Active Evidence-Seeking and Diagnostic Reasoning in Large Language Models for Clinical Decision Support

Chen Zhan[1†], Xihe Qiu[1*†], Xiaoyu Tan[2†], Xibing Zhuang[3†], Gengchen Ma[1], Yue Zhang[1], Shuo Li[4], Peifeng Liu[5], Xiaoxiao Ge[6*], Liang Liu[6*], Lu Gan[7*]

[1]School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, 201620, China.
[2]Department, Tencent Youtu Lab, Shanghai, 200233, China.
[3]Department of Oncology, Jinshan Hospital, Fudan University, Shanghai, 201508, China.
[4]Departments of Biomedical Engineering, and Computer and Data Science,Case Western Reserve University Cleveland, 44106, Ohio, USA.
[5]State Key Laboratory of Systems Medicine for Cancer, Shanghai Cancer Institute, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, 200032, China.
[6]Integrative Clinical Research Ward, Clinical Medicine Research Institute, Zhongshan Hospital, Fudan University, Shanghai, 200032, China.
[7]Department of Medical Oncology, Zhongshan Hospital, Fudan University, Shanghai, 200032, China.

*Corresponding author(s). E-mail(s): qiuxihe1993@gmail.com; tinagxx.zs@163.com; liu.liang1@zs-hospital.sh.cn; gan.lu@zs-hospital.sh.cn;
Contributing authors: M325124527@sues.edu.cn; txywilliam1993@outlook.com; 19111210024@fudan.edu.cn; m320123320@sues.edu.cn; m325123614@sues.edu.cn; shuo.li11@case.edu; lpf@sjtu.edu.cn;
†These authors contributed equally to this work.

Abstract

Large language models perform well on static medical examinations, yet clinical diagnosis often requires iterative evidence gathering under uncertainty. Building on prior interactive evaluation efforts, we introduce an OSCE-inspired standardized patient simulator and a controlled, reproducible benchmark for active diagnostic inquiry. Across 468 cases and 15 models in our protocol, we observe that multi-turn evidence seeking reduces diagnostic accuracy by 12.75% and lowers supporting-evidence quality by 24.36% relative to full-context evaluation; error analyses associate these drops with premature diagnostic closure and inefficient questioning. Together, these results suggest that static full-context benchmarks may overestimate performance in interactive evidence-seeking settings, motivating complementary interactive assessment for safer clinical decision support.

Keywords: Large language models, Clinical decision support, Standardized patient simulation, Interactive diagnostic reasoning

1 Introduction

In recent years, general-purpose large language models (LLMs) have demonstrated strong capabilities in medical knowledge retrieval and reasoning [1–3]. Results from static medical benchmarks such as MedQA-USMLE, PubMedQA, and MedMCQA suggest that models ranging from ChatGPT to medically specialized variants such as Med-PaLM 2 can meet or exceed passing thresholds on standardized licensing-style examinations [4–9]. These achievements have motivated growing interest in using LLMs for Clinical Decision Support Systems. However, many widely used evaluation settings remain largely passive and full-context: complete case information, including patient history, symptoms, and test results, is provided to the model simultaneously before it generates a response [10, 11]. This setup is well-suited for assessing summary interpretation given complete information, but it does not directly test how models decide what evidence to request when information is initially sparse. As a result, it can blur the distinction between using provided facts and actively reasoning under uncertainty in iterative clinical workflows.

In practice, clinical diagnosis is a dynamic, multi-turn interactive process governed by the hypothetico-deductive model [12]. Clinicians typically do not receive all relevant data points instantaneously, especially early in a workup. Instead, they must actively formulate initial hypotheses and seek evidence through successive exchanges involving history taking, physical examinations, and laboratory testing to dynamically confirm or refute differential diagnoses [13]. Medical education often uses the Objective Structured Clinical Examination (OSCE) to assess clinicians during such active reasoning processes [14, 15]. By utilizing standardized patient simulations, the OSCE evaluates key competencies including active inquiry, information acquisition strategies, and the ability to synthesize evolving data [16]. This creates a partial mismatch between real diagnostic workflows and many existing evaluation setups: static information feeds can underrepresent the role of active evidence acquisition and iterative decision-making.

Several recent studies have moved toward more clinically grounded evaluations by introducing interactive inquiry or multi-step diagnostic procedures [17–19]. For example, MediQ proposed a simulated clinical dialogue framework to assess information gathering through question-answer interactions, highlighting that under incomplete initial information even large models may struggle to generate efficient and clinically meaningful follow-up questions [1, 20]. Other efforts structure evaluation into staged diagnosis such as preliminary diagnosis, differential diagnosis, and final diagnosis to better reflect sequential decision making. Nonetheless, these approaches involve practical trade-offs, and it is not always straightforward to isolate information acquisition strategies from diagnostic reasoning performance: performance can be influenced by open-ended dialogue variability, prompting choices, and scoring criteria, which can make it difficult to determine whether errors stem from suboptimal evidence seeking or from evidence synthesis [21–24]. In addition, work examining conversational quality and interpretability further suggests that static question-answer benchmarks do not adequately characterize performance during realistic clinical interactions.

To better approximate authentic diagnostic processes, some initiatives have explored system designs that support cyclical inquiry and dialogue management [25]. For instance, the Articulate Medical Intelligence Explorer, known as AMIE, studied an optimized conversational diagnostic system to evaluate performance across multiple dimensions including history collection, diagnostic accuracy, and communication skills [26]. Another representative work, Ask Patients with Patience, structured multi-turn inquiry into iterative dialogues driven by medical guidelines [27, 28]. While these studies demonstrate the value of dialogue-oriented designs, they are often optimized for specific systems, and their evaluation protocols may not directly provide a standardized, reusable benchmark. These systems often rely on open-ended conversations, which can complicate reproducibility and objective grading without human intervention [29]. Consequently, there remains a need for benchmarks that combine controlled information release, reproducibility, and objective scoring-especially for separating evidence acquisition from evidence synthesis for systematically evaluating model capabilities in active evidence seeking and clinical reasoning within a controlled and reproducible environment that can isolate specific failure modes [30, 31].

To complement existing static and dialogue-based evaluations, we introduce an OSCE-inspired interactive framework, ROUNDS-Bench. This framework incorporates a Standardized Patient Simulator and controls information release through a multi-turn inquiry design, providing evidence only when the model submits logical clinical requests [32, 33]. Compared with open-ended chat evaluations, this request–response protocol improves controllability and reproducibility. It resembles common query workflows (e.g., electronic health record retrieval), enabling measurement of questioning efficiency and diagnostic performance across turns during both the information acquisition and reasoning phases separately[18].

Using 468 diverse cases covering six major clinical systems, we designed a dual-task evaluation structure to quantify the performance gap between full-context and active inquiry settings. Task 1, Full-Context Diagnosis, establishes the upper bound of model performance under conditions of complete information, whereas Task 2, Active Evidence Seeking, requires the model to autonomously drive a multi-turn diagnostic

process starting with only the chief complaint. A series of systematic evaluations on current state-of-the-art Large Language Models, including GPT-4o and Qwen [34], revealed a significant capability gap. When transitioning from the passive diagnostic task to the active diagnostic task, diagnostic accuracy decreases by 12.75% on average in our setting. Supporting-evidence quality decreases by 24.36% under active inquiry. This gap matters because high scores under full-context evaluation may not fully reflect whether a model can elicit and cite discriminative evidence during interactive diagnostic workflows. Specifically, models may output correct diagnoses without consistently eliciting, selecting, or citing the key evidence needed to justify the decision.

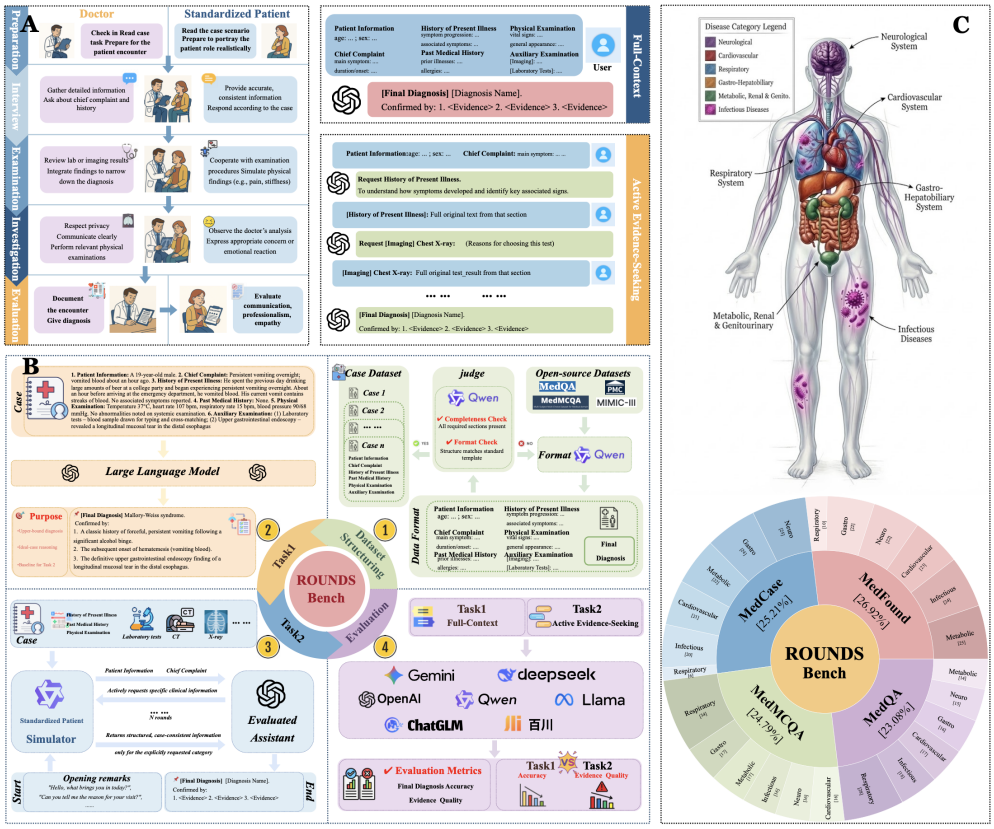

Fig. 1: **Overview of the ROUNDS-Bench framework for evaluating active clinical reasoning. A** Paradigm Comparison. Contrasts the authentic OSCE clinical workflow with the passive Full-Context baseline and the proposed Active Evidence-Seeking interactive protocol. **B** Technical Pipeline. Illustrates the workflow integrating dataset structuring, dual-task evaluation (Task 1 vs. Task 2), and multi-dimensional grading. **C** Dataset Stratification. Anatomical visualization of the six clinical systems covered (top) and the distribution of 468 cases across four heterogeneous medical corpora (bottom).

This study not only precisely quantifies this gap but also provides a detailed analysis of typical failure modes in active diagnosis, such as premature diagnostic closure and inefficient information seeking. By identifying these specific weaknesses, we offer critical evaluation benchmarks and optimization directions for building safer and more clinically effective medical AI agents. Through this work, we highlight the limitations of existing medical AI evaluation paradigms and introduce a reproducible interactive evaluation protocol. Our goal is to propel future model evaluation to transition from static question-answering to authentic, dynamic interaction, thereby ensuring usability and safety in clinical decision support applications.

