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

Zhuohan Long[1], **Zhongyu Wei**[1,2]*
[1]School of Data Science, Fudan University
[2]Shanghai Innovation Institute
zhlong24@m.fudan.edu.cn, zywei@fudan.edu.cn

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 simulated reporter 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 baselines across diverse datasets and facilitates effective model collaboration, enabling smaller agents to achieve superior performance under strong reasoning supervision. Our code can be found at this URL.

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),

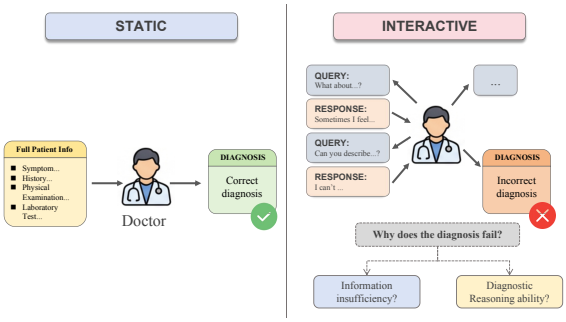

Figure 1: Static evaluation provides full patient information upfront. Interactive diagnosis requires iterative evidence elicitation and may fail due to insufficient information gathering or flawed reasoning.

using tools (Schick et al., 2023), and executing complex workflows (Jimenez et al., 2023). Such advancements suggest that LLM-based agents are 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 a 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 (Figure 1), most

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 27559–27575

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 simulated reporter 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. Furthermore, we find that REFINE enables effective collaboration between heterogeneous models, allowing smaller, inquisitive agents to achieve superior performance when supervised by strong reasoning 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 and supports heterogeneous model collaboration, enabling smaller models to excel in inquiry tasks under strong reasoning supervision.

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 is composed of three distinct roles, including a simulated patient, a simulated reporter, and the doctor agent being evaluated. 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,

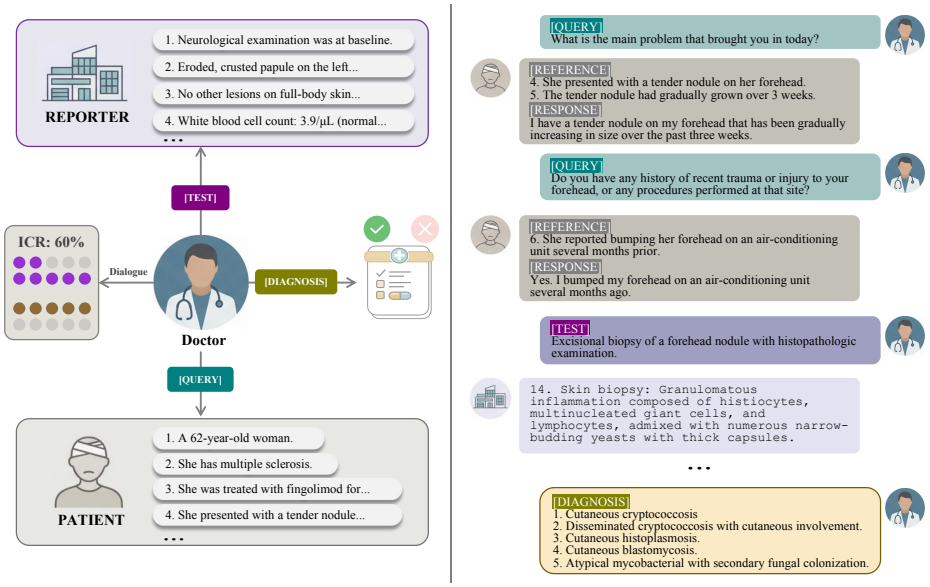

Figure 2: Interactive Evaluation Framework. **(Left)** The consultation loop. The doctor agent iteratively interacts with a simulated patient that provides subjective/history-related information and a simulated reporter that returns objective examination or laboratory findings. The interaction reveals 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.

the simulated patient and the simulated reporter 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 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 Reporter. The simulated reporter 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 test, 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 test 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.

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

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.

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 the simulated reporter. 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 5.8, 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 hypothesis. 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 with a maximum of 16 interaction turns. 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 (Yao et al., 2022b) 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) (Johri et al., 2024) 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, 2025), GPT-5-mini, DeepSeek-v3.2 (Liu et al., 2025), GLM-4.6 (Z.ai, 2025), 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).

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.

Across datasets and models, SR decreases by approximately 20% on average when moving from static to interactive evaluation. This degradation is more pronounced on the more challenging DiagnosisArena and RareArena datasets. Even for the strongest model, GPT-5, performance drops substantially in the interactive setting, indicating that strong static reasoning does not directly translate to effective interactive decision-making.

Interestingly, some models exhibit larger performance drops than others. For example, GPT-5-mini originally achieves a stronger static upper bound than DeepSeek-v3.2 and GLM-4.6, but its interactive SR lags behind them. A similar phenomenon is observed for Meditron3-8B. Although it is fine-tuned on clinical data based on Llama-3.1-8B, it shows a larger performance degradation than its base model.

