Strong Reasoning Isn't Enough: Evaluating Evidence Elicitation in Interactive Diagnosis

Anonymous ACL submission

Abstract

Interactive medical consultation requires an agent to proactively elicit missing clinical evidence under uncertainty. Yet existing evaluations largely remain static or outcome-centric, neglecting the evidence-gathering process. In this work, we propose an interactive evaluation framework that explicitly models the consultation process using a simulated patient and a measurement module grounded in atomic evidences. Based on this representation, we introduce Information Coverage Rate (ICR) to quantify how completely an agent uncovers necessary evidence during interaction. To support systematic study, we build EviMed, an evidence-based benchmark spanning diverse conditions from common complaints to rare diseases, and evaluate 10 models with varying reasoning abilities. We find that strong diagnostic reasoning does not guarantee effective information collection, and this insufficiency acts as a primary bottleneck limiting performance in interactive settings. To address this, we propose REFINE, a strategy that leverages diagnostic verification to guide the agent in proactively resolving uncertainties. Extensive experiments demonstrate that REFINE consistently outperforms baseline methods across diverse models and datasets, achieving superior information coverage and diagnostic accuracy.

1 Introduction

Large Language Models (LLMs) have achieved remarkable progress in recent years, evolving from passive language processors to autonomous agents (Ahn et al., 2022; Liu et al., 2023). Beyond text generation, these agents demonstrate increasing capabilities in interacting with external environments (Zhou et al., 2023; Yao et al., 2022a), using tools (Schick et al., 2023), and executing complex workflows (Jimenez et al., 2023). Such advancements suggest that LLM-based agents are

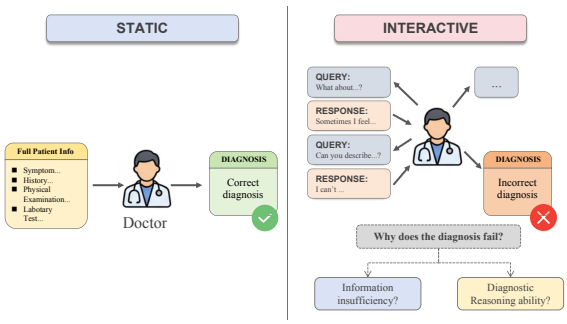

Figure 1: Static evaluation provides full patient information upfront. Interactive diagnosis requires iterative evidence elicitation and may fail due to insufficient information gathering or flawed reasoning.

becoming proficient at following user instructions to accomplish multi-step tasks.

However, many real-world decision-making scenarios extend beyond simple instruction following. In these settings, the agent is not provided with all necessary context upfront. Instead, the agent must actively identify missing information and acquire it through interaction before making a decision. Consequently, the quality of the final outcome hinges heavily on the agent's ability to gather information effectively under uncertainty.

Medical consultation represents a quintessential instance of such information-seeking scenarios. In clinical practice, diagnosis is an interactive evidence-gathering process rather than an one-shot prediction task (Meyer et al., 2021). Key evidence, including symptoms, medical history, and examinations, must be actively elicited through patient inquiry or clinical testing. Therefore, an effective medical agent must proactively ask relevant questions and decide when sufficient evidence has been collected to support a reliable diagnosis.

Despite this interactive nature (Fig. 1), most existing evaluations (Nori et al., 2023; Chen et al., 2023; Liévin et al., 2024; Wu et al., 2024; Singhal et al., 2025) focus on static settings where all

patient information is provided to the LLM in advance. While recent studies have begun to explore interactive diagnosis, they still primarily assess the final diagnostic accuracy. This outcome-oriented evaluation overlooks the evidence-elicitation process, leaving it unclear whether the agent can efficiently and systematically gather the information required for a correct diagnosis.

In this work, we argue that information collection ability should be treated as a first-class evaluation target for medical agents. To this end, we propose an interactive evaluation framework that explicitly models the consultation process using a simulated patient and a measurement module by leveraging the generative capabilities of language models (Park et al., 2023; Du et al.). We specifically represent all clinical information within these modules as atomic evidences, which are defined as minimal and self-contained units of facts. This granular representation enables us to introduce a new metric, the Information Coverage Rate (ICR). Unlike traditional success rates, ICR explicitly measures the proportion of necessary evidence successfully revealed by the agent, providing a fine-grained assessment of its active inquiry capabilities.

To support this evaluation framework, we construct EviMed, a new benchmark for interactive medical consultation. We transform existing medical datasets into the evidence-based format through an automated construction process. The resulting benchmark covers a diverse range of scenarios, spanning from common medical inquiries to challenging rare disease diagnoses that require extensive evidence accumulation.

We evaluate 10 LLMs of varying diagnostic reasoning ability on EviMed, revealing a performance disparity between static and interactive settings. On average, diagnostic success rates drop by approximately 20% when agents are required to actively collect information, with even sharper declines observed in rare disease scenarios. Moreover, our results show that strong diagnostic reasoning alone does not guarantee effective information acquisition. Insufficient information collection during interaction appears to be the main bottleneck underlying performance degradation.