Table 1: **Baseline characteristics of the virtual patient cohort in ROUNDS-Bench.** Summary of the 468 standardized clinical cases, stratified by data source (MedQA, MedMCQA, MedFound, and MedCaseReasoning) and six major organ systems to ensure balanced representation and prevent prevalence bias.

Characteristic	**No. of Cases** ($N = 468$)	**Percentage** (%)
Data Source		
MedQA (USMLE-style)	108	23.1
MedMCQA (Indian medical exam)	116	24.8
MedFound (clinical data derived from MIMIC-III)	126	26.9
MedCaseReasoning (PMC Case Reports)	118	25.2
Clinical System Category		
Cardiovascular System	78	16.7
Respiratory System	78	16.7
Gastro-Hepatobiliary System	78	16.7
Neurological System	78	16.7
Infectious Diseases	78	16.7
Metabolic, Renal & Genitourinary	78	16.7

Note: Percentages may not sum to 100% due to rounding. The dataset was strictly stratified to ensure equal representation ($N = 78$) across all six clinical system categories to prevent prevalence bias during evaluation.

2 Results

In this section, we present our main findings from the evaluation of 15 state-of-the-art LLMs using the ROUNDS-Bench framework. We begin with an overview of the virtual patient cohort and evaluation protocol, followed by a comparative analysis of the capability gap between passive Full-Context diagnosis and Active Evidence-Seeking. Finally, we examine the critical safety risks associated with evidence quality and dissect the key determinants driving performance variability, including scaling laws, clinical system differences, and data source complexity.

2.1 The ROUNDS-Bench Framework and Virtual Patient Cohort

To complement static benchmark settings and quantify interactive evidence acquisition under controlled information release, we developed the ROUNDS-Bench interactive evaluation framework. Distinct from the Full-Context paradigm typified by MedQA or USMLE, ROUNDS-Bench integrates a high-fidelity Standardized Patient Simulator (SP-sim) designed to reproduce the key characteristic of progressive disclosure found in authentic clinical diagnosis. Specifically, the SP-sim incorporates a rigorous Information Gating Mechanism: initiating with only basic demographics and a chief complaint, models must unlock subsequent evidence, such as symptom details, physical exam findings, and lab results, through clinically logical History Taking and Test Requesting. This mechanism mimics the information asymmetry and iterative evidence-seeking process of real-world doctor-patient interactions (detailed in Methods 4.2). By precluding the shortcut of reading the full case summary, this design compels the evaluation to focus squarely on the model's ability to formulate queries, select tests, and dynamically update hypotheses under uncertainty.

To ensure clinical representativeness and cross-domain comparability, we constructed a strictly stratified Virtual Patient Cohort:

- **Clinical Diversity and Balance:** As shown in Table 1, the cohort comprises 468 independent cases, precisely equilibrated across six core clinical systems: Gastro-Hepatobiliary, Neurological, Infectious Diseases, Metabolic, Renal & Genitourinary, Cardiovascular, and Respiratory, with each system accounting for 16.7% ($N = 78$) of the dataset. This balanced sampling strategy significantly mitigates the impact of disease spectrum bias, ensuring fair comparisons across identical task difficulties and distributions.
- **Data Heterogeneity:** Cases were synthesized and structured from four complementary, high-quality corpora to ensure heterogeneity: MedQA (USMLE-style), MedMCQA (high-difficulty Indian medical exams), MedFound (clinical data derived from MIMIC-III), and MedCaseReasoning (PMC Case Reports)[4, 6, 35–37]. This multi-source fusion spans a broad clinical spectrum from typical common presentations to atypical complex conditions, reducing the risk of domain overfitting and enhancing robustness to diverse clinical expressions.

Leveraging this framework, we implemented a paired dual-task experimental design to isolate and quantify the capability variance between passive reading and active acquisition workflows:

- **Task 1 (Full-Context):** Models receive the complete electronic health record instantaneously, establishing the theoretical upper bound of static diagnostic proficiency.
- **Task 2 (Active Evidence-Seeking):** Models are provided only with the Chief Complaint and must autonomously drive a multi-turn diagnostic dialogue to progressively acquire evidence. By comparing performance on the same cases across these two settings, we precisely characterize the performance degradation when

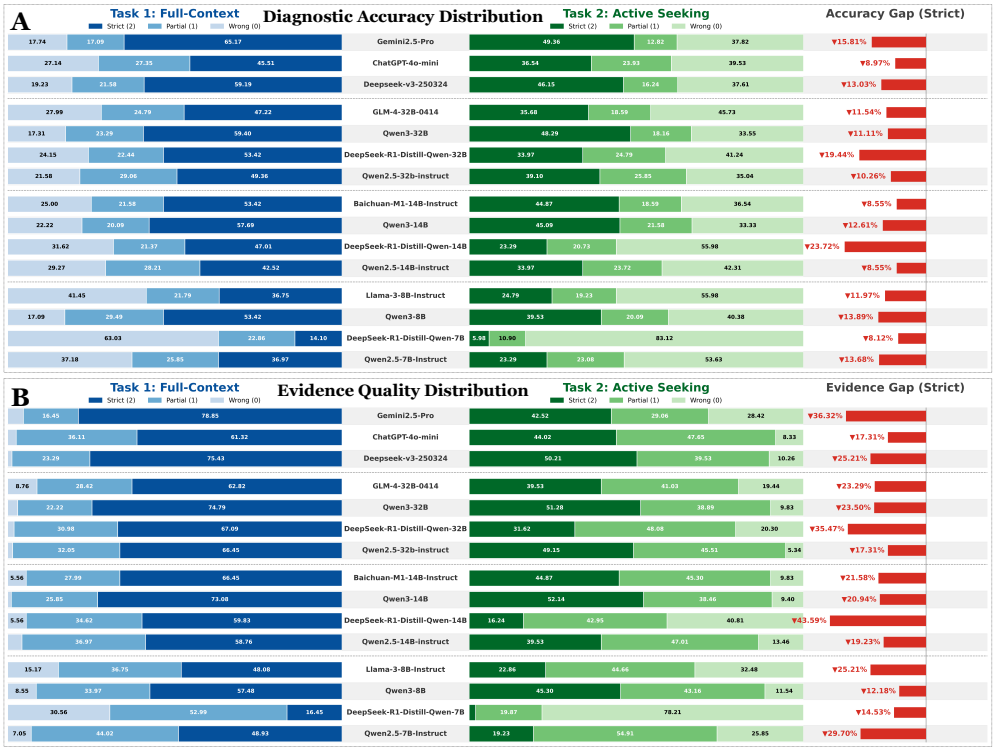

Fig. 2: **Distribution of diagnostic outcomes and quantification of the capability gap.** The figure presents a paired analysis of **A** Diagnostic Accuracy and **B** Evidence Quality across 15 LLMs. **Center:** Models are listed centrally, stratified by parameter size (separated by dashed lines). **Left (Blue panels):** Performance distribution in the static Task 1 (Full-Context) setting. **Right (Green panels):** Performance distribution in the dynamic Task 2 (Active Evidence-Seeking) setting. Darker shades indicate strict correctness, while lighter shades represent partial correctness or errors. Note the visible expansion of the Wrong (lightest shade) segments in Task 2. **Far Right (Red bars):** The Performance Drop (Δ) column explicitly visualizes the decline in strict accuracy/evidence quality when transitioning from Task 1 to Task 2.

transitioning from static, full-info diagnosis to active, evidence-based diagnosis, a metric we define as the Capability Gap (Fig. 1).

2.2 Significant Capability Gap in Active Diagnosis

The transition from the static baseline (Task 1) to the dynamic inquiry setting (Task 2) was associated with a consistent and substantial deterioration in diagnostic performance across all evaluated models. This finding systematically shows a consistent performance drop across models in our interactive setting.

Quantitative Performance Degradation. As evidenced in Table 2 and Figure 2,

Table 2: **Leaderboard of clinical reasoning performance across 15 LLMs stratified by parameter scale.** Models are hierarchically ranked within four tiers based on parameter density. This leaderboard contrasts the static performance ceiling (Task 1: Full-Context) with the interactive reasoning capability (Task 2: Active Seeking). The gap metric quantifies the performance erosion inherent in active inquiry, while "StrictEQ" serves as a rigorous audit of evidence grounding. Best-in-class performers within each tier are highlighted in bold, identifying the most clinically efficient architectures under varying computational constraints.

Model Family	**Model Name**	**Task 1: Full-Context** *Exact Accuracy (%)*	**Task 2: Active Seeking** *Exact Accuracy (%)*	**The Gap** *(T2 − T1)*	**Evidence Quality** *StrictEQ (T2, %)*
Proprietary / Large Models (> 32B)					
Gemini	Gemini-2.5-Pro	**65.2**	**49.4**	-15.8	42.5
DeepSeek	DeepSeek-v3-250324	59.2	46.2	-13.0	**50.2**
OpenAI	ChatGPT-4o-mini	45.5	36.5	**-9.0**	44.0
Medium Models (32B)					
Qwen	Qwen3-32B	**59.4**	**48.3**	-11.1	**51.3**
Qwen	Qwen2.5-32B-Instruct	49.4	39.1	**-10.3**	49.2
ZhipuAI	GLM-4-32B	47.2	35.7	-11.5	39.5
DeepSeek	DeepSeek-R1-Distill-Qwen-32B	53.4	33.9	-19.5	31.7
Small Models (14B)					
Qwen	Qwen3-14B	**57.7**	**45.1**	-12.6	**52.1**
Baichuan	Baichuan-M1-14B	53.4	44.9	**-8.6**	44.9
Qwen	Qwen2.5-14B-Instruct	42.5	34.0	**-8.6**	39.5
DeepSeek	DeepSeek-R1-Distill-Qwen-14B	47.0	23.3	-23.7	16.2
Very Small Models (7B-8B)					
Qwen	Qwen3-8B	**53.4**	**39.5**	-13.9	**45.3**
Meta	Llama-3-8B-Instruct	36.8	24.8	-12.0	22.9
Qwen	Qwen2.5-7B-Instruct	37.0	23.3	-13.7	19.2
DeepSeek	DeepSeek-R1-Distill-Qwen-7B	14.1	6.0	**-8.1**	1.9

the average Exact Accuracy across the full model cohort declined by approximately 12.75% when transitioning from Task 1 to Task 2. This trend demonstrates a consistent deterioration across diverse architectures when the cognitive load shifts from passively receiving complete case details to actively planning inquiry paths and acquiring evidence.

Proprietary Models. Even proprietary models that excelled in the static baseline failed to maintain their superiority in the active diagnostic setting. For instance, Gemini-2.5-Pro, which achieved an Exact Accuracy of 65.20% in Task 1 among the higher scores observed in the full-context setting in our evaluation, saw its performance retract to 49.36% in Task 2, an absolute decline of 15.84 percentage points. Similarly, ChatGPT-4o-mini, despite a moderate static baseline (45.51%), suffered further degradation to 36.54% (a drop of 8.97 percentage points), exhibiting significant capability erosion.

