

Quantifying the Reasoning Abilities of LLMs on Clinical Cases

Corresponding Author: Professor Weidi Xie

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

High-level Summary

This paper is a dataset paper. The paper proposes a new benchmark for medical reasoning in LLMs focusing on evaluating the quality of the reasoning processes. The authors claim that current benchmarks focus on final answers such as diagnostic accuracies, but disregard the reasoning process to reach those answers. The proposed benchmark MedR-Bench contains 1.4k real-world clinical cases. Each case contains patient information, a reasoning chain simulating the reasoning of a physician to analyze the patient's case and reach a diagnostic, and the final diagnosis or treatment plan.

They propose an agentic evaluation system based on LLMs to evaluate the medical reasoning performance on an LLM. They evaluate multiple both open and closed weights LLMs and analyze that they perform well on simple tasks such as diagnostics when all information is available but struggle with complex tasks such as recommending new tests or planning the treatment.

Contributions

The contributions of this paper are:

1. A medical reasoning benchmark for LLMs
2. Analysis of the medical reasoning performance of multiple popular LLMs.

Summary of Strengths

The paper covers an important evaluation domain: medical reasoning.

The paper proposes a new task that has been unexplored as far as I know, examination recommendation, and can be useful to develop future medical assistants.

Summary of Weaknesses

1. The main contribution of the paper is not novel. There are multiple medical reasoning benchmarks targeting the same or very similar problems and are not discussed in the paper. I believe a more in-detail comparison of these benchmarks with the proposed benchmark is needed to clarify the novelty and contributions.

- [1] proposes a chain of thought dataset for medical diagnosis. Each chain simulates the reasoning process of a physician to reach a diagnosis based on the patient's symptoms.

- [2] proposed an evaluation protocol for chain of thoughts in the medical domain. This protocol is not discussed in the paper, despite targeting the same problem. Similar to the submitted paper, [2] checks for correct reasoning steps and correct recalling of information.

- MedMCQA (citation 7 in the paper) and MEDQA (citation 6) indeed focus on evaluating the final outputs, but they also provide explanations of the answers. These explanations are reasoning chains that aim to reach a diagnostic or treatment plan based on some symptoms. Therefore, they can be used to analyze medical reasoning as in [2].

2. Missing baselines. Most of the baselines are general-domain LLMs, which are fine, but we also miss LLMs for the medical domain. For example, DiagnosisGPT [2] is trained to generate medical CoTs, so, in my opinion, it is a must have baseline for this paper. Baichuan is the only medical domain LLM used, but it has some issues detailed in the next point.

- <https://huggingface.co/FreedomIntelligence/DiagnosisGPT-34B>

- <https://huggingface.co/FreedomIntelligence/DiagnosisGPT-6B>

3. The baselines may be contaminated. The authors acknowledge that some models do not disclose their cut-off date and therefore it is possible that there is data contamination. Why not use other models such as DiagnosisGPT for the medical domain instead of Baichuan? Instead of Quen-QwQ, LLaMA 3.X. Using LLMs with guarantees that were not trained on the benchmark should be a priority.

4. The benchmark misses the human performance upper bound. For example, [2] provides a human upper bound. This upper bound is important to contextualize how far these LLMs are from the application in clinics.

[1] Chen, Junying et al. "CoD, Towards an Interpretable Medical Agent using Chain of Diagnosis." *ArXiv* abs/2407.13301 (2024).

[2] Liévin, Valentin, et al. "Can large language models reason about medical questions?." *Patterns* 5.3 (2024).

Questions and Comments

The evaluation protocol requires searching for documents on the internet using some API. Wouldn't this make the evaluation not reproducible? Another person next year may try to evaluate the same model on your benchmark and get different results because of this internet search. Websites are not static, they may change or may be down and this could affect the evaluation, right?

What are the generation parameters? (e.g., sampling temperature, seed, etc)

Why does performance on rare diseases not decrease? Is there any intuition from physicians behind this?

Why is there a page break in page 9?

In Examination Recommendation in 4.2, it is written that you build a patient agent with Prompt 7. However, that prompt states that the LLM is a medical expert providing guidance to a junior physician, not a patient. A better explanation of the role of the LLM in prompt 7 is needed.

Are the human evaluations assessing the correctness of the evaluation system conducted by physicians? Assessing the correctness requires domain knowledge, so I think it should be conducted by physicians.

Overall Clarity of the Paper

- Although the paper is in general clear, there are some details about the benchmark and baselines missing to interpret the results (section 2) and they are only provided at the end (section methods). This makes it difficult to interpret the results. For example, Baichuan is an LLM in the medical domain.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The authors introduce MedR-Bench, an LLM evaluation platform, including more than 1000 structured patient reports. Including common and rare disorder regarding different organ systems or clinical disciplines

The reasoning tasks refer to core tasks including examinations, diagnosis and treatment recommendation

They also introduce a comprehensive Reasoning Evaluator, a novel automated system that validates free-text reasoning processes calculating metrics for efficiency, factuality, and completeness.

Major issues

There is a clear lack of novelty and clinical relevance as there are papers published comparing LLMs with and/or without reasoning, including DeepSeek models for clinical tasks:

1: <https://www.nature.com/articles/s41591-025-03727-2>

2:<https://www.nature.com/articles/s41591-025-03726-3>

A number of different methodological ideas like 1: utilizing rare and frequent diseases, 2: including the three core tasks: examination, diagnosis and treatment planning have been applied (though with a different focus) in a previous evaluation regarding non-reasoning models by Sandmann et al., 2024 on <https://www.nature.com/articles/s41467-024-46411-8>. The authors should reference this and provide more related work as the manuscript in its current form currently express these steps as innovations established by them.

Evaluating medical output of LLMs with automatic LLM-based evaluation is scalable but still highly questionable as it lacks expert-validation. Even though ground-truth references are used, each individual LLM output is different and subtle semantic word-differences should be evaluated by medical experts in order to judge the individual output of LLMs. At least the reasoning evaluator should have been validated with expert-based separate evaluation process previously. Otherwise it is hard to estimate the clinical accuracy of presented evaluation findings.

The authors' reasoning and criteria for defining or classifying common and rare diseases were not clearly stated. If they only took information from medline plus this could be only vague information. Classification should be grounded on disease prevalences or incidences, citing the exact sources, as the numbers vary across different regions.

Minor:

The term "real-world" cases in the title and main text is confusing as it could imply real clinical cases from EHR or hospital systems. Instead, the work includes PMC-listed case reports, which only include highly processed and shortened/removed information. Thus it only includes information, which were deemed relevant for the publication of the case report and not necessarily relevant for the actual care of that patient. For the sake of LLM evaluation this approach is fine, but I strongly recommend to change the wording to avoid misleading information.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

High-level Summary

This paper is a dataset paper. The paper proposes a new benchmark for medical reasoning in LLMs focusing on evaluating the quality of the reasoning processes. The authors claim that current benchmarks use synthetic or exam-style questions and focus on final answers, but disregard the reasoning process to reach those answers (e.g., diagnosis). The proposed benchmark MedR-Bench contains 1.4k real-world clinical cases. Each case contains patient information, a reasoning chain simulating the reasoning of a physician to analyze the patient's case, propose examinations, reach a diagnostic, and a treatment plan.

The paper proposes an agentic evaluation system based on LLMs to evaluate the medical reasoning performance on an LLM. They evaluate multiple both open and closed weights LLMs and analyze that they perform well on simple tasks such as diagnostics when all information is available but struggle with complex tasks such as recommending new tests or planning the treatment.

The models solve the task in a pipeline. First, they solve the examination recommendation task with 2 settings: one-turn and free-turn. Then, they use these outputs as inputs of the second task: diagnostic decision making. For the final task, treatment planning, they don't not use prior responses, but instead, they use the ground-truth labels (oracle setting)

Contributions

The contributions of this paper are:

1. A medical reasoning benchmark for LLMs
2. Analysis of the medical reasoning performance of multiple popular LLMs.

Summary of Strengths

The paper covers an important evaluation domain, medical reasoning, covering the full spectrum of a clinical case: examination recommendations, diagnosis, and treatment plan.

Summary of Weaknesses

One of the key aspects of the paper is to evaluate the reasoning abilities of LLMs on clinical cases. However, in the case of reasoning models, they do not do this. They use the final answer (in the format of a reasoning chain) as the "reasoning" of the model. This is effectively a summary of the reasoning and not the real "reasoning." While I understand that o3 does not provide the reasoning chains, Deepseek does.

The implementation of the efficiency metric is biased, and especially affects reasoning models. The implementation

analyzes whether each reasoning step provides valuable information or not and this is decided by another LLM. This has the following problems.

A model X can generate 3 extremely verbose reasoning steps, but model Y can generate 4 much shorter reasoning steps. However, the verbosity of model A wouldn't be penalized because overall, the reasoning steps contribute to the reasoning. In other words, the implementation of the metrics is biased towards assigning the category "reasoning" to most steps, disregarding their verbosity (especially when evaluating the summaries of the thinking part of reasoning models, e.g., deepseek-r1, see weakness #1).

Since the authors evaluate the efficiency of the summary/final answer (see point #1) and not the thinking part of reasoning models, it is natural that there are no repetitions and redundancies, making the efficiency practically 100% according to their definition. However, this is far from the truth. Table 2 shows that Deepseek R1 671B and OpenAI o3-mini are the most efficient models (the evaluation is based on summaries of their real reasoning steps). However, Biachuan M1 has better examination recommendations than these models (Table 1) and is significantly smaller than them, hence it's more efficient. Efficiency should be calculated as task performance/cost, and cost should be the number of generated tokens, GPU time, or monetary cost, which are objective metrics.

Another problem with the implementation of the metrics is that the authors state that the reasoning quality in treatment planning is generally strong (L. 295) based on the high factuality scores (>95%). However, the accuracy of all models is below 30% (Table 6), so there is an apparent contradiction between reasoning factuality, reasoning correctness, and final performance, which may indicate that the evaluation of factuality or reasoning quality is not accurate or does not represent what the name suggests.

The results of the free-turn setting in recommended examinations (lines 192-204) give potentially illogical results, or at least they are not adequately explained. DiagnoseGPT achieves a 1-turn recall of 22.70% while the free-turn has a lower recall of 12.38%. How is it possible that adding more conversation turns asking for more examinations drops the recall? The first turn should be the same in both cases, so the recall in free-turn should be equal to or larger than. From the appendix, prompt 8 (1-turn) and prompt 10 (free-turn) are extremely similar, so I don't understand how the first turn of free-turn can lead to worse results than the first turn in 1-turn setting.

The results in diagnostic decision-making also exhibit inconsistent patterns. Deepseek achieves 71.79% acc in 1-turn setting and 76.18% in free-turn. However, this large improvement is inconsistent with the results from the prior (and dependent) step examination recommendation. The authors claim in L.199 that performance drops in free-turn and explain that it's due to repetitive loops, etc. Since the diagnostic decision depends on examination recommendation (it's a pipeline), there should be error propagation, so the diagnostic decision in free-turn should be equal or worse than in 1-turn. Intuitively, a worse quality of examinations should give a worse diagnosis, but the results seem otherwise and it's not explained.

The evaluation of the reasoning process is composed of efficiency, factuality, and completeness. However, S2.4 (analysis of the reasoning process for diagnosis) lacks an analysis of the efficiency and completeness. These analyses are claimed to be a major component of this paper, but are missing here. Similar problem in S2.5, where the analysis is superficial.