To address these challenges, we propose REFINE (Reasoning-Enhanced Feedback for INformation Elicitation). It employs a Diagnosis Verifier to examine whether a generated Diagnosis hypothesis is fully grounded in the collected evidences, guiding the agent to resolve uncertainties. Extensive experiments demonstrate that this strategy effectively mitigates the performance degradation in interactive settings, yielding substantial improvements in both information coverage and diagnostic success rates across diverse models.

Overall, our contributions are threefold:

(1) We propose an interactive evaluation framework for evidence collection and introduce the Information Coverage Rate (ICR) metric, which formalizes information collection as a measurable objective grounded in atomic evidence units.

(2) We construct EviMed, a comprehensive diagnostic benchmark, and conduct a systematic evaluation of 10 models with varying diagnostic capabilities, revealing that strong diagnostic reasoning does not guarantee effective information collection, which emerges as a key performance bottleneck.

(3) We introduce REFINE, a feedback-driven strategy that leverages diagnostic verification to guide the interaction. It effectively mitigates the performance degradation in interactive settings, yielding substantial improvements in both information coverage and diagnostic success rates across diverse models.

2 Interactive Evidence Collection Evaluation Framework

In real clinical settings, relevant information is often incomplete and distributed across multiple sources. A clinician must determine what information is missing, how to obtain it, and when the collected evidence is sufficient to support a diagnosis. To model this interactive evidence collection process, we construct a medical consultation environment concentrating on evidence collection shown in Figure 2.

The environment includes three roles, namely a simulated patient, a simulated measurement, and a doctor agent under evaluation. The consultation proceeds in multiple turns. At each turn, the doctor agent chooses an action, including asking the patient a question, requesting a clinical examination, or issuing a diagnosis. In response, the simulated patient and the measurement component return information grounded in the case record.

2.1 Roles

We define the three roles and their modeling assumptions below.

Simulated Patient. The simulated patient maintains the patient's personal information in the form

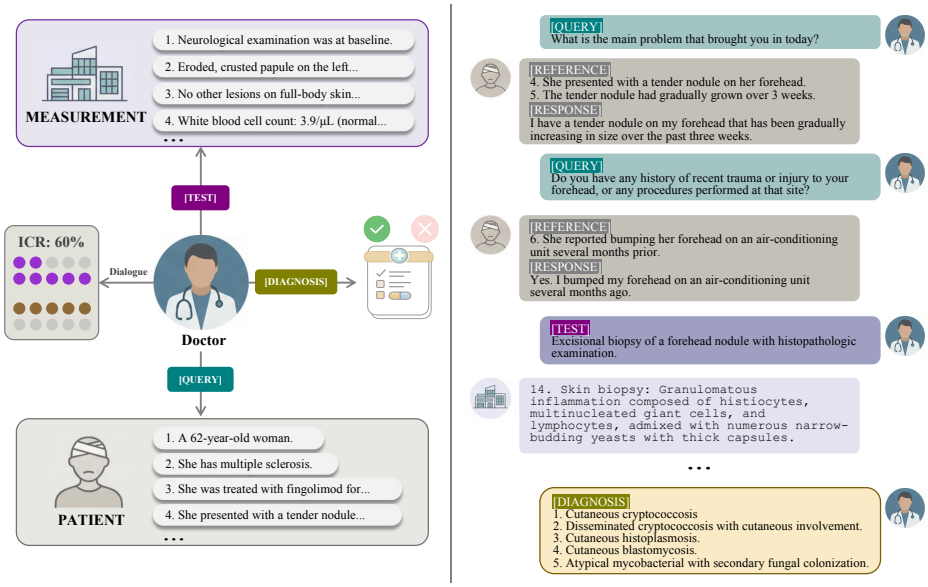

Figure 2: Interactive Evaluation Framework. **(Left)** The consultation loop. The doctor agent iteratively interacts with a simulated patient that provides subjective/history-related information and a simulated measurement module that returns objective examination or laboratory findings. The interaction reveals atomic evidences used for evaluation, where we track *Information Coverage Rate (ICR)* to measure how many relevant evidences have been collected during the dialogue, and assess diagnostic *Success Rate* based on the final diagnosis. **(Right)** An example consultation trajectory showing queries, evidence-grounded responses, test requests/results, and the final diagnosis.

of atomic evidences, covering symptoms, history, and related clinical details. It follows an evidence selection mechanism similar to the Fact-Select approach in Li et al. (2024), the patient selects the most relevant evidences for a given query and generates a natural language response grounded in them. Each response is supported by at most two evidences. If the query is unrelated to any evidence, the patient explicitly indicates uncertainty.

Simulated Measurement. The simulated measurement component maintains clinical examination results and laboratory test findings. For each request, it returns one or more relevant evidences as objective observations. These results are provided directly without natural language generation. This component evaluates whether the agent can select appropriate examinations and utilize objective clinical evidence.