Open-Weights Models. The capability gap was even more pronounced within open-weights architectures. For example, Qwen3-32B declined from 59.40% in Task 1 to 48.29% in Task 2 (Δ=11.11%). Most notably, the reasoning-distilled model DeepSeek-R1-Distill-Qwen-32B experienced the largest performance decline, decreasing from 53.42% to 33.92%, a reduction of 19.50 percentage points. These results collectively indicate that significant performance bottlenecks persist for both proprietary and open-weights models when facing tasks requiring autonomous inquiry planning and evidence acquisition.

Error Distribution Shift. The chart in Figure 2 further elucidates the shifting error patterns underlying this performance decline. Crucially, cases deemed Strictly Correct in Task 1 did not merely degrade into Partially Correct (ambiguous) diagnoses in Task 2; rather, a substantial proportion shifted directly to the Wrong category. This phenomenon, visualized by the significant expansion of the light green region, implies that in dynamic environments with incomplete information, models do not adopt a more conservative or cautious decision-making strategy. Instead, they are prone to constructing erroneous causal chains based on fragmented or insufficient evidence, leading to a complete deviation from the actual disease course. This error distribution shift highlights that the challenge in active diagnosis is not merely a drop in overall accuracy, but a larger share of cases shift directly from strictly correct to wrong in Task 2, rather than to partially correct, suggesting that errors under incomplete information are not consistently mitigated by more conservative outputs.

2.3 Evidence Quality and Safety Risks: Hallucinated Reasoning

We observe that the decline in evidence quality (StrictEQ) is larger than the decline in diagnostic accuracy. StrictEQ extends beyond mere diagnostic correctness; it evaluates whether the model can provide a complete, traceable chain of evidence, comprising both key positive and negative findings, derived from actual information acquired during the interaction. As such, StrictEQ serves as a direct proxy for the interpretability and auditability of the decision-making process. Consequently, the observation of a pronounced decoupling between diagnostic accuracy and evidence quality constitutes a distinct safety signal.

Hallucinated Reasoning. As illustrated by the prominent red bars in the right

panel of Figure 2, the Evidence Gap (the StrictEQ differential between Task 1 and Task 2) is consistently wider than the Accuracy Gap across the majority of models. In other words, as models transition from static to dynamic environments, their ability to retrieve evidence decays significantly faster than their ability to output correct diagnostic labels. This suggests that in a substantial proportion of cases, models arrive at the correct diagnosis via some correct diagnoses are produced without retrieving or citing all discriminative supporting evidence available under our interaction constraints. We operationally define this mode of correct outcome with missing evidentiary basis as Hallucinated Reasoning, effectively getting the right answer for the wrong (or non-existent) reasons.

Case Analysis. Gemini-2.5-Pro exemplifies this risk. While it maintained a diagnostic accuracy of 49.36% in Task 2, its StrictEQ score lagged significantly at 42.52%. This inversion implies that in approximately 7% of cases, the model produced "correct diagnoses" that failed to meet strict evidentiary standards. In these instances, the model's final output aligned with the reference standard, yet the dialogue history lacked the complete evidence set required to substantiate that conclusion. For Clinical Decision Support Systems (CDSS), such unsubstantiated inferences are inherently resistant to physician auditing, posing a risk of amplifying clinical errors if blindly trusted.

Safety Benchmark. In contrast, DeepSeek-v3-250324 demonstrated superior clinical rigor. Although its diagnostic accuracy (46.15%) was marginally lower than Gemini's, its StrictEQ reached 54.30%, the highest among all evaluated models. This indicates a unique characteristic where evidence retrieval performance surpasses diagnostic accuracy. Even in cases where DeepSeek-v3-250324 missed the precise reference diagnosis, it frequently succeeded in systematically acquiring and presenting key clinical evidence. This evidence-leading (as opposed to guess-leading) trait is vital for CDSS safety. It not only mitigates the latent risks of lucky guesses but also provides a solid, auditable foundation for human clinicians, thereby establishing a more robust trust boundary in human-AI collaboration.

2.4 Determinants of Performance: Scaling Laws and Reasoning Architecture

After establishing the existence of a substantial capability gap in active diagnosis, we next examined how model scale, reasoning paradigm, and the alignment between diagnostic accuracy and evidentiary support jointly shape performance (Fig.3).

Scaling behavior in active diagnosis. As shown in Fig.3 A, Task 2 strict diagnostic accuracy exhibits a monotonic, approximately log-linear improvement with increasing parameter size within each model family. Larger models benefit from enhanced multi-turn state tracking and long-context retention, which partially buffer the degradation observed when transitioning from passive to active diagnosis. For example, within the Qwen3 family, strict accuracy in Task 2 rises steadily from the 7B to the 32B variant, with a concomitant reduction in variance across cases. A similar trend is observed for the Qwen2.5 series. These scaling curves indicate that classical scaling laws extend beyond static benchmarks to confer robustness in interactive settings; however, even the largest models still experience a substantial active–passive gap, suggesting that

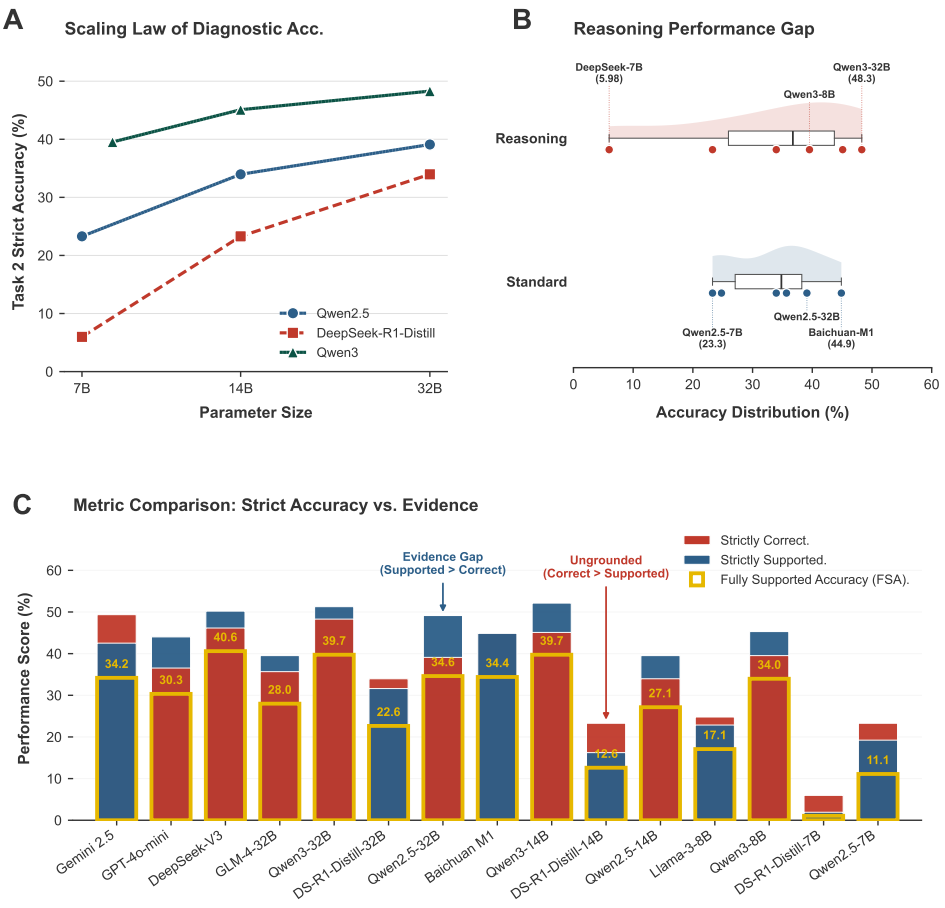

Fig. 3: **Scaling behavior, reasoning architectures, and diagnosis–evidence alignment in Task 2 active diagnosis. A** Strict diagnostic accuracy increases approximately log-linearly with model size within each family, but a residual capability gap remains. **B** Reasoning-enhanced models are generally more robust than size-matched instruction-tuned baselines, while small distilled reasoning models perform poorly in active inquiry. **C** For each model, red bars show strict diagnostic accuracy, blue bars Strict Evidence Quality, and gold bars Fully Supported Accuracy (FSA); the shaded "Ungrounded" region marks correct diagnoses lacking fully grounded evidence, whereas the "Evidence-leading" region highlights models whose evidence quality exceeds their diagnostic accuracy.

increased capacity alone is insufficient to close the deficit.

Reasoning architectures and robustness. Fig.3 B contrasts Task 2 performance between reasoning-enhanced models and size-matched standard instruction-tuned baselines. At medium and large scales, reasoning-oriented architectures such as Qwen3-32B achieve some of the highest strict accuracies in active inquiry, indicating that native long-horizon reasoning can translate into improved diagnostic planning under uncertainty. In contrast, distilled reasoning models at smaller scales, such as DeepSeek-R1-Distill-7B and 14B, display pronounced fragility: their Task 2 accuracies cluster near the bottom of the distribution despite exhibiting competitive performance in static settings. This asymmetry suggests that reasoning capabilities distilled from static Chain-of-Thought traces primarily optimize step-by-step deduction within fully observed contexts and do not consistently generalize to proactive evidence seeking. In other words, these models are proficient at reasoning over given information but struggle to decide what to ask next when information is missing.

Alignment between diagnosis and evidence. Beyond raw accuracy, Fig.3 C analyzes how often diagnostic predictions are backed by fully grounded evidence. For each model, the red bar denotes strict diagnostic accuracy in Task 2, the blue bar denotes Strict Evidence Quality (proportion of cases with evidence score $s_{\text{eq}}^{(i)} = 2$), and the bright gold overlay corresponds to Fully Supported Accuracy (FSA), defined as the proportion of cases where both the diagnosis and the supporting evidence achieve the maximum score ($s_{\text{acc}}^{(i)} = 2$ and $s_{\text{eq}}^{(i)} = 2$;). The upper shaded region labeled "Ungrounded" highlights cases in which strict diagnostic accuracy exceeds evidence quality, indicating correct labels that are not fully supported by acquired evidence. Conversely, the "Evidence-leading" region marks cases where evidence quality exceeds strict accuracy, reflecting models that collect coherent evidence but still fail to commit to the correct diagnosis. Clinically, the most desirable pattern is a high FSA with minimal ungrounded mass. In our cohort, only a small subset of models, such as DeepSeek-v3-250324 and Qwen3-32B, begin to approximate this evidence-leading profile, whereas many others display sizable ungrounded segments, reinforcing the safety concerns around hallucinated or unjustified reasoning in active diagnostic workflows.

2.5 System-Specific Performance Variability

Performance in active inquiry tasks exhibited pronounced domain-specific variability, a phenomenon intrinsically linked to the distinct diagnostic logic and epistemic structures underlying different clinical specialties (Fig. 4).

Resilient Domains. Most models demonstrated robust performance and minimal capability gaps within the Respiratory and Cardiovascular systems. We attribute this resilience to the fact that common pathologies in these fields, such as pneumonia and coronary artery disease, typically follow highly standardized symptom-sign-test-diagnosis mapping pathways. Pathognomonic features, such as specific radiation patterns of chest pain or descriptions of rales possess high textual expressivity and are readily accessible via semi-structured history taking. Consequently, models can effectively capture key diagnostic signals through standard inquiry workflows, keeping performance degradation within a controllable range.