The human performance of treatment planning is surprisingly low (36% in Table 6). Yet, there is no analysis of the reasons behind it. According to S4.1, there is a different human analysis of the quality of the dataset that claims 96% of the treatment cases can pass human evaluation. Therefore, I wonder why that low 36% accuracy. Are many cases failing because only a specialist can solve them and the anotator was not specialised in that branch of medicine?

One of the novelty claims of the paper is the use of real clinical cases, while prior benchmarks use synthetic or medical exam-style questions (lines 372-279). However, this is not entirely true. Citation 27 (DiagnosisGPT) evaluates on DxDiag, which is a dataset with real conversions between patients and physicians.

Questions and Comments

Line 283: precision of recommended ... do you mean accuracy?

In line 185 you state that Deepseek achieves the highest recall at 43% demonstrating the ability to identify most examinations. This is less than half, so it's not most of them.

I believe it would be good to evaluate the models for the treatment planning using their responses from the prior tasks to "simulate the entire patient care journey" (line 11), but I guess the results would be too low due to error accumulation and that's why the authors use the ground truths.

Please indicate that your factuality metric evaluates the factuality of the reasoning steps that marked as "reasoning." Defining it as correct steps/efficiency is a bit odd

Overall Clarity of the Paper

The paper is in general clear.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The authors have significantly revised the manuscript according to the reviewer's issues. I commend the authors for huge amount of re-analysis and re-writing of the manuscript.

The reviewer finds the revised work relevant and impactful, especially due the involvement of human expert analysis and inclusion of further models. The reviewer is unsure regarding the true significance and novelty within the scope of Nature Communications in the light of related work.

Minor points:

- Why is the related work listed in the Discussion section? It should belong to the introduction?
- The authors have correctly removed "real-world" from their title but still write "Use of Real Clinical Cases" or using terms as "authentic". The authors should consider to change this as the cases were processed and do represent clinical reality. Further this should be underlined as a limitation in the discussion section.
- The authors mention in their rebuttal letter: "While we share certain methodological elements with [7], including the categorization of rare versus frequent diseases and focus on three core clinical tasks (examination recommendation, diagnostic decision-making, and treatment planning)". The reviewer suggests to explicitly mention these certain methodological elements while citing the source, ideally in the methods and/or related work.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)

Overall, this revision is an improvement of the manuscript and attachments.

Regarding Reviewers 1 comments:

- I agree with the authors that disentangling the chain of thoughts process would be valuable but not a must-have.
- Regarding the term efficiency, I understand that there might be some misunderstanding and that the LLM-driven metrics can pose underestimated risks. I suggest to adapt or add to the naming of the metrics e.g. in Figure 1: "Metrics on Reasoning" -> "LLM-powered Metrics on Reasoning". Please also consider this in the main-text
- The authors have appropriately updated their explanations regarding free-turn settings in the new manuscript.
- All other issues raised require some more explanations, which the authors have provided by adding to the manuscript or supplement.

The authors have added some more related work and cited them properly. However, the three critical stages of examination, diagnosis and treatment should reference to the prior work, where these stages were originally used in evaluating LLMs, in the Method section -> Under 4.2 Evaluation Framework:

"Three critical patient

stages are considered: examination recommendation, diagnostic decision-making, and treatment planning (e.g. based on or similar to [Sandmann et al.])"

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Point-point Respond to Referees: NCOMMS-25-22367A

We sincerely thank all reviewers for their valuable comments to improve the quality of our manuscripts. In the following, we aim to resolve the raised concerns point-by-point. In addition, we have also uploaded the revised manuscript, with **RED** to highlight the changes. Generally we have made the following changes:

- **More LLM baselines:** Following the reviewer’s suggestion, we have incorporated two medical domain LLMs—DiagnoseGPT[2] and MedGemma[3]—into our benchmark, with a clearly defined cut-off date. This addition expands the breadth of medical baselines in our evaluation.
- **Human reference:** We introduce new human reference standards in diagnosis and treatment, reflecting the performance of physicians with 5-year of clinical experience. This enhances our MedR-bench, providing a better showcase of the clinical capabilities of current LLMs.
- **Human evaluation:** We conducted a thorough human evaluation to assess the reliability of our cases and proposed agentic metrics, involving six physicians with five years of clinical experience. The evaluation was carried out in two stages: (i) case quality assessment, and (ii) agentic metric quality assessment. Our results demonstrate that 100% of diagnostic cases and 96% of treatment cases are rated as clinically adequate, and our agentic metrics exhibited higher consistency with manual majority vote results compared to individual physicians.
- **Efforts on metric reproducibility:** To evaluate the effects of Internet changes, we recalculated our metrics. The consistent results demonstrate the robustness and reliability of our proposed agentic metrics over time, with minimal impact from Internet updates. Additionally, all code, metric calculation procedures, and **retrieved knowledge snapshots** are openly available for full reproducibility.
- **More case studies on rare diseases:** For rare diseases, we conducted in-depth consultations with Dr. Yanjie Fan, a rare disease expert specializing in rare disease research, to provide insightful explanations for the robust performance of LLMs in this domain. The key insight is that diagnosing rare diseases is not necessarily more complex in reasoning, but rather challenging for physicians due to limited exposure and familiarity. In contrast, LLMs—trained on vast medical corpora—do not pose such limitation. We have included more case studies to illustrate these findings in greater detail.
- **Thorough and comprehensive discussions:** To address concerns regarding the lack of related work and novelty, we have conducted a comprehensive review of prior works and clearly highlighted our unique contribution.

Reviewer 1

Question 1.1: The main contribution of the paper is not novel. There are multiple medical reasoning benchmarks targeting the same or very similar problems, and are not discussed in the paper. I believe a more detailed comparison of these benchmarks with the proposed benchmark is needed to clarify the novelty and contributions.

- [2] proposes a chain of thought dataset for medical diagnosis. Each chain simulates the reasoning process of a physician to reach a diagnosis based on the patient’s symptoms.
- [4] proposed an evaluation protocol for chain of thoughts in the medical domain. This protocol is not discussed in the paper, despite targeting the same problem. Similar to the submitted paper, [4] checks for correct reasoning steps and correct recalling of information.
- MedMCQA (citation 7 in the paper) and MEDQA (citation 6) indeed focus on evaluating the final outputs, but they also provide explanations of the answers. These explanations are reasoning chains that aim to reach a diagnostic or treatment plan based on some symptoms. Therefore, they can be used to analyze medical reasoning as in [4].

Answer: Thanks for raising this valuable feedback. We believe our work is fundamentally novel from three aspects: **testing cases**, **evaluation tasks**, and **metrics**, which are completely different from the given references, as detailed below.

- **Use of real clinical cases:** **MedRBench** is built upon the case reports from real clinical patient (curated from PubMed Central Open Access (PMC-OA)). This is in contrast to prior benchmarks, for example, [2] is based on synthetic reverse-engineering of clinical encyclopedias, MedQA/MedMCQA are based on exam-style multiple-choice questions (MCQs). They are undoubtedly valuable, but often lack the nuance, ambiguity, and diversity inherent in clinical narratives. **MedRBench** captures the variability and complexity of real patient encounters, which we argue is critical for evaluating the LLM performance in clinically relevant settings.
- **Comprehensive clinical task coverage:** **MedRBench** evaluates models across the full spectrum of clinical reasoning: from **examination recommendation** to **diagnosis formulation** and **treatment planning**. This multi-stage structure offers an evaluation for holistic and diverse tasks, unlike previous works that typically focus on a narrow slice of the clinical workflow, for example, [2] on inquiry chains, or [4] on reasoning errors in answering MCQs for a limited subset of cases. Our benchmark thus enables a more robust and realistic assessment of LLMs’ potential to assist with full-cycle clinical decision-making.
- **Agentic and scalable reasoning metrics:** Critically, we introduce a novel agentic evaluation framework to measure reasoning quality in free-text outputs, a long-standing challenge in the field. Prior methods either rely on handcrafted scoring rules [2] or on labor-intensive manual annotations [4], which limit their scalability and generalizability. Our framework decomposes model reasoning into structured steps, classifies them as reasoning, citation, repetition, *etc.*, and quantifies their alignment with factual references by leveraging web-scale information retrieval. This approach enables scalable, reproducible, and domain-adaptable evaluation of reasoning fidelity, marking a methodological advancement.

We have also revised the main text accordingly, including an extra “Related Work and Novelty” paragraph in the discussion section (**Lines 359-393**).

Question 1.2: Missing baselines. Most of the baselines are general-domain LLMs, which are fine, but we also miss LLMs for the medical domain. For example, DiagnoseGPT is trained to generate medical CoTs, so, in my opinion, it is a must have baseline for this paper. Baichuan is the only medical domain LLM used, but it has some issues detailed in the next point.

Answer: We appreciate the reviewer’s suggestion. In response, we incorporate additional baselines into our revised experiments, including **DiagnoseGPT** [2] and **MedGemma** [3], both of which are designed for medical tasks. The updated performance of these models is shown in Respond Letter Table 3, 5, 7, and 8 for different clinical tasks respectively. For fair comparisons, we adopt the same prompts as other models.

- DiagnoseGPT performs competitively on the **Oracle Diagnosis** task, achieving an accuracy of 79.62%, on par with strong general-domain LLMs (Respond Letter Table 5). However, its performance drops significantly under free-form interaction settings, with accuracy falling to 39.60%. This discrepancy highlights a key limitation: DiagnoseGPT is specifically optimized for single-turn dialogues and struggles in multi-turn scenarios. In our trials, even when explicitly prompted to request additional information, the model tends to finalize diagnoses prematurely in the first turn, resulting in incomplete reasoning and suboptimal final generations. For MedGemma, it achieves 84.22% in oracle diagnosis comparable to other commercial models, like OpenAI-o3-mini (83.91%), as shown in Respond Letter Table 5. Similarly, its performance declines in the 1-turn and free-turn interaction settings, achieving accuracies of 61.23% and 64.23%, respectively, which lag behind the latest LLMs such as Gemini-2.0-FT and DeepSeek-R1. This underscores the importance of evaluating models in realistic, multi-turn clinical contexts and avoiding overgeneralization of their capabilities beyond their intended design.
- Regarding reasoning quality, as reported in Respond Letter Table 4, 6 and 8. DiagnoseGPT demonstrates strong reasoning performance. In the Oracle Diagnosis task (Respond Letter Table 6), it significantly outperforming general-domain baselines such as Qwen-QwQ, with **Efficiency** at 90.42%, **Factuality** at 93.72%, and **Completeness** at 79.20%. Notably, MedGemma exhibits even superior performance as one of the latest medical-specific LLMs, achieving comparable results to DeepSeek-R1 despite having only 27B parameters, with scores of 96.78% for **Efficiency**, 95.96% for **Factuality**, and 87.72% for **Completeness**. These results confirm that medical-specific LLMs, when trained on curated domain

corpora, can generate more factually accurate, comprehensive, and structured clinical reasoning.

We have revised the main manuscript to include these new baselines and insights (including the related tables and lines in the result section). We believe these additions further strengthen our conclusions and provide a more balanced and comprehensive evaluation landscape.

Question 1.3: The baselines may be contaminated. The authors acknowledge that some models do not disclose their cut-off date and therefore it is possible that there is data contamination. Why not use other models such as DiagnosisGPT for the medical domain instead of Baichuan? Instead of Qwen-QwQ, LLaMA 3.X. Using LLMs with guarantees that were not trained on the benchmark should be a priority.