Doctor Agent. The doctor agent is the model under evaluation. During the consultation, it decides its next action among asking a question, requesting a measurement, or issuing a diagnosis. This requires the agent to determine whom to interact with, how to formulate queries, and when to terminate the interaction. Therefore, the consultation process is modeled as a sequential decision-making problem under incomplete information.

2.2 Information Coverage Rate

To evaluate the agent's ability to collect relevant information, we propose Information Coverage Rate (ICR). ICR measures the proportion of evidences that are successfully collected by the agent through interaction. It reflects how thoroughly the agent explores the evidence space required for diagnosis.

Formally, let E denote the set of all relevant evidences for a given case. Let $\hat{E}$ denote the set of evidences collected by the agent during the consultation. ICR is defined as

$$\text{ICR} = \frac{|\hat{E} \cap E|}{|E|}.$$

As both patient responses and measurement results are grounded in atomic evidences, ICR is directly computable from the evidence revealed during interaction. Together with diagnostic success rate, ICR provides a complementary view by separating evidence collection completeness from final diagnostic correctness.

3 EviMed Benchmark Construction

Most existing medical datasets present complete case narratives without explicit atomic evidence structure. This makes it difficult to support selective evidence disclosure and compute information

collection coverage. To address this gap, we construct EviMed, an evidence-based benchmark for interactive medical consultation evaluation.

3.1 Source Datasets

EviMed integrates five complementary data sources covering general medicine, specialty diagnosis, complex multi-specialty reasoning, rare diseases, and real-world clinical records. We sample two hundred cases from each source, yielding the EviMed-1K benchmark, which spans a wide range of diagnostic settings. The five data sources are described as follows:

AgentClinic-MedQA (Schmidgall et al., 2024) is adapted from USMLE-style medical examination cases and rewritten into consultation-oriented scenarios. It covers a wide range of common diseases and clinical conditions. We use this source to evaluate general diagnostic reasoning.

Derm (Johri et al., 2024) focuses on dermatological diagnosis and emphasizes fine-grained descriptions. It contains both publicly available cases and clinician-authored cases with similar structures. We include the full set to evaluate detailed symptom inquiry in a specialized domain.

DiagnosisArena (Zhu et al., 2025b) is constructed from real-world case reports published in top-tier medical journals. The cases require complex diagnostic reasoning across multiple clinical specialties. We use it to assess information collection in challenging diagnostic scenarios.

RareArena (Zhao et al., 2025) is built from publicly available case reports in PubMed Central (PMC) and covers a wide range of rare diseases, which involve limited prior knowledge and ambiguous symptom presentations. We sample cases according to disease frequency to encourage diversity across different levels of rarity.

ClinicalBench (Yan et al., 2024) is derived from real electronic medical records containing both structured and unstructured information. It covers cases from multiple clinical departments and a broad set of disease categories. We sample cases to ensure coverage across disease types.

3.2 Automatic Construction

For each selected case, we transform the original case description into an evidence-based representation. We separate patient basic information from examination-related information, and then decompose each part into non-overlapping atomic evidences, where each evidence corresponds to a minimal and self-contained clinical fact. This conversion is performed automatically using GPT-5-mini.

Table 1: Statistics of EviMed-1K.

Dataset	Data Size	Avg Pat. Evi.	Avg Exam Evi.
AgentClinic-MedQA	200	14.87	12.73
Derm	200	7.31	2.87
DiagnosisArena	200	8.04	12.82
RareArena	200	17.11	17.72
ClinicalBench	200	21.76	21.14

After construction, each case is associated with a set of patient evidences and a set of examination evidences. These evidences serve as the information sources accessed during interaction by the simulated patient and simulated measurement components. Table 1 summarizes the benchmark statistics, including the average number of patient evidences and examination evidences per case in each source dataset. The number of atomic evidences varies across sources, reflecting differences in case complexity and information density.

In Section 5.6, we verify that the automatic construction process preserves diagnostic information.

4 REFINE: Feedback-Driven Evidence Collection

In interactive medical consultation, the agent must make diagnostic decisions under incomplete and evolving evidence. This setting introduces two tightly coupled challenges. First, the agent must determine which information to elicit next in order to efficiently reduce the diagnostic uncertainty. Second, it must decide when the accumulated evidence is sufficient to support a reliable diagnosis, rather than terminating the consultation prematurely.

To address these challenges, we propose Reasoning-Enhanced Feedback for INformation Elicitation (REFINE), a feedback-driven strategy for evidence collection. As illustrated in Figure 3, REFINE consists of an Information Collector, an Evidence Organizer, a Diagnosis Reasoner and a Diagnosis Verifier. The Information Collector interacts with the consultation environment across multiple turns. At each turn, it assesses whether the currently collected information is sufficient, decides whether to continue acquiring evidence, or terminates the interaction to make a diagnosis. When the collector stops, the Evidence Organizer consolidates the collected findings into a structured evidence summary.