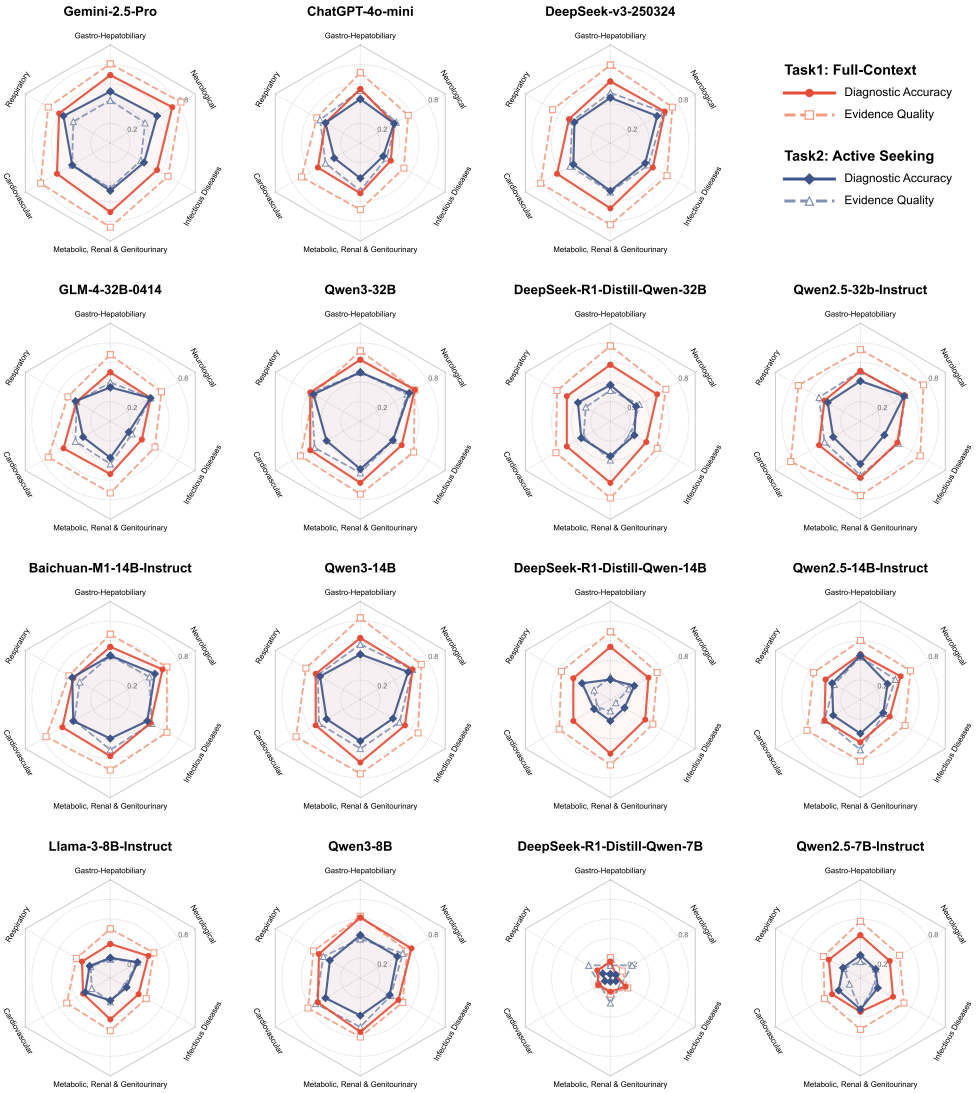

Fig. 4: **Specialty-specific performance and diagnostic fragility across clinical systems.** Paired analysis of diagnostic accuracy and evidence quality for each model across six clinical specialties.

Vulnerable Domains. Conversely, models displayed significant fragility in Neurological and Metabolic Renal cases, suffering the steepest declines in both accuracy and evidence quality. This vulnerability stems from specific clinical exigencies: **Neurological System:** Diagnosis relies heavily on nuanced, structured physical examination, including grading of muscle power and reflexes, assessment of ataxia, cranial nerve

deficits, and sensory level localization, and the systematic exclusion of pertinent negative signs. In a text-only interface devoid of multi-modal sensory input, models struggle to execute the precise verification required for anatomical localization. **Metabolic & Renal Systems:** These diagnoses necessitate the interpretation of complex combinatorial logic regarding laboratory values, including patterns of electrolyte imbalance, acid–base parameters, and longitudinal trends in renal function. This demands not only numerical sensitivity but also sophisticated longitudinal integration of multi-parametric data. Failure to proactively and precisely request specific biochemical markers often leads to misjudgment of disease severity or etiology.

Implication. We observe particularly low performance for some models in neurological differential diagnosis under our text-only protocol. This suggests that domains requiring fine-grained physical exam localization or multi-parameter lab interpretation may be more challenging in this setting, motivating cautious deployment and targeted optimization.

2.6 Impact of Data Source Complexity on Diagnostic Performance

Beyond inherent variations in clinical systems, the provenance and textual characteristics of case data serve as critical determinants of active diagnostic difficulty. Stratified analysis of Task 2 results reveals a stark performance divergence when models process standardized examination questions versus authentic real-world case reports. (Fig. 5).

Higher performance on standardized exam-style cases than on case-report-style records. In cases derived from MedQA (USMLE-style) and MedMCQA, the majority of models maintained elevated diagnostic accuracy in active inquiry tasks (averaging $> 50\%$). We attribute this to the highly distilled nature of exam data, which exhibits typical textbook-like characteristics: standardized symptom descriptions, short logical chains, and minimal irrelevant noise. Furthermore, given that models likely encountered similar schemas during pre-training, they can often lock onto diagnoses via pattern matching rather than de novo clinical reasoning. This simplification effect leads to a systematic inflation of evaluation results on standardized benchmarks.

Precipitous Decline in Real-World Data. In sharp contrast, performance across all models suffered a marked decline when evaluating real-world case reports from MedCaseReasoning (average accuracy $< 35\%$). These datasets mirror the high-entropy nature of actual practice, featuring long-context unstructured narratives with convoluted timelines, such as multiple visits and evolving disease courses, and a plethora of irrelevant confounders, including detailed social history, family history, or incidental findings. In such complex contexts, models must possess the high-order capability to separate signal from noise within redundant text, a challenge far exceeding that of extracting keywords from finite options in standardized tests.

Quantitative Evidence and Attribution. Even DeepSeek-R1, possessing the strongest reasoning capabilities, exhibited a performance gap of approximately 20 percentage points between MedQA and MedCase scenarios. This discrepancy elucidates a critical bottleneck: while current LLMs excel at processing structured knowledge, they struggle profoundly with information extraction and noise filtering amidst the

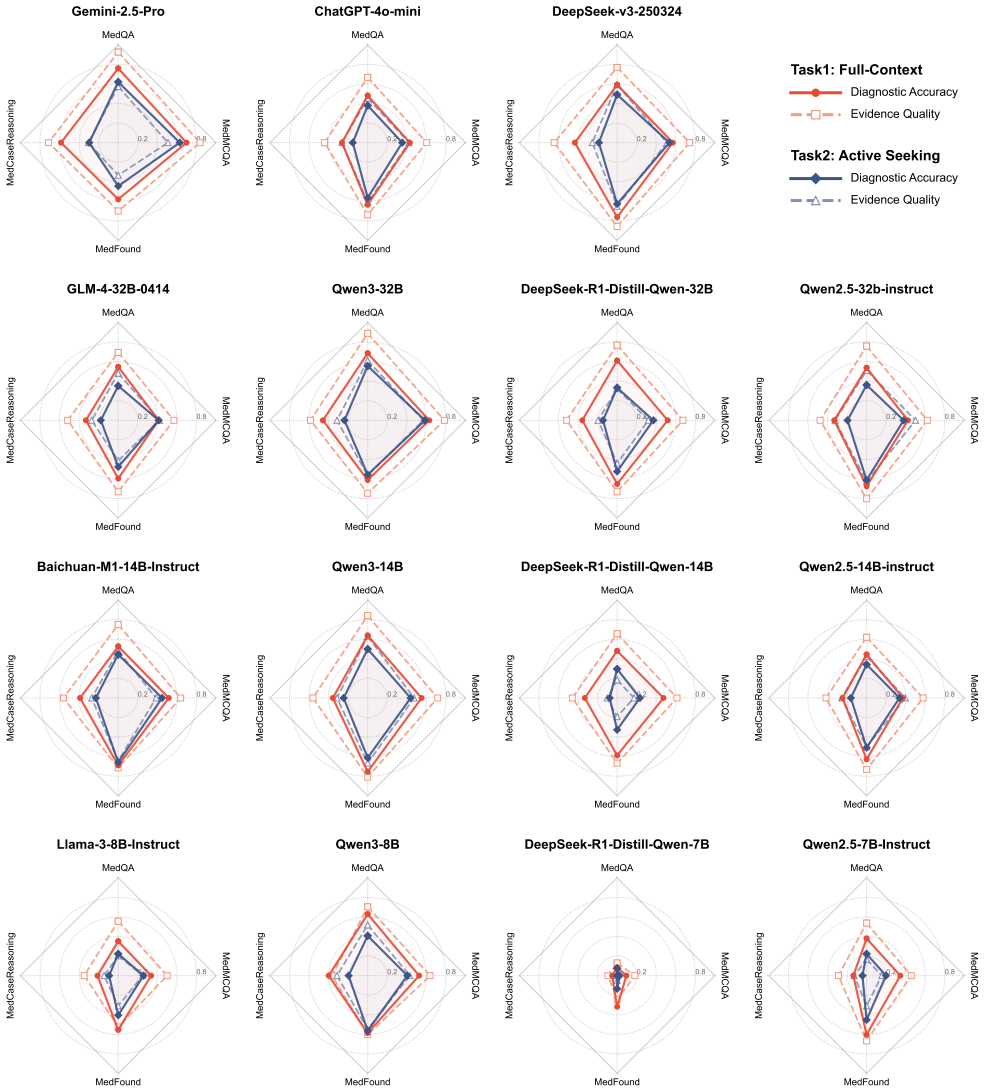

Fig. 5: **Impact of data source complexity and information density on active reasoning.**

unstructured complexity of real clinical records. These results are consistent with the view that performance on exam-style benchmarks may not fully characterize evidence extraction and noise filtering in more heterogeneous case records, and they motivate interactive evaluation as a complementary setting.

3 Discussion

Performance gap between full-context diagnosis and active evidence seeking. Using the ROUNDS-Bench framework, we quantify the performance gap between full-context medical question answering and interactive diagnostic inquiry, showing that strong static performance does not reliably carry over to staged, information-sparse settings. Our results highlight a limitation of prevailing full-context evaluations: models that perform well when all evidence is provided upfront often degrade when they must decide what to ask for and how to update hypotheses across turns. he paired decline in diagnostic accuracy and evidence quality suggests caution in treating licensing-style, full-context scores as stand-alone proxies for interactive clinical decision support behavior. More broadly, these findings underscore that static medical QA and interactive diagnostic workups impose different operational demands on models. Interactive diagnosis additionally requires purposeful information acquisition, uncertainty-aware hypothesis management, and evidence synthesis over multiple turns.