Answer: We thank the reviewer for this important observation. As mentioned in **Question 1.2**, we have now included DiagnoseGPT and MedGemma, two medical LLMs with known data cut-off dates, in our experiments. Their results are clearly reported in the revised manuscript throughout the revised “Results” section.

We recognize the importance of using models with verifiable training data to avoid contamination. However, there is a trade-off between transparency and evaluating state-of-the-art performance, as many cutting-edge models lack detailed training disclosures. We initially included Baichuan and Qwen-QwQ due to their strong performance and recency, which reflect the models clinicians are more likely to use.

Importantly, these models did not significantly outperform MedGemma, which has a clear cutoff, and some are even worse. This suggests that our benchmark is not substantially impacted by data leakage, even when newer models are evaluated.

To address this concern, we have revised the text under “Limitations” in the manuscript (**Lines 397-401**) to distinguish between models with known and unknown training data provenance. We agree that future work should prioritize cleaner benchmarks using private clinical data and models with transparent training histories to ensure fairness and reliability.

Question 1.4: The benchmark misses the human performance upper bound. For example, [4] provides a human upper bound. This upper bound is important to contextualize how far these LLMs are from the application in clinics.

Answer: Thank you for the insightful comment. We have included a human reference standard for diagnosis and treatment tasks, to provide a more comprehensive context for evaluating large language models.

Specifically, we invited six physicians (each with five years of clinical experience) from Xin Hua Hospital, affiliated with Shanghai Jiao Tong University School of Medicine, to independently perform **oracle diagnosis** and **treatment tasks**. During this process, the physicians were allowed to use online resources but were restricted from accessing the original case reports. We recorded their average accuracy to better reflect the relative performance of current clinical LLMs.

In **Respond Letter Table 5**, for the oracle diagnosis task, human physicians achieve an average accuracy of 76.67% across all diseases and 73.33% for rare diseases. Interestingly, recent LLMs generally outperform individual physicians in diagnostic accuracy. We attribute this to the broad and systematic medical knowledge LLMs acquired at training, in contrast to the limitations of individual human expertise across multiple specialties. Therefore, the performance of a single physician should be viewed as a reference point rather than a definitive upper bound of human capability. Since all benchmark cases were successfully resolved by expert teams, producing definitive diagnoses and guideline-compliant treatment plans, the true human upper bound corresponds to solving all cases correctly within a multidisciplinary clinical workflow. This comparison is further discussed in detail in **Lines 252-258** of the revised manuscript.

In contrast, for the treatment tasks, human physicians still outperformed LLMs. As illustrated in **Respond Letter Table 8**, the physicians achieved 36.67% accuracy across all diseases and 30% for rare diseases. These results underscore the significant challenges that current LLMs face beyond diagnosis. This finding is discussed in **Lines 293-294**.

Question 1.5: The evaluation protocol requires searching for documents on the internet using some API. Wouldn't this make the evaluation not reproducible? Another person next year may try to evaluate the same model on your benchmark and get different results because of this internet search. Websites are not static, they may change or may be down and this could affect the evaluation, right?

Answer: Thank you for highlighting this point. We carry out the following steps to ensure reproducibility:

- Medical knowledge tends to be **factual**, *i.e.*, unlikely to be affected by web updates. To validate the robustness of our evaluation, we re-ran the reasoning factuality metric around 4 months later. As shown in **Response Letter Table 1**, the results demonstrate the overall scores remain highly consistent, over 90% of individual reasoning steps received the same true/false judgment. This suggests that web content changes have minimal impact on our metrics.
- We have open-sourced all evaluation code and released comprehensive snapshots of retrieved webpages and supporting knowledge used during evaluation, enabling others to reproduce our results under identical conditions. The code is available at <https://github.com/MAGIC-AI4Med/MedRBench> and the knowledge snapshots can be accessed at <https://huggingface.co/datasets/Henrychur/MedRBench-Knowledge-Snapshots>.

Finally, we emphasize that our main contribution is the design of the agentic evaluation protocol. As LLMs and web content evolve, our framework remains applicable and adaptable, offering a scalable way to align model reasoning with up-to-date medical knowledge. We have added these results in **Lines 536-543**.

Respond Letter Table 1 | Comparison of Factuality Scores in Two Evaluation Rounds. “Original” refers to the results obtained during the submission in February 2025, while “Latest” represents the results recalculated four months later in June 2025. The “Consistency” line represents the proportion of reasoning steps that received the same judgment across the two metric calculation procedures.

Times	Date	Examination Recommendation	Diagnostic Decision-making		Treatment Planning
		1-turn	1-turn	oracle	
Original	2025-2	96.79	87.15	95.03	94.59
Latest	2025-6	97.55	87.77	95.03	95.60
Consistency(%)	-	97.13	93.20	96.20	95.30

Question 1.6: What are the generation parameters? (e.g., sampling temperature, seed, etc)

Answer: For all evaluated models in the paper, we set the default temperature to 1.0, except for DeepSeek-R1, for which we use a temperature of 0.6, as their official documentation claims that 0.6 yields the best performance. When using large language models in the evaluation system as metric, we set the temperature to 0.0 across the board to minimize randomness. The random seed is set to 42. We have revised the manuscripts according, adding the details in **Supplementary Section A.2: “Generation Parameters”**.

Question 1.7: Why does performance on rare diseases not decrease? Is there any intuition from physicians behind this?

Answer: Thank you for this insightful question. To better understand this phenomenon, we consulted with a rare disease specialist with over 10 years of clinical experience at Xin Hua Hospital, Dr. Yanjie Fan, who has now been added as a co-author of the manuscript. Together, we conducted a thorough review of the rare disease cases in our dataset.

The key insight is that diagnosing rare diseases is **NOT** inherently more complex in terms of reasoning. Instead, the primary challenge for human physicians lies in limited exposure and familiarity due to the low incidence of such conditions. In contrast, LLMs trained on large-scale medical corpora are not constrained by case frequency or clinical exposure. As long as a rare disease follows a recognizable diagnostic pathway-such as guideline-driven decision-making or the presence of specific test results-LLMs can perform comparably to how they do on more common diseases.

This observation aligns with our empirical findings: recent LLMs demonstrate broad coverage of medical knowledge, enabling them to effectively diagnose many rare conditions. It supports our claim that LLMs possess a strong foundational understanding of medicine in **Lines 348-352** of the main manuscript.

To further illustrate this point, we have added case studies in **Supplementary Section A.1.4**: “Rare Disease Cases with Specific Tests”, including Figure 2. In one such case, both DeepSeek-V3 and OpenAI-o3-mini successfully diagnosed adult nephroblastoma—a rare renal malignancy—by recognizing its hallmark histopathological pattern.

Question 1.8: Why is there a page break in page 9?

Answer: Thanks for the careful review. We originally aimed to split the method section for better readability. In the revision, we have removed the page break.

Question 1.9: In Examination Recommendation in 4.2, it is written that you build a patient agent with Prompt 7. However, that prompt states that the LLM is a medical expert providing guidance to a junior physician, not a patient. A better explanation of the role of the LLM in prompt 7 is needed.

Answer: Thank you for pointing this out. We agree that the term “patient agent” may have been misleading. To clarify, this module is responsible for extracting and presenting patient information based on the case report—it does not simulate a patient’s speech or behavior.

To accurately align queries with matched lab event results, the agent should still be prompted as a medical expert. This ensures the interaction process can effectively simulate a scenario where a patient completes the recommended examinations and presents their professional clinical reports to a clinician.

We have explained this more in the revised manuscripts in the method section (**Lines 461-467**).

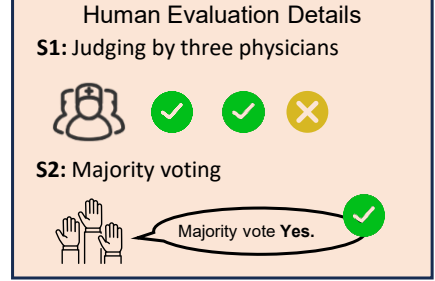
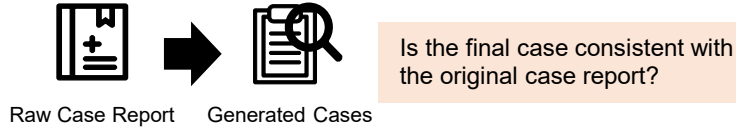
Question 1.10: Are the human evaluations assessing the correctness of the evaluation system conducted by physicians? Assessing the correctness requires domain knowledge, so I think it should be conducted by physicians.

Answer: Thank you for raising this concern. Our human evaluations were conducted in collaboration with six licensed physicians from Xinhua Hospital, Shanghai Jiao Tong University School of Medicine.

The evaluation involved two stages as illustrated in **Respond Letter Figure 1**:

- **Case quality assessment:** Physicians reviewed a random set of 100 generated cases. Each case was independently evaluated by three physicians, with final labels determined via majority vote. Results show that 100% of diagnostic and 96% of treatment cases were rated clinically adequate, confirming the reliability of our case set.
- **Metric quality assessment:** Physicians also assessed the correctness of our agentic evaluation metrics. We measured agreement between: (i) LLM-based scores and physician majority votes, and (ii) individual physicians and the majority vote. High consistency across both diagnosis and treatment (see **Response Letter Table 2**) supports the validity of our automated evaluation method. Additionally, we report the “Precision” and “Recall” scores, treating the physician majority vote as the ground truth and the agentic evaluation as predictions. These scores can further indicate the failure modes that automatic agentic evaluation metrics are prone to exhibit.

a. Case Quality Checking

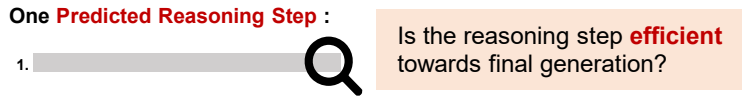


b. Agentic Metric Quality Checking

✓ Accuracy:



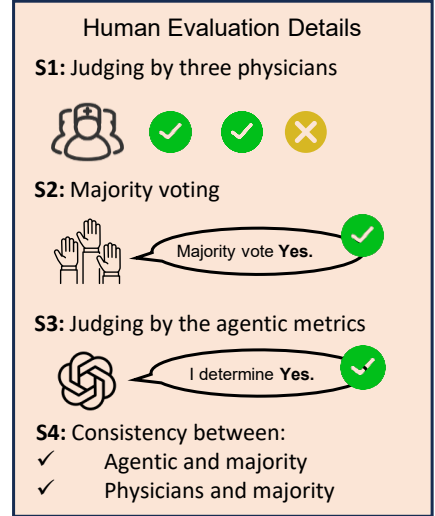
✓ Efficiency:



✓ Factuality:



✓ Completeness:



Respond Letter Figure 1 | The demonstration of human evaluation pipeline. We evaluate our automated pipeline from two key perspectives: case quality and agentic metric quality. **a** Case quality is assessed by having three physicians review whether the final patient case aligns with the raw case report, with the majority vote determining the final accuracy. **b** For agentic metric quality, we evaluate the model’s final predictions and reasoning steps based on accuracy and reasoning step-wise efficiency, factuality, and completeness, respectively. Physicians rate whether the predictions are correct compared to the ground truth or assess each provided reasoning step for whether it is correct, efficient, and covered (the referenced reasoning step is given). Then, we similarly obtain the majority vote results and calculate consistency between them and the agentic results. For reference, we also calculate the agreement with the majority vote to better demonstrate whether our agentic rating procedure is reliable compared to human beings.

