judge, and it is straightforward to design evaluation metrics to calculate the performance of different LLMs. However, their decision-making process cannot be effectively explained. The number of options in a multiple-choice question significantly affects the evaluation of model performance. If a multiple-choice question has only 4 options, then even if the model randomly selects, it will have a 25% accuracy rate. Some work recognizes these problems and therefore expands the number of options to 10 to reduce the accuracy of random selection (Wang et al., 2024; Zuo et al., 2025). This makes the multiple-choice questions closer to the short-answer question setting and clinical environment.

4.2 Short Answer Questions

Compared to multiple-choice questions, short-answer questions have a larger or even infinite solution space. Sometimes, this type of question is also considered a special type of closed-ended question. For example, some questions require output of disease names or disease codes (Zhu et al., 2025; Ma et al., 2024; Soroush et al., 2024; Xu et al., 2025b). The current ICD system has about 70,000 disease codes. Some questions need to output a numerical value, such as the dosage of a patient's medication (Khandekar et al., 2024). Compared to open-ended questions, their correctness is easier to verify. For example, if we give a medication dosage, we can easily compare it with the groundtruth to determine whether it is the correct answer. In open-ended questions, sentences and paragraphs of different styles may be the correct answer.

4.3 Open-Ended Questions

An open-ended question is a question that cannot be answered with a "yes" or "no" response or with a static response. Open-ended questions are phrased as a statement that requires a longer answer. Open questions can be knowledge-based questions without clinical backgrounds (Singhal et al., 2023a). They can also be questions involving complex clinical scenarios, such as medical ethics issues (Balas et al., 2024). In addition, open-ended questions can often be generated directly from closed-ended questions, such as removing the options of closed-ended questions (Sandeep Nachane et al., 2024). Some generative tasks, such as summarizing electronic health records, can also be viewed as open-ended questions.

4.4 Summary

We categorize the clinical question settings into four major types. To our knowledge, the first two settings are often found in medical qualification exams, while the latter two settings are closer to clinical settings. Figure 3 shows the clinical background and clinical settings used by existing mainstream datasets/benchmarks.

5 Methods for Clinical Decision-Making

In this section, we will introduce methods used in clinical decision-making. Although most of these methods were applied to clinical settings with precise information and multiple-choice questions, they still have implications for real-world clinical decision-making. We can categorize them into two types: training-time techniques and test-time techniques. Training-time techniques fundamentally change the model weights to achieve better reasoning capabilities. Test-time techniques guide and optimize the model's output during reasoning without modifying the model itself.

5.1 Training-time Techniques

In this paper, we divide training-time techniques into two categories: supervised fine-tuning (SFT) and reinforcement learning (RL). Although in some works, reinforcement learning is also considered a form of supervised fine-tuning (Dong et al., 2023). We believe that the primary difference between them is that SFT utilizes explicit groundtruth as a supervisory signal, whereas reinforcement learning can employ "soft constraints" such as reward functions and human preferences, and does not require groundtruth mandatorily (Wang et al., 2025).

Supervised fine-tuning (SFT). Recent work has explored strategies for fine-tuning large language models in multiple stages using high-quality synthetic data (Yu et al., 2025a). There are also some works exploring strategies for supervised fine-tuning in multimodal backgrounds, which is one of the rich Information backgrounds (Li et al., 2025a; Chen et al., 2024b). In the context of incomplete information (generally multi-round doctor-patient dialogue), supervised fine-tuning using answers to multiple-choice questions or short-answer