Cognitive interpretation: from pattern completion to hypothesis-driven inquiry. From a cognitive perspective, the degradation may reflect a mismatch between next-token prediction optimized for pattern completion and the hypothesis-driven inquiry required in clinical workups. In static, full-context settings, models can often map a complete symptom/evidence set to a diagnosis by leveraging distributional regularities learned during training. By contrast, clinical diagnosis typically follows a hypothetico-deductive process in which clinicians generate hypotheses, seek discriminative evidence, and revise beliefs as new information emerges. Our results suggest that many models are less reliable in the strategic layer of this process, often exhibiting inefficient questioning and premature diagnostic closure. One interpretation is that models capture many medical entities and associations, but still struggle with procedural decision-making such as selecting high, yield next questions/tests and using them to update a differential diagnosis over time.

Safety implications: evidence gaps and reduced auditability in interactive workflows. A key safety-relevant observation is "hallucinated reasoning," where diagnostic labels can appear correct even when the interaction fails to elicit sufficient supporting evidence. Across many cases, models produced correct labels despite not having acquired the key positive and negative findings needed to justify those conclusions. Although such outcomes can improve scores on static benchmarks, in clinical decision support they reduce auditability and can elevate risk by weakening a clinician's ability to verify the rationale. A system operating on probabilistic intuition rather than a verifiable chain of evidence is inherently untrustworthy for physician oversight. One possible explanation is that common optimization targets emphasize final-answer correctness more than evidentiary sufficiency and traceable support in multi-turn settings. To improve safety, future work could explicitly penalize unsupported "correct" answers and optimize for evidentiary sufficiency and traceability alongside accuracy.

Determinants: model scale, reasoning scaffolds, and planning for inquiry. Our analysis of model scale and reasoning-related training signals provides insight into factors associated with improved performance under active inquiry. While larger models appear more resilient in longer interactions, scaling alone does not eliminate the

gap between full-context and staged-information settings. By comparison, models with stronger multi-step reasoning behavior (e.g., CoT-style traces or RL-aligned variants) showed better stability in maintaining evidence quality across turns. These patterns suggest that beyond knowledge capacity, planning, deciding what to ask next and when enough information has been gathered is a major bottleneck. Active evidence seeking requires maintaining an evolving state representation and selecting prospective queries, which is closer to agentic decision-making than single-pass next-token prediction over a fixed context.

Limitations. Limitations. This study has limitations that are inherent to simulation-based evaluation. First, while the standardized patient simulator improves reproducibility, it abstracts away sociopsychological factors present in real encounters, which may limit direct generalization to practice. Second, the interaction is text-only and does not capture multimodal evidence (e.g., imaging or waveform data) that is central to many clinical domains.Third, a single-reference diagnosis may not fully reflect real-world comorbidity and diagnostic uncertainty. Finally, because parts of the grading rely on LLM-based evaluation, additional validation with expert clinician adjudication would strengthen the conclusions.

ROUNDS-Bench provides a complementary evaluation perspective that emphasizes model behavior under staged information release, in addition to what models know under full-context conditions. The observed performance decline suggests that progress on static QA should be paired with methods that improve interactive evidence seeking, auditability, and safety constraints in workflow-like settings. Future research could explore reinforcement learning from clinician feedback to reward evidence-based reasoning, discourage low-yield inquiries, and promote efficient information acquisition. In addition, evaluation could incorporate stopping criteria that test whether models can decide when sufficient information has been gathered before committing to a diagnosis. Improving dynamic inquiry and decision thresholds is likely necessary for medical AI to move from passive assistance toward more reliable participation in clinical workflows.

4 Methods

This section elaborates on the development of ROUNDS-Bench, the interaction mechanics of the Standardized Patient (SP) simulator, the dual-task evaluation protocol, and the multidimensional metrics used for assessment. Additionally, all related prompt templates and data cleaning scripts are included.

4.1 Data Curation and Stratification Pipeline

To construct a high-quality benchmark that bridges clinical representativeness with the precise evaluation of active inquiry capabilities, we designed and implemented a rigorous Three-Stage Construction Pipeline. This pipeline encompasses the entire lifecycle of data curation, from the aggregation of heterogeneous multi-source data to multi-stage structural cleaning and clinical system stratification.

1. Multi-source Aggregation. To maximize data diversity and minimize domain bias arising from single sources, we aggregated four high-quality medical corpora,

covering the full spectrum from standardized examinations to authentic clinical narratives:

MedQA (USMLE) & MedMCQA: Representing standardized medical licensing examinations, providing expert-reviewed, typical clinical presentations.

MedFound & MedCaseReasoning: Providing long-context clinical case reports closer to real-world scenarios. These datasets feature unstructured timelines, redundant information, and atypical symptoms, serving to evaluate model robustness in high-noise environments.

2. Multi-stage Filtering and Structuring. Raw data typically exist as unstructured text blocks or contain non-diagnostic questions. To adapt this data for the SP-sim's Gating Mechanism, we developed a processing protocol comprising three strict stages:

Stage I: Diagnostic-only Filtering. To rigorously isolate diagnostic reasoning from treatment or prognostic tasks, we implemented a dual-filtering mechanism: **Diagnosis-type Filter:** Determines whether the question requires inferring a disease from clinical manifestations, explicitly precluding best next step in management or pathophysiological mechanism queries. **Diagnosis-term Filter:** Verifies that candidate answers correspond to distinct Disease Entities, rather than symptoms or test names. Cases were retained only if they passed both checks.

Stage II: Schema-constrained Structuring. Leveraging DeepSeek-v3, we restructured the retained free text into structured records comprising six standardized sections. We utilized a prompt with hard constraints (see Appendix A, Fig. A2) to enforce the following rules: **Section Delineation:** (1) Patient Info, (2) Chief Complaint, (3) History of Present Illness, (4) Past Medical History, (5) Physical Exam, and (6) Auxiliary Exam. **Prevention of Data Leakage:** Explicit prohibition of including the final diagnosis or any suggestive conclusions within the input sections. **Null Value Handling:** Information not present in the source text must be explicitly marked as None, strictly prohibiting the model from generative completion to mitigate hallucination risks.

Stage III: Structural Validation and Integrity Check. We introduced an independent, lightweight Validator to audit the structured data. The validator performs a binary assessment on two dimensions: **Completeness:** Verifying the presence of all six required sections. **Faithfulness:** Verifying that structured content is grounded entirely in the source text, checking for confabulated details. Only cases passing this validation were admitted to the final repository, ensuring data veracity and consistency.

3. Clinical System Stratification. To eliminate the impact of class imbalance on evaluation metrics, we adopted a Rule-based System Categorization strategy rather than random sampling. Using medical ontology logic, cleaned cases were automatically mapped to six core clinical systems: Cardiovascular, Respiratory, Gastro-Hepatobiliary, Neurological, Infectious Diseases, and Metabolic, Renal & Genitourinary. The final ROUNDS-Bench comprises 468 cases, strictly controlled at 78 cases per system ($N = 78$). This balanced stratified design (see Table 1) ensures equitable comparison of model performance across different clinical specialties and supports granular, system-specific analysis.

4.2 Standardized Patient (SP) Simulator Implementation

A core methodological innovation of this study is the development of a high-fidelity Standardized Patient Simulator (SP-sim) powered by Large Language Models. Unlike traditional static benchmarks, the SP-sim is specifically engineered to simulate the information asymmetry inherent in authentic clinical encounters: the physician is denied full access to information at the outset and must instead acquire evidence incrementally through sequential inquiry and test ordering. To ensure the controllability and reproducibility of the interaction, we eschewed open-ended, free-form chatbot paradigms in favor of a strictly controlled Request-Response Engine. In this system, the doctor model initiates structured requests, and the SP-sim returns finite, gated information based on predefined rules.

1. Simulator Architecture and Base Model. The SP-sim is driven by the Qwen2.5-32B-Instruct model. This selection was motivated by the model's stability in instruction following, enabling it to rigorously execute our gating rules without generating boundary-crossing responses or hallucinations. To maximize determinism and experimental reproducibility, we fixed the inference temperature at 0.0 and locked the random seed, ensuring that identical inputs yield strictly identical outputs from the SP-sim.

2. Initialization and Opening Strategy. To prevent doctor models from overfitting to a single opening pattern, we designed 15 diverse opening templates for the SP-sim (see Appendix B, Fig. B6), ranging from "What brings you in today?" to "Please describe any discomfort." At Turn 0, regardless of the doctor model's opening approach, the SP-sim is programmed to release only two sections from the structured record: (1) Patient Information and (2) Chief Complaint. All other clinical evidence modules, History of Present Illness, Past Medical History, Physical Exam, and Auxiliary Exams, remain in a Hidden State, completely invisible to the doctor model. These sections are unlocked only when the doctor model issues compliant queries in subsequent turns.

3. Information Gating Mechanism. The Information Gating Mechanism is the core logic of the SP-sim, designed to strictly control the pace of evidence disclosure while maintaining realism. In each turn, the simulator parses the doctor model's questions in real-time and makes decisions based on natural language understanding and matching rules.

Allowed Query Scopes. The SP-sim responds exclusively to explicit queries targeting specific modules: (1) HPI, (2) PMH, (3) Physical Exam, and (4) specific categories of Auxiliary Exams. Generalized queries falling outside these scopes, for example "Tell me everything", do not trigger information release, thereby preventing models from bypassing the sequential inquiry process via unrestricted omniscient queries.

Intent Recognition and Verbatim Retrieval. When the doctor model issues a compliant query, such as "Please auscultate the chest", the SP-sim locates the corresponding module in the hidden structured data and executes the following strategy:

Hit: If a corresponding entry exists in the structured record, for instance when the Ground Truth records "bibasilar crackles", the SP-sim performs Verbatim Retrieval, returning the raw text without semantic rewriting or summarization. This design

ensures that the information exposed in Task 2 is semantically and linguistically identical to the full case record in Task 1, maintaining strict consistency. **Miss / Negative Feedback:** If the queried test was not performed or the information is absent from the record, the SP-sim returns a standardized negative response, such as "This test was not performed yet" or "The patient denies this symptom". This fixed-format response serves as a guardrail against hallucination, preventing the SP-sim from fabricating false positives while clearly signaling to the doctor model that the information is absent or negative.

4. Leakage Prevention and Termination Criteria.

Leakage Prevention. To prevent models from acquiring information beyond the task scope via prompt injection or leading questions, the SP-sim is embedded with high-priority system-level anti-leakage instructions. Regardless of how the doctor model guides the conversation (including direct requests for the diagnosis), the SP-sim is constrained to never reveal the Final Diagnosis or provide suggestive conclusions during intermediate turns, ensuring the independence and rigor of the evaluation.

Termination Criteria. The interaction proceeds until one of the following conditions is met: **Doctor-Initiated Termination.** The doctor model deems the evidence sufficient and voluntarily outputs the [Final Diagnosis] tag. In our protocol, this marks the decision point for diagnostic judgment. **Maximum Turn Limit Reached.** If the dialogue reaches the preset maximum (Max Turns = 10), the SP-sim ceases to provide further information, and the doctor model is forced to generate a final diagnosis based on currently available evidence. This constraint prevents infinite inquiry loops and incorporates the ability to judge "when to stop" into the evaluation dimensions.