Respond Letter Table 2 | Comparison between agentic pipeline and real physicians in evaluating final generation and reasoning processes. “Consistency” quantifies the alignment between agentic evaluation results or individual physician results and the final manual majority rating results. “Precision” indicates the fraction of model-predicted positive cases that are true positives based on manual majority voting. “Recall” measures the fraction of actual positive cases that are successfully detected by the model.

Metric	Patient Stages	Agentic vs. Majority			Individual vs. Majority
		Precision (%)	Recall (%)	Consistency (%)	Consistency (%)
Accuracy	Diagnostic Decision-Making	97.78	95.65	94.00	92.00
	Treatment Planning	93.75	71.43	86.00	85.33
Efficiency	Diagnostic Decision-Making	97.67	97.67	95.56	94.83
	Treatment Planning	96.15	98.04	94.34	93.68
Completeness	Diagnostic Decision-Making	99.19	92.79	91.78	91.78
	Treatment Planning	99.01	84.75	86.33	84.67
Factuality	Diagnostic Decision-Making	98.84	98.84	97.75	96.24
	Treatment Planning	98.95	96.91	96.00	92.00

As shown in **Response Letter Table 2**, our evaluation system demonstrates high consistency with physicians' majority votes across most tasks—approximately 95% agreement. Notably, across all evaluated metrics, the consistency between our automated evaluation system and the physician majority vote is comparable to that between an individual physician and the majority vote. For example, our system achieves 94.00% diagnostic accuracy compared to 92.00% for individual physicians, and 86.00% treatment accuracy compared to 85.33%. These results suggest that our agentic metrics are on par with human evaluators in terms of reliability.

To further understand the automated evaluation process, we analyze precision and recall. We find that the agentic system exhibits high precision (93.75%) but relatively lower recall (71.43%) in treatment accuracy, indicating that while the evaluator is generally accurate when it identifies a response as correct, it tends to be overly conservative. This leads to lower recall scores, suggesting that the system prioritizes reliability over inclusiveness in its judgments.

We have clarified this in the manuscript to emphasize that the evaluation framework was validated with physician oversight, and that its performance aligns closely with human standards (**Lines 559-593**).

Question 1.11: Although the paper is in general clear, there are some details about the benchmark and baselines missing to interpret the results (section 2) and they are only provided at the end (section methods). This makes it difficult to interpret the results. For example, Baichuan is an LLM in the medical domain.

Answer: Thanks for the constructive feedback. We have revised the manuscript accordingly. Specifically, we summarize the key aspects of the benchmark design and baseline model characteristics earlier in the Results section in **Lines 68-89**. Additionally, we have added cross-references between the Results and Methods sections to further improve clarity, in **Line 90**.

Reviewer 2

Reviewer 2 co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer 3

Question 3.1: There is a clear lack of novelty and clinical relevance as there are papers [6, 8] published comparing LLMs with and/or without reasoning, including DeepSeek models for clinical tasks. A number of different methodological ideas like 1: utilizing rare and frequent diseases, 2: including the three core tasks: examination, diagnosis and treatment planning have been applied (though with a different focus) in a previous evaluation regarding non-reasoning models by [7]. The authors should reference this and provide more related work as the manuscript in its current form currently express these steps as innovations established by them.

Answer: Thank you for highlighting this point. We are aware of the two excellent papers published on Nature Medicine [6, 8], April 23, 2025. However, note that, this is about two months after our submission. Therefore, we believe the criticism on novelty is invalid, on the contrary, this precisely reflects the quality of our paper, matching those on Nature Medicine. We have added [6, 8] as concurrent work and discuss the differences in “Related Work and Novelty” **Lines 360-363**.

While we share certain methodological elements with [7], including the categorization of rare versus frequent diseases and focus on three core clinical tasks (examination recommendation, diagnostic decision-making, and treatment planning), our work presents several key innovations: (i) our evaluation specifically targets the reasoning process quality rather than solely final generation accuracy; (ii) we have developed an agentic evaluation system for automated assessment, addressing the scalability limitations of physician-dependent evaluations used in previous studies. Overall, our benchmark is specifically designed to evaluate reasoning capabilities in complex clinical scenarios, which represents a novel contribution to the field. We also have

included [7] as our related work in “Related Work and Novelty” (**Lines 357-385**).

As clarified in **Q1.1**, our work differs in key aspects:

- **Dataset:** We construct a high-quality benchmark with structured patient cases that include both diagnostic and therapeutic decisions, each supported by detailed professional reasoning mined from clinical case reports.
- **Evaluation scope:** Our benchmark spans the full clinical workflow—from initial examination to diagnosis and treatment planning—covering reasoning tasks that others often omit or treat separately.
- **Evaluation method:** We propose a scalable, automated agentic evaluation framework that objectively assesses multi-step clinical reasoning with high agreement to expert physician judgment.

We have updated the manuscript to explicitly discuss overlapping and differentiating elements with recent work (see Related Work, **Lines 359-393**).

Question 3.2: Evaluating medical output of LLMs with automatic LLM-based evaluation is scalable but still highly questionable as it lacks expert-validation. Even though ground-truth references are used, each individual LLM output is different, and subtle semantic word-differences should be evaluated by medical experts in order to judge the individual output of LLMs. At least the reasoning evaluator should have been validated with expert-based separate evaluation process previously. Otherwise it is hard to estimate the clinical accuracy of presented evaluation findings.

Answer: Thank you for raising the constructive suggestion. We have conducted human evaluation in collaboration with six licensed physicians, each with over 5 years of clinical experience from Xinhua Hospital, Shanghai Jiao Tong University School of Medicine.

The evaluation consisted of two stages, as shown in Response Letter Figure 1 and detailed in **Q1.10**. In summary, we evaluate two key aspects: case quality and metric quality. In case quality, we show that 100% of diagnostic and 96% of treatment cases were rated clinically adequate, confirming the reliability of our case set. In metrics, as demonstrated in Response Letter Table 2, all agentic metrics can achieve higher consistency with majority vote results compared to individual physicians, demonstrating their reliability and clinical soundness.

We have added these human evaluation results and discussion in the method section (**Lines 559-593**).

Question 3.3: The authors’ reasoning and criteria for defining or classifying common and rare diseases were not clearly stated. If they only took information from medline plus this could be only vague information. Classification should be grounded on disease prevalences or incidences, citing the exact sources, as the numbers vary across different regions.

Answer: Thanks for the suggestion. We have explained this in **Lines 397-401** of the method section.

To identify rare diseases, we use the Orphanet Rare Disease Ontology (ORDO¹) [9], which provides a rare disease list based on established prevalence statistics. Although ORDO follows the European Union’s definition of rare diseases, it has become a widely adopted ontology framework for research worldwide.

Our rare disease identification process employs a rigorous two-step approach to ensure accuracy and reliability. First, we conduct rule-based screening using SciSpacy [5] to extract all relevant UMLS [1] Concept Unique Identifiers (CUIs) from each patient case. Cases are retained only if any of the extracted CUIs are listed in ORDO. Second, we utilize GPT-4o to confirm that the case is truly focused on the identified rare disease items, ensuring its relevance.

¹<http://www.ebi.ac.uk/ols4/ontologies/ordo>

Question 3.4: The term “real-world” cases in the title and main text is confusing as it could imply real clinical cases from EHR or hospital systems. Instead, the work includes PMC-listed case reports, which only include highly processed and shortened/removed information. Thus, it only includes information, which were deemed relevant for the publication of the case report and not necessarily relevant for the actual care of that patient. For the sake of LLM evaluation this approach is fine, but I strongly recommend to change the wording to avoid misleading information.

Answer: Thank you very much for the constructive suggestions. We have replaced “real-world” with “clinical” to better emphasize that the cases are derived from the clinical pipeline presented in case reports, rather than focusing on their origin in real-world settings.

Respond Letter Table 3 | Results for recommended examinations on both “all diseases” and “rare diseases” in 1-turn and free-turn settings. The 0.95 confidence interval (CI) is reported in brackets.

Method	Model Size	All Diseases				Rare Diseases			
		1-turn		Free-turn		1-turn		Free-turn	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Close-source Reasoning LLMs									
OpenAI-o3-mini	–	33.75	38.47	33.57	38.22	33.73	38.25	32.94	37.24
		(32.21, 35.29)	(36.85, 40.10)	(31.96, 35.19)	(36.59, 39.85)	(31.60, 35.86)	(36.03, 40.47)	(30.77, 35.11)	(35.04, 39.45)
Gemini-2.0-FT	–	22.77	43.12	24.09	39.88	21.93	42.96	22.73	39.66
		(21.42, 24.12)	(41.41, 44.83)	(22.54, 25.63)	(38.21, 41.55)	(20.21, 23.64)	(40.56, 45.36)	(20.72, 24.75)	(37.30, 42.01)
Open-source Reasoning LLMs									
DeepSeek-R1	671B	31.50	43.61	32.23	40.67	30.92	43.73	31.64	40.18
		(29.97, 33.03)	(41.92, 45.30)	(30.58, 33.89)	(38.92, 42.42)	(28.81, 33.03)	(41.42, 46.04)	(29.36, 33.93)	(37.77, 42.59)
Qwen-QwQ	32B	24.43	39.90	25.39	40.83	24.08	39.44	25.47	40.79
		(23.19, 25.67)	(38.22, 41.57)	(24.16, 26.63)	(39.15, 42.51)	(22.36, 25.80)	(37.15, 41.74)	(23.76, 27.18)	(38.51, 43.08)
Baichuan-M1	14B	41.78	37.88	41.99	36.99	41.58	38.09	42.06	37.73
		(39.91, 43.64)	(36.31, 39.45)	(40.11, 43.87)	(35.40, 38.59)	(39.05, 44.12)	(36.01, 40.18)	(39.57, 44.56)	(35.58, 39.89)
DiagnoseGPT	34B	30.73	22.70	17.56	12.38	29.83	21.52	16.66	10.95
		(28.74, 32.73)	(21.14, 24.25)	(15.75, 19.37)	(11.04, 19.37)	(27.15, 32.52)	(19.40, 23.64)	(14.19, 19.14)	(9.27, 12.64)
MedGemma	27B	22.65	39.38	22.20	34.24	21.79	38.66	21.40	33.95
		(21.25, 24.04)	(37.62, 41.13)	(20.73, 23.68)	(32.43, 36.04)	(19.90, 23.67)	(36.22, 41.11)	(19.43, 23.37)	(31.49, 36.41)

Respond Letter Table 4 | Results for examination recommendation reasoning on both “all diseases” and “rare diseases” in the 1-turn settings. For DeepSeek-R1, we assess its two types of reasoning, *i.e.*, those presented in the formal answer part and the default thinking part, recorded as “XX /xx” respectively. The 0.95 confidence interval (CI) is reported in brackets.