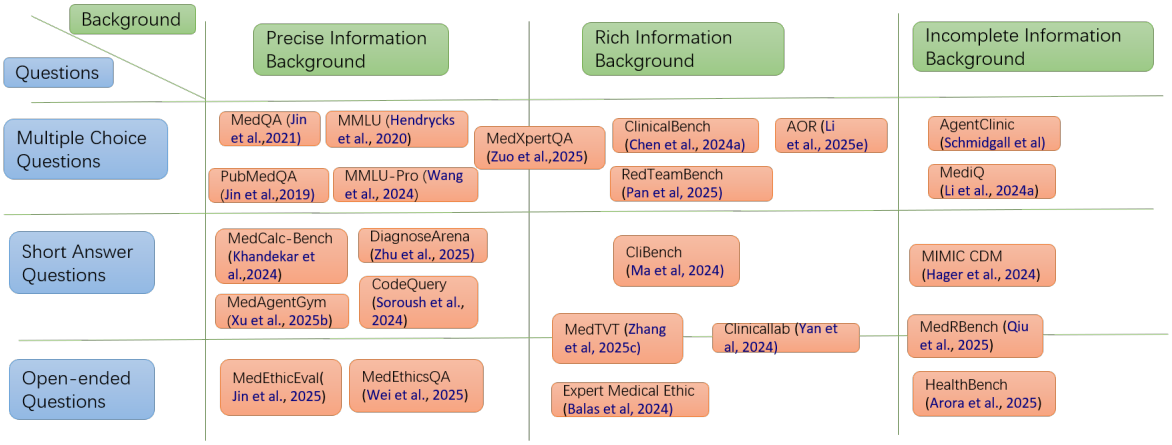

Figure 3: The clinical background and clinical question settings of mainstream medical datasets/benchmarks. Note that some datasets/benchmarks may have multiple settings.

questions as supervisory signals has also begun to emerge (Li et al., 2025b; Ding et al., 2025). These works have shown the effectiveness of SFT in incomplete information backgrounds and multiple-choice questions.

Reinforcement learning (RL). In the development of large language models, many reinforcement learning methods have emerged, such as proximal policy optimization (PPO) (Schulman et al., 2017), direct preference optimization (DPO) (Rafailov et al., 2023), and group relative policy optimization (GRPO) (Shao et al., 2024). Currently, DPO and GPRO have begun to be applied to clinical environment settings due to their advantages, such as being lightweight and having no need for value networks. For example, GPRO is used for optimization and alignment in multimodal clinical backgrounds (Xu et al., 2025a), and DPO is used for optimization and alignment in multi-round dialogues (Li et al., 2025b; Ding et al., 2025).

5.2 Test-time Techniques

Common test-time techniques include Chain-of-Thought (CoT) (Wei et al., 2022), retrieval-augmented generation (RAG) (Lewis et al., 2020), multi-agent systems, etc.

Chain-of-Thought (CoT). CoT facilitates problem-solving by guiding the model through a step-by-step reasoning process by using a coherent series of logical steps. There are many ways to decompose a medical task into different steps. For example, in clinical diagnostic tasks, some works decompose the task into the doctor's underlying reasoning process (Wu et al., 2023; Chen et al., 2024c; Zhou et al., 2025a). In the medical dialogue generation task, the task is decomposed into simple subtasks (He et al., 2024).

Retrieval-augmented generation (RAG). RAG is a technique that enhances large language models (LLMs) by retrieving relevant information from external knowledge sources before generating a response, making answers more accurate, timely, and grounded in verifiable facts. Early work directly searched databases such as PubMed or Wikipedia (Xiong et al., 2024). But now, some works have first constructed these databases into knowledge graphs, which greatly improves efficiency when searching in LLM (Wu et al., 2025, 2024b; Wen et al., 2023).

Multi-agent systems. Multi-agent systems consist of multiple intelligent, autonomous agents that collaborate, coordinate, and interact in a shared environment to solve complex problems. There are many works that use multi-agent collaboration to solve medical problems (Kim et al., 2024; Ke et al., 2024; Zhang et al., 2025b). In addition, multi-agent systems have unique advantages in simulating doctor-patient interactions. Using doctor agents and patient agents can simulate real medical scenarios. Currently, there are many works on simulating real medical scenarios based on multi-agent systems, ultimately improving the diagnostic capabilities of doctor agents after doctor-patient interactions (Fan et al., 2024; Li et al., 2024a)

5.3 Summary

We have introduced a variety of training-time and test-time techniques. However, these techniques are not mutually exclusive and can be used in combination. For example, in multi-stage learning, SFT can be used in the first stage and RL in the second stage (Ding et al., 2025). We can also use multi-

agent systems, RAG, and CoT at the same time during the testing phase.

6 Evaluation and Benchmarking

In this section, we introduce metrics for evaluating the performance of LLMs in clinical decision-making tasks.