4.3 Evaluation Framework

To precisely quantify the impact of active evidence-seeking on diagnostic performance, we designed a rigorous within-subject controlled dual-task experiment. The core objective of this protocol is to methodologically isolate and contrast model capabilities in static knowledge retrieval versus dynamic information acquisition and reasoning, while holding the underlying clinical case content constant.

Task 1: Full-Context Diagnosis (Static Baseline.) Serving as the control condition, Task 1 evaluates the model's diagnostic upper bound under conditions of Information Transparency.

- **Input:** The model receives the fully structured and cleaned case record $\mathcal{X}_{\text{full}}$ instantaneously. This includes all available history, physical examination details, and auxiliary test results, exposing all diagnostic evidence in a single pass.
- **Output Requirements:** The model is instructed to directly generate a [Final Diagnosis] and concurrently list three key evidence items supporting that diagnosis. Evidence can be drawn from any section without sequential constraints.
- **Objective:** To establish the Theoretical Performance Upper Bound achievable under idealized conditions with zero information acquisition barriers, serving as a reference baseline for measuring performance loss in the active setting.

Task 2: Active Evidence-Seeking (Dynamic Evaluation). Task 2 simulates the entropy reduction process inherent in authentic clinical practice, transitioning

from high uncertainty to diagnostic convergence. The model assumes the role of a physician, tasked with actively constructing the patient profile from an initial state of information scarcity through multi-turn interaction.

- **Initialization:** The model is provided only with Patient Demographics and the Chief Complaint, such as "Male, 45, chest pain for 3 hours". All information regarding HPI, PMH, Physical Exam, and Auxiliary Exams remains in a Hidden State, completely invisible to the model.
- **Discrete Action Space ($\mathcal{A}$):** In each turn, the model must execute a single action from a predefined set:
  - **History Taking**: Request [History of Present Illness] or [Past Medical History].
  - **Physical Exam**: Request [Physical Examination] to retrieve vital signs and physical findings.
  - **Diagnostic Testing**: Issue specific orders via Request [Test Name], such as Request [Chest X-ray].
  - **Termination**: Output [Final Diagnosis] to conclude the interaction.
- **Interaction Constraints and Process Control:** To simulate Resource Constraints and Time Pressure in clinical practice, we impose strict limitations:Single-Step Execution: Only one query is permitted per turn. This constraint compels the model to perform Clinical Triage, prioritizing queries based on diagnostic yield rather than batch-requesting tests to evade decision planning.Turn Cap: The interaction is strictly limited to 10 turns. If this limit is reached, the model is forced to render a diagnosis based on currently available evidence.No Repetition: Redundant requests for previously acquired information or tests marked Not Performed are prohibited to penalize inefficient operational loops.
- **Output Specification & Evidence Grounding:** Non-Terminal Turns: Models must output in the format (Action): (Rationale), explicitly stating the clinical reasoning behind the chosen action to enhance the interpretability of the inquiry process. Evidence Grounding Rule: In the final diagnostic phase, the three supporting evidence items $\hat{E}$ must be derived strictly from content actually released by the SP-sim during the Interaction History. Citing information that never appeared in the dialogue, such as hallucinated results or data visible only in Task 1, constitutes Hallucinated Evidence and results in a penalty to the Evidence Quality score.

4.4 Evaluation Metrics

To achieve objective, quantifiable, and reproducible assessment within the context of unstructured natural language generation, we established a Deterministic Evaluation Protocol based on the LLM-as-a-Judge paradigm. This protocol centers on two core endpoints: (1) Diagnostic Accuracy and (2) Reasoning Evidence Quality.

1. Evaluator Setup & Protocol. To ensure consistency across different models and runs, we employed DeepSeek-v3 as a fully independent Arbitrator, implementing strict parameter controls on its inference behavior: **Deterministic Inference:** All evaluation calls were executed with a temperature of 0.0 and Top-p of 1.0. This eliminates sampling stochasticity, ensuring that identical inputs yield strictly consistent

scores across multiple evaluations. **Blinded Evaluation:** The arbitrator operates in a blinded manner, accessing only the model's output (diagnosis and evidence) and the Gold Standard. Although DeepSeek-v3 is also included as one of the evaluated models, the evaluator is used in a strictly separate capacity with frozen weights, standardized prompts, and blinded access to model identities, thereby minimizing potential bias toward specific architectures or stylistic patterns.**Standardized Prompts:** We designed structured scoring prompts to guide the arbitrator based on predefined scales, rather than allowing free-form commentary, to maximize the standardization of the scoring process.

2. Exact Diagnostic Accuracy. We prioritize clinical precision over mere semantic proximity. For each case i, the model-generated final diagnosis $\hat{y}^{(i)}$ is semantically compared against the gold standard $y^{*(i)}$ using a 3-Level Scoring Scale to derive the score $s_{acc}^{(i)}$:

- **2 (Strictly Correct):** The diagnosis implies Semantic Equivalence with the gold standard, for example Myocardial Infarction vs. Heart Attack, and exhibits precise matching of the Disease Subtype.
- **1 (Partially Correct):** The diagnosis falls within the correct Disease Category or family but lacks specificity, for example Meningitis vs. Viral Meningitis, or presents an incorrect subtype within the correct lineage.
- **0 (Incorrect):** The diagnosis is factually wrong, unrelated, or consists only of non-diagnostic entities, such as symptoms.

Primary Metric: We report Exact Diagnostic Accuracy (ExactAcc), defined as the proportion of cases in the total sample N achieving a score of 2:

$$\text{ExactAcc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[s_{acc}^{(i)} = 2]. \tag{1}$$

This metric intentionally excludes "partially correct" instances, establishing a rigorous standard for diagnostic precision aligned with clinical requirements.

3. Evidence Quality (StrictEQ). To quantify the risk of "Hallucinated Reasoning", where models "guess" the correct diagnosis without reliable evidentiary support, we introduced the Strict Evidence Quality (StrictEQ) metric. This metric specifically evaluates whether the model can construct a complete, traceable chain of evidence based only on information actually acquired during the interaction. For each case, the model outputs three evidence items $\hat{E} = \{e_1, e_2, e_3\}$. The assessment involves two constraints:

Grounding Check: In Task 2, we first extract the set of all information fragments actually retrieved by the model during T turns from the SP-sim interaction log $\mathcal{S}_T$. We then verify whether each claimed evidence item e_k can be semantically matched within this history:

$$\forall \hat{e}_k \in \hat{E},\ \hat{e}_k \in \text{SemanticMatch}(S_T). \tag{2}$$

If an evidence item e_k cannot be sourced from the interaction log, meaning the test was never requested, the sign was never disclosed, or the model fabricated a "new fact", it is classified as "Hallucinated Evidence." This step strictly distinguishes between

citations grounded in real interaction and confabulated details.
Scoring Criteria: Upon passing the grounding check, evidence is further scored based on logical support strength $s_{\text{eq}}^{(i)} \in \{0, 1, 2\}$: At the dataset level, we summarize strict evidence performance using the Strict Evidence Quality (StrictEQ) metric, defined as the proportion of cases with maximally supported evidence:

$$\text{StrictEQ} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\big[s_{\text{eq}}^{(i)} = 2\big]. \tag{3}$$

- **2 (Strictly Supported):** All three evidence items are grounded in the interaction history; and the evidence provides clear, positive clinical support for the diagnosis, such as pathognomonic signs or exclusionary negatives.
- **1 (Partially Supported):** Only 1–2 items are valid, or the logical link to the diagnosis is weak/tangential.
- **0 (Unsupported):** All evidence is hallucinated, contradicts the history, or lacks a causal relationship with the diagnosis.

Fully Supported Accuracy (FSA). While ExactAcc captures how often the final diagnosis is strictly correct, it says nothing about whether the accompanying evidence is reliable. To jointly require both perfect diagnosis and perfect evidentiary support within the same task, we define the Fully Supported Accuracy (FSA) metric. At the case level, we refer to $s_{\text{acc}}^{(i)} = 2$ as a strictly correct diagnosis and $s_{\text{eq}}^{(i)} = 2$ as strictly supported evidence. FSA is then defined as the proportion of cases in the total sample N that satisfy both conditions simultaneously:

$$\text{FSA} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\big[s_{\text{acc}}^{(i)} = 2 \ \wedge \ s_{\text{eq}}^{(i)} = 2\big]. \tag{4}$$

By construction, FSA is a joint metric that extends beyond accuracy-only measures: it captures the subset of model predictions that are not only strictly correct but also fully supported by grounded evidence. Applied separately to Task 1 and Task 2, FSA enables us to quantify how many clinically safe, well-justified decisions are preserved when moving from static full-context evaluation to active evidence-seeking.

4.5 Evaluated Models and Implementation Details

To comprehensively benchmark the performance of the current LLM ecosystem in active clinical reasoning, we selected a representative cohort of 15 models, stratified by Parameter Scale, Access Mode, and Reasoning Architecture. Our evaluation spans the full spectrum from lightweight edge models to hyperscale cloud-based models:

- **Proprietary SOTA Models:** This tier establishes the performance benchmark and includes Gemini-2.5-Pro, ChatGPT-4o-mini, and DeepSeek-v3-250324. These models represent the current upper bound of generalist intelligence and large-scale architectural optimization[38–40].

- **Reasoning-Enhanced Models:** A core focus of our study is the evaluation of models featuring native Chain-of-Thought (CoT) and long-horizon reasoning capabilities. This category includes the Qwen3 series (8B, 14B, 32B) and DeepSeek-R1-Distill (7B, 14B, 32B) variants. Unlike standard models, these reasoning models are optimized via Reinforcement Learning (RL) to internalize exploratory reasoning paths, allowing us to assess whether native CoT improves diagnostic inquiry [41, 42].
- **Generalist Instruction-Tuned Models:** To provide a controlled comparison, we included the Qwen2.5 series (7B, 14B, 32B), Llama-3-8B-Instruct, GLM-4-32B, and Baichuan-M1-14B. This enables an empirical analysis of the Reasoning Premium, the performance gain attributable to reasoning specific tuning over standard instruction-following [43–46].

Acknowledgements. This work was supported by the National Natural Science Foundation of China (82303695), Shanghai Municipal Natural Science Foundation (23ZR1425400), and Shanghai Soft Science Project (25692114700), 2025 Annual Special Program for the "Public Benefit" Science Popularization and Innovation Project of Xuhui District, xhkp-HM-2025037, the Clinical Research Project of the Shanghai Municipal Health Commission (20214Y0468), Zhongshan Hospital Management Science Foundation Project 2024ZSGL09 and Zhongshan Hospital Medical Humanities and Ideological Education Research Project 2024YXRWSZ-007.

Code availability

The code used to generate the results in this study is publicly available at https://github.com/Leonard-zc/ROUNDS-Bench.

Author contributions

Chen Zhan, Xihe Qiu, Xiaoyu Tan and Xibing Zhuang contributed equally to this work. Chen Zhan, Xihe Qiu, Xiaoyu Tan and Xibing Zhuang conceived and designed the study. Chen Zhan and Gengchen Ma conducted the experiments and collected the data. Chen Zhan and Yue Zhang analyzed the data. Chen Zhan, Xihe Qiu and Xiaoyu Tan drafted the manuscript. Shuo Li and Peifeng Liu provided critical feedback. Xihe Qiu, Xiaoxiao Ge, Liang Liu and Lu Gan supervised and administered the project, acquired funding, and revised the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Liao, Y. *et al.* Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495* (2024).