Method	Model Size	All Diseases		Rare Diseases	
		Efficiency	Factuality	Efficiency	Factuality
Close-source Reasoning LLMs					
OpenAI-o3-mini	–	95.17	95.89	92.73	95.90
		(94.47, 95.87)	(95.18, 96.61)	(91.16, 94.29)	(94.93, 96.87)
Gemini-2.0-FT	–	95.42	98.75	84.22	98.98
		(94.79, 96.05)	(98.38, 99.13)	(82.62, 85.83)	(98.52, 99.43)
Open-source Reasoning LLMs					
DeepSeek-R1	671B	98.59 /89.75	96.79 /95.38	95.96 /88.91	97.02 /95.10
		(98.17, 99.00)	(96.17, 97.41)	(94.83, 97.08)	(96.17, 97.87)
Qwen-QwQ	32B	86.53	94.87	76.34	94.53
		(85.34, 87.71)	(94.09, 95.64)	(74.02, 78.67)	(93.45, 95.62)
Baichuan-M1	14B	90.66	96.87	84.37	97.06
		(89.66, 91.66)	(96.35, 97.39)	(82.26, 86.48)	(96.36, 97.76)
DiagnoseGPT	34B	80.63	94.49	81.53	94.07
		(78.39, 82.87)	(93.35, 95.63)	(78.43, 84.62)	(92.33, 95.80)
MedGemma	27B	90.88	97.23	90.79	97.65
		(90.16, 91.60)	(96.72, 97.74)	(89.73, 91.85)	(96.99, 98.32)

Respond Letter Table 5 | Results for disease diagnosis in the 1-turn, free-turn, and oracle evaluation settings on both “all diseases” and “rare diseases”. The 0.95 confidence interval (CI) is reported in brackets.

Method	Model Size	All Diseases			Rare Diseases		
		1-turn	Free-turn	Oracle	1-turn	Free-turn	Oracle
		Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
Human Reference							
5-year physician	-	-	-	76.67	-	-	73.33
Close-source Reasoning LLMs							
OpenAI-o3-mini	-	64.99 (61.97, 68.02)	67.19 (64.21, 70.17)	83.91 (81.58, 86.24)	63.75 (59.48, 68.01)	65.99 (61.78, 70.19)	85.54 (82.43, 88.65)
Gemini-2.0-FT	-	68.55 (65.60, 71.49)	68.76 (65.81, 71.70)	86.83 (84.69, 88.98)	66.40 (62.20, 70.59)	68.43 (64.31, 72.56)	88.60 (85.78, 91.41)
Open-source Reasoning LLMs							
DeepSeek-R1	671B	71.79 (68.93, 74.64)	76.18 (73.47, 78.88)	89.76 (87.84, 91.68)	70.67 (66.63, 74.71)	77.60 (73.90, 81.30)	91.04 (88.51, 93.57)
Qwen-QwQ	32B	63.74 (60.69, 66.79)	74.71 (71.95, 77.47)	85.06 (82.80, 87.32)	64.15 (59.90, 68.41)	73.93 (70.03, 77.83)	86.15 (83.09, 89.21)
Baichuan-M1	14B	66.77 (63.78, 69.76)	61.23 (58.14, 64.33)	84.43 (82.13, 86.73)	67.21 (63.04, 71.38)	60.69 (56.36, 65.03)	88.39 (85.55, 91.23)
DiagnoseGPT	34B	54.44 (51.28, 57.60)	39.60 (36.50, 42.71)	79.62 (77.07, 82.18)	52.34 (47.91, 56.78)	35.44 (31.19, 39.68)	79.63 (76.06, 83.21)
MedGemma	27B	61.23 (58.14, 64.33)	64.23 (61.18, 67.27)	84.22 (81.91, 86.54)	60.69 (56.36, 65.03)	65.17 (60.94, 69.40)	86.35 (83.31, 89.40)

Respond Letter Table 6 | Results for diagnostic decision-making reasoning in the 1-turn and oracle evaluation settings on “all diseases”. For **DeepSeek-R1**, we assess its two types of reasoning, *i.e.*, those presented in the formal answer part and the default thinking part, recorded as “XX /xx” respectively. The 0.95 confidence interval (CI) is reported in brackets.

Method	Model Size	1-turn reasoning			Oracle reasoning		
		Efficiency	Factuality	Completeness	Efficiency	Factuality	Completeness
Close-source Reasoning LLMs							
OpenAI-o3-mini	–	91.59	83.15	50.87	94.33	94.94	75.42
		(90.37, 92.81)	(81.29, 85.01)	(48.60, 53.13)	(93.61, 95.05)	(94.23, 95.64)	(74.14, 76.70)
Gemini-2.0-FT	–	83.77	87.17	54.45	95.89	98.23	83.28
		(82.56, 84.97)	(85.52, 88.83)	(52.20, 56.69)	(95.35, 96.43)	(97.84, 98.62)	(82.17, 84.40)
Open-source Reasoning LLMs							
DeepSeek-R1	671B	95.86/88.24	87.15/85.90	54.88/67.59	97.17/89.73	95.03/96.31	78.27/90.79
		(95.02, 96.71)	(85.54, 88.76)	(52.77, 56.99)	(96.65, 97.70)	(94.34, 95.71)	(77.17, 79.37)
Qwen-QwQ	32B	76.97	88.14	66.94	71.20	84.02	79.97
		(75.30, 78.64)	(86.48, 89.80)	(63.83, 70.04)	(69.46, 72.94)	(82.23, 85.80)	(78.11, 81.83)
Baichuan-M1	14B	82.91	88.62	53.43	92.80	96.84	75.11
		(81.34, 84.48)	(87.23, 90.01)	(51.17, 55.68)	(92.09, 93.51)	(96.35, 97.32)	(73.87, 76.36)
DiagnoseGPT	34B	80.99	89.14	25.44	90.42	93.72	79.20
		(78.78, 83.19)	(87.58, 90.70)	(23.39, 27.50)	(89.12, 91.73)	(92.87, 94.56)	(77.91, 80.48)
MedGemma	27B	90.22	88.23	49.91	96.78	95.96	87.72
		(89.42, 91.01)	(87.02, 89.44)	(47.61, 52.21)	(96.33, 97.24)	(95.41, 96.51)	(86.69, 88.74)

Respond Letter Table 7 | Results for diagnostic decision-making reasoning in the 1-turn and oracle evaluation settings on “rare diseases”. For **DeepSeek-R1**, we assess its two types of reasoning, *i.e.*, those presented in the formal answer part and the default thinking part, recorded as “XX /xx” respectively. The 0.95 confidence interval (CI) is reported in brackets.

Method	Model Size	1-turn reasoning			Oracle reasoning		
		Efficiency	Factuality	Completeness	Efficiency	Factuality	Completeness
Close-source Reasoning LLMs							
OpenAI-o3-mini	–	92.73	81.28	49.83	94.80	95.02	76.44
		(91.16, 94.29)	(78.58, 83.99)	(46.59, 53.07)	(93.83, 95.77)	(94.14, 95.89)	(74.74, 78.14)
Gemini-2.0-FT	–	84.22	85.93	54.06	96.45	98.39	84.30
		(82.62, 85.83)	(83.53, 88.33)	(50.85, 57.28)	(95.74, 97.16)	(97.89, 98.89)	(82.75, 85.85)
Open-source Reasoning LLMs							
DeepSeek-R1	671B	95.96/88.91	84.76/84.70	54.10/67.54	97.61/90.33	94.75/96.83	79.00/91.14
		(94.83, 97.08)	(82.30, 87.22)	(51.05, 57.15)	(96.95, 98.27)	(93.77, 95.74)	(77.52, 80.48)
Qwen-QwQ	32B	76.34	89.05	66.53	72.25	84.30	80.70
		(74.02, 78.67)	(86.79, 91.31)	(62.14, 70.92)	(69.83, 74.68)	(81.83, 86.77)	(78.16, 83.24)
Baichuan-M1	14B	84.37	88.90	53.11	93.94	96.91	76.05
		(82.26, 86.48)	(87.06, 90.73)	(49.95, 56.27)	(93.07, 94.81)	(96.26, 97.56)	(74.31, 77.78)
DiagnoseGPT	34B	81.82	88.09	25.01	91.87	93.57	79.02
		(78.80, 84.85)	(85.85, 90.33)	(22.17, 27.85)	(90.18, 93.56)	(92.45, 94.69)	(77.20, 80.83)
MedGemma	27B	89.87	88.02	50.18	96.99	95.72	88.77
		(88.72, 91.02)	(86.31, 89.72)	(46.89, 53.47)	(96.41, 97.57)	(94.95, 96.48)	(87.46, 90.08)

Respond Letter Table 8 | Results for treatment planning on both “all diseases” and “rare diseases”. We calculate metrics for both the final generation and reasoning processes. For **DeepSeek-R1**, we have assessed its two types of reasoning, *i.e.*, those presented in the formal answer part and the default thinking part, recorded as “XX /xx” respectively. The 0.95 confidence interval (CI) is reported in brackets.

Method	Model Size	All Diseases				Rare Diseases			
		Treatment Plans	Reasoning Processes			Treatment Plans	Reasoning Processes		
		Accuracy	Efficiency	Factuality	Completeness	Accuracy	Efficiency	Factuality	Completeness
Human Reference									
5-year physician	-	36.67	-	-	-	30.00	-	-	-
Close-source Reasoning LLMs									
OpenAI-o3-mini	-	27.03 (23.09, 30.98)	94.67 (93.90, 95.44)	96.77 (96.06, 97.48)	70.63 (68.55, 72.71)	23.17 (16.65, 29.70)	95.06 (93.79, 96.33)	96.81 (95.51, 98.11)	69.86 (66.33, 73.38)
Gemini-2.0-FT	-	25.66 (21.80, 29.52)	93.66 (92.82, 94.49)	96.96 (96.34, 97.59)	75.89 (73.81, 77.98)	23.78 (17.20, 30.37)	94.41 (93.07, 95.74)	96.68 (95.57, 97.78)	77.10 (73.68, 80.51)
Open-source Reasoning LLMs									
DeepSeek-R1	671B	30.51 (26.43, 34.58)	95.25 /88.93 (94.43, 96.08)	94.59 /95.93 (93.63, 95.56)	68.08 /80.48 (65.97, 70.20)	27.27 (20.41, 34.14)	95.37 /89.19 (93.92, 96.82)	95.22 /95.01 (93.62, 96.47)	68.28 /81.05 (64.57, 71.99)
Qwen-QwQ	32B	20.89 (17.12, 24.66)	84.76 (83.42, 86.10)	94.40 (93.44, 95.36)	77.66 (75.36, 79.96)	19.61 (13.25, 25.97)	83.31 (81.08, 85.54)	94.05 (92.00, 96.10)	78.74 (74.85, 82.63)
Baichuan-M1	14B	30.65 (26.57, 34.72)	88.47 (87.37, 89.57)	96.56 (95.80, 97.31)	70.18 (68.04, 72.32)	30.30 (23.22, 37.39)	87.87 (85.96, 89.77)	95.97 (94.62, 97.31)	69.56 (65.91, 73.21)
DiagnoseGPT	34B	17.74 (14.37, 21.12)	89.21 (86.87, 91.55)	92.86 (91.46, 94.26)	53.86 (50.80, 56.93)	14.55 (9.11, 19.98)	89.33 (85.61, 93.05)	91.47 (88.87, 94.06)	52.25 (46.91, 57.59)
MedGemma	27B	25.05 (21.22, 28.88)	96.53 (95.84, 97.23)	95.62 (94.66, 96.59)	71.70 (68.91, 74.49)	21.95 (15.55, 28.35)	96.91 (95.77, 98.05)	95.14 (93.30, 96.98)	70.79 (65.76, 75.82)