6.1 Effectiveness

For closed-ended questions, accuracy is an easy-to-use metric to measure effectiveness, which is calculated as follows:

$$Accuracy = a_c/a_t \tag{1}$$

a_c is the number of samples predicted correctly, a_t is the number of all samples. It reflects the exact match or multiple-choice score on known-answer questions. On some datasets, such as MedQA and MMLU, LLMs such as Med-PaLM2 (Singhal et al., 2025) have reached or exceeded the expert level. However, the clinical background and answer settings of these datasets are relatively simple. Recent datasets have increased the length of text in clinical contexts, added data from other modalities such as images and time series data, and increased the number of multiple-choice questions. These changes have made multiple-choice questions more complex, and even the most advanced model GPT-4o cannot achieve an accuracy of 50% (Zuo et al., 2025).

6.2 Efficiency

Efficiency is crucial for complete and redundant information backgrounds and incomplete information backgrounds. Doctors or clinical decision makers must extract key information from complex and chaotic information; ask patients accurate questions, and require patients to undergo appropriate examinations to ensure the timeliness and effectiveness of clinical decisions. Some recent studies have explored methods for evaluating efficiency. They mainly use the LLM-as-a-judge method to evaluate the proportion of key questions or key dialogues in the doctor-patient dialogue (Qiu et al., 2025). Specifically, this work uses high-performance LLM (such as GPT-4o) to divide the reasoning step into N sub-steps: $r_1, r_2, ...r_N$, For each step, a binary variable e_i is used to indicate whether it is on the correct reasoning path. $e_i = 1$ means that the step refers to steps that provide additional insights and contribute to the final decision, and does not repeat the previous steps; otherwise, it is 0. Efficiency is calculated according to the following formula:

$$Efficiency = \frac{1}{N}\sum_{i=1}^{N} e_i \tag{2}$$

6.3 Explainability

In medical decisions, the decision maker or patient needs to give a reasonable explanation for the answer rather than simply giving an answer. This is especially important in short-answer questions and open-ended questions, especially open-ended questions. There are currently two main ways to evaluate LLM explainability: manual evaluation and LLM-as-a-judge.

Manual evaluation often develops a series of scoring rules based on the task, such as using a Likert scale (Balas et al., 2024), and finally averages the scores of multiple people to get the final score (Jin et al., 2025). In some studies, such as *Healthbench* (Arora et al., 2025), these scoring settings are more complex, requiring not only the accuracy of the answers, but also the order in which different contents appear.

LLM-as-a-judge generally uses LLMs with stronger reasoning ability for scoring. It can be expressed by the following formula (Gu et al., 2024):

$$\mathcal{E} = \mathcal{P}_{\mathcal{LLM}}(x \oplus \mathcal{C}) \tag{3}$$

$\mathcal{E}$ is the final evaluation obtained from the whole LLM-as-a-Judge process, it can be scores or other metrics. $\mathcal{P}_{\mathcal{LLM}}$ is the large language model that use to evaluate. x is the data which waiting to be evaluated, it is generally the output of a large language model in a clinical case. $\mathcal{C}$ The context for the input x, which is often prompt template. Essentially, it is a replacement for human evaluation. Therefore, some of the rules used in human evaluation can be used as context for LLM-as-a-judge; $\mathcal{C}$ is often the scoring criterion specified by the user. A common evaluation method is the reasoning process is broken down into several parts and scored based on whether the answer matches each part (Qiu et al., 2025; Wei et al., 2025). For example, if the answer is divided into 4 steps and the full score of each step is 1 point, when LLM evaluates the answer, it is considered that 2 steps have 1 point, 1 step has 0.5 points, and 1 step has no points, so the final score is 2.5 points.

7 Open Issues

In this section, we summarize some of the current status in clinical decision-making research and raise some open issues.

7.1 Creating Datasets and Benchmarks relevant to Real Clinical Background

At present, most of the work is still limited to the clinical environment with precise information backgrounds and multiple-choice questions. Although the background settings are becoming more and more complex and the multiple-choice options are increasing, they still do not break out of this fundamental framework (Zuo et al., 2025). An important leap from a complete information context to a real clinical context with redundant information is that most redundant information is real laboratory test data, which is difficult to synthesize artificially. It is necessary to create a multi-task clinical benchmark full of open-ended questions.