[2] Idan, D. & Einav, S. Primer on large language models: an educational overview for intensivists. *Critical Care* **29**, 238 (2025).

[3] Sandmann, S., Riepenhausen, S., Plagwitz, L. & Varghese, J. Systematic analysis of chatgpt, google search and llama 2 for clinical decision support tasks. *Nature communications* **15**, 2050 (2024).

[4] Jin, D. *et al.* What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* **11**, 6421 (2021).

[5] Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W. & Lu, X. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146* (2019).

[6] Pal, A., Umapathi, L. K. & Sankarasubbu, M. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering (2022). URL https://arxiv.org/abs/2203.14371. arXiv:2203.14371.

[7] Singhal, K. *et al.* Toward expert-level medical question answering with large language models. *Nature Medicine* **31**, 943–950 (2025).

[8] Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).

[9] Brin, D. *et al.* Comparing chatgpt and gpt-4 performance in usmle soft skill assessments. *Scientific Reports* **13**, 16492 (2023).

[10] Liu, F. *et al.* A medical multimodal large language model for future pandemics. *NPJ Digital Medicine* **6**, 226 (2023).

[11] Bedi, S. *et al.* Testing and evaluation of health care applications of large language models: a systematic review. *Jama* (2025).

[12] Coderre, S., Mandin, H., Harasym, P. H. & Fick, G. H. Diagnostic reasoning strategies and diagnostic success. *Medical education* **37**, 695–703 (2003).

[13] Elstein, A. S., Shulman, L. S. & Sprafka, S. A. *Medical problem solving: An analysis of clinical reasoning* (Harvard University Press, 1978).

[14] Harden, R. M., Stevenson, M., Downie, W. W. & Wilson, G. Assessment of clinical competence using objective structured examination. *Br Med J* **1**, 447–451 (1975).

[15] Khan, K. Z., Ramachandran, S., Gaunt, K. & Pushkar, P. The objective structured clinical examination (osce): Amee guide no. 81. part i: an historical and theoretical perspective. *Medical teacher* **35**, e1437–e1446 (2013).

[16] Barrows, H. S. An overview of the uses of standardized patients for teaching and evaluating clinical skills. aamc. *Academic medicine* **68**, 443–51 (1993).

[17] Yao, Z. *et al.* Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. *arXiv preprint arXiv:2410.01553* (2024).

[18] Jiang, Y. *et al.* Medagentbench: a virtual ehr environment to benchmark medical llm agents. *NEJM AI* **2**, AIdbp2500144 (2025).

[19] Williams, C. Y., Miao, B. Y., Kornblith, A. E. & Butte, A. J. Evaluating the use of large language models to provide clinical recommendations in the emergency department. *Nature communications* **15**, 8236 (2024).

[20] Li, S. *et al.* Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems* **37**, 28858–28888 (2024).

[21] Chen, S. *et al.* Meddialog: a large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329* **3** (2020).

[22] Saley, V. V., Saha, G., Das, R. J., Raghu, D. *et al.* Meditod: An english dialogue dataset for medical history taking with comprehensive annotations. *arXiv preprint arXiv:2410.14204* (2024).

[23] Tsoukalas, A., Albertson, T., Tagkopoulos, I. *et al.* From data to optimal decision making: a data-driven, probabilistic machine learning approach to decision support for patients with sepsis. *JMIR medical informatics* **3**, e3445 (2015).

[24] von Kleist, H., Zamanian, A., Shpitser, I. & Ahmidi, N. Evaluation of active feature acquisition methods for time-varying feature settings. *Journal of Machine Learning Research* **26**, 1–84 (2025).

[25] Markus, A. F., Kors, J. A. & Rijnbeek, P. R. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics* **113**, 103655 (2021).

[26] Tu, T. *et al.* Towards conversational diagnostic artificial intelligence. *Nature* 1–9 (2025).

[27] Zhu, J., Pan, J., Liu, Y., Liu, F. & Wu, J. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning (2025). URL https://arxiv.org/abs/2502.07143. arXiv:2502.07143.

[28] Walker, L. *Artificial narrow intelligence-driven diagnostics: impacts, inequities, and policy imperatives in global healthcare.* Ph.D. thesis, Technische Universität Wien (2024).

[29] Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* **36**, 46595–46623 (2023).

[30] Ke, Y. H. *et al.* Clinical and economic impact of a large language model in perioperative medicine: a randomized crossover trial. *npj Digital Medicine* **8**, 462 (2025).

[31] Griot, M., Hemptinne, C., Vanderdonckt, J. & Yuksel, D. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications* **16**, 642 (2025).

[32] Gaber, F. *et al.* Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine* **8**, 263 (2025).

[33] Luo, M.-J. *et al.* A large language model digital patient system enhances ophthalmology history taking skills. *NPJ Digital Medicine* **8**, 502 (2025).

[34] Hurst, A. *et al.* Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[35] Liu, X. *et al.* A generalist medical language model for disease diagnosis assistance. *Nature medicine* **31**, 932–942 (2025).

[36] Wu, K. *et al.* Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports (2025). URL https://arxiv.org/abs/2505.11733. arXiv:2505.11733.

[37] Johnson, A. E. *et al.* Mimic-iii, a freely accessible critical care database. *Scientific data* **3**, 1–9 (2016).

[38] DeepSeek-AI, A. L. *et al.* Deepseek-v3 technical report, 2024. *URL https://arxiv.org/abs/2412.19437* (2024).

[39] Team, G. *et al.* Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[40] Comanici, G. *et al.* Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[41] Guo, D. *et al.* Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[42] Yang, A. *et al.* Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[43] GLM, T. *et al.* Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793* (2024).

[44] Wang, B. *et al.* Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671* (2025).

[45] Grattafiori, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[46] Hui, B. *et al.* Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186* (2024).

Appendix A Extended dataset construction

This appendix provides the exact prompt templates and examples used in the three-stage data curation pipeline described in Section 4.1. We include only implementation details that are necessary for replication and omit conceptual descriptions that already appear in the main text.

A.1 Stage I: diagnosis-type and diagnosis-term filtering

To isolate items that genuinely require diagnostic reasoning, we first apply a pair of lightweight filters that remove questions about mechanisms, treatment, prognosis, or next best step decisions. A1

Prompt 1 — Diagnosis-Type Filter

You are a medical board exam assistant. Please determine whether the following question is a diagnosis-type question.
A diagnosis-type question requires the examinee to determine the most likely diagnosis or disease based on the clinical presentation. It does not ask about etiology, treatment, mechanisms, or next steps.
Please answer with only one word: "Yes" or "No".
Question stem: `<question_text>`
Options: `<options_text>`
Answer:

Prompt 2 — Diagnosis-Term Filter

You are a medical expert. Please determine whether the following terms are medical diagnoses (names of diseases or conditions), rather than symptoms, tests, mechanisms, or treatments.
Please answer with only one word: "Yes" or "No".
Options: `<options_text>`

Fig. A1: Side-by-side prompts used for diagnostic-only filtering (Stage 1).

A.2 Stage II: schema-constrained structuring

For all retained items, we convert the free-text vignette into a six-section clinical record using a schema-constrained prompt:A2

A.3 Stage III: structural validation and fabrication check

A separate validation prompt audits the structured record for completeness and faithfulness:A3

Prompt: Schema-Constrained Structuring

You are a clinical documentation expert. Convert any form of clinical vignette or patient record into a standardized medical record format. Do NOT infer or include any diagnosis.
Use this exact structure:
1. Patient Information
- [Sex, Age] (or "None")

2. Chief Complaint
- [Primary symptom] + [Duration] (or "None")

3. History of Present Illness
- Progression: [Chronological illness course]
- Accompanying symptoms: [Comma-separated symptoms]
(Use "None" if not available)

4. Past Medical History
- [Relevant history] (or "None")

5. Physical Examination
- Vital signs: [Temperature, HR, RR, BP...] (or "None")
- [System-specific findings] (or "None")

6. Auxiliary Examination
- (1) Imaging test: [Findings] (or "None")
- (2) Laboratory tests: [Key abnormal results] (or "None")
- (3).....

Repeat back only the structured output.
Please convert the following clinical vignette or case into a structured medical record.
Do NOT include any diagnosis results. Fill "None" for missing fields. The text may be a clinical note or exam question.

Case: `question_text`

Please strictly follow this format, but leave out the final diagnosis:
`1.Patient Information`
`2.Chief Complaint`
`3.History of Present Illness`
`4.Past Medical History`
`5.Physical Examination`
`6.Auxiliary Examination`

Fig. A2: Schema-constrained prompt used to convert free-text vignettes into a six-section clinical record (no diagnosis). The schema enforces fixed headers, no leakage, and explicit "None" for missing fields.

A.4 Illustrative structured case (no diagnosis).

Figure A4 presents an example record after structuring: demographics, a concise chief complaint with timeline, focused HPI, vitals and exam, and multi-modal auxiliary studies (imaging, laboratory, ECG, hemodynamics, biopsy). No diagnostic label is included at this stage, preserving the intended separation between *evidence* and *diagnosis* for both Task 1 (full-context) and Task 2 (active evidence-seeking).

A.5 System-level categorization.

Finally, diseases are mapped to six clinical systems using a transparent rule-based prompt (Fig. A5). The instruction returns a JSON tuple {`primary_diagnosis`,

Prompt: Structure & Fabrication Check

You are a critical evaluator of medical documentation quality.
Your task is to determine whether the following structured medical record is a faithful and properly formatted transformation of the original clinical case description.
Evaluation Criteria:

1. The record must contain all 6 required sections with their respective content:
   - Patient Information
   - Chief Complaint
   - History of Present Illness
   - Past Medical History
   - Physical Examination
   - Auxiliary Examination

   Note: Minor variations in section titles (e.g., spacing, punctuation, casing) are acceptable as long as the structure is clearly preserved.
2. The structured record must **not fabricate any content**. All included details must be:
   - Explicitly stated in the original case, **or**
   - Clearly implied with no assumptions beyond clinical description.
3. If any section lacks source information, using `"None"` is acceptable.

Original Case: `original_text`
Structured Medical Record: `formatted_record`
Please assess strictly but reasonably.
Answer only with `yes` (fully valid) **or** `no` (any fabrication, omission, or structural failure).

Fig. A3: Validation prompt for auditing the structured record against the original text: all six sections present, no fabrication ("None" allowed), and faithful formatting. Output is a strict yes/no.

`category`}, aligning each case to one of the six systems or `Other` when appropriate. This categorization underpins the balanced distribution reported in Table 1 and supports per-system analyses in the experiments.

Appendix B Interaction and evaluation prompts

B.1 Opening utterances.

To initiate the encounter, the clinician uses one of **15** templated openings (Fig. B6); these variants reduce prompt overfitting to any particular phrasing. Upon the first opening, the SP-sim must return **exactly** the `1.Patient Information` and `2.Chief Complaint` sections from the structured record produced in the previous stage, without any inference or additional details.

