nature communications

Article

https://doi.org/10.1038/s41467-025-64769-1

Quantifying the reasoning abilities of LLMs on clinical cases

Received: 24 March 2025

Accepted: 25 September 2025

Published online: 06 November 2025

Check for updates

Pengcheng Qiu[1,2,5], Chaoyi Wu[1,2,5], Shuyu Liu[1], Yanjie Fan[3], Weike Zhao[1,2], Zhuoxia Chen[4], Hongfei Gu[4], Chuanjin Peng[4], Ya Zhang[1,2], Yanfeng Wang[1,2] ✉ & Weidi Xie[1,2] ✉

Recent advances in reasoning-enhanced large language models (LLMs) show promise, yet their application in professional medicine, especially the evaluation of their reasoning process, remains underexplored. We present MedR-Bench, a benchmark of 1453 structured patient cases with reference reasoning derived from clinical case reports, spanning 13 body systems and 10 specialties across common and rare diseases. Our evaluation framework covers three stages of care: examination recommendation, diagnostic decision-making, and treatment planning. To assess reasoning quality, we develop the Reasoning Evaluator, an automated scorer of written reasoning along efficiency, factual accuracy, and completeness. We evaluate seven state-of-the-art reasoning LLMs. Here we show that current models exceed 85% accuracy on simple diagnostic tasks when sufficient examination results are available, but performance drops on examination recommendation and treatment planning. Reasoning is generally factual, yet critical steps are often missing. Open-source models are closing the gap with proprietary systems, highlighting potential for more accessible, equitable clinical AI.

Large language models (LLMs) have advanced significantly in recent years, with systems such as OpenAI-o1[1] and DeepSeek-R1[2] demonstrating remarkable reasoning capabilities. These models have excelled in structured problem-solving and logical inference, achieving notable success in fields like mathematics and programming[2–4]. However, their application in the medical domain—a field defined by complexity, high stakes, and the need for contextual understanding—remains underexplored.

Existing medical LLM benchmarks[5–15] focus on evaluating final generation accuracy using exam-style questions. Some studies have begun to move beyond exam-style evaluations, such as ref. 16, which assesses LLMs on authentic clinical tasks, and ref. 17, which benchmarks advanced models like DeepSeek-R1 on real clinical challenges. However, most efforts still fall short in systematically evaluating reasoning quality, a critical aspect of clinical LLMs. Clearer reasoning processes are invaluable for medical human-AI interactions, as they enable clinicians to trust and effectively follow the recommendations provided. A few recent benchmarks[18–21] have explored the reasoning abilities of LLMs, while they often rely on synthetic or conversational data instead of real clinical cases and typically lack scalable, automated metrics for assessing reasoning processes. This gap limits a comprehensive understanding of the reliability and utility of reasoning LLMs in clinical settings.

To address this, we propose *MedR-Bench*, the first benchmark specifically designed to evaluate the medical reasoning capabilities of state-of-the-art LLMs. MedR-Bench includes 1453 clinical cases spanning 13 body systems and 10 disorder types, with 656 cases dedicated to rare diseases. Unlike existing benchmarks, MedR-Bench emphasizes not only the correctness of final diagnoses or treatment plans but also the transparency, coherence, and factual soundness of the reasoning processes behind them. Inspired by prior works[22,23], the benchmark is constructed from clinical case reports in the PMC Open Access

[1]Shanghai Jiao Tong University, Shanghai, China. [2]Shanghai Artificial Intelligence Laboratory, Shanghai, China. [3]Xin Hua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China. [4]China Mobile Communications Group Shanghai Co., Ltd., Shanghai, China. [5]These authors contributed equally: Pengcheng Qiu, Chaoyi Wu. ✉e-mail: wangyanfeng622@sjtu.edu.cn; weidi@sjtu.edu.cn

Subset[24], reorganized into structured patient cases using GPT-4o. Each case consists of (i) detailed patient information (e.g., chief complaint, medical history), (ii) a structured reasoning process derived from case discussions, and (iii) the final diagnosis or treatment plan, reflecting practical clinical complexity. By incorporating diverse and challenging cases, including rare conditions, MedR-Bench serves as a comprehensive testbed for assessing the reasoning capabilities of LLMs in clinical environments.

To evaluate LLMs, we propose a framework spanning three critical clinical stages: examination recommendation, diagnostic decision-making, and treatment planning, capturing the entire patient care trajectory. Examination recommendation evaluates the model's ability to suggest relevant clinical assessments and iteratively gather necessary information. Diagnostic decision-making tests the model's ability to derive precise diagnoses based on patient history, examination findings, lab tests, and imaging findings. Finally, treatment planning assesses the model's ability to recommend appropriate interventions, such as monitoring strategies, medications, or surgical options, grounded in diagnostic conclusions and patient context.

To quantify performance, we develop an evaluation system to assess both reasoning quality and final outputs. For reasoning evaluation, we introduce the *Reasoning Evaluator*, an automated agentic system that validates free-text reasoning processes using web-scale medical resources and performs cross-referencing. It calculates LLM-powered reasoning metrics for *efficiency*, *factuality*, and *completeness*. For final outputs, we adopt standard metrics such as accuracy, precision, and recall. Using MedR-Bench, we evaluate seven reasoning-enhanced LLMs—OpenAI-o3-mini, Gemini-2.0-Flash Thinking, DeepSeek-R1, Qwen-QwQ, Baichuan-M1, DiagnoseGPT, MedGemma—providing a comparative analysis of their strengths and limitations across various clinical stages.

Our findings reveal that current clinical LLMs perform well on relatively simple tasks, such as generating accurate diagnoses when sufficient information is available, achieving over 85% accuracy. However, they struggle with complex tasks, such as examination recommendation and treatment planning. In terms of reasoning quality, LLMs exhibit strong factual accuracy, with nearly 90% of reasoning steps being correct, but omissions in critical reasoning steps are common, indicating a need for improved reasoning completeness. For rare diseases, while these cases remain challenging, models generally show consistent performance across reasoning and prediction tasks, suggesting a robust understanding of medical knowledge across case types.

Encouragingly, our findings suggest that open-source models, such as *DeepSeek-R1*, are steadily closing the gap with proprietary systems like *OpenAI-o3-mini*, underscoring their potential to drive accessible and equitable healthcare innovations, motivating continued efforts in their development. All codes, data, assessed model responses, and the evaluation pipeline are fully open-source in MedR-Bench.

Results

In this section, we present our main findings. We begin with an overview of *MedR-Bench*, followed by an analysis of results across the three key stages: examination recommendation, diagnostic decision-making, and treatment planning. In Supplementary A.1, we provide qualitative case studies.

LLMs models for evaluation

This study utilizes a range of models with varying versions, sizes, cut-off dates for training data, and release dates. For closed-source models, we accessed their APIs directly, while for open-source models, we downloaded the model weights and conducted local inference. The details are presented below.

- *OpenAI-o3-mini*: this is a closed-source model with the version identifier o3-mini-2025-01-31. Its model size is not disclosed. The cut-off date for training data is October 2023, and it was officially released in January 2025.
- *Gemini-2.0-FT*: this is a closed-source model, identified by the version Gemini-2.0-flash-thinking-exp-01-21. Similar to OpenAI-o3-mini, the model size is not disclosed. Its cut-off date for training data is June 2024, and it was officially released in January 2025.
- *DeepSeek-R1*: this is an open-source model with the version identifier deepseek-ai/DeepSeek-R1. It is a large-scale model with 671 billion parameters (671B). The cut-off date for training data is not disclosed, and it was released in January 2025.
- *Qwen-QwQ*: this is an open-source model with the version identifier Qwen/QwQ-32B-Preview. It has 32 billion parameters (32B). The cut-off date for training data is not disclosed, and the model was released in November 2024.
- *Baichuan-M1*: unlike the previously mentioned LLMs designed for general domains, this is an open-source medical-specific model with the version identifier baichuan-inc/Baichuan-M1-14B-Instruct. It has 14 billion parameters (14B), with no disclosed cut-off date for training data. The model was released in January 2025.
- *DiagnoseGPT*: DiagnoseGPT is a series of medical LLMs specifically developed for diagnosis. In our evaluation, we deploy the FreedomIntelligence/DiagnosisGPT-34B locally for assessment, which was released in July 2024.
- *MedGemma*: MedGemma is a variant of Gemma 3, which is optimized for the medical domain by Google DeepMind. It has 27 billion parameters. Its base model Gemma 3's cut-off date for training data is August 2024, and it was officially released in May 2025.

A more detailed introduction to these LLMs is provided in Section "LLM baselines."

Introduction of MedR-Bench

Our proposed *MedR-Bench* comprises three key components: (1) structured patient cases, (2) a versatile evaluation framework spanning three stages, and (3) a comprehensive set of evaluation metrics.

Patient cases. Leveraging the case reports from the PMC Open Access Subset[24], we compiled a dataset of 1453 patient cases published after July 2024 to ensure a fair and robust assessment across all models based on their cut-off date for training data. These are divided into two subsets: *MedR-Bench-Diagnosis* with 957 diagnosis-related cases, and *MedR-Bench-Treatment* with 496 treatment-related cases. As illustrated in Supplementary Fig. 1, all cases are systematically organized into the following elements:

- *Case Summary*: documents key patient information. For diagnosis cases, this includes basic patient demographics (e.g., age, sex), chief complaint, history of present illness, past medical history, family history, physical examination, and ancillary tests (e.g., lab and imaging results). For treatment cases, additional factors such as allergies, social history, and diagnostic results are included, as these influence treatment decisions. Any missing information in the raw case reports is recorded as "not mentioned."
- *Reasoning Processes*: summarized from the discussion sections of case reports, this captures the logical steps used to reach a diagnosis or formulate a treatment plan. For diagnosis cases, the reasoning focuses on methods like differential diagnosis. For treatment cases, it emphasizes treatment goals and the rationale behind the chosen interventions.
- *Diagnosis or Treatment Results*: directly extracted from the raw case reports. For diagnosis, this includes identified diseases. For treatment, it consists of free-text descriptions of the recommended interventions.