7.2 Filtering Redundant Information

The real clinical environment has to deal with a lot of redundant information. Long contexts already pose a challenge to the reasoning of LLMs (Liu et al., 2025), and redundant information may further degrade the reasoning performance of LLMs (Rajeev et al., 2025). Therefore, it is crucial to guide LLMs to filter out redundant information. In the context of incomplete information, this question can be further extended to: how to get the doctor's agent to ask a "good" question (Li et al., 2025b; Fang et al., 2025), this means that doctors need to ask questions that are closely related to the illness the patient may be diagnosed with.

7.3 Multi-agent systems suitable for clinical environment

Multi-agent systems have been used in medical question answering and have proven effective. In data-rich clinical environments, however, multi-agent systems offer unique advantages. This is because the context in clinical settings is often longer. By splitting clinical tasks into subtasks, each processing only a portion of the data, multi-agent systems can avoid the issues caused by the long context of a single agent.

In addition, multi-agent systems are a good way to simulate real clinical environments. Multi-agent dialogue is crucial for simulating the doctor-patient interaction process in an incomplete information environment. There is currently little research on patient agents in doctor-patient dialogues. Some work has studied the bias issues of patient agents (Schmidgall et al., 2024), but there is insufficient research on whether patient agents can effectively retrieve datasets and accurately answer doctors' questions.

7.4 Designing New Training Paradigms for Open-ended Questions

In "Closed-end questions" such as multiple-choice questions and short-answer questions, whether it is supervised fine-tuning or reinforcement learning, the existence of standardized groundtruth makes it easier to set the loss function/reward. However, in open-ended questions, there may be multiple answers that meet the requirements, and there is no standard groundtruth for the answers. How to design the loss function or reward is a problem that needs to be solved. Some label-free reinforcement learning methods may be the key to solving this problem (Zhou et al., 2025b).

7.5 New Modeling Strategies for Evaluation

Currently, there are few evaluation metrics for efficiency, and there are no evaluation metrics specifically designed for multi-round doctor-patient dialogues. In terms of explainability, human evaluation is too expensive to evaluate clinical open-ended questions, while using LLM-as-a-judge faces issues of reliability, robustness, and hallucinations (Gu et al., 2024). Design efficiency evaluation metrics and design a better LLM-as-a-judge method to evaluate clinical questions, especially open-ended questions, as a possible future research direction.

8 Conclusion

This systematic review analyzes the current status and challenges of LLMs in clinical decision-making. We first divide clinical decision-making tasks based on the two dimensions: clinical background and clinical questions. Then, we further explore the current status of methods and evaluation metrics from these two dimensions. Finally, we summarize the challenges faced in clinical decision-making.

9 Limitations

Our study still has some limitations. First, our study may have overlooked some recent preprints. Sec-

ond, our study provides a classification perspective on clinical decision-making. However, there may still be other classification perspectives on clinical decision-making. For example, clinical decision-making can be classified by task type.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. 2025. Longhealth: A question answering benchmark with long clinical documents. *Journal of Healthcare Informatics Research*, pages 1–17.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Michael Balas, Jordan Joseph Wadden, Philip C Hébert, Eric Mathison, Marika D Warren, Victoria Seavilleklein, Daniel Wyzynski, Alison Callahan, Sean A Crawford, Parnian Arjmand, and 1 others. 2024. Exploring the potential utility of ai large language models for medical ethics: an expert panel evaluation of gpt-4. *Journal of medical ethics*, 50(2):90–96.

Vahid Balazadeh, Michael Cooper, David Pellow, Atousa Assadi, Jennifer Bell, Mark Coatsworth, Kaivalya Deshpande, Jim Fackler, Gabriel Funingana, Spencer Gable-Cook, and 1 others. 2025. Red teaming large language models for healthcare. *arXiv preprint arXiv:2505.00467*.

Maggi Banning. 2008. A review of clinical decision making: models and current research. *Journal of clinical nursing*, 17(2):187–195.

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

Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019.

Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. 2024a. Clinicalbench: Can llms beat traditional ml models in clinical prediction? *arXiv preprint arXiv:2411.06469*.