Structured Case Example

1. Patient Information
- Male, 44

2. Chief Complaint
- Chills for 3 days and arthralgias in the knees and hips (preceded by several days of unproductive cough and headache)

3. History of Present Illness
- Progression: Unproductive cough and headache preceded chills and arthralgias. One week before presentation, he was treated with a macrolide antibiotic and an NSAID.
- Accompanying symptoms: Cough, Headache, Chills, Arthralgias

4. Past Medical History
- Smoking history (None otherwise)

5. Physical Examination
- Vital signs: Temperature 38.5 °C, Heart Rate 113/min, Blood Pressure 126/64 mmHg, Oxygen Saturation 98% on room air
- Findings: No pericardial rub or crackles; epigastric tenderness

6. Auxiliary Examination
- (1) Imaging test: Chest radiograph showed mild peribronchial cuffing. Transthoracic echocardiography revealed preserved LV function, a 9-mm pericardial effusion, and slight IVC dilation. Coronary CT excluded obstructive disease. Cardiac MRI demonstrated myocardial edema with multifocal subepicardial and subendocardial late gadolinium enhancement and pericardial inflammation.
- (2) Laboratory tests: WBC $13.4 \times 10^3/\mu$L, CRP 16.9 mg/dL, ESR 95 mm/h, high-sensitivity troponin T 656.2 ng/L; differential count showed no eosinophilia. Blood cultures, serology, and PCR for pathogens negative; vasculitis-associated autoantibodies absent.
- (3) Electrocardiogram: Sinus tachycardia with first-degree AV block (PQ 210 ms).
- (4) Right heart catheterization: Cardiac Index 1.65 L/min/m^2, mean PCWP 34 mmHg, LVEDP 29 mmHg.
- (5) Endomyocardial biopsy: Eight specimens obtained from the left ventricle.

Fig. A4: Illustrative example of a structured case produced by the schema in Fig. A2. The record contains demographics, chief complaint with timeline, focused HPI, exam with vitals, and auxiliary studies. No diagnosis is included.

B.2 Standardized patient simulator (SP-sim)

The standardized patient simulator described in Section 4.2 is instantiated with the following system prompt:B7

B.3 Doctor-agent configuration and action templates

We run **15** instruction-tuned LLMs as the doctor agent . To ensure strict comparability, all models share deterministic decoding and identical prompt scaffolds (Figs. B8 B9).

Endpoints and scoring.

We report two endpoints: *Final Diagnosis Accuracy* and *Evidence Quality*. Accuracy is scored on a 3-level scale {0,1,2}: **0** incorrect (different disease family), **1** partially correct (right family, wrong subtype/term), **2** fully correct (including synonyms/-subtypes). Evidence Quality is also scored on {0,1,2}: **2** all three evidence items are supported and consistent; **1** one–two items supported or partially reasonable; **0** unsupported/contradictory. Figures B10 and B11 show the prompts used to elicit *single-integer* decisions.

Prompt: System-Level Categorization

You are a medical assistant with extensive clinical knowledge. Your task is to classify each disease name into one of the following six major clinical system categories. If a disease does not clearly belong to any of these categories, label it as **"Other"**.

Use the following classification rules:

1. Cardiovascular System
- Includes: Acute coronary syndrome, heart failure, arrhythmias (e.g., atrial fibrillation), hypertensive emergencies, aortic dissection, pericarditis, etc.

2. Respiratory System
- Includes: Asthma, COPD, pneumonia, pulmonary embolism, spontaneous pneumothorax, hemoptysis-related diseases, etc.

3. Gastro-Hepatobiliary System
- Includes: Upper or lower GI bleeding, appendicitis, cholecystitis, pancreatitis, liver cirrhosis and complications, inflammatory bowel disease, abdominal pain, diarrhea, etc.

4. Neurological System
- Includes: Ischemic stroke, TIA, seizures/epilepsy, subarachnoid hemorrhage, headaches, dizziness/vertigo, migraine, CNS infections, etc.

5. Infectious Diseases
- Includes: Bacterial meningitis, urinary tract infections, community or hospital-acquired pneumonia, skin and soft tissue infections, early sepsis, tropical diseases (e.g., dengue, malaria), etc.

6. Metabolic, Renal & Genitourinary System
- Includes: Diabetes mellitus (DKA, HHS), hypoglycemia, thyroid disorders (hyper/hypothyroidism), electrolyte disorders (e.g., hyponatremia, hyperkalemia), acute kidney injury, kidney stones, urinary tract diseases, etc.

If a disease does not fit into any of the above six categories, classify it as: **Other**

Please return your answer in JSON format, like this:
`{"primary_diagnosis": "Pneumonia", "category": "Respiratory System"}`

Fig. A5: Rule-based categorization prompt that maps diseases to one of six clinical systems (or `Other`) and returns a JSON tuple {`primary_diagnosis`, `category`}. This supports the balanced per-system distribution in Table 1.

Opening Utterances (15 Templates)

- "Hello, what brings you in today?"
- "Can you tell me the reason for your visit?"
- "Let's begin — have you noticed anything unusual lately?"
- "What symptoms have you been experiencing?"
- "What seems to be the problem today?"
- "What kind of health concerns are you having recently?"
- "Please describe any discomfort or issues you've been having."
- "Is there anything in particular that's been bothering you?"
- "Tell me what's going on — when did it start?"
- "Are you here because of any new or ongoing symptoms?"
- "What made you decide to come see a doctor today?"
- "We'll start with your concerns — what would you like to discuss?"
- "Have you experienced any recent changes in your health?"
- "Can you walk me through your current symptoms?"
- "What's the main reason for your appointment today?"

Fig. B6: Fifteen interchangeable opening utterances used by the clinician to start the encounter. On the first turn, the SP-sim responds with `1.Patient Information` and `2.Chief Complaint` verbatim from the structured case.

SP-sim Policy Prompt (Patient Role)

You are simulating a real patient during a clinical consultation. You will **ONLY** answer based on this record, and **ONLY** respond to the doctor's most recent statement.

—— **Case Record** ——
`formatted_record`

Opening Behavior:
When the consultation begins, the doctor will ask a general opening question (e.g., "What brings you in today?" or "Where do you feel unwell?").
In your first response, you must present the following two sections **exactly as written in the case record**:
- **1.Patient Information**
- **2.Chief Complaint**

After the opening, respond ONLY when the doctor requests a specific module or test.

Allowed Module Responses:
- [History of Present Illness]
- [Past Medical History]
- [Physical Examination]

Allowed Examination Requests (you will recognize them like this):
- [Laboratory Tests: ...]
- [Imaging Studies: ...]
- [Functional Tests: ...]
- [Specialized Panels: ...]

How to respond:

- Only respond if the requested test **is present in the case record**.
- Return the exact predefined result from the record.
- If a test is **not included in the case**, reply with: `"This test was not performed yet."`
- **Do NOT** invent any additional information.
- **Do NOT** include or hint at the final diagnosis at any time.

Response Format:
`[Module Name]: Full original text from that section`
`[test_name]: Full original test_result from that section`

Fig. B7: Policy prompt for the standardised patient driven by Qwen2.5-32B-Instruct. The simulator is reactive, releases only requested modules/tests, echoes verbatim from the record, and never reveals the diagnosis.

Task 1 — Full-Context Clinician Prompt

You are an experienced senior clinician at a top-tier tertiary hospital. Your task is to carefully analyze structured patient records and provide the most accurate final diagnosis based on the clinical information.

You must follow strict clinical reasoning and adhere to the output format and diagnostic criteria described below.

Output Format
- When reaching a final diagnosis, **YOU MUST** start the response with:

```
[Final Diagnosis] [Diagnosis Name]. Confirmed by:
1. ...
2. ...
3. ...
```

- You **MUST** include the exact tag `[Final Diagnosis]` with brackets — do not rephrase, omit, or replace it.

Fig. B8: Clinician prompt for **Task 1** (full context). The agent reads the complete structured record and must output `[Final Diagnosis]` followed by exactly three evidential items, using the mandatory tag verbatim.

Task 2 — Active Evidence-Seeking Clinician Prompt

You are an experienced senior clinician at a top-tier tertiary hospital. Your goal is to gather only the necessary clinical information to reach a diagnosis.

You may choose **ONE** action per turn from:

- `[History of Present Illness]`
- `[Past Medical History]`
- `[Physical Examination]`
- Request `[Laboratory Tests: test_name]`
- Request `[Imaging Studies: test_name]`
- Request `[Functional Tests: test_name]`
- Request `[Specialized Panels: test_name]`
- `[Final Diagnosis]`

Your output format should be:
`(one of the actions): (your reasoning process for choosing the next action)`

Output Format for Final Diagnosis:

- When reaching a final diagnosis, **YOU MUST** start the response with:
  `[Final Diagnosis] [Diagnosis Name]. Confirmed by:`
  `1. ...`
  `2. ...`
  `3. ...`
- You **MUST** include the exact tag `[Final Diagnosis]` with brackets — do not rephrase, omit, or replace it.

Note:

- To improve diagnostic efficiency, please perform tests only when necessary for diagnosis.
- You may only request **one specific test per turn**.
- **Do NOT** repeat tests or other modules.
- When confident, issue a `[Final Diagnosis]`.
- You must complete the diagnosis within a maximum of **10 turns**.
- If a test or module returns `"This test was not performed yet."`, you **MUST NOT** request it again.
- If your previous request failed (e.g., returned `"This test was not performed yet."`), consider alternatives or proceed to diagnosis if sufficient.
- For each turn, check previous requests before proposing the next action.
- **Do not re-request** the same test, even if it previously failed.

Fig. B9: Clinician prompt for **Task 2** (active evidence-seeking). Each turn, the agent selects exactly one action (module or a specific test), avoids repetitions and unavailable tests, and stops with `[Final Diagnosis]` within 10 turns. For non-terminal turns it outputs `(action): (brief rationale)`; final outputs must include the mandatory tag and three supporting items.

Accuracy Evaluation Prompt (0/1/2)

You are a senior medical expert tasked with evaluating whether the AI model's diagnosis is accurate. Score according to the criteria below:
[Scoring Criteria]

- **0**: Completely incorrect (a different disease category)
- **1**: Partially correct (correct disease category, but wrong subtype)
- **2**: Completely correct (including synonyms or correct subtype)

[Output Requirements]
Return only a single integer score (0, 1, or 2). **Do NOT** include any other text.

Fig. B10: Prompt used by the external evaluator to score **Final Diagnosis Accuracy** on a 0/1/2 scale. The evaluator returns a *single integer* only. This prompt is used identically for **Task 1** and **Task 2**.

Evidence Quality Evaluation Prompt (0/1/2)

You are a medical evidence expert tasked with evaluating how well the diagnostic evidence provided by the model matches the clinical case information. Please assign a score based on the following criteria:
[Scoring Criteria]

- **2**: All 3 evidence points are clearly supported in the case and fully consistent with the diagnosis
- **1**: 1–2 evidence points are supported, or there is partial reasonable inference
- **0**: Evidence points contradict the case or are not supported at all

[Output Requirement]
Only return a single integer score (0, 1, or 2). **DO NOT** include any other text.
[Important]

- Only evaluate the model's evidence points (usually 1–3)
- Evidence must be specific, verifiable, and clearly stated

Fig. B11: Prompt used by the external evaluator to score **Evidence Quality** on a 0/1/2 scale. The evaluator checks that the model's three evidence items are *verbatim supported* by the case and consistent with the diagnosis, and returns a *single integer* only. This prompt is used identically for **Task 1** and **Task 2**.