Given the organized evidence summary, the Diagnosis Reasoner produces a diagnostic hypothe-

sis. The Diagnosis Verifier then checks whether the hypothesis is fully supported by the available evidence. If the verifier detects the evidence is insufficient, it provides explicit feedback identifying missing information and unresolved uncertainties, which is sent back to the Information Collector to resume the interaction phase and guide subsequent evidence acquisition steps.

This loop continues until the verifier finds that the hypothesis is sufficiently supported by collected evidence or the interaction reaches a maximum turn limit. This design separates an internal hypothesis used for probing the evidence state from the final diagnostic output. As a result, the feedback specifies what to collect next, and the absence of critical evidence gaps provides a natural criterion for when to stop.

5 Experiments

5.1 Setup

We evaluate different models and methods under the interactive evidence collection framework. We consider the following methods for comparison:

Upper Bound uses a static full-information setting where all patient information and examination results are provided upfront. We prompt the model to generate intermediate reasoning before producing the final diagnosis, establishing a performance upper bound where active information acquisition is not required.

Baseline represents a standard interactive setting where a single doctor agent interacts directly with the environment. At each turn, the model determines whether to ask a question, request a specific examination, or terminate the session to output a final diagnosis.

ReAct augments the baseline by enforcing an explicit Thought-Act cycle during the consultation. Before taking any external action, the agent must generate a reasoning trace to analyze the current clinical state and justify its next move, thereby improving decision-making.

Summarized-Conversation (SC) decouples information gathering from diagnosis. It first conducts a full multi-turn consultation to collect evidence, then summarizes the entire interaction history into a structured format. The final diagnosis is produced solely based on this consolidated summary rather than turn-level context.

REFINE is our proposed feedback-driven strategy designed to optimize evidence collection. It utilizes a diagnostic verification mechanism to assess the sufficiency of collected information, providing the agent with explicit feedback to guide subsequent inquiry steps and proactively resolve remaining uncertainties.

We evaluate a diverse set of language models spanning different scales and domain specializations. The evaluated models include GPT-5 (OpenAI, 2025b), GPT-5-mini, DeepSeek-v3.2 (Liu et al., 2025), GLM-4.6 (OpenAI, 2025a), Qwen2.5-72B (Hui et al., 2024), Qwen2.5-32B, Qwen2.5-7B, Qwen2.5-3B, Llama-3.1-8B-Instruct (Dubey et al., 2024), and Meditron3-8B (Sallinen et al., 2025).

We set the maximum interaction turns to 16 and report diagnostic success rate (SR) and Information Coverage Rate (ICR). Following the original benchmarks, DiagnosisArena, RareArena, and ClinicalBench are differential diagnosis tasks where the model outputs five diseases ranked by likelihood and SR is computed as top-1 accuracy. AgentClinic-MedQA and Derm require a single diagnosis and SR is standard accuracy.

5.2 Static vs. Interactive Evaluation

We compare static and interactive evaluation to assess whether strong full-information reasoning performance transfers to realistic consultations that require evidence acquisition. Specifically, we report the Success Rate under the static full-information upper bound setting and report both ICR and SR under the interactive baseline. Results are summarized in Table 2, and Figure 4 provides a compact cross-model view of the relationship between ICR and SR under interactive evaluation.

Across datasets and models, even the strongest model GPT-5, performance drops substantially when moving from static to interactive evaluation. On average, SR decreases by about 20 percentage points, and the gap is more pronounced on the challenging DiagnosisArena and RareArena datasets.

We further observe differences across different models. GPT-5-mini achieves stronger upper bound than DeepSeek-v3.2 and GLM-4.6, yet its interactive ICR is comparatively low, and correspondingly its SR lags behind DeepSeek-v3.2 and GLM-4.6 in multiple datasets. A similar trend is observed between Meditron3-8B and Llama-3.1-8B. Despite being fine-tuned on clinical data, Meditron exhibits a larger performance degradation in the interactive setting. These cases indicate that static medical resoning ability does not automatically translate into effective inquiry strategies.

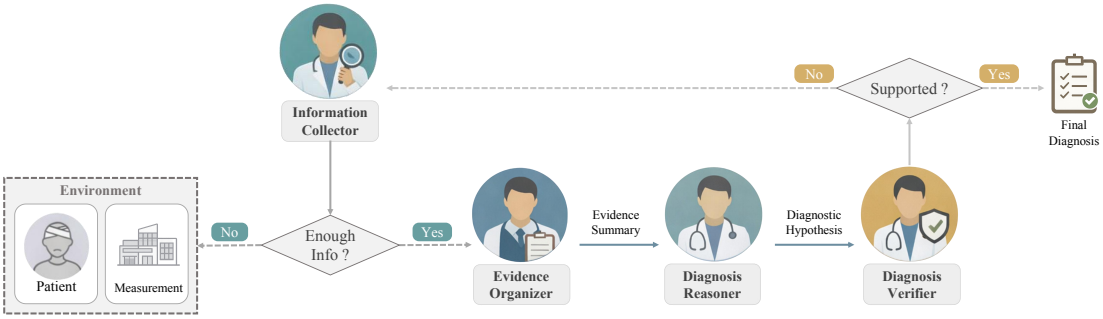

Figure 3: Overview of the REFINE Strategy.