Patient Case			
Case ID	PMC11492433		
Category	- Body system: Kidneys and Urinary System - Disorders and conditions: Cancers - Rare Disease-related		
Case Summary	Patient Information: - 48-year-old male - Chief Complaint: Persistent left abdominal pain - History of Present Illness: Six-month history of progressively worsening left abdominal pain, with associated unintentional weight loss and increasing fatigue. - Past Medical History: None noted (non-contributory) - Personal History: None provided - Family History: None mentioned - Physical Examination: No specific physical abnormalities beyond the chief complaint of abdominal pain. Ancillary Tests: - Pre-operative laboratory tests: Mild anemia (Hb 10.5 g/dL) without other hematologic or metabolic abnormalities. Renal and liver function tests were normal. - Imaging: CT scan revealed a large (12 cm), heterogeneous mass in the left kidney with solid and cystic components, irregular borders, enlarged regional peri-hilar and left aortic lymph nodes (1.2 cm), and suspected local extension into perinephric fat. No distant metastases identified. - Histopathology: Post-nephrectomy histology confirmed nephroblastoma with a triphasic pattern (blastemal, stromal, and epithelial components). Regional lymph nodes exhibited metastatic involvement. Pathologic staging: pT2b N2 (local extension to the peri-hilar region with metastatic node involvement). No perinephric fat invasion or thrombus present. Surgical margins were negative.		
Reasoning Processes	Differential diagnosis: 1. Renal Cell Carcinoma (RCC) - RCC is the most common adult renal cancer and closely resembles nephroblastoma on imaging, presenting as a heterogeneous renal mass with possible lymph node involvement. However, histopathological analysis confirmed a triphasic nephroblastoma pattern and specific immunohistochemical markers, which ruled out RCC as a diagnosis. 2. Transitional Cell Carcinoma - Renal tumors involving the renal pelvis could suggest transitional cell carcinoma. However, imaging and histopathology findings localized the tumor strictly within the kidney and demonstrated features inconsistent with transitional cell carcinoma, excluding it as a possibility. 3. Adult Nephroblastoma/Wilms' Tumor - The patient's presentation, including a large renal mass, triphasic histopathologic pattern, and lymph node involvement, closely matched the typical features of rare adult nephroblastoma. Diagnosis was definitively confirmed based on the histological triphasic pattern and pathologic staging. 4. Other Rare Renal Neoplasms (e.g., Sarcoma, Lymphoma) - Rare possibilities such as sarcoma or lymphoma were considered based on the patient's age and presentation. However, histopathological findings definitively confirmed nephroblastoma, ruling out a mesenchymal or lymphoid origin.		
Diagnosis Results	Adult nephroblastoma (Wilms' tumor)		
Generation			
	Ground Truth	DeepSeek-R1	OpenAI-o3-mini
Diagnostic Decision-Making	Reasoning: <Step 1> Renal Cell Carcinoma (RCC) was considered because it is the most common adult renal cancer and closely resembles nephroblastoma on imaging, with a heterogeneous renal mass and possible lymph node involvement. <Step 2> Renal Cell Carcinoma (RCC) was excluded due to histopathological confirmation showing a triphasic nephroblastoma pattern and specific immunohistochemical markers consistent with nephroblastoma. <Step 3> Transitional Cell Carcinoma was considered because renal tumors involving the renal pelvis could suggest this diagnosis. <Step 4> Transitional Cell Carcinoma was excluded as imaging and histopathology localized the tumor strictly within the kidney and demonstrated features inconsistent with this carcinoma. <Step 5> Adult Nephroblastoma/Wilms' Tumor was considered due to the patient's presentation of a large renal mass, histopathologic triphasic pattern, and lymph node involvement, which matched typical features of adult nephroblastoma. <Step 6> Adult Nephroblastoma/Wilms' Tumor was confirmed based on the histological triphasic pattern and pathologic staging. <Step 7> Other rare renal neoplasms, such as sarcoma or lymphoma, were considered due to the age and presentation of the patient. <Step 8> Other rare renal neoplasms were excluded as histopathology definitively confirmed nephroblastoma rather than a mesenchymal or lymphoid origin. <Step 9> The diagnosis of adult nephroblastoma (Wilms' tumor) was confirmed post-nephrectomy by histopathology with the characteristic triphasic pattern, including blastemal, stromal, and epithelial components. <Step 10> The absence of fat invasion, negative surgical margins, and lymph node involvement supported the pathological staging as pT2b N2. Disease Diagnosis: Adult nephroblastoma (Wilms' tumor)	Reasoning: <step 1> The patient's persistent left abdominal pain, weight loss, fatigue, and imaging findings of a large heterogeneous renal mass with local extension and lymphadenopathy strongly suggest a malignant renal neoplasm. <step 2> The histopathologic confirmation of a triphasic pattern (blastemal, stromal, epithelial) is diagnostic of nephroblastoma (Wilms tumor), a tumor classically seen in children but rarely in adults. <step 3> Regional lymph node involvement (N2) and local extension (pT2b) confirm advanced disease without distant metastases. The absence of perinephric fat invasion or thrombus and negative surgical margins indicate localized resectability. <step 4> The combination of clinical, imaging, and histologic findings rules out other renal malignancies (e.g., renal cell carcinoma) and confirms nephroblastoma with metastatic lymph node involvement. Disease Diagnosis: Nephroblastoma (Wilms tumor) with regional lymph node metastasis (pT2b N2).	Reasoning: <step 1> The patient is a 48-year-old male presenting with persistent left abdominal pain, weight loss, and increasing fatigue over the past six months. These symptoms raise suspicion for an abdominal mass or a neoplastic process in the kidney or nearby organs. <step 2> The CT scan reveals a 12 cm heterogeneous mass in the left kidney with solid and cystic components, irregular borders, and regional lymphadenopathy. This suggests a malignant process such as a renal cell carcinoma, though a differential diagnosis must include other renal neoplasms. <step 3> The histopathology report confirms a nephroblastoma (Wilms' tumor) with a triphasic pattern involving blastemal, stromal, and epithelial components. Nephroblastoma is a type of kidney cancer primarily seen in children, but it can also occur in adults, though rarely. <step 4> The pT2b N2 staging indicates that the tumor is locally advanced (perinephric fat extension without invasion) and has regional lymph node involvement, but no distant metastases are observed. The patient's anemia and systemic symptoms are likely related to the cancer's local effects and possibly the regional metastasis. Disease Diagnosis: Nephroblastoma (Wilms' tumor), stage pT2b N2.
Metrics			
Accuracy	\	1	1
Efficiency	\	0.75	0.75
Factuality	\	1	1
Recall	\	0.8	0.8

Respond Letter Figure 2 | A case of oracle diagnosis on rare diseases.

References

- [1] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.
- [2] Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*, 2024.
- [3] Google. Medgemma hugging face. <https://huggingface.co/collections/google/medgemma-release-680aade845f90bec6a3f60c4>, 2025. Accessed: [Insert Date Accessed, e.g., 2025-05-20].
- [4] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5(3), 2024.
- [5] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.
- [6] Sarah Sandmann, Stefan Hegselmann, Michael Fujarski, Lucas Bickmann, Benjamin Wild, Roland Eils, and Julian Varghese. Benchmark evaluation of deepseek large language models in clinical decision-making. *Nature Medicine*, pages 1–1, 2025.
- [7] Sarah Sandmann, Sarah Riepenhausen, Lucas Plagwitz, and Julian Varghese. Systematic analysis of chatgpt, google search and llama 2 for clinical decision support tasks. *Nature Communications*, 15(1):2050, 2024.
- [8] Mickael Tordjman, Zelong Liu, Murat Yuce, Valentin Fauveau, Yunhao Mei, Jerome Hadjadj, Ian Bolger, Haidara Almansour, Carolyn Horst, Ashwin Singh Parihar, et al. Comparative benchmarking of the deepseek large language model on medical tasks and clinical reasoning. *Nature medicine*, pages 1–1, 2025.
- [9] Steffanie S Weinreich, R Mangon, JJ Sikkens, ME En Teeuw, and MC Cornel. Orphanet: a european database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, 152(9):518–519, 2008.

Point-point Respond to Referees: NCOMMS-25-22367B

We sincerely thank all reviewers for their valuable comments to improve the quality of our manuscripts. In the following, we aim to resolve the raised concerns point-by-point. In addition, we have also uploaded the revised manuscript, with **RED** to highlight the changes.

Reviewer 1

Question1.1 One of the key aspects of the paper is to evaluate the reasoning abilities of LLMs on clinical cases. However, in the case of reasoning models, they do not do this. They use the final answer (in the format of a reasoning chain) as the “reasoning” of the model. This is effectively a summary of the reasoning and not the real “reasoning.” While I understand that o3 does not provide the reasoning chains, Deepseek does.

Answer:

Thank you for the thoughtful feedback. In the context of large language models (LLMs), reasoning typically refers to their ability to perform logical deductions, solve problems, and understand context beyond simple surface-level text generation [5]. Several prior surveys [5, 6, 2] consistently characterize the reasoning processes exhibited in final outputs-often facilitated by chain-of-thought prompting-as instances of LLM-based reasoning.

Therefore, we respectfully disagree with the critique that ‘in the case of reasoning models, they do not do this’. We argue that, regardless of the output format-whether reasoning is integrated directly into the final answer or explicitly demarcated using special tokens such as “<think>”-such outputs should be considered valid forms of reasoning, so long as the content is coherent and aligns with general common sense.

Moreover, we would like to clarify that for DeepSeek-R1, we **did** evaluate its inherent reasoning processes. As described in our original submission (**Lines 723-736**), DeepSeek-R1 was the only model for which we were able to extract additional reasoning steps explicitly marked with special tokens. For this model, we reported two distinct scores: one for the reasoning present in the final generated answer, and the other for the inherent reasoning chain extracted separately. For the sake of fair comparison across models, our primary analyses focused on the reasoning reflected in the final answer. Nonetheless, in all reasoning-related tables, we consistently present both scores in the format “XX/xx,” with detailed explanations provided in the corresponding table captions. We hope this clarification addresses your concerns.

Question1.2 The implementation of the efficiency metric is biased, and especially affects reasoning models. The implementation analyzes whether each reasoning step provides valuable information or not and this is decided by another LLM. This has the following problems.

A model X can generate 3 extremely verbose reasoning steps, but model Y can generate 4 much shorter reasoning steps. However, the verbosity of model A wouldn’t be penalized because overall, the reasoning steps contribute to the reasoning. In other words, the implementation of the metrics is biased towards assigning the category “reasoning” to most steps, disregarding their verbosity (especially when evaluating the summaries of the thinking part of reasoning models, e.g., deepseek-r1, see weakness #1).

Since the authors evaluate the efficiency of the summary/final answer (see point #1) and not the thinking part of reasoning models, it is natural that there are no repetitions and redundancies, making the efficiency practically 100% according to their definition. However, this is far from the truth. Table 2 shows that Deepseek R1 671B and OpenAI o3-mini are the most efficient models (the evaluation is based on summaries of their real reasoning steps). However, Biachuan M1 has better examination recommendations than these models (Table 1) and is significantly smaller than them, hence it’s more efficient. Efficiency should be calculated as task performance/cost, and cost should be the number of generated tokens, GPU time, or monetary cost, which are objective metrics.

Answer:

Thank you for the thoughtful feedback on the **efficiency** metric. We believe there may be a misunderstanding.

Our efficiency metric is designed to capture **semantic efficiency**-whether a reasoning step contributes meaningful insight toward the final diagnosis-rather than computational cost. From the user’s perspective, this means the step offers a potentially helpful diagnostic hint. Therefore, we do not penalize verbosity: even verbose reasoning may provide unique, valuable insights, for example, some medical-related common sense.