Additionally, each case is categorized by "body system" and "disorders and conditions" following the taxonomy from MedlinePlus (https://medlineplus.gov/healthtopics.html). We further utilize Orphanet Rare Disease Ontology (ORDO)(http://www.ebi.ac.uk/ols4/ontologies/ordo)[25] to identify rare diseases among the cases. This allows MedR-Bench-Diagnosis and MedR-Bench-Treatment to be further split to create rare disease subsets containing 491 and 165 cases, respectively. Case distributions are detailed in the Methods section, with patient case examples provided in Supplementary A.1.

Evaluation settings. To evaluate LLMs' clinical capabilities, we developed a framework covering three stages of the patient care journey: examination recommendation, diagnostic decision-making, and treatment planning, as shown in Fig. 1a (more detailed demonstrations are shown in Supplementary Fig. 8). Below, we summarize these components (see Section "Evaluation framework" in the "Methods" for full implementation details).

Examination recommendation. This setting simulates a scenario where a patient first visits a hospital, and LLMs are tasked with recommending examination items such as lab tests or imaging studies, iteratively gathering information to aid diagnosis or treatment. Using the *MedR-Bench-Diagnosis*, the case summaries—excluding ancillary test results—serve as input, while the ancillary test events are used as for ground-truth reference. Similar to previous works[14,26,27], we initialize an LLM-powered agent to play the role of the patient. The assessed clinical LLM can interact with it by recommending relevant examination items, and the agent provides corresponding results.

To evaluate performance, we define two sub-settings: (i) 1-turn examination recommendation: LLMs can query examination results in a single round of interaction; (ii) Free-turn examination recommendation: LLMs can query information through multiple rounds until sufficient information is gathered for subsequent decisions.

Diagnostic decision-making. This setting evaluates whether LLMs can deliver accurate diagnoses based on the given patient information. Using the *MedR-Bench-Diagnosis*, case summaries serve as input, while the recorded diagnoses serve as the ground truth.

We define three sub-settings based on the availability of examination information: (i) diagnostic decision after 1-turn examination recommendation: LLMs use the limited information gathered from the 1-turn setting; (ii) diagnostic decision after free-turn examination recommendation: LLMs use more comprehensive information from the free-turn setting; (iii) oracle diagnosis: LLMs have access to all ground-truth examination evidence, representing the easiest setting.

Treatment planning. This setting evaluates LLMs' ability to propose suitable treatment plans. Using the *MedR-Bench-Treatment*, case summaries—including diagnostic results—serve as input, with the practical treatment plan as the reference. Unlike diagnosis, only the oracle setting is used, where LLMs are provided with all ground-truth patient data, for example, basic patient information, ancillary tests, and ground-truth diagnostic results. This reflects the challenges of treatment planning, which is sufficiently challenging, as suggested by our results.

Evaluation metrics. We designed six metrics to objectively evaluate the performance of LLMs, focusing on both their reasoning processes and final outputs, as illustrated in Fig. 1b. Notably, for DeepSeek-R1, it will have two potential reasoning parts, one presented in the formal part and the other presented in the default thinking part (please refer to "Methods" "LLM baselines" for more detailed explanations). By default, in figures, we report the former for fair comparison. In tables, we report LLM-powered reasoning metrics for both, recorded as "XX /xx," where the former denotes the reasoning part in the formal answer part, and the latter denotes the marked thinking part. Below, we briefly introduce these metrics, with more detailed explanations provided in Section "Evaluation metrics."

For reasoning processes, which are primarily expressed in free text and pose significant evaluation challenges[11,12,28,29], we developed a LLM-based system called the *Reasoning Evaluator*. This system decomposes, structures, and verifies reasoning steps. It identifies effective versus repetitive steps and evaluates their alignment with medical knowledge or guidelines by referencing online medical resources. If ground-truth reasoning references are available, the system further assesses whether all relevant steps have been included. Please refer to the Method for more details.

Based on this pipeline, we define the following LLM-powered reasoning metrics:

- *Efficiency*: evaluates whether each reasoning step contributes new insights toward the final answer rather than repeating or rephrasing previous results.
- *Factuality*: assesses whether effective reasoning steps are consistent with established medical guidelines or knowledge. Similar to a "precision" score, it calculates the proportion of factually correct steps among all predicted effective reasoning steps.
- *Completeness*: measures how many reasoning steps explicitly marked in the raw case report are included in the generated content. Analogous to "recall," it computes the proportion of mentioned reasoning steps among all ground-truth steps. While raw case reports may omit some steps, those included are considered essential reasoning evidence.

These three LLM-powered metrics work together to comprehensively assess the quality of the reasoning process. Efficiency evaluates whether each step offers a potential direction for reasoning, while factuality assesses whether the reasoning aligns with established medical knowledge. Completeness, from another perspective, measures whether the model's reasoning process covers all necessary analytical steps.

On the final generation, for example, recommended examinations, diagnosed diseases, and treatment plans, the following metrics are used:

- *Accuracy*: evaluates whether the final answer (both diagnosis and treatment) explicitly matches the ground-truth provided in the raw case reports.
- *Precision and Recall*: used for examination recommendation, where LLMs generate a list of recommended examinations for a given patient case. These metrics are calculated by comparing the generated examination list with the ground-truth ancillary test list recorded in the case report.

Results in examination recommendation

This section presents the main evaluation results for examination recommendations, as illustrated in Fig. 1c, d. Detailed results for the recommended examinations are summarized in Supplementary Table 1, while the results for the reasoning processes are provided in Supplementary Table 2.

Analysis on recommended examinations. In the 1-turn setting, as shown in Supplementary Table 1, DeepSeek-R1 achieves the highest recall at 43.61%, demonstrating its ability to identify the most relevant examinations. Gemini-2.0-FT follows closely with a recall of 43.12%. Qwen-QwQ and MedGemma rank in the middle, while OpenAI-o3-mini, Baichuan-M1, and DiagnoseGPT perform sub-optimally.

For *precision*, Baichuan-M1 outperforms other models with a score of 41.78%, indicating better alignment with medical scenarios and the ability to recommend relevant examinations. In contrast, Gemini-2.0-FT and MedGemma score the lower precision at 22.77% and 22.65% respectively, suggesting frequent recommendations of irrelevant examinations.

https://doi.org/10.1038/s41467-025-64769-1

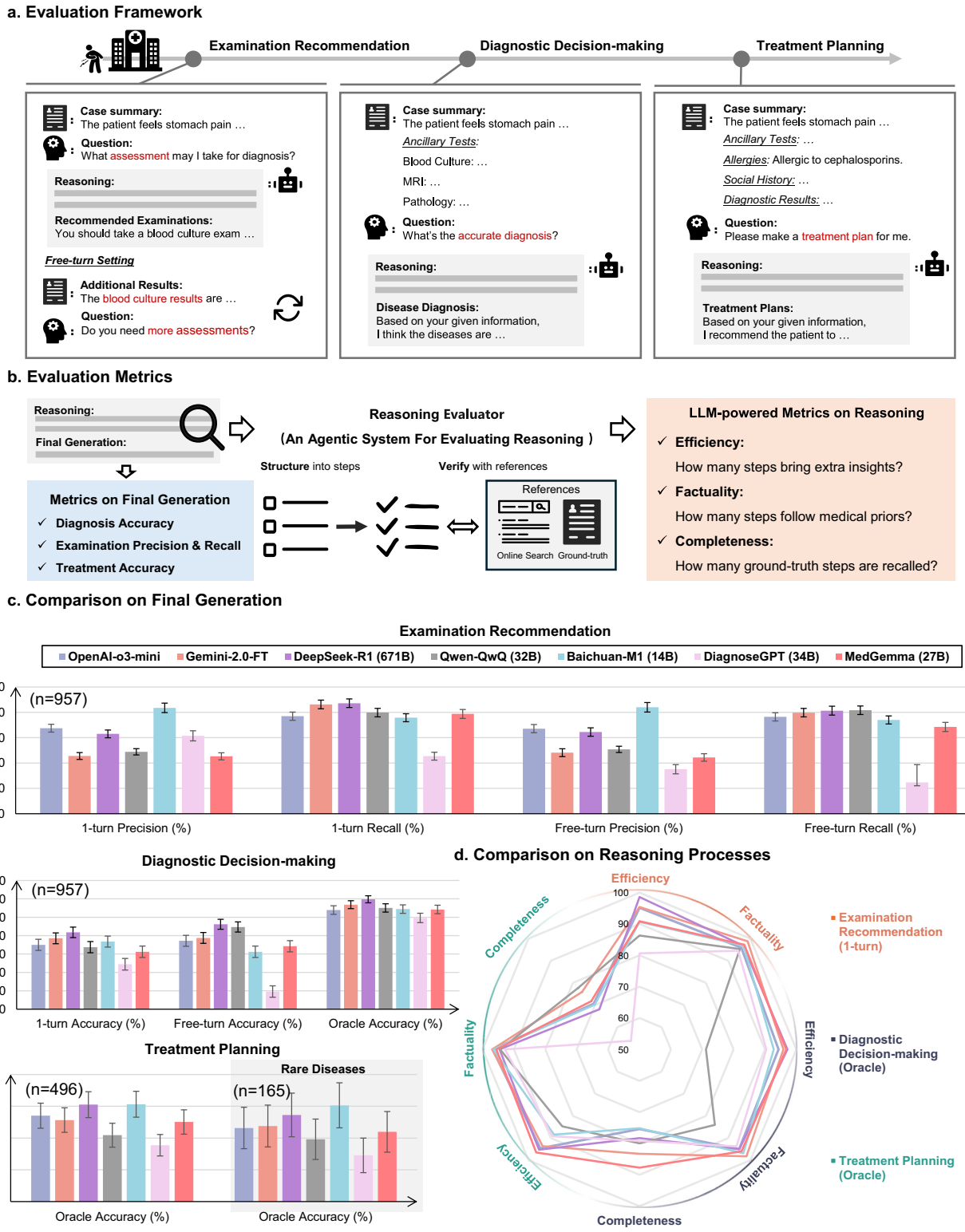

Fig. 1 | Overview of our main evaluation pipeline and results. a Our evaluation framework across three critical patient stages. **b** The LLM-powered metrics for reasoning processes and final generations using our Reasoning Evaluator. **c** The performance of seven LLMs on examination recommendation, diagnostic decision-making, and treatment planning. Notably, for treatment planning, we include a comparison on rare disease cases. For other settings, as the rare disease results show minimal variation compared to all cases, we omit them here and provide them in the supplementary tables. **d** The qualities of reasoning processes, with results for rare cases also provided in the supplementary tables. For examination recommendation, 1-turn reasoning results are plotted, and for diagnostic decision, oracle reasoning results are plotted. Error bars show two-sided 95% z-based confidence intervals for the mean across cases; *n* denotes the number of independent patient cases.