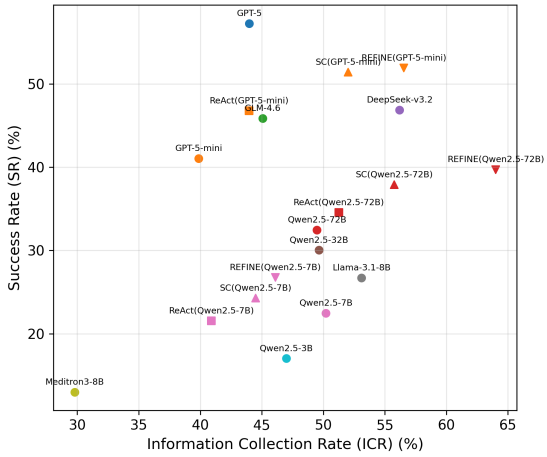

Figure 4: Relationship between average Information Coverage Rate and Success Rate under interactive evaluation, with values averaged over the five datasets.

The Qwen2.5 series presents a insight regarding model scaling. While these models exhibit strong ICR, increasing model size can sometimes lead to degraded performance in information collection. The benefits of scaling are mainly manifested in improved SR, indicating that larger models excel at reasoning over already collected evidence rather than at discovering new information.

5.3 Strategy Comparison

We compare interaction strategies to assess their effects on ICR and SR. We report results for GPT-5-mini, Qwen2.5-72B, and Qwen2.5-7B in Table 3.

From the results, ReAct improves both ICR and SR for stronger models such as GPT-5-mini and Qwen2.5-72B. However, for the weaker model Qwen2.5-7B, ReAct decreases both ICR and SR. This may reflect increased difficulty under longer multi-turn trajectories (Laban et al., 2025), which is also suggested by Section 5.5.

SC more consistently improves ICR and SR across the evaluated models. One possible explanation is that separating information collection from the final diagnosis helps the agent stay focused on evidence acquisition. In addition, producing the diagnosis from a structured summary may mitigate reasoning degradation in long conversations, which could contribute to better diagnosis accuracy.

REFINE achieves the highest ICR across most datasets and models, and the coverage gains are typically accompanied by higher SR. The improvements are especially pronounced on Diagnosis-Arena and RareArena, where evidence is sparse or ambiguous. These results support reasoning-based feedback as an effective mechanism for aligning information collection with diagnostic needs.

5.4 Information Coverage in Successful vs. Failed Diagnosis

We examine the association between diagnostic success and information coverage using outcome-conditioned ICR distributions.

We study DiagnosisArena and RareArena, two challenging rare-disease benchmarks that typically require extensive information collection. For each dataset, we aggregate successful and failed cases from the three models used in Section 5.3 and present the distributions of their Information Coverage Rate. We compare three interaction strategies, Baseline, ReAct, and REFINE. Figure 5 summarizes the resulting ICR distributions.

Across both datasets and all strategies, successful cases consistently exhibit higher ICR than failed ones. These observations indicate that insufficient information coverage is commonly associated with diagnostic failure, supporting ICR as a process-level indicator of diagnostic quality in interactive consultation.

Table 2: Static Upper Bound vs. interactive Baseline evaluation. UB denotes SR under static evaluation. ICR and SR are metrics under interactive evaluation framework. For SR, the subscript indicates the percentage decrease relative to UB. **Bold** values denote the best performance under each metric.