We appreciate the suggestion for considering computational cost. However, such metrics do not reflect the semantic value of reasoning. As shown in prior prompt tuning work [4, 1, 3], LLMs can generate unreadable reasoning yet high accuracy with minimal tokens, highlighting the disconnect between token efficiency and reasoning quality.

We acknowledge that our efficiency metric only assesses whether a reasoning step offers a plausible direction, therefore, we pair it with a factuality metric, which treats a step as ‘incorrect’, if it lacks medical validity, as relevant clinical guidelines would not support it.

We have clarified this in the revised manuscript (**Lines 175-178**). Regarding the inclusion of the “summary/final answer”, please refer to **Q1.1**. We believe this aligns with our goals: a model that summarizes its reasoning clearly in a final answer is, by our definition, semantically efficient and user-friendly.

In summary, our focus remains on user-perceived semantic efficiency, not computational cost.

Question1.3 Another problem with the implementation of the metrics is that the authors state that the reasoning quality in treatment planning is generally strong (L. 295) based on the high factuality scores (>95%). However, the accuracy of all models is below 30% (Table 6), so there is an apparent contradiction between reasoning factuality, reasoning correctness, and final performance, which may indicate that the evaluation of factuality or reasoning quality is not accurate or does not represent what the name suggests.

Answer:

Thank you for the comments. We have to clarify that the quality of reasoning depends on **both** factuality and completeness. Thus, high factuality only refers to that the model’s reasoning processes are free from factual errors or hallucinations. However, the reasoning completeness in treatment is low with only around 70%, meaning that 30% of critical reasoning aspects are missed for each case. This lack of completeness leads to low accuracy, particularly in treatment planning, where recommended plans are expected to address all treatment aspects. Missing reasoning steps often result in incomplete or poorly considered plans, ultimately reducing overall accuracy. We have clarified this point more explicitly in the revised main text (**Lines 380-390**)

Question1.4 The results of the free-turn setting in recommended examinations (lines 192-204) give potentially illogical results, or at least they are not adequately explained. DiagnoseGPT achieves a 1-turn recall of 22.70% while the free-turn has a lower recall of 12.38%. How is it possible that adding more conversation turns asking for more examinations drops the recall? The first turn should be the same in both cases, so the recall in free-turn should be equal to or larger than. From the appendix, prompt 1 (1-turn) and prompt 2 (free-turn) are extremely similar, so I don’t understand how the first turn of free-turn can lead to worse results than the first turn in 1-turn setting.

Answer:

Thank you for raising this insightful point. While the prompts in the 1-turn and free-turn settings may appear similar at first glance, there is a fundamental difference in their design and intent. In the 1-turn setting (Prompt 1), the instruction explicitly asks the model to recommend additional examinations. As a result, examination suggestions are directly elicited from the model in every case. Furthermore, in the conclusion section, the model is instructed to ‘give a preliminary conclusion if possible, or summarize the current findings’-thus emphasizing summary over final diagnosis.

In contrast, the free-turn setting (Prompt 2) grants the model greater autonomy to decide whether to request further information, recommend examinations, or proceed directly to a diagnosis. Correspondingly, the conclusion prompt states: “if you do not require additional information, please provide a final conclusive diagnosis”. This structure encourages the model to behave more flexibly and emulate realistic diagnostic decision-making.

This difference in instruction design is critical, as it can lead models to omit examination queries entirely in the free-turn setting. This issue is particularly pronounced for DiagnoseGPT, which is specifically optimized for diagnostic decision-making and tends to deprioritize examination recommendations. In our evaluation of 957 diagnostic cases, DiagnoseGPT assumed that the available information was sufficient in the first turn in 288 cases-resulting in a recall of 0 for recommended examinations in these instances. An illustrative failure case is provided in Figure 1.

This behavior contributes to a lower average recall for examination suggestions in the free-turn setting. To address potential confusion and clarify this point, we have included a detailed explanation in the revised manuscript (**Lines 204-212**). We also add a clear description of the difference between two prompts in **Line 568-579**.

Question1.5 The results in diagnostic decision-making also exhibit inconsistent patterns. Deepseek achieves 71.79% acc in 1-turn setting and 76.18% in a free-turn. However, this large improvement is inconsistent with the results from the prior (and dependent) step examination recommendation. The authors claim in L.199 that performance drops in free-turn and explain that it’s due to repetitive loops, etc. Since the diagnostic decision depends on examination recommendation (it’s a pipeline), there should be error propagation, so the diagnostic decision in free-turn should be equal or worse than in 1-turn. Intuitively, a worse quality of examinations should give a worse diagnosis, but the results seem otherwise, and it’s not explained.

Answer:

Thank you for your thoughtful comments. For examination recommendation, we use precision and recall metrics for evaluation. In the free-turn setting, we observe that LLMs often tend to enter repetitive loops, which prevents the final recall performance from improving, even with more available interactive turns. Moreover, for some models, recall may even decline because, in the free-turn setting, models are also required to **decide** when to stop. In certain cases, they may stop prematurely, sometimes as early as the beginning, without making additional queries (see **Q1.4** for a more detailed explanation).

However, whether this error propagates to affect the final diagnostic accuracy depends on its severity. For instance, in the case of DiagnoseGPT, whose examination-wise performance drops significantly (by more than 10%) in the free-turn setting (see **Q1.4** for detailed reasons), its diagnosis-wise performance also declines notably, from 54.44% to 39.60%. In contrast, for the remaining models, which perform similarly or experience only a slight drop (within a 5% recall margin) between the 1-turn and free-turn settings for examination recommendations, they generally achieve higher final diagnosis-wise scores as the free-turn setting serves as a form of **inference scaling**. Specifically, during free-turn interactions, while the models may not propose additional critical examinations and may repeatedly suggest redundant ones, they can still revisit and reinterpret previously obtained results. This allows for longer and more detailed reasoning chains, providing the models with more opportunities to justify and refine their decisions, as well as to compensate for earlier mistakes.

We have included more analysis of this phenomenon in the revised manuscript (**Lines 253-264**).

Question1.6 The evaluation of the reasoning process is composed of efficiency, factuality, and completeness. However, S2.4 (analysis of the reasoning process for diagnosis) lacks an analysis of the efficiency and completeness. These analyses are claimed to be a major component of this paper, but are missing here. Similar problem in S2.5, where the analysis is superficial.

Answer:

Thank you for your valuable feedback. In response, we have expanded the analyses of efficiency and completeness in both Section 2.4 (diagnostic reasoning) and Section 2.5 (treatment planning).

For **Section 2.4**, the revised analysis reveals that strong general-purpose, such as DeepSeek-R1, achieve high efficiency but may sacrifice completeness, while more verbose models like Qwen-QwQ obtain the highest completeness scores at the cost of efficiency. We also observe that specialized models such as DiagnoseGPT tend to prioritize factuality and efficiency, but their completeness lags behind generalist or verbose models.

These findings hold consistently across both all-disease and rare-disease cohorts, with completeness generally limited by the availability of examination data in the 1-turn setting.

For **Section 2.5**, our analysis shows that most models maintain high efficiency and factuality in treatment reasoning, but completeness varies more widely, reflecting differences in reasoning style and model design. Notably, Qwen-QwQ again achieves the highest completeness via exhaustive reasoning, whereas specialized models favor brevity. These trends persist in rare disease scenarios, indicating that the efficiency-completeness trade-off is robust to case rarity.

All newly added analyses and discussions can be found in the revised manuscript (**Lines 288-327, 353-390**).

Question1.7 The human performance of treatment planning is surprisingly low (36% in Table 6). Yet, there is no analysis of the reasons behind it. According to S4.1, there is a different human analysis of the quality of the dataset that claims 96% of the treatment cases can pass human evaluation. Therefore, I wonder why that low 36% accuracy. Are many cases failing because only a specialist can solve them and the annotator was not specialized in that branch of medicine?

Answer:

Your hypothesis correctly identifies a key challenge: treatment planning is significantly more complex than diagnosis. It often requires multidisciplinary collaboration and must account for a wide range of potential clinical scenarios to develop an appropriate and comprehensive plan.

Another contributing factor is the stricter evaluation criteria for treatment accuracy. While a diagnosis is considered correct if it matches the ground truth, treatment plans must address multiple critical aspects to be deemed complete. As a case shown in Figure 2, omitting even one essential component can render the entire plan incorrect. These factors help explain the lower accuracy observed in individual physicians’ treatment planning. A more detailed analysis of human performance in this area is included in the revised manuscript (**Lines 344-351**).

Question1.8 One of the novelty claims of the paper is the use of real clinical cases, while prior benchmarks use synthetic or medical exam-style questions (lines 372-279). However, this is not entirely true. Citation 27 (DiagnosisGPT) evaluates on DxDiag, which is a dataset with real conversions between patients and physicians.

Answer:

Thank you for pointing this out. We acknowledge that DxDiag, as used in DiagnosisGPT, is derived from MedDialog, which contains physician-patient conversations. However, it is important to note that MedDialog is primarily collected from **online** medical consultation websites, rather than from real-world clinical centers, which may lack the context of actual clinical practice. In contrast, our dataset is based on real clinical cases sourced from authentic case report from pubmed central.

Moreover, the key novelty of our work lies in evaluating the LLM’s “**reasoning ability**” on clinical cases, rather than simply utilizing clinical cases as a dataset. We point out that case reports are a perfectly suitable clinical case collection serving for LLM reasoning evaluation, as: (1) they come from clinical practice (2) most of them provide reasoning references and detailed analyses in the raw discussion sections. We believe this is crucial, as it allows for a more rigorous and clinically relevant evaluation of LLMs’s reasoning ability, more closely reflecting real-world medical practice. We have clarified this distinction more clearly in the revised manuscript (**Lines 464-472**).

Reviewer 2

Reviewer 2 co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer 3

Question3.1 Why is the related work listed in the Discussion section? It should belong to the introduction?

Answer:

Thank you for the suggestion. Our intention was to provide a more comprehensive discussion of related work in the Discussion section. In response to your comment, we have now also included a concise summary of related work in the Introduction (**Lines 18-27**) to improve the structure of the paper.

Question3.2 The authors have correctly removed “real-world” from their title but still write “Use of Real Clinical Cases” or using terms as “authentic”. The authors should consider to change this as the cases were processed and do represent clinical reality. Further this should be underlined as a limitation in the discussion section.

Answer:

Thank you for the valuable suggestion. We recognize that while case reports provide rich and clinically relevant information that closely approximates real clinical scenarios, the automated processing and restructuring of these cases in our benchmark may introduce certain imperfections or deviations from actual clinical practice. In response, we have revised our manuscript to avoid terms such as “real clinical cases” and “authentic”, and have updated our wording throughout to more accurately reflect the nature of our dataset. Furthermore, we have explicitly acknowledged this as a limitation in the Discussion section (**Lines 493-496**)

Question3.3 The authors mention in their rebuttal letter: “While we share certain methodological elements with [7], including the categorization of rare versus frequent diseases and focus on three core clinical tasks (examination recommendation, diagnostic decision-making, and treatment planning)”. The reviewer suggests to explicitly mention these certain methodological elements while citing the source, ideally in the methods and/or related work.

Answer:

Thank you for your suggestion. We agree that [7] is highly relevant and of significant clinical value, and indeed shares several methodological elements with our study. In response, we have now explicitly detailed these shared elements in our revised manuscript (**Lines 448-453**).