In the free-turn setting, where models are allowed unlimited queries, no significant improvements are observed in either precision or recall across all models. Missed examinations remain unrecovered, even with additional turns, and performance even declines in some cases.

For general LLMs, OpenAI-o3-mini achieves a recall of 38.22% in the free-turn setting, slightly lower than its 1-turn recall of 38.47%. Similarly, DeepSeek-R1 drops from 43.61% in the 1-turn setting to 40.67% in the free-turn setting.

For medical LLMs, DiagnoseGPT achieves a recall of only 12.38%, which is substantially lower than its 1-turn recall of 22.70%. To explain, we first notice that general LLMs tend to enter repetitive query loops in the free-turn setting, which hinders further improvement in recall performance, even with more available interactive turns. Additionally, different from the 1-turn setting, which we force LLMs to query examinations at least for one turn in the instruction, in the free-turn setting, models are required to self-determine when to stop querying. Therefore, in certain cases, they may terminate the process prematurely, without making any more queries, resulting in their recall scores dropping in this setting. This problem is particularly severe for DiagnoseGPT due to its training being tailored to diagnosis only instead of examination recommendations, leading to a significant decline in its performance.

Overall, current LLMs still face significant challenges in handling multi-turn dialog effectively, which limits the utility of the free-turn setting and underscores the difficulties these models encounter when dynamically generating appropriate queries during extended clinical interactions.

Finally, when analyzing performance on rare diseases (Supplementary Table 1), we find that most models maintain comparable performance to that across common diseases.

Analysis on reasoning processes. At the reasoning level, we focus primarily on the 1-turn setting, as the free-turn setting involves extended reasoning processes that grow with the number of turns. Notably, *completeness* cannot be calculated in this context because raw case reports rarely document the reasoning behind the selection of specific examinations.

As shown in Supplementary Table 2, the results on *efficiency* reveal that DeepSeek-R1 achieves the highest score at 98.59%, demonstrating its ability to produce concise and relevant reasoning steps. In contrast, Qwen-QwQ performs sub-optimal, with an efficiency score of just 86.53%. This may be attributed to its training objective of "reflecting deeply"[30], likely causing it to generate excessive attempts, ultimately reducing its efficiency. DiagnoseGPT performs worst on this metric due to its failure in interactive examination querying and tendency to repeatedly summarize the patient's situation rather than reasoning through the next step.

For *factuality*, most LLMs perform well, achieving scores close to 95%. Among them, Gemini-2.0-FT emerges as the most reliable model in examination recommendation, with a factuality score of 98.75%. However, it is notable that none of the models achieve perfect factuality (100%) in their reasoning processes, underscoring the need to carefully verify critical reasoning steps in practical medical applications.

When analyzing reasoning on rare diseases (Supplementary Table 2), we observe consistent trends with those for common diseases, suggesting the robustness of LLMs across common and rare cases.

Results in diagnostic decision-making

This section presents the results for diagnostic decision-making, analyzing performance on both the final output and reasoning levels. Figure 1c and Supplementary Table 3 show the diagnostic decision-making accuracy for both all diseases and rare diseases. Figure 1d and Supplementary Table 4 present the results for diagnostic reasoning processes across all diseases. Supplementary Table 5 further shows the results of the reasoning process specifically for rare diseases.

Analysis on disease diagnosis. As shown in Fig. 1c and Supplementary Table 3, we evaluate diagnostic performance across three settings: *1-turn*, *free-turn*, and *oracle*. Notably, for oracle diagnosis, we also introduce a new human reference benchmark. Six physicians with 5 years of clinical experience from Xin Hua Hospital, affiliated with Shanghai Jiao Tong University School of Medicine, were invited to independently perform an oracle diagnosis task with the help of online searching (restricting access to the original case reports). Their average performance is recorded to provide a meaningful human baseline.

In the 1-turn setting, DeepSeek-R1 achieves the highest diagnostic accuracy (71.79%), demonstrating its ability to gather relevant information and produce accurate diagnoses. Gemini-2.0-FT follows with an accuracy of 68.55%. These results highlight the correlation between active information collection and diagnostic precision. Baichuan-M1 and OpenAI-o3-mini rank in the middle, while Qwen-QwQ, MedGemma, and DiagnoseGPT perform less effectively, consistent with their results in examination recommendation.

In the free-turn setting, where models can iteratively query additional information, most models show improved diagnostic accuracy. For instance, DeepSeek-R1 increases its accuracy from 71.79% (1-turn) to 76.18%, and OpenAI-o3-mini improves from 64.99% to 67.19%, even though they do not demonstrate higher examination recommendation recall in the free-turn setting. This pattern can be primarily attributed to the increased number of reasoning tokens generated in the free-turn setting. Specifically, during free-turn interactions, models may not necessarily propose more critical examinations, but they can revisit and reinterpret previously recommended examinations. This enables the formation of longer and more elaborate reasoning chains, which not only provide the model with greater opportunity to justify and refine its decisions but also enhance its ability to compensate for earlier errors, serving as a form of inference scaling. However, if the model performs too poorly in free-turn examination recommendations, we also observe that errors tend to propagate. For instance, DiagnoseGPT drops significantly in free-turn diagnosis accuracy, dropping from 54.44% to 39.60%. This trend aligns with its markedly reduced performance in examination recommendation tasks, where its accuracy drops by more than 10% in the free-turn setting compared to the one-turn setting.

In the oracle setting, where all crucial diagnostic information is provided, all models achieve significantly higher accuracy. For example, DeepSeek-R1 improves from 76.18% in the free-turn setting to 89.76%, followed by Gemini-2.0-FT. MedGemma, OpenAI-o3-mini, Qwen-QwQ, and Baichuan-M1 also perform well, achieving accuracies above 83%. DiagnoseGPT performs worst in this task, with an accuracy of 79.62%. These results emphasize the importance of identifying and recommending relevant examinations to support accurate diagnoses. Interestingly, recent LLMs generally outperform individual physicians in diagnostic accuracy. We attribute this to the broad and systematic medical knowledge LLMs acquired during 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.

Performance on rare diseases is consistent with that on common ones. On the one hand, this further demonstrates the robustness of these models in challenging scenarios. Through pretraining on large medical corpora, they have encountered rare conditions that are difficult for ordinary physicians to master. On the other hand, we

conducted a thorough case study and found that many rare diseases have specific diagnostic tests, which are provided in the auxiliary test results, significantly reducing the difficulty of such tasks. The primary challenge in diagnosing rare diseases lies in proposing the appropriate specific test as early as possible. We also include a case study for this situation, as illustrated in Supplementary A.1.4.

Analysis on reasoning processes. In the 1-turn diagnostic setting, as shown Supplementary Table 4, in where reasoning builds on incomplete examinations, most models—except Qwen-QwQ—show a decline in factuality compared to the oracle setting. This suggests that missing examinations increases the likelihood of hallucinated reasoning.

Delving deeper into factuality, specialized models like DiagnoseGPT achieve the highest score of 89.14%, outperforming generalists such as Qwen-QwQ (88.14%) and Baichuan-M1 (88.62%), possibly due to its targeted optimization for diagnostic tasks that enhances reliability under uncertainty. DeepSeek-R1 follows closely at 87.15%, demonstrating robust performance.

When it comes to efficiency and completeness, DeepSeek-R1 demonstrates superior efficiency, achieving a score of 95.86%, and outperforms closed-source models such as OpenAI-o3-mini (91.59%) and Gemini-2.0-FT (83.77%). This advantage is likely attributable to its 671B parameters, which facilitate concise pattern recognition even when information is limited. In contrast, smaller models like Baichuan-M1 (82.91%) and Qwen-QwQ (76.97%) encounter greater challenges in efficiency. Notably, Qwen-QwQ adopts a verbose style to compensate for incomplete information, which comes at the expense of brevity. However, this verbosity enhances completeness; Qwen-QwQ achieves the highest completeness score (66.94%), as generating more detailed responses helps retrieve ground-truth evidence under uncertainty. Conversely, specialized models such as DiagnoseGPT prioritize accuracy, resulting in high factuality (89.14%) but lower completeness (25.44%). MedGemma strikes a balance, with moderate efficiency (90.22%) and completeness (49.91%).

Overall, completeness scores in the 1-turn diagnostic setting remain lower than those in the oracle setting, highlighting the constraints imposed by missing examinations on comprehensive reasoning. This is expected, as limited examination data increase the risk that LLMs overlook necessary reasoning steps due to the absence of prior information.

In the oracle setting, where all essential examination results are provided, models generally exhibit improved performance across metrics due to the availability of complete data.

Delving into factuality, closed-source models lead with Gemini-2.0-FT achieving the highest score of 98.23%, followed closely by OpenAI-o3-mini at 94.94%, reflecting their strong ability to avoid hallucinations when all information is present. Among open-source models, Baichuan-M1 (96.84%) and DeepSeek-R1 (95.03%) perform robustly, while Qwen-QwQ lags at 84.02%, possibly due to its verbose tendencies introducing unnecessary inferences. Compared to the 1-turn setting, factuality improves significantly.