Model	ClinicalBench			Derm			DiagnosisArena			MedQA			RareArena		
	UB (%)↑	ICR (%)↑	SR (%)↑	UB (%)↑	ICR (%)↑	SR (%)↑	UB (%)↑	ICR (%)↑	SR (%)↑	UB (%)↑	ICR (%)↑	SR (%)↑	UB (%)↑	ICR (%)↑	SR (%)↑
GPT-5	64.0	35.2	47.0$_{(-27\%)}$	**93.5**	54.9	**75.0**$_{(-20\%)}$	**75.0**	55.4	**47.0**$_{(-37\%)}$	**97.5**	31.3	**78.0**$_{(-20\%)}$	**75.0**	42.7	**37.0**$_{(-51\%)}$
GPT-5-mini	**67.5**	31.3	46.5$_{(-31\%)}$	92.0	57.8	58.5$_{(-36\%)}$	68.5	53.1	16.0$_{(-77\%)}$	92.0	38.8	40.5$_{(-56\%)}$	68.0	35.7	15.5$_{(-77\%)}$
DeepSeek-v3.2	63.0	**46.7**	**51.0**$_{(-19\%)}$	80.5	**77.0**	63.5$_{(-21\%)}$	56.5	63.4	18.0$_{(-68\%)}$	92.5	**47.8**	71.0$_{(-23\%)}$	61.5	**49.1**	24.5$_{(-60\%)}$
GLM-4.6	60.5	32.6	46.5$_{(-23\%)}$	78.5	59.1	61.0$_{(-22\%)}$	52.0	54.5	12.5$_{(-76\%)}$	86.0	35.2	51.5$_{(-40\%)}$	59.0	36.4	22.5$_{(-62\%)}$
Qwen2.5-72B	64.5	41.2	49.0$_{(-24\%)}$	75.5	67.6	61.0$_{(-19\%)}$	32.0	57.8	10.5$_{(-67\%)}$	75.5	43.4	58.0$_{(-23\%)}$	32.0	46.1	14.0$_{(-56\%)}$
Qwen2.5-32B	60.0	42.1	45.0$_{(-25\%)}$	76.5	69.0	63.0$_{(-18\%)}$	20.0	43.4	8.0$_{(-60\%)}$	76.5	45.3	60.5$_{(-21\%)}$	29.0	40.0	6.0$_{(-79\%)}$
Qwen2.5-7B	52.0	34.4	42.0$_{(-19\%)}$	60.5	62.3	45.0$_{(-26\%)}$	13.0	41.3	8.0$_{(-38\%)}$	60.5	43.1	48.5$_{(-20\%)}$	19.5	37.1	5.0$_{(-74\%)}$
Llama-3.1-8B	37.0	41.3	34.5$_{(-7\%)}$	67.0	66.1	50.5$_{(-25\%)}$	23.0	41.3	6.0$_{(-57\%)}$	67.0	40.7	52.0$_{(-22\%)}$	23.0	35.7	9.0$_{(-61\%)}$
Meditron3-8B	42.5	33.1	32.0$_{(-25\%)}$	68.0	38.6	53.0$_{(-22\%)}$	13.0	28.8	2.5$_{(-76\%)}$	68.0	35.2	48.5$_{(-29\%)}$	25.5	28.2	2.5$_{(-90\%)}$
Qwen2.5-3B	38.0	38.2	27.5$_{(-28\%)}$	49.5	52.7	37.0$_{(-25\%)}$	9.5	30.0	5.5$_{(-42\%)}$	49.5	39.6	36.5$_{(-26\%)}$	9.5	29.1	2.0$_{(-79\%)}$

Table 3: Comparison of representative models under different interactive strategies across datasets.

Dataset	Model	Baseline		ReAct		SC		REFINE	
		ICR (%)↑	SR (%)↑	ICR (%)↑	SR (%)↑	ICR (%)↑	SR (%)↑	ICR (%)↑	SR (%)↑
ClinicalBench	GPT-5-mini	31.3	46.5	33.5	**49.5**	41.5	49.0	**43.5**	**49.5**
	Qwen2.5-72B	40.8	43.5	38.7	42.0	43.2	49.5	**51.1**	**53.0**
	Qwen2.5-7B	**41.0**	35.5	29.8	33.0	32.5	35.0	35.3	**38.5**
Derm	GPT-5-mini	51.7	57.5	59.5	62.0	70.9	**66.5**	**76.7**	66.0
	Qwen2.5-72B	67.5	43.0	71.2	45.0	77.9	**45.5**	**80.8**	44.0
	Qwen2.5-7B	**71.9**	20.0	63.5	23.0	62.8	25.0	67.1	**28.5**
DiagnosisArena	GPT-5-mini	47.2	23.0	53.9	29.5	60.8	39.5	**64.6**	**42.0**
	Qwen2.5-72B	55.5	10.0	58.9	12.5	63.4	15.5	**73.8**	**18.5**
	Qwen2.5-7B	53.1	8.0	45.1	3.5	**53.6**	10.5	50.8	**13.0**
MedQA	GPT-5-mini	33.0	61.0	33.2	67.0	40.2	**69.5**	**45.5**	68.0
	Qwen2.5-72B	41.7	54.0	39.5	59.0	43.6	62.0	**53.9**	**64.5**
	Qwen2.5-7B	**43.5**	43.0	35.8	45.0	36.0	44.5	38.6	**45.5**
RareArena	GPT-5-mini	35.7	15.5	39.3	24.0	46.2	30.5	**51.8**	**32.0**
	Qwen2.5-72B	41.6	10.5	47.4	13.0	50.1	15.5	**59.9**	**17.0**
	Qwen2.5-7B	**41.2**	5.0	29.8	2.5	37.0	5.5	38.2	**7.5**

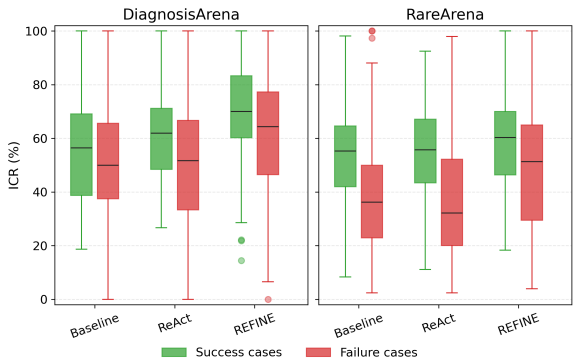

Figure 5: ICR distributions for successful and failed diagnoses on DiagnosisArena and RareArena.