Prompt

Prompt 1. Prompt for asking information under 1-turn examination recommendation setting.

Please thoroughly examine the patient case summary presented below. Your objective is to perform a detailed diagnostic analysis utilizing all available information. Note that due to the potentially limited details, the preliminary diagnosis may encompass several possible conditions. Should you ascertain that the provided data is inadequate for a definitive conclusion, please enumerate any additional diagnostic tests or information that would be necessary. However, if you can deduce a conclusive diagnosis, please proceed to provide it. Too many requests for information are also inappropriate.

[Patient Case Summary without Ancillary Test]

Guidelines:

Evaluate the patient's symptoms, medical history, and all pertinent details from the case summary.

Formulate differential diagnoses based on your analysis.

If the information is not sufficient for a conclusive diagnosis, specify the further tests or details required.

Always following the response format in each turn of the dialogue:

Chain of Thought:

Please sort out your thinking process step by step, with each logical step in a separate paragraph, and use a format such as <step 1> to label each step.

<step 1> Specific thinking content of this step

<step 2> Specific thinking content of this step

...

<step n> Specific thinking content of this step

Conclusion:

Give a preliminary conclusion if possible, or summarize the current findings.

Additional Information Required:

[Indicate if further information is needed by specifying the required tests or data. If a conclusive diagnosis has been made and no additional information is necessary, only output "Not required." directly without any other words in this section.]

For example:

Not required.

or

1. Laboratory tests: details

2. Imaging: details

Prompt

Prompt 2. Prompt for the first turn under the free-turn examination recommendation setting.

Please thoroughly examine the patient case summary presented below. Your objective is to perform a detailed diagnostic analysis utilizing all available information. Note that due to the potentially limited details, the preliminary diagnosis may encompass several possible conditions. Should you ascertain that the provided data is inadequate for a definitive conclusion, please enumerate any additional diagnostic tests or information that would be necessary. However, if you can deduce a conclusive diagnosis, please proceed to provide it. Too many requests for information are also inappropriate.

[Patient Case Summary without Ancillary Test]

Guidelines:

Evaluate the patient's symptoms, medical history, and all pertinent details from the case summary.

Formulate differential diagnoses based on your analysis.

If the information is not sufficient for a conclusive diagnosis, specify the further tests or details required.

Always following the response format in the following dialogue:

Chain of Thought:

Please sort out your thinking process step by step, with each logical step in a separate paragraph, and use a format such as <step 1> to label each step.

<step 1> Specific thinking content of this step

<step 2> Specific thinking content of this step

...

<step n> Specific thinking content of this step

Additional Information Required:

Indicate if further information is needed by specifying the required tests or data. If a conclusive diagnosis has been made and no additional information is necessary, only output "Not required." directly without any other words in this section.

For example:

Not required.

or

1. Laboratory tests: details

2. Imaging: details

Conclusion:

If do not require additional information, please provide a final conclusive diagnosis.

Otherwise, summarize the current findings.

Patient Case		
Case ID	PMC11347276	
Case Summary	Patient Information: - 10-year-old female. - Chief Complaint: Abdominal pain for 3 days. - History of Present Illness: Large pelvic mass identified through abdominal ultrasound. - Past Medical History: None. - Personal History: None. - Family History: No known history of tumors or genetic diseases. - Physical Examination: Palpable large, hard, low-motion mass in pelvic cavity. Ancillary Tests: - Abdominal CT: Slightly low-density pelvic mass with unclear boundary to sigmoid colon, showing uneven enhancement on contrast-enhanced CT, suggesting a possible malignancy. - 18F-FDG PET/CT: Increased uptake of 18F-FDG in lesion but no other hot spots. - Histopathology: Spindle-shaped tumor cells with scattered inflammatory cells. Immunohistochemical markers positive for SMA, ALK, CD117, and vimentin, negative for CK, Desmin, and Dog-1.	
Diagnosis Results	Inflammatory Myofibroblastic Tumor (IMT) in sigmoid colon	
Generation		
Turn Round	1-turn	Free-turn
1	Reasoning: <step 1> Given the symptoms and information provided, the chief complaint of abdominal pain for 3 days in a 10-year-old female patient, alongside the history of a large pelvic mass identified through abdominal ultrasound, we can deduce potential causes. <step 2> The palpable large, hard, low-motion mass in the pelvic cavity further narrows down our differential diagnoses. <step 3> The absence of known history of tumors or genetic diseases in the family history reduces the likelihood of hereditary conditions but does not completely rule out genetic predispositions. <step 4> Taking into consideration the age of the patient and the nature of the symptoms, we must carefully evaluate for both congenital and acquired causes of pelvic masses in this age group. <step 5> The provided information lacks detailed specifics such as the exact location of the mass (other than pelvic), its size, texture, and any associated symptoms (e.g., changes in urinary or bowel habits, weight loss) that could help further refine our differential diagnoses. Recommended Examinations: - Detailed imaging results (MRI or CT scan) of the pelvic mass. - Blood tests to check tumor markers. - Further pediatric examinations to rule out any congenital anomalies. Conclusion: Based on the symptom information and history provided, the preliminary differential diagnosis may include conditions related to the presence of a large pelvic mass such as ovarian or uterine tumors, cysts, or other growths. Further imaging studies and possibly a biopsy would be necessary to confirm the diagnosis.	Reasoning: <step 1> Given the symptoms and information provided, the chief complaint of abdominal pain for 3 days in a 10-year-old female patient, alongside the history of a large pelvic mass identified through abdominal ultrasound, we can deduce potential causes. <step 2> The palpable large, hard, low-motion mass in the pelvic cavity further narrows down our differential diagnoses. <step 3> The absence of known history of tumors or genetic diseases in the family history reduces the likelihood of hereditary conditions but does not completely rule out genetic predispositions. <step 4> Taking into consideration the age of the patient and the nature of the symptoms, we must carefully evaluate for both congenital and acquired causes of pelvic masses in this age group. Recommended Examinations: Not required. Conclusion: The patient's symptoms and findings are consistent with a diagnosis of ovarian teratoma, a type of germ cell tumor. The recommended approach involves surgical excision of the tumor, with care taken to preserve the contralateral ovary and to prevent complications such as torsion or rupture. The surgical technique may involve excising the tumor along with the affected ovary or attempting to preserve the ovarian tissue on the affected side. It is crucial to have a thorough preoperative and intraoperative assessment to prevent severe injury to surrounding organs. Additional tests and imaging may be needed to confirm the diagnosis and plan the surgical approach.
2	Reasoning: <step 1> The patient is a 10-year-old female presenting with abdominal pain for the past 3 days and has a large pelvic mass identified through an abdominal ultrasound. <step 2> The physical examination reveals a palpable large, hard, low-motion mass in the pelvic cavity. <step 3> The detailed imaging results (MRI or CT scan) of the pelvic mass suggest a possible malignancy, with a low-density mass that has an unclear boundary with the sigmoid colon and uneven enhancement on contrast-enhanced imaging. <step 4> Blood tests to check tumor markers are requested but there is no information available on the results yet. <step 5> Further pediatric examinations to rule out any congenital anomalies are also requested but no information is available. Conclusion: Based on the provided symptoms, history, and imaging findings, the differential diagnosis leans towards a possible malignancy involving the pelvic region, considering the presence of a large pelvic mass with characteristics suggestive of malignancy on imaging studies. Without information on tumor markers and further pediatric examinations, the diagnosis cannot be definitively concluded but should prompt consideration of a neoplastic process.	NO ROUND 2
Metrics		
Precision	0.33	0
Recall	0.33	0

Respond Letter Figure 1 | A Case Study of DiagnoseGPT: Comparison Between One-Turn and Free-Turn Settings. In the one-turn setting, DiagnoseGPT is prompted directly to specify the necessary examinations, which leads it to request more tests. In contrast, under the free-turn setting, the model must decide whether to request additional information or make a final diagnostic decision. In this case, DiagnoseGPT assumed that the available information was sufficient and proceeded to diagnosis without requesting further examinations in the first turn, resulting in both recall and precision being zero.

	Diagnostic Decision-making	Treatment Planning
Groudtruth answer	Metastatic pulmonary adenocarcinoma with EGFR L858R mutation, presenting as advanced non-small cell lung cancer (NSCLC) with brain and bone metastases (T4N3M1c staging).	Perform embolization of the type 2 endoleak using a micropuncture introducer set and a triple-coaxial system via the left femoral artery . Use a mixture of N-butyl-2-cyanoacrylate and iodized oil for embolization.
Predicted answer	Metastatic pulmonary adenocarcinoma (Stage IV, T4N3M1c) with *EGFR L858R* mutation, brain and bone metastases.	Repeat embolization of the left 3rd lumbar artery via the left deep iliac circumflex artery using a left femoral artery approach .
Judgement	Correct	Incorrect

Respond Letter Figure 2 | Comparison of diagnostic and treatment evaluation standards. The diagnosis is judged correct based on accurate identification of the primary disease and metastatic pattern. In contrast, the treatment is marked incorrect, as the predicted plan, although both involve embolization, differs significantly from the ground truth in terms of target vessel selection and procedural technique-highlighting the stricter criteria for treatment planning evaluation.

References

- [1] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332*, 2021.
- [2] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.
- [3] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- [4] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.
- [5] Avinash Patil and Aryan Jadon. Advancing reasoning in large language models: Promising methods and approaches. *arXiv preprint arXiv:2502.03671*, 2025.
- [6] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, 2023.
- [7] Sarah Sandmann, Sarah Riepenhausen, Lucas Plagwitz, and Julian Varghese. Systematic analysis of chatgpt, google search and llama 2 for clinical decision support tasks. *Nature Communications*, 15(1):2050, 2024.

Point-point Respond to Referees: NCOMMS-25-22367C

We sincerely thank all reviewers for their valuable comments to improve the quality of our manuscripts. In the following, we aim to resolve the raised concerns point-by-point. In addition, we have also uploaded the revised manuscript.

Reviewer 3

Question1.1 Regarding the term efficiency, I understand that there might be some misunderstanding and that the LLM-driven metrics can pose underestimated risks. I suggest to adapt or add to the naming of the metrics e.g. in Figure 1: “Metrics on Reasoning” -> “LLM-powered Metrics on Reasoning”. Please also consider this in the main-text

Answer:

Thank you for your thoughtful suggestion. We have revised the caption of Figure 1 to use "LLM-powered Metrics on Reasoning" in place of "Metrics on Reasoning" as you recommended. We have also updated the main text accordingly, consistently referring to these as "LLM-powered metrics" to avoid any potential misunderstanding.

Question1.2 The authors have added some more related work and cited them properly. However, the three critical stages of examination, diagnosis and treatment should reference to the prior work, where these stages were originally used in evaluating LLMs, in the Method section -> Under 4.2 Evaluation Framework: “Three critical patient stages are considered: examination recommendation, diagnostic decision-making, and treatment planning (e.g. based on or similar to [1])”

Answer:

Thank you for your valuable suggestion. We have revised the Method section to include an explicit reference to the prior work, as you recommended. Specifically, we now state: “Three critical patient stages are considered: examination recommendation, diagnostic decision-making, and treatment planning, similar to recent work [1].”

References

- [1] Sarah Sandmann, Stefan Hegselmann, Michael Fujarski, Lucas Bickmann, Benjamin Wild, Roland Eils, and Julian Varghese. Benchmark evaluation of deepseek large language models in clinical decision-making. *Nature Medicine*, pages 1–1, 2025.