Building on factuality, efficiency is notably high in this setting, with DeepSeek-R1 topping the list at 97.17%, outperforming even closed-source models like Gemini-2.0-FT (95.89%) and OpenAI-o3-mini (94.33%). This reflects DeepSeek-R1's streamlined processes with complete data, while smaller models like Baichuan-M1 (92.80%) maintain solid efficiency, but Qwen-QwQ struggles at 71.20% due to its lengthy outputs.

Shifting to completeness, MedGemma excels at 87.72%, benefiting from its medical specialization to retrieve comprehensive evidence, closely followed by Gemini-2.0-FT (83.28%). Qwen-QwQ achieves 79.97%, where its verbosity proves advantageous by covering more ground-truth elements, though this comes at the expense of efficiency and factuality. In contrast, smaller models like Baichuan-M1 (75.11%) show moderate completeness.

Notably, for rare diseases as shown in Supplementary Table 5, the performance remains consistent, and the factuality of most LLMs does not decline. Efficiency trends mirror the all-diseases setting, with DeepSeek-R1 again leading in both 1-turn (95.96%) and oracle (97.61%) evaluations, though Qwen-QwQ's efficiency remains low (76.34% in 1-turn and 72.25% in oracle). Completeness shows similar patterns, with Qwen-QwQ (66.53% in 1-turn) and MedGemma (88.77% in oracle) performing strongly, indicating that rarity does not significantly impair models' ability to generate comprehensive reasoning when examinations are available.

Results in treatment planning

This section presents the results of treatment planning. The overall findings are illustrated in Fig. 1c (final generation) and Fig. 1d (reasoning processes), with detailed results provided in Supplementary Table 6.

Analysis on treatment plans. In treatment planning, similar to oracle diagnosis, we also introduce a human reference benchmark. We observe that the precision of recommended treatment plans is significantly lower than the accuracy of diagnostic outputs. Among the models, Baichuan-M1 and DeepSeek-R1 achieve the highest accuracy at 30.65% and 30.51%, respectively. These results underline the increased complexity of treatment planning compared to diagnosis, emphasizing the need for further development of LLMs.

Unlike diagnosis, where rare cases do not impact performance, treatment planning shows a notable decline in precision for rare diseases across general models. For instance, OpenAI-o3-mini drops from 27.03% to 23.17%, and DeepSeek-R1 decreases from 30.51% to 27.27%. This highlights a persistent gap in therapeutic knowledge for rare conditions. In contrast, Baichuan-M1 maintains stable performance, with precision only slightly decreasing from 30.65% to 30.30%, demonstrating the effectiveness of its medical knowledge enhancement. Regarding the human baseline, the accuracy for all diseases is 36.67%, which is significantly higher than that of current LLMs, indicating that current LLMs still lack sufficient capability in treatment planning. However, it is important to note that the human baseline (36.67%) is still far from ideal. In our evaluation, this baseline reflects the performance of a single physician with 5 years of experience working independently. Treatment planning is inherently more challenging than diagnosis, often requiring multidisciplinary input and consideration of various clinical scenarios. In addition, the evaluation criteria for treatment are stricter: while diagnosis is considered correct if it matches the ground truth, a treatment plan must comprehensively address multiple key aspects. Missing even one critical component renders the plan incorrect. These factors collectively contribute to the relatively low accuracy observed in both human and model performance. A more detailed case demonstration is provided in Supplementary A1.5.

Analysis on reasoning processes. As shown in Supplementary Table 6, reasoning quality in treatment planning is generally strong.

Delving into factuality, most models demonstrate robust performance across both "all diseases" and "rare diseases," with scores typically above 94%. In the "all diseases" setting, closed-source models like Gemini-2.0-FT lead with the highest score of 96.96%, closely followed by OpenAI-o3-mini at 96.77% and open-source models such as Baichuan-M1 at 96.56%. DeepSeek-R1 achieves 94.59%, while specialized models like DiagnoseGPT (92.86%) and Qwen-QwQ (94.40%) are slightly lower but still reliable. For "rare diseases," patterns remain similar, with OpenAI-o3-mini topping at 96.81% and Gemini-2.0-FT at 96.68%, followed by Baichuan-M1 at 95.97%. Overall, factuality shows minimal degradation in rare diseases, suggesting that models handle less common conditions without increased hallucination.

Shifting to efficiency, most models excel in producing concise reasoning, with scores often exceeding 90%, highlighting their

capability to generate streamlined treatment plans without unnecessary verbosity. For "all diseases," MedGemma stands out with the highest efficiency of 96.53%, followed by DeepSeek-R1 at 95.25% and OpenAI-o3-mini at 94.67%. Closed-source models generally perform well, with Gemini-2.0-FT at 93.66%, while open-source ones vary: Baichuan-M1 achieves 88.47%, but Qwen-QwQ lags at 84.76%, likely due to its tendency for verbose outputs. In "rare diseases," trends persist, with MedGemma again leading at 96.91% and DeepSeek-R1 at 95.37; Qwen-QwQ remains the lowest at 83.31%. This consistency across disease types indicates that efficiency is robust to rarity.

Finally, for completeness, models vary more widely, with scores ranging from 50% to nearly 80%. In "all diseases," Qwen-QwQ achieves the highest completeness of 77.66%, benefiting from its verbose reasoning that covers more ground-truth elements, followed by Gemini-2.0-FT at 75.89% and MedGemma at 71.70%. DeepSeek-R1 scores 68.08%, while specialized models like DiagnoseGPT lag at 53.86%, prioritizing brevity over exhaustiveness. For "rare diseases," Qwen-QwQ again leads at 78.74%, with Gemini-2.0-FT at 77.10% and DeepSeek-R1 at 68.28%; DiagnoseGPT remains low at 52.25%. Completeness shows slight improvements in rare diseases for some models, suggesting that verbosity aids in addressing uncertainties in uncommon cases. However, this metric often trades off with efficiency and factuality, as seen in Qwen-QwQ's high completeness at the expense of lower scores in the other two areas, whereas balanced models like MedGemma achieve moderate completeness without sacrificing overall reasoning quality.

Considering that the final accuracy of treatment planning remains below 30%, it becomes evident that the current reasoning processes, while generally concise and exhibiting reduced hallucinations (though not entirely eliminating them), are still insufficient to ensure high-quality treatment recommendations. This highlights the inherent complexity of treatment planning: even when reasoning is streamlined and hallucinations are mitigated, omissions of critical reasoning steps often lead to incomplete or incorrect treatment plans. Model completeness remains around 70%, indicating substantial room for improvement in covering all necessary aspects. Compared to diagnosis, treatment planning accuracy is much more sensitive to missing reasoning steps. In diagnostic tasks, models may occasionally reach correct conclusions even if some reasoning elements are overlooked. In contrast, for treatment planning, missing a key step typically has a direct and negative impact on the final recommendation, resulting in plans that lack essential components. These findings underscore the need for further advancements in both the depth and completeness of reasoning chains.

Discussion

In this study, we evaluate the latest reasoning-enhanced LLMs in the medical domain, focusing on both final outputs and the underlying reasoning processes. Unlike previous work on medical LLMs evaluation[5–15], our approach places greater emphasis on quantifying the quality of reasoning. The key contributions of this study are as follows:

A diverse evaluation dataset on clinical patient cases with reasoning references. We introduce *MedR-Bench*, a dataset of 1453 structured patient cases derived from published case reports. It spans 13 medical body systems and 10 disorder specialties, covering both common and rare diseases for diagnosis and treatment planning. Unlike existing multiple-choice datasets, MedR-Bench closely mirrors practical medical practice. Furthermore, each case is enriched with reasoning evidence extracted from high-quality case reports, enabling a rigorous evaluation of reasoning processes.

A versatile evaluation framework covering three critical patient stages. Our benchmark assesses LLM performance across three key stages of patient care: examination recommendation, diagnostic decision-making, and treatment planning. This framework replicates a typical clinical workflow, providing insights into areas where LLMs perform well and identifying gaps in their ability to support clinical decision-making.

A set of objective metrics from multiple perspectives. We adopt a multi-dimensional set of metrics to assess LLM performance. Beyond evaluating the accuracy of final outputs, we introduce the LLM-powered *Reasoning Evaluator*, a system designed to quantitatively measure the quality of free-text reasoning. Using an automated verification mechanism, this system ensures that reasoning is supported by authoritative medical evidence or aligns with reference ground-truth reasoning.

The following findings summarize the performance of LLMs on MedR-Bench:

LLMs demonstrate strong diagnostic performance with sufficient examinations. State-of-the-art reasoning LLMs demonstrate strong diagnostic capabilities when presented with sufficient, well-structured information (Supplementary Table 5). These models excel at synthesizing medical examination results from different specialists to produce clear and accurate diagnoses. While occasional mistakes occur, the overall results are encouraging and highlight the potential for integrating LLMs into clinical workflows. This represents a promising step toward their integration into medical practice.

Examination recommendation and treatment planning remain challenging. Despite their diagnostic success, LLMs struggle with recommending additional examinations to gather necessary diagnostic clues (Supplementary Table 1). This limitation is critical, as such recommendations are central to accurate medical decision-making. Similarly, treatment planning poses significant challenges, with performance in this area lagging notably. This shortfall is likely due to the fact that the oracle diagnosis setting closely resembles multiple-choice medical question-answering datasets[6,7], which have been widely optimized. This suggests that while LLMs have mastered basic medical knowledge and can synthesize examination results, they are not yet aligned with the dynamic hospital environment. These gaps underscore the need for human oversight in clinical applications and highlight key areas for future improvement.

Reasoning capabilities of LLMs in medicine remain inadequate. Our benchmark evaluates reasoning quality through LLM-powered metrics such as efficiency, factuality, and completeness. While most models achieve high efficiency (over 90%, except for Qwen-QwQ), indicating that their reasoning steps contribute meaningfully to decisions, factuality scores reveal occasional errors. Such mistakes, though tolerable in general contexts, pose risks in clinical settings where over-reliance on LLM outputs could lead to harm. Completeness is particularly concerning, with scores between 70% and 80%, reflecting frequent omissions of critical reasoning steps essential for clinical decision-making. Overall, the reasoning capabilities of current LLMs are barely satisfactory and require substantial improvement to meet the demands of clinical reliability and accuracy.