5.5 Effect of Interaction Budget

We analyze how interaction budget affects ICR and SR. We vary the maximum number of interaction turns while keeping other settings fixed. We conduct this study on DiagnosisArena with Qwen2.5-72B, comparing Baseline and REFINE. Figure 6 summarizes the results.

At low budgets, both strategies improve quickly in both ICR and SR. Both metrics rise sharply in the first few turns. As the budget grows, the incremental gains taper off. ICR typically reaches its plateau slightly later than SR. Comparing the two strategies, REFINE sustains higher ICR and SR throughout the range of budgets and saturates later than Baseline.

These results indicate that early turns are most effective for acquiring the evidence needed for diagnosis. After most relevant evidence is collected, additional interaction yields limited benefit and may increase reasoning burden.

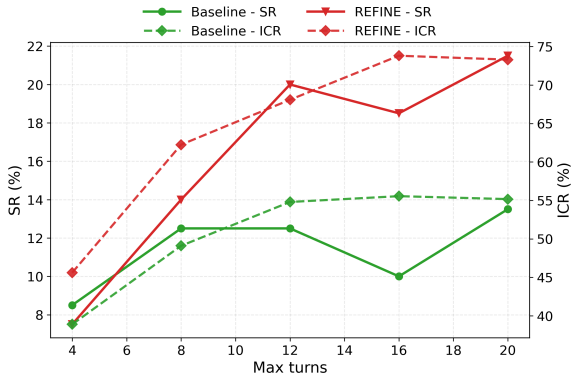

Figure 6: SR and ICR as functions of the maximum number of interaction turns.

Table 4: Diagnostic success rate under static evaluation using original case descriptions and concatenated evidences.

Dataset	Ori.	Concat.
DiagnosisArena	63.0	62.5
RareArena	68.0	69.5
MedQA	92.0	94.5
ClinicalBench	67.5	67.5
Derm	84.0	82.5

5.6 Evidence Construction Sanity Check

We conduct a sanity check to examine whether essential diagnostic information is preserved after evidence construction. To ensure essential information is preserved, we compare the diagnostic Success Rate (SR) of the original case descriptions against the concatenated constructed evidences in a static evaluation. Table 4 shows that The performance differences between original cases and concatenated evidences are small, indicating the information loss introduced by this process is negligible.

6 Related Work

Task-Oriented Agents Early research primarily focused on tool utilization in static scenarios. In these settings, agents are required to decompose specific user instructions and invoke appropriate APIs or search engines to execute actions (Yao et al., 2022b; Qin et al., 2023; Patil et al., 2024).

Later work (Deng et al., 2023; Wang et al., 2023; Yao et al., 2022a; Zhou et al., 2023) moved to dynamic environments that require multi-turn interaction, such as web navigation and database manipulation. Agents must track dialogue state and plan over long horizons to complete tasks reliably.

Another line studies robustness under user-specified policies and evolving constraints in realistic workflows, including retail customer service and flight booking (Yao et al., 2024; Barres et al., 2025). These evaluations prioritize constraint compliance and adaptability during interaction.

Most of the above benchmarks assume an instruction-following paradigm where the user states intent and supplies sufficient information. In many real-world settings, users cannot provide complete information upfront, so agents must form hypotheses and elicit missing evidence through inquiry (Zhu et al., 2025a; Mukherjee et al., 2024). Focusing on the medical consultation setting, our work introduces an interactive evaluation framework that requires agents to proactively gather information throughout the consultation process.

Medical Agent Evaluation Medical LLM evaluation has largely relied on static question answering datasets with complete case descriptions, testing knowledge retention and diagnostic reasoning (Jin et al., 2021; Chen et al., 2025; Wang et al., 2024; Fansi Tchango et al., 2022). Multi-agent collaboration can improve reasoning, but it typically remains within full-information inputs and does not require selective evidence discovery (Kim et al., 2024; Tang et al., 2024; Wang et al., 2025).

To better reflect clinical practice, recent work simulates doctor patient encounters where agents interact with patients to gather symptoms and request examinations or tests (Schmidgall et al., 2024; Fan et al., 2025; Johri et al., 2024; Almansoori et al., 2025; Bao et al., 2025). These environments introduce interaction structure and information asymmetry compared with static benchmarks.

7 Conclusion

In this work, we revisit medical agent evaluation by shifting the focus from static prediction to interactive evidence collection. We establish a fine-grained evaluation framework grounded in atomic evidences and construct EviMed to systematically measure the agent's active inquiry capabilities. Our analysis reveals a critical bottleneck: even models with strong reasoning capabilities often fail to collect sufficient information, leading to a significant performance gap between static and interactive settings. Ultimately, this work provides a valuable resource for assessing autonomous clinical decision-making and offers a scalable path toward bridging the gap between static knowledge retention and interactive diagnostic reasoning.