Jiawei Chen, Dingkang Yang, Yue Jiang, Yuxuan Lei, and Lihua Zhang. 2024b. Miss: A generative pretraining and fine-tuning approach for med-vqa. In *International Conference on Artificial Neural Networks*, pages 299–313. Springer.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024c. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Hongxin Ding, Baixiang Huang, Yue Fang, Weibin Liao, Xinke Jiang, Zheng Li, Junfeng Zhao, and Yasha Wang. 2025. Promed: Shapley information gain guided reinforcement learning for proactive medical llms. *arXiv preprint arXiv:2508.13514*.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*.

Yue Fang, Yuxin Guo, Jiaran Gao, Hongxin Ding, Xinke Jiang, Weibin Liao, Yongxin Xu, Yinghao Zhu, Zhibang Yang, Liantao Ma, and 1 others. 2025. Toward better ehr reasoning in llms: Reinforcement learning with expert attention guidance. *arXiv preprint arXiv:2508.13579*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob

Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and 1 others. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Yuhong He, Yongqi Zhang, Shizhu He, and Jun Wan. 2024. Bp4er: Bootstrap prompting for explicit reasoning in medical dialogue generation. *arXiv preprint arXiv:2403.19414*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. 2024. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. *arXiv preprint arXiv:2410.04585*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Haoan Jin, Jiacheng Shi, Hanhui Xu, Kenny Q Zhu, and Mengyue Wu. 2025. Medethiceval: Evaluating large language models based on chinese medical ethics. *arXiv preprint arXiv:2503.02374*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, and 1 others. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature medicine*, 31(1):77–86.

Yu He Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. 2024. Enhancing diagnostic accuracy through multi-agent conversations: using large language models to mitigate cognitive bias. *arXiv preprint arXiv:2401.14589*.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad Safranek, Abid Anwar, Andrew Zhang, and 1 others. 2024. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and 1 others. 2024a. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Qingqiu Li, Zihang Cui, Seongsu Bae, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, Junjun He, and 1 others. 2025a. Aor: Anatomical ontology-guided reasoning for medical large multimodal model in chest x-ray interpretation. *arXiv preprint arXiv:2505.02830*.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Jonathan S Ilgen, Yulia Tsvetkov, and Maarten Sap. 2025b. Aligning llms to ask good questions a case study in clinical reasoning. *arXiv preprint arXiv:2502.14860*.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024b. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024b. Large language models are poor clinical decision-makers: A comprehensive benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13696–13710, Miami, Florida, USA. Association for Computational Linguistics.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, and 1 others. 2025. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and 1 others. 2023. Benchmarking large language models on cmexam–a comprehensive chinese medical exam dataset. *arXiv preprint arXiv:2306.03030*.

Mingyu Derek Ma, Chenchen Ye, Yu Yan, Xiaoxuan Wang, Peipei Ping, Timothy S Chang, and Wei Wang. 2024. Clibench: A multifaceted and multigranular evaluation of large language models for clinical decision making. *arXiv preprint arXiv:2406.09923*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Jiazhen Pan, Bailiang Jian, Paul Hager, Yundi Zhang, Che Liu, Friedrike Jungmann, Hongwei Bran Li, Chenyu You, Junde Wu, Jiayuan Zhu, and 1 others. 2025. Beyond benchmarks: Dynamic, automatic and systematic red-teaming agents for trustworthy medical language models. *arXiv preprint arXiv:2508.00923*.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Quantifying the reasoning abilities of llms on real-world clinical cases. *arXiv preprint arXiv:2503.04691*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Meghana Rajeev, Rajkumar Ramamurthy, Prapti Trivedi, Vikas Yadav, Oluwanifemi Bamgbose, Sathwik Tejaswi Madhusudan, James Zou, and Nazneen Rajani. 2025. Cats confuse reasoning llm: Query agnostic adversarial triggers for reasoning models. *arXiv preprint arXiv:2503.01781*.

Konstantinos I Roumeliotis and Nikolaos D Tselikas. 2023. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Saeel Sandeep Nachane, Ojas Gramopadhye, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. 2024. Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering. *arXiv e-prints*, pages arXiv–2403.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales,

Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023b. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. 2024. Large language models are poor medical coders—benchmarking of medical code querying. *NEJM AI*, 1(5):AIdbp2300040.