LLMs maintain robust performance on rare diseases, despite challenges. MedR-Bench includes cases involving rare diseases, which are inherently more difficult. While performance in treatment planning for these cases is weaker, the decline is mild, and LLMs maintain consistent accuracy in other tasks. This robustness suggests that current LLMs possess a strong foundational understanding of medical knowledge, even for rare disease domains, underscoring their potential applicability in diverse clinical scenarios.

The gap between open-source and closed-source LLMs is narrowing. Encouragingly, the latest open-source models, such as DeepSeek-R1, are approaching the performance of closed-source LLMs in medical tasks. Open-source models offer significant advantages for clinical applications, including local deployment to safeguard patient privacy and mitigate risks of data leakage. Their accessibility also reduces

reliance on proprietary systems, fostering broader adoption of LLM-driven solutions in medicine while avoiding monopolization of medical resources.

Related work and novelty. Traditional evaluations of medical LLMs predominantly rely on the accuracy in multiple-choice question (MCQ) exams, such as MedQA[6], MedMCQA[7], and PubmedQA[5]. As attention shifts toward real-world clinical applications, recent studies have broadened evaluation scopes.

Recent work[16] evaluates LLMs on clinically relevant tasks, categorizing diseases as rare or common, and focuses on three core clinical activities: examination recommendation, diagnostic decision-making, and treatment planning. The evaluation is conducted manually by physicians. Building on this, the work[17] further assesses advanced models such as DeepSeek-R1 on a range of clinical challenges, demonstrating the advancement of state-of-the-art open-source LLMs.

As reasoning becomes a more explicit focus in recent LLMs, new benchmarks and studies have been introduced to evaluate reasoning capabilities[18–21]. While these advancements provide significant value to the community, certain limitations persist.

First, these benchmarks predominantly focus on exam-style[19,20] questions or sourcing from online conversational cases[18], these sources typically do not fully capture the depth, complexity, and structured reasoning present in real-world clinical practice.

Second, most evaluations[18–21] lack scalable, reasoning-oriented automated metrics. Many[18–20] rely on final multiple-choice or diagnosis accuracy as an indirect measure of reasoning quality. While recent concurrent work[21] employs human ratings to evaluate reasoning, this approach, though reliable, is too costly for timely monitoring of the rapid advancements in clinical reasoning LLMs.

To address these gaps, our paper distinguishes itself in testing cases, evaluation tasks, and metrics:

- *Use of Clinical Cases*: unlike prior benchmarks that rely heavily on synthetic data or medical exam-style MCQs[5–7,19,20], MedR-Bench is built upon authentic patient case reports curated from PubMed Central Open Access (PMC-OA), which are from clinical practice. We point out that clinical case reports are a great treasure for assessing clinical reasoning LLMs as: (1) all of them are collected from clinical practices, (2) most of them provide reasoning references and detailed analyses in the raw discussion sections, serving as reasoning-wise ground truth reference. The utilization of case reports as reasoning-focused LLM benchmark allows for a more rigorous and clinically relevant evaluation of LLMs's reasoning ability, more closely reflecting real-world medical practice.
- *Comprehensive Clinical Task Coverage*: MedR-Bench also evaluates models across the full spectrum of clinical reasoning: from examination recommendation to diagnosis formulation and treatment planning. This multi-stage structure offers a holistic and task-diverse evaluation, supplementing prior works with additional clinical tasks. Our benchmark thus enables a more robust and realistic assessment of LLMs' potential to assist with full-cycle clinical decision-making.
- *Agentic and Scalable LLM-powered Reasoning Metrics*: critically, we introduce an 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[18] or on labor-intensive manual annotations[19], which limit their scalability and generalizability. Our framework systematically decomposes model outputs into structured reasoning steps, classifies them (e.g., 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 significant methodological advancement.

Limitations

Finally, we must acknowledge several limitations of our work. First, although we have considered the most recently published case reports, we cannot fully guarantee that all cases were excluded from the training procedures of all models, as some of the latest LLMs do not disclose their training data cut-off time. Among the evaluated models, DiagnoseGPT, MedGemma, and Gemini-2.0-FT explicitly guarantee that all MedR-Bench cases were excluded from their training data. Other models show no suspicious performance gains compared to these two, suggesting that data leakage is not a significant issue in our evaluation. However, to mitigate data leakage risks, we still recommend that future benchmarks incorporate more private data from local clinical centers and consider more clean LLM baselines. Second, although sources of case reports provide rich and clinically meaningful data that reflect real clinical scenarios, the cases in our benchmark are automatically restructured by LLMs. This process may introduce minor imperfections and inevitably results in some differences from actual clinical practice.

Third, the LLM-based reasoning metrics we designed aim to automatically and objectively quantify the LLM's performance in a scalable manner; however, they cannot fully replace human verification, which, though costly, remains essential.

To address these limitations, we have released all code, evaluation cases, and model responses for the community to access and refine. We encourage clinicians to engage in reviewing and validating LLM-generated responses to further advance research in this domain.

Methods

This study used only publicly available, de-identified data and did not involve any experiments on human participants or animals. Therefore, ethical approval was not required. This section describes the development of MedR-Bench, including the data curation pipeline, the three-stage evaluation framework, and the implementation of evaluation metrics via the LLM-powered Reasoning Evaluator. All text prompts used are provided in the Supplementary Materials and referenced as *Prompt* X, where *X* denotes the corresponding prompt number.

Data curation

As illustrated in Fig. 2a, case reports were collected from the PMC-OA Subset[24], focusing on articles labeled as "case reports." To minimize potential data leakage, we excluded papers published before July 2024, aligning with the training data cut-off date of OpenAI-o3-mini and Gemini-2.0-FT. While other models did not disclose their cut-off dates, their release timeline (near January 2025) and comparable performance suggest this cut-off date is adequate for analysis. This filtering yielded 3817 raw case reports.

To ensure relevance, case reports unrelated to diagnosis or treatment, such as those focused on medical education, were excluded using GPT-4o[31] (gpt-4o-2024-11-20) with *Prompt* 1. Relevant reports were reformatted into structured patient cases using GPT-4o. Diagnosis-related cases included sections on "differential diagnosis processes" and "final diagnosis explanations" (*Prompt* 2), while treatment-related cases included "treatment objectives" and "comprehensive rationale" (*Prompt* 3).

Case quality human evaluation. To ensure the quality of the generated cases, we conducted a thorough human evaluation in collaboration with 6 licensed physicians, each with approximately 5 years of experience, from "Xin Hua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine," as shown in Fig. 3a. Specifically, a random sample of 100 generated cases was selected for evaluation. Each medical case is assessed by *three* physicians independently, and their majority vote results are adopted as the final decision. For

https://doi.org/10.1038/s41467-025-64769-1

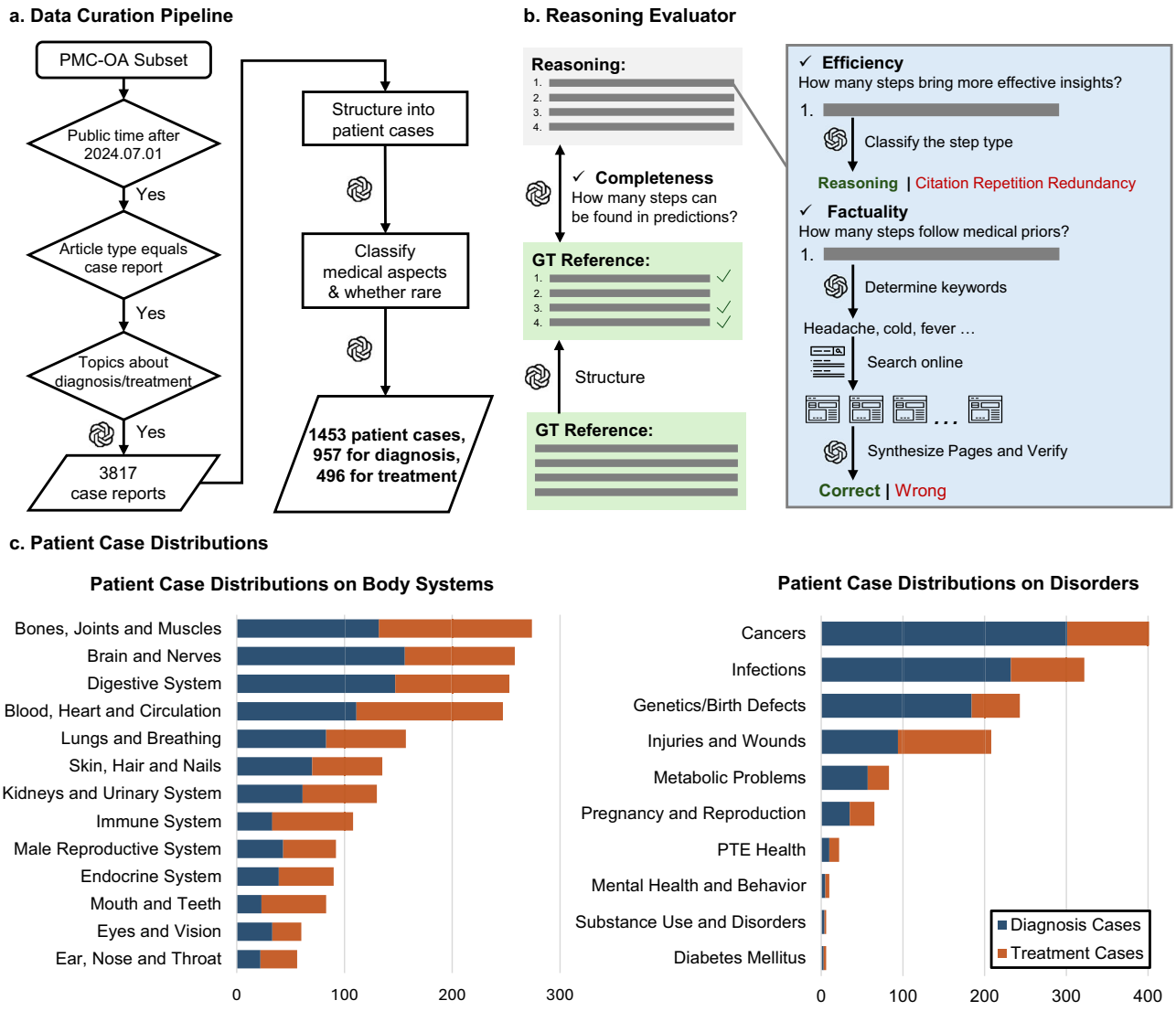

Fig. 2 | Overview of our data curation pipeline, Reasoning Evaluator, and final patient case distributions. a The data curation pipeline. We start with the original case reports from the PMC-OA subset, then filter and reorganize them into structured patient cases for testing. **b** The Reasoning Evaluator quantitatively measures reasoning quality across three dimenstions: efficiency, factuality, and completeness. External search engines are employed to assist the agent in more accurately evaluating the correctness of the provided reasoning steps. **c** The distribution of patient cases across different medical aspects.

diagnosis, they verify that whether the symptom presentations and diagnostic pathways align with clinical practice. For treatment cases, they assess whether the proposed treatment options adhere to current medical guidelines and standards. The results demonstrated that 100% of diagnostic cases and 96% of treatment cases can pass human evaluation, thereby validating the overall quality of the dataset.