8 Limitations

Our evaluation is conducted in a controlled interactive simulator. This setting may not match the distribution of real patient narratives, clinician behaviors, and institutional constraints. Simulation can reduce ambiguity and compress the space of plausible follow up trajectories, which can shift the optimal elicitation strategy. Prior work on simulated patients and multi agent clinical simulators similarly highlights remaining gaps between multi turn interaction and curated or non interactive settings (Holderried et al., 2024; Fan et al., 2025; Almansoori et al., 2025). Accordingly, our results should be read as comparative performance within this environment rather than a direct measure of clinical readiness.

We model clinical information as a set of atomic evidences to enable systematic scoring. This abstraction omits graded severity, temporal evolution, and dependencies across findings. Clinical facts also exhibit substantial semantic variability across documentation styles and contexts. Recent studies on feature and concept extraction from clinical notes suggest that fine grained clinical signal recovery remains challenging and can be sensitive to annotation and modeling choices (Abumelha et al., 2025). As a result, higher atomic coverage may not always correspond to clinically sufficient information gathering.

ICR is defined with respect to a case specific relevant evidence set E. In practice, multiple evidence subsets can justify the same diagnosis, and experts may disagree on what is necessary versus merely supportive. Inter rater reliability in related clinical evaluation settings varies across attributes and rubrics, indicating that a single reference set can encode subjective decisions (Holderried et al., 2024). Future work should report agreement statistics, test sensitivity to alternative definitions of E, and consider softer relevance modeling for borderline evidences.

References

Manal Abumelha, Abdullah Al-Malaise Al-Ghamdi, Ayman Fayoumi, and Mahmoud Ragab. 2025. Medical feature extraction from clinical examination notes: Development and evaluation of a two-phase large language model framework. *JMIR Medical Informatics*, 13:e78432.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, and 1 others. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. 2025. Medagentsim: Self-evolving multi-agent simulations for realistic clinical interactions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 362–372. Springer.

Zhijie Bao, Qingyun Liu, Xuan-Jing Huang, and Zhongyu Wei. 2025. Sfmss: Service flow aware medical scenario simulation for conversational data generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4586–4604.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. tau^2-bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.

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

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114.

Z Du, L Zheng, R Hu, Y Xu, X Li, Y Sun, W Chen, J Wu, H Cai, and H Ying. Llms can simulate standardized patients via agent coevolution. arxiv 2024. *arXiv preprint arXiv:2412.11716*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: A new

dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35:31306–31318.

Friederike Holderried, Christian Stegemann-Philipps, Anne Herrmann-Werner, Teresa Festl-Wietek, Martin Holderried, Carsten Eickhoff, Moritz Mahling, and 1 others. 2024. A language model–powered simulated patient with automated feedback for history taking: Prospective study. *JMIR Medical Education*, 10(1):e59213.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

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

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. Craft-md: A conversational evaluation framework for comprehensive assessment of clinical llms. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

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

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

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

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, and 1 others. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Ashley ND Meyer, Traber D Giardina, Lubna Khawaja, and Hardeep Singh. 2021. Patient and clinician experiences of uncertainty in the diagnostic process: current understanding and future directions. *Patient Education and Counseling*, 104(11):2606–2615.

Subhabrata Mukherjee, Paul Gamble, Markel Sanz Ausin, Neel Kant, Kriti Aggarwal, Neha Manjunath, Debajyoti Datta, Zhengliang Liu, Jiayuan Ding, Sophia Busacca, and 1 others. 2024. Polaris: A safety-focused llm constellation architecture for healthcare. *arXiv preprint arXiv:2403.13313*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI. 2025a. Glm-4.6. Available at https://z.ai/blog/glm-4.6.

OpenAI. 2025b. Gpt-5. Available at https://openai.com/index/introducing-gpt-5/.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2024. Gorilla: Large language model connected with massive apis. *Advances in Neural Information Processing Systems*, 37:126544–126565.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Alexandre Sallinen, Antoni-Joan Solergibert, Michael Zhang, Guillaume Boyé, Maud Dupont-Roc, Xavier Theimer-Lienhard, Etienne Boisson, Bastien Bernath, Hichem Hadhri, Antoine Tran, and 1 others. 2025. Llama-3-meditron: An open-weight suite of medical llms based on llama-3.1. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

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

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

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, and 1 others. 2024. Cmb: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.

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

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

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022a. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. tau-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022b. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Zhiyu Zhao and 1 others. 2025. Rarearena: A comprehensive rare disease diagnostic dataset with nearly 50,000 patients covering more than 4000 diseases. https://github.com/zhao-zy15/RareArena.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, and 1 others. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Jiayuan Zhu, Jiazhen Pan, Yuyuan Liu, Fenglin Liu, and Junde Wu. 2025a. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2846–2857.

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