Carl Thompson and Dawn Dowding. 2002. Clinical decision making and judgement in nursing. *(No Title)*.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Wenxuan Wang, Zizhan Ma, Meidan Ding, Shiyi Zheng, Shengyuan Liu, Jie Liu, Jiaming Ji, Wenting Chen, Xiang Li, Linlin Shen, and 1 others. 2025. Medical reasoning in the era of llms: A systematic review of enhancement techniques and applications. *arXiv preprint arXiv:2508.00669*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jianhui Wei, Zijie Meng, Zikai Xiao, Tianxiang Hu, Yang Feng, Zhijie Zhou, Jian Wu, and Zuozhu Liu. 2025. Medethicsqa: A comprehensive question answering benchmark for medical ethics evaluation of llms. *arXiv preprint arXiv:2506.22808*.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024a. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Cheng-Kuang Wu, Wei-Lin Chen, and Hsin-Hsi Chen. 2023. Large language models perform diagnostic reasoning. *arXiv preprint arXiv:2307.08922*.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, and 1 others. 2025. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024b. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Jie Yang, and 1 others. 2024c. Medjourney: Benchmark and evaluation of large language models over patient clinical journey. *Advances in Neural Information Processing Systems*, 37:87621–87646.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, and 1 others. 2024a. Me-llama: Foundation large language models for medical applications. *Research square*, pages rs–3.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and 1 others. 2024b. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Huihui Xu, Yuanpeng Nie, Hualiang Wang, Ying Chen, Wei Li, Junzhi Ning, Lihao Liu, Hongqiu Wang, Lei Zhu, Jiyao Liu, and 1 others. 2025a. Medground-r1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization. *arXiv preprint arXiv:2507.02994*.

Jie Xu, Lu Lu, Xinwei Peng, Jiali Pang, Jinru Ding, Lingrui Yang, Huan Song, Kang Li, Xin Sun, Shaoting Zhang, and 1 others. 2024. Data set and benchmark (medgpteval) to evaluate responses from large language models in medicine: evaluation development and validation. *JMIR Medical Informatics*, 12(1):e57674.

Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Xiangru Tang, Hang Wu, May D Wang, Peifeng Ruan, Donghan Yang, Tao Wang, and 1 others. 2025b. Medagentgym: Training llm agents for code-based medical reasoning at scale. *arXiv preprint arXiv:2506.04405*.

Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, and 1 others. 2020. An interpretable mortality prediction model for covid-19 patients. *Nature machine intelligence*, 2(5):283–288.

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan Zhao, Yixin Zhang, Renjun Zhang, and 1 others. 2024. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. *arXiv preprint arXiv:2406.13890*.

Hongzhou Yu, Tianhao Cheng, Ying Cheng, and Rui Feng. 2025a. Finemedlm-o1: Enhancing the medical reasoning ability of llm from supervised fine-tuning to test-time training. *arXiv preprint arXiv:2501.09213*.

Zhihao Yu, Yujie Jin, Yongxin Xu, Xiaoyun Zhang, and Yasha Wang. 2025b. Intellicare: Improving healthcare analysis with patient-level knowledge from large language models. *Knowledge-Based Systems*, page 114191.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, and 1 others. 2023. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, and 1 others. 2025a. Llmeval-med: A real-world clinical benchmark for medical llms with physician validation. *arXiv preprint arXiv:2506.04078*.

Weitong Zhang, Mengyun Qiao, Chengqi Zang, Steven Niederer, Paul M Matthews, Wenjia Bai, and Bernhard Kainz. 2025b. Multi-agent reasoning for cardiovascular imaging phenotype analysis. *arXiv preprint arXiv:2507.03460*.

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

Brian Nlong Zhao, Zilong Wang, XINYANG JIANG, Xufang Luo, Yifan Yang, Bo Li, Javier Alvarez-Valle, Matthew P Lungren, Dongsheng Li, and Lili Qiu. Benchmark dataset for radiology report generation with instructions and contexts.

Junjie Zhou, Yingli Zuo, Shichang Feng, Peng Wan, Qi Zhu, Daoqiang Zhang, and Wei Shao. 2025a. Pathcot: Chain-of-thought prompting for zero-shot pathology visual reasoning. *arXiv preprint arXiv:2507.01029*.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025b. Evolving language models without labels: Majority drives selection, novelty promotes variation. *arXiv preprint arXiv:2509.15194*.

Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. Diagnosisarena: Benchmarking diagnostic reasoning for large language models. *arXiv preprint arXiv:2505.14107*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.