Patient case classification. To ensure comprehensive coverage of patient cases in our evaluation dataset, each case is classified based on medical aspects and its relevance to rare diseases. For medical aspects, we adopt the "Body System" and "Disorders and Conditions" taxonomies from MedlinePlus[32], as outlined on their "Health Topics" page. Cases that do not fit into any predefined category are classified as "others." Using *Prompt* 4, GPT-4o categorizes cases into body system classes based on the primarily affected body part, while *Prompt* 5 assigns cases to disorder categories based on the associated diseases.

To identify rare disease cases, we use the ORDO[25]. First, `Scispacy`[33] is employed to extract all associated UMLS[34] Concept Unique Identifiers (CUIs) from the patient case. If any CUIs match those listed in ORDO, the case undergoes further verification using GPT-4o at the free-text level with *Prompt* 6 to confirm explicit mention of rare diseases. Cases passing both steps are classified as rare disease-related; otherwise, they are labeled as unrelated to rare diseases.

Consequently, all patient cases are categorized into three dimensions: "Body System," "Disorders and Conditions" (abbreviated as "Disorder"), and rare disease relevance. In total, 1453 clinical patient cases are included in *MedR-Bench*, comprising 957 diagnosis cases and 496 treatment cases. Among these, 491 diagnosis cases and 165 treatment cases are related to rare diseases. The distribution of cases across medical aspects is shown in Fig. 2a. Detailed patient cases, along with reference category labels, are provided in Supplementary A.1.

Evaluation framework

In this section, we introduce the implementation details of our evaluation framework. Three critical patient stages are considered: examination recommendation, diagnostic decision-making, and treatment planning, similar to recent work[17],

Examination recommendation. In this stage, inspired by prior works[14,26,27], we evaluate the ability of LLMs to dynamically interact with

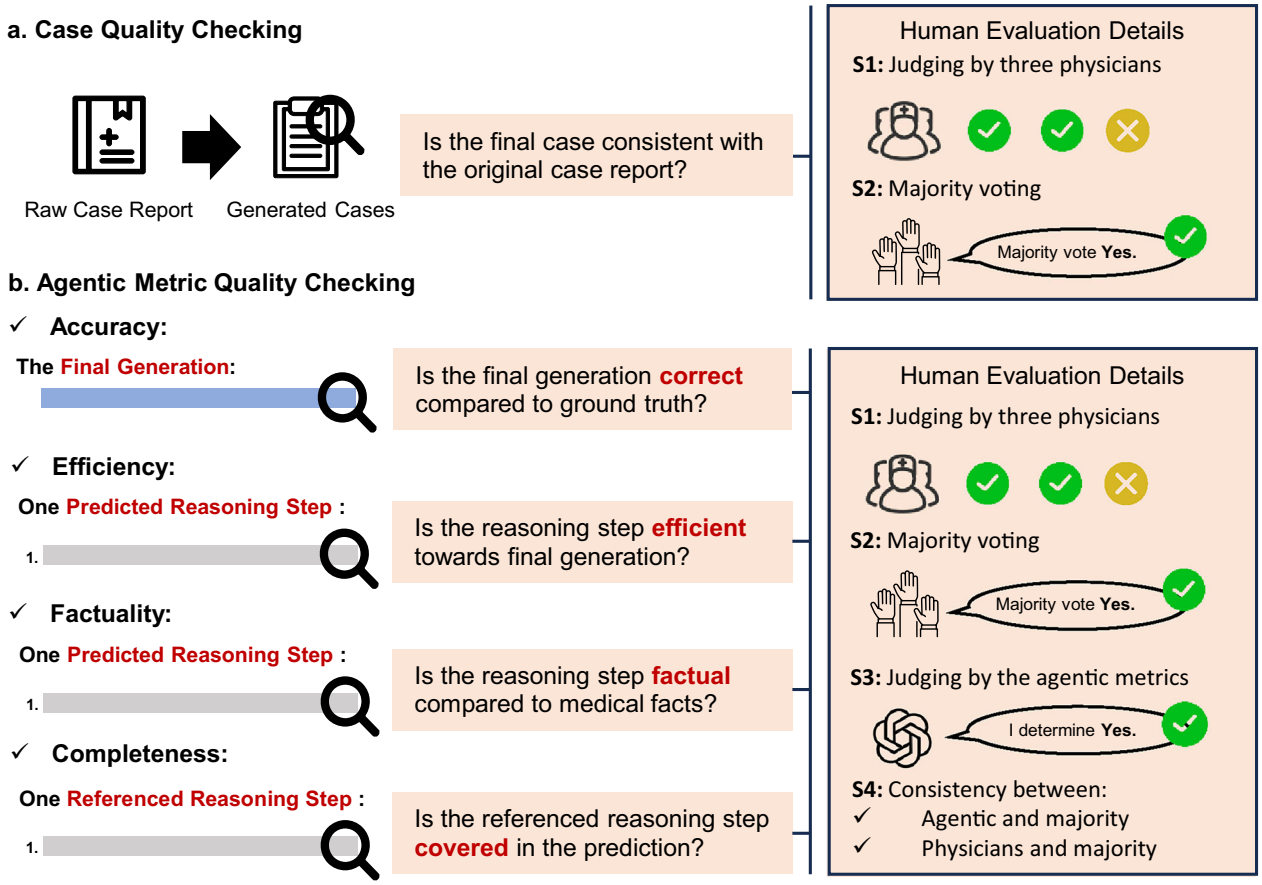

Fig. 3 | The demonstration of the human evaluation pipeline. a Our assessment of case quality, in which three physicians independently review whether each final patient case aligns with the corresponding raw case report. The final accuracy is determined by the majority vote. **b** Our evaluation of agentic metric quality. We assess 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 rating 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.

patients and actively recommend necessary examinations for clinical decision-making. To achieve this, we build a patient agent using GPT-4o with *Prompt* 7, where {case} represents a specific patient case from *MedR-Bench*.

The patient agent is responsible for providing examination results upon request, simulating a scenario in which a patient presents their available test reports to a clinician. Notably, the term "patient" here refers to the returned information about the patient, rather than implying that the agent itself is simulating a patient's responses. Specifically, *Prompt* 7 is designed to instruct the LLM to act as a medical expert to perform exact information extraction, ensuring that the agent provides accurate and clinically appropriate responses grounded in the case information, simulating the scenario in which a patient completes the recommended examinations and presents their professional clinical reports to a clinician.

During evaluation, clinical LLMs are presented with a patient case summary, excluding details about ancillary tests, and tasked with interacting with the patient agent to gather the required information for an accurate diagnosis. The interaction follows one of two protocols: 1-turn examination recommendation or a free-turn examination recommendation. In each turn, LLMs may request additional examinations, such as imaging or lab tests, simulating clinical workflows. If the requested examination is not available in the patient case, the patient agent will respond with: "There is no relevant ancillary test information available for this request."

Under the 1-turn protocol, LLMs are prompted to request essential additional information based on the patient case using *Prompt* 8. The instruction explicitly directs the model to recommend additional examinations. Consequently, examination suggestions are elicited from the model in every case. Moreover, in the conclusion section, the model is instructed to "give a preliminary conclusion if possible, or summarize the current findings," thereby placing greater emphasis on summarization rather than providing a definitive diagnosis.

In the free-turn protocol, LLMs are first prompted with *Prompt* 10 to input the patient case summary. This instruction 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 design encourages the model to behave more flexibly and to emulate realistic diagnostic decision-making. For subsequent turns, they are prompted using *Prompt* 11 to decide whether the available information is sufficient to make a clear diagnosis.

Diagnostic decision-making. In this stage, we assess the LLM's diagnostic capabilities across different settings, ordered by increasing critical information availability: (i) diagnosis after 1-turn examination recommendation: The LLMs are prompted to provide a final diagnosis by integrating the basic patient case information with the additional details obtained during the 1-turn examination recommendation stage,

using *Prompt* 9. (ii) diagnosis after free-turn examination recommendation: In this scenario, the LLMs diagnose based on examination information gathered during free-turn interactions, where they determine that the available information is sufficient. To prevent infinite loops, the maximum number of turns is capped at five. If this limit is reached, the LLM is required to make a diagnosis based on the information collected up to that point. (iii) oracle diagnosis: In this setting, the LLMs are provided with the full ground-truth patient information, including all auxiliary tests, and are prompted to make a diagnosis using *Prompt* 12.

Treatment planning. In this stage, we provide the LLMs with the complete patient information, including the final diagnosis result, to generate recommendations for the preferred treatment plans using *Prompt* 13. Specifically, for each patient case in MedR-Bench, the complete case summary is provided as input (oracle evaluation), and the LLMs are instructed to perform treatment planning.

Evaluation metrics

In this section, we provide a detailed explanation of the implementation of various evaluation metrics.

To begin with, *at the reasoning level*, we introduce *Reasoning Evaluator*, an agentic system powered by GPT-4o, designed to objectively assess the quality of free-text reasoning, as shown in Fig. 2b. Formally, let the predicted reasoning process be denoted as $\hat{\mathcal{R}} = \{\hat{r}_1, \hat{r}_2, \cdots, \hat{r}_N\}$, where each $\hat{r}_i$ represents a reasoning step generated by the original assessed LLMs. The system begins by evaluating the effectiveness of each reasoning step, classifying each step into one of four categories: {citation, repetition, redundancy, reasoning}.

- *Citation* refers to steps that solely restate or cite information directly from the input.
- *Repetition* refers to steps that merely restate conclusions already made in earlier reasoning steps.
- *Redundancy* denotes steps that do not contribute meaningfully to the final decision and are irrelevant to the reasoning process.
- *Reasoning* refers to steps that provide additional insights and contribute to the final decision.

Formally, this classification can be formulated as $e_i = \mathcal{A}(\hat{r}_i|P_e)$, where $e_i \in \{0, 1\}$ indicates whether a given step is effective, and P_e represents the prompt (*Prompt* 15) used to instruct GPT-4o.

Afterward, the agentic system evaluates the factuality of each effective reasoning step by verifying its consistency with external medical knowledge or established guidelines. Specifically, the system first generates a series of search keywords for each effective reasoning step, which is formulated as:

$$\mathcal{K} = \mathcal{A}(\hat{r}_i|P_k), \quad \text{if } e_i = 1, \tag{1}$$

where $\mathcal{K}$ denotes the search keyword set and P_k represents the related prompts (*Prompt* 16). By interacting with external search engine tools, including Google (www.google.com), Bing (www.bing.com), or DuckDuckGo (www.duckduckgo.com), we can retrieve the Top-3 recommended online pages. The system will then summarize their information as the environment response, formulated as $\text{Response} = \mathcal{A}(\text{Search}(\mathcal{K})|P_s)$, where $\text{Search}(\cdot)$ represents the search APIs and P_s is the prompt used for summarization. Finally, the agentic system determines the correctness of each step based on the summarized response:

$$c_i = \begin{cases} 0, & \text{if } e_i = 0, \\ \mathcal{A}(\hat{r}_i|\text{Response}, P_c), & \text{if } e_i = 1. \end{cases} \tag{2}$$

Similarly, here, P_c is the prompt (*Prompt* 17) used to evaluate whether the model output is consistent with the searched factual information or contradicts it.

Next, if the ground truth reasoning evidence $\mathcal{R}$ is provided, the agentic system will be employed to compare it against the prediction. It evaluates how many steps of the ground truth reasoning evidence can be found within the prediction $\hat{\mathcal{R}}$. We first decompose $\mathcal{R}$ into multiple steps as $\{r_1, r_2, \ldots, r_M\} = \mathcal{A}(\mathcal{R}|P_d)$ using *Prompt* 14. Then, we prompt the system with *Prompt* 18 to determine whether each step can be found in the prediction using P_f:

$$f_i = \mathcal{A}(r_i, \hat{\mathcal{R}}|P_f). \tag{3}$$

Based on the results obtained from the agentic reasoning judgment process, the following reasoning-related LLM-powered metrics can be derived:

- *Efficiency*: This metric evaluates the extent to which reasoning steps contribute additional insights toward the final answer, rather than merely repeating previous results or invoking irrelevant reasoning content. The efficiency score is defined as:

$$\text{Efficiency} = \frac{1}{N}\sum_{i=1}^{N} e_i, \tag{4}$$

- *Factuality*: In this metric, we focus on evaluating the factual accuracy of reasoning steps. This can be analogous to Precision scores. Based on the results of the *Reasoning Evaluator*, we calculate the proportion of steps that adhere to established medical knowledge or guidelines among all effective steps:

$$\text{Factuality} = \frac{\sum_{i=1}^{N} c_i}{\sum_{i=1}^{N} e_i}, \tag{5}$$

- *Completeness*: This metric assesses the extent to which reasoning steps outlined in raw case reports are reflected in the generated content. It is analogous to Recall scores and is calculated as:

$$\text{Completeness} = \frac{1}{M}\sum_{i=1}^{M} f_i. \tag{6}$$

Here, the calculation of the factuality score relies heavily on external Internet searches, which may raise concerns about reproducibility due to the rapid evolution of online content. However, we have to emphasize that our results remain stable, as our approach focuses on retrieving key medical knowledge that tends to be factual, i.e., unlikely to be affected by web updates. To demonstrate it, we conduct repeat evaluations at two different time points (2025-02 and 2025-06), 4 months apart, and observed that the consistency rates for key factual reasoning metrics remained high, as illustrated in Supplementary Table 7. We also have open-sourced all evaluation data, code, and snapshots of referenced medical information to enhance reproducibility.

Additionally, at the final generation level, e.g., examination recommendation, disease diagnosis, treatment planning, we adopt several classical metrics to quantify performance:

- *Accuracy*: this metric is a binary metric. It directly compares whether the final answer clearly matches the ground truth provided in the raw case reports. Since medical terminologies often have synonyms, we utilize GPT-4o to verify whether the predicted

results are equivalent to the ground truth. For accurate diagnosis, we employ the prompt described in *Prompt* 19. In contrast, treatment planning is more complex than accurate diagnosis, as even the same disease can have multiple treatment pathways. To address this complexity, we first extract keywords from patient cases using *Prompt* 16. Subsequently, we use a search engine to gather relevant information and make a judgment based on both the retrieved information and the ground-truth treatment plan, as described in *Prompt* 20.

- *Precision and Recall*: these metrics are employed in the context of examination recommendation. They compare the recommended examination list generated by the LLM against the ground-truth practical list using list-wise precision and recall scores. Since the LLM's queries are presented in free-text format, we first utilize GPT-4o to summarize and reorganize them into a structured list using *Prompt* 21. Subsequently, we use *Prompt* 18 to evaluate the hit rate.

Human evaluation of the agentic metrics. To further evaluate the reliability of these LLM-based agentic metrics, we conduct human evaluations on the metrics as shown in Fig. 3b. Similar to that for case quality, here, we ask the same six physicians to check the generated predictions. For *outcome evaluation*, physicians independently assess the accuracy of the diagnostic and treatment planning final generations. For *reasoning process evaluation*, physicians evaluate each reasoning step based on three dimensions: efficiency, factuality, and completeness, across both diagnosis and treatment tasks, detailed as follows:

- *Accuracy*: we randomly sampled a total of 100 cases for evaluation, including 50 cases for diagnostic decision-making and 50 cases for treatment planning. Physicians were provided with the case ground truth and the LLM's prediction, and they were tasked with determining the correctness of the model's final generation predictions.
- *Efficiency*: we randomly sampled 146 reasoning steps for diagnostic decision-making and 139 reasoning steps for treatment planning. Physicians were provided with the case, the current reasoning step, and the preceding reasoning steps. They were required to classify whether the current step was efficient (categorized as "Reasoning") or not.
- *Completeness*: we randomly sampled 90 reasoning steps for diagnostic decision-making and 106 reasoning steps for treatment planning. Physicians were given the case, one essential ground-truth reasoning step, and the LLM's predicted reasoning process. They were tasked with verifying whether the essential step was included in the LLM's reasoning process.
- *Factuality*: we randomly sampled 89 reasoning steps for diagnostic decision-making and 100 reasoning steps for treatment planning. Physicians were given the case, and a single reasoning step was generated by the LLM. They were required to determine whether the step contained any medical factual errors.

During sampling, we preserved the complete sequence of reasoning steps for each case. After collecting the human evaluation results, the majority vote is determined as the final outcome. Finally, we assess the consistency of the agentic results with the majority vote, alongside the consistency of individual physicians. The results are illustrated in Supplementary Table 8. Generally, our LLM-powered agentic evaluation metrics achieve high consistency (around 95%) with the physicians' majority vote on most tasks. It is worth highlighting that, across all metrics, our evaluation system demonstrates higher consistency compared to that between an individual physician and the majority vote. For instance, it achieves 94.00% versus 92.00% in diagnostic accuracy and 86.00% versus 85.33% in treatment accuracy. This demonstrates that our agentic metrics are comparable to those of humans.

To better indicate the evaluation bias in the agentic system, we analyze the precision and recall metrics. We find that the agentic evaluation system exhibits high precision but relatively low recall, e.g., 93.75% precision and 71.43% recall in treatment accuracy and 99.01% precision and 84.75% recall in treatment completeness. This indicates that while the agentic evaluator is generally accurate it tends to be overly conservative, thus judges more answers as incorrect, resulting in lower recall scores.

LLM baselines

In our *MedR-Bench*, we evaluated seven mainstream reasoning LLM series:

- *OpenAI-o3-mini*[35]: the `o3-mini` is the latest LLM developed by OpenAI and is widely regarded as the most powerful LLM currently available. Compared to OpenAI's previous model, GPT-4o, its most notable feature is its enhanced reasoning ability, or, in other words, its capability to "think" before answering. We evaluated the model version `o3-mini-2025-01-31` using the official API.
- *Gemini-2.0-Flash-Thinking (FT)*[36]: the `Gemini-2.0-Flash-Thinking` is an experimental "thinking" LLM developed by Google. It exhibits stronger reasoning capabilities in its responses compared to its predecessor, the Gemini-2.0 Flash Experimental model. This model is characterized by its explicit "thinking process" prior to generating answers. We evaluated the model version `gemini-2.0-flash-thinking-exp-01-21` using the official API.
- *DeepSeek-R1*[2]: DeepSeek-R1 is a 671B-parameter LLM developed by the DeepSeek company. It is an open-source model and is regarded as achieving performance comparable to OpenAI's o1. Similar to o1, it is a reasoning LLM, capable of producing explicit "thinking" outputs. In our evaluation, we use the model weights from Huggingface (https://huggingface.co/deepseek-ai/DeepSeek-R1), `deepseek-ai/DeepSeek-R1`, and deploy it locally.
- *Qwen-QWQ*[30]: Qwen-QwQ is a 32B-parameter experimental research model developed by the Qwen Team. Similar to OpenAI-o1 and DeepSeek-R1, it is also focused on advancing LLM reasoning capabilities. We use the model weights from Huggingface (https://huggingface.co/Qwen/QwQ-32B-Preview), `Qwen/QwQ-32B-Preview` and deploy it locally for evaluation.
- *Baichuan-M1*[37]: Baichuan-M1 is a 14B-parameter medical-specific LLM developed by the Baichuan company. Unlike the previously mentioned models, which are designed for general domains, Baichuan-M1 is a specialized medical LLM. We use the model weights from Huggingface (https://huggingface.co/baichuan-inc/Baichuan-M1-14B-Instruct), `baichuan-inc/Baichuan-M1-14B-Instruct` deployed locally for evaluation.
- *DiagnoseGPT*[18]: DiagnoseGPT is a series of medical LLMs specifically developed for diagnostic tasks based on Yi-34B-Base[38], featuring a step-by-step reasoning chain to enhance interpretability. In our evaluation, we deploy the 34B-parameter model weight from Huggingface (https://huggingface.co/FreedomIntelligence/DiagnosisGPT-34B), `FreedomIntelligence/DiagnosisGPT-34B`, locally for assessment.
- *MedGemma*[39]: MedGemma is a variant of Gemma 3, which is optimized for the medical domain by Google DeepMind. In our evaluation, we use the 27B-parameter model weight from Huggingface (https://huggingface.co/google/medgemma-27b-text-it), `google/medgemma-27b-text-it`, locally for assessment.

Notably, during evaluation, there are two ways to obtain a model's reasoning responses. One approach is to use the model's default marked "thinking parts." For instance, in the case of *DeepSeek-R1*, its

responses always consist of two distinct parts: a thinking part and a formal answer part, separated by the special tokens "<think>" and "</think>." The output format of *OpenAI-o3-mini* follows the same structure. While it seems natural to consider the thinking part as reasoning, *OpenAI-o3-mini* omits this by default, and other models, such as *Qwen-QWQ*, *Baichuan-M1*, *DiagnoseGPT*, *MedGemma*, and *Gemini-2.0-Flash-Thinking (FT)*, do not make such a distinction between the reasoning and answer parts. Thus, to standardize reasoning evaluation across all models, we employ the second approach to obtain reasoning: prompting them with "summarize the reasoning step-by-step" to explicitly instruct them to generate reasoning responses. For *DeepSeek-R1*, this approach results in two potential reasoning outputs: the reasoning response generated within the formal answer part and an additional thinking part marked by the special tokens.

By default, in figures, we report the former for fair comparison. In tables, we report LLM-powered reasoning metrics for both, recorded as "XX/xx," where the former denotes the reasoning part in the formal answer part, and the latter denotes the marked thinking part.

Statistics and reproducibility

No statistical method was used to predetermine sample size.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The MedR-Bench dataset in this study has been deposited in https://github.com/MAGIC-AI4Med/MedRBench/tree/main/data/MedRBenchwith *CC BY-NC-SA* license. The models' responses during evaluation are available in https://github.com/MAGIC-AI4Med/MedRBench/tree/main/data/InferenceResultswith *CC BY-NC-SA* license. The snapshot of the website and supporting knowledge during evaluation are available in https://huggingface.co/datasets/Henrychur/MedRBench-Knowledge-Snapshotswith *CC BY-NC-SA* license. Source data are provided with this paper.

Code availability

Source codes of this paper are released in https://github.com/MAGIC-AI4Med/MedRBenchwith *CC BY-SA* license, cited as ref. 40.

References

1. Jaech, A. et al. Openai o1 System Card. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2412.16720 (2024).
2. Guo, D. et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638 (2025).
3. Zhong, T. et al. Evaluation of openai o1: opportunities and challenges of AGI. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2409.18486 (2024).
4. Phan, L. et al. Humanity's last exam. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2501.14249 (2025).
5. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. & Lu, X. PubMedQA: A dataset for biomedical research question answering. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (eds Inui, K., Jiang, J., Ng, V. & Wan, X.) 2567–2577 (Association for Computational Linguistics, 2019).
6. Jin, D. et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, 6421 (2021).
7. Pal, A., Umapathi, L. K. & Sankarasubbu, M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proc. Conference on Health, Inference, and Learning* 248–260 (PMLR, 2022).
8. Wu, C. et al. Towards evaluating and building versatile large language models for medicine. *npj Digit. Med.* **8**, 58 (2025).
9. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
10. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
11. Wu, C. et al. PMC-LLAMA: toward building open-source language models for medicine. *J. Am. Med. Inform. Assoc.* **31**, 1833–1843 (2024).
12. Qiu, P. et al. Towards building multilingual language model for medicine. *Nat. Commun.* **15**, 8384 (2024).
13. Xie, Y. et al. A preliminary study of o1 in medicine: Are we closer to an AI doctor? Preprint at *arXiv* https://doi.org/10.48550/arXiv.2409.15277 (2024).
14. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
15. Lamparth, M. et al. Moving beyond medical exam questions: a clinician-annotated dataset of real-world tasks and ambiguity in mental healthcare. Preprint at https://doi.org/10.48550/arXiv.2502.16051 (2025).
16. Sandmann, S., Riepenhausen, S., Plagwitz, L. & Varghese, J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nat. Commun.* **15**, 2050 (2024).
17. Sandmann, S. et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat. Med.* **31**, 2546–2549 (2025).
18. Chen, J. et al. CoD, Towards an Interpretable Medical Agent using Chain of Diagnosis. In Findings of the Association for Computational Linguistics: ACL 2025 (ed. Che, W. et al.), 14345–14368 (Association for Computational Linguistics, Vienna, Austria, 2025).
19. Liévin, V., Hother, C. E., Motzfeldt, A. G. & Winther, O. Can large language models reason about medical questions? *Patterns* **5**, 100943 (2024).
20. Zuo, Y. et al. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. In Forty-second International Conference on Machine Learning (ICML 2025).
21. Tordjman, M. et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nat. Med.* **31**, 2550–2555 (2025).
22. Zhao, Z., Jin, Q., Chen, F., Peng, T. & Yu, S. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Sci. Data* **10**, 909 (2023).
23. Wu, C. et al. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nat Commun* **16**, 7866 (2025).
24. National Library of Medicine. PMC open access subset. https://pmc.ncbi.nlm.nih.gov/tools/openftlist/ (2003).
25. Weinreich, S. S., Mangon, R., Sikkens, J. J., Teeuw, M. E. & Cornel, M. C. Orphanet: a European database for rare diseases. *Ned. Tijdschr. Geneeskd.* **152**, 518–519 (2008).
26. Johri, S. et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* **31**, 77–86 (2025).
27. Liao, Y. et al. Automatic interactive evaluation for large language models with state aware patient simulator. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2403.08495 (2024).
28. Zhao, W. et al. RaTEScore: a metric for radiology report generation. In *Proc. 2024 Conference on Empirical Methods in Natural Language Processing* (eds Al-Onaizan, Y., Bansal, M., and Chen, Y.-N.) 15004–15019 (Association for Computational Linguistics, 2024).
29. Calamida, A. et al. Radiology-aware model-based evaluation metric for report generation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2311.16764 (2023).

30. Qwen Team. QwQ: reflect deeply on the boundaries of the unknown. Accessed 27 February 2025 (2024).
31. OpenAI. Hello GPT-4o. Accessed 27 February 2025 (2025).
32. National Library of Medicine (US). MedlinePlus. [updated Jun 24; cited 2020 Jul 1] (2020).
33. Neumann, M., King, D., Beltagy, I. & Ammar, W. ScispaCy: fast and robust models for biomedical natural language processing. In *Proc. 18th BioNLP Workshop and Shared Task* 319–327 (Association for Computational Linguistics, 2019).
34. Bodenreider, O. The Unified Medical Language System (umls): integrating biomedical terminology. *Nucleic Acids Res.* **32**(suppl_1), D267–D270 (2004).
35. OpenAI. OpenAI o3 mini. Accessed 23 February 2025 (n.d.).
36. Gemini Team et al. Gemini: a family of highly capable multimodal models. *arXiv* https://doi.org/10.48550/arXiv.2312.11805 (2023).
37. Wang, B. et al. Baichuan-M1: pushing the medical capability of large language models. *arXiv* https://doi.org/10.48550/arXiv.2502.12671 (2025).
38. Young, A. et al. Yi: Open foundation models by 01.AI. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2403.04652 (2024).
39. Google. Medgemma hugging face. https://huggingface.co/collections/google/medgemma-release-680aade845f90bec6a3f60c4 Accessed 20 May 2025 (2025).
40. Qiu, P. et al. Quantifying the reasoning abilities of LLMs on clinical cases. https://doi.org/10.5281/zenodo.17046070 (2025).

Acknowledgements

This work is supported by the National Key R&D Program of China (No. 2022ZD0160702) and the Scientific Research Innovation Capability Support Project for Young Faculty (ZY-GXQNJSKYCXNLZCXM-I22).

Author contributions

All listed authors clearly meet the ICMJE 4 criteria. P.Q. and C.W. contribute equally to this work. Y.W. and W.X. are the corresponding authors. Specifically, P.Q., C.W., S.L., Y.F., W.Z., Z.C., H.G., C.P., Y.Z., Y.W., and W.X. all make contributions to the conception or design of the work, and P.Q. and C.W. further perform acquisition, analysis, or interpretation of data for the work. In writing, P.Q. and C.W. draft the work. S.L., Y.F., W.Z., Z.C., H.G., C.P., Y.Z., Y.W., and W.X. review it critically for important intellectual content. All authors approve of the version to be published and agree to be accountable for all aspects of the work to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-64769-1.

Correspondence and requests for materials should be addressed to Yanfeng Wang or Weidi Xie.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.