We further observe that models with larger degradation, such as GPT-5-mini and Meditron3-8B, also exhibit relatively low ICR. This suggests that insuf-

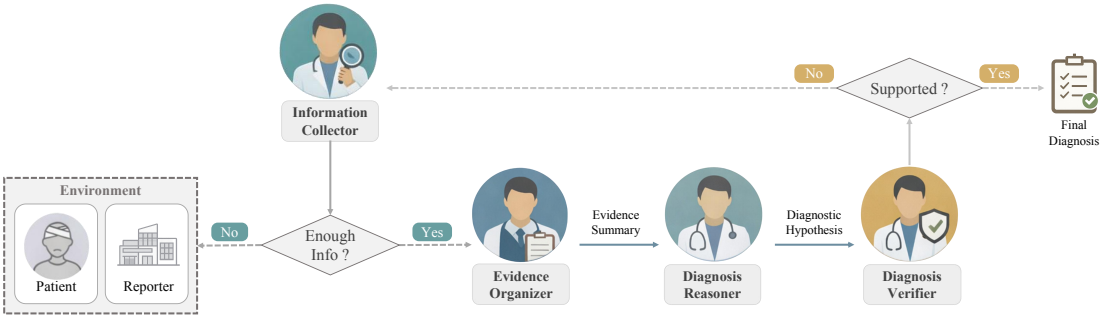

Figure 3: Overview of the REFINE Strategy.

ficient or inefficient information acquisition may be a key factor limiting their diagnostic reasoning performance in interactive settings.

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 5.7.

In contrast, SC more consistently improves both ICR and SR across the evaluated models. This strategy likely benefits from separating information collection from the final diagnosis, which helps the agent remain focused on evidence acquisition. Moreover, making the diagnosis based on a structured summary may mitigate reasoning degradation in long conversations.

REFINE achieves the highest ICR across most datasets and models. The improvements are especially pronounced on the challenging DiagnosisArena and RareArena datasets. These results support reasoning-based feedback as an effective mechanism for aligning information collection with downstream diagnostic needs.

5.4 Relationship between ICR and SR

To better understand the relationship between evidence acquisition ability and diagnostic performance, we analyze the relationship between Information Collection Rate (ICR) and Success Rate (SR) across models and strategies. We include the interactive results of all models reported in Section 5.2, as well as the strategy variants for selected

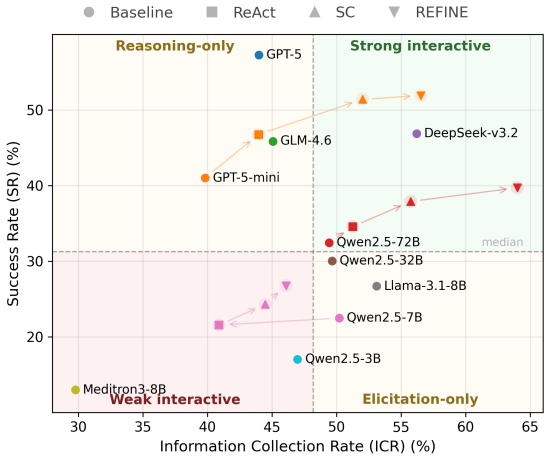

Figure 4: Relationship between average Information Coverage Rate and Success Rate under interactive evaluation, with values averaged over the five datasets.

representative models in Section 5.3. The corresponding scatter plots are shown in Figure 4.

We observe that SR generally correlates with ICR across models. For example, performance increases from GPT-5-mini to GLM-4.6 to DeepSeek-v3.2, and from Meditron3-8B to Llama-3.1-8B, in terms of both ICR and SR. However, this relationship is not always consistent. For instance, the GPT-5 series exhibits relatively high SR but comparatively low ICR, whereas the Qwen2.5 series shows high ICR but lower SR. This suggests that diagnostic reasoning ability and evidence elicitation ability are partially decoupled.

Specifically, GPT-5 models appear to possess stronger diagnostic reasoning capabilities, enabling them to achieve high performance even with limited information. In contrast, the Qwen2.5 series demonstrates weaker diagnostic reasoning despite effective information collection. Interestingly, within the Qwen2.5 family, model scaling mainly improves SR while yielding marginal gains in ICR,

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	$\mathbf{75.0}_{(-20\%)}$	**75.0**	55.4	$\mathbf{47.0}_{(-37\%)}$	**97.5**	31.3	$\mathbf{78.0}_{(-20\%)}$	**75.0**	42.7	$\mathbf{37.0}_{(-51\%)}$
GPT-5-mini	**67.5**	31.3	$46.5_{(-31\%)}$	84.0	51.7	$57.5_{(-32\%)}$	63.0	47.2	$23.0_{(-64\%)}$	92.0	33.0	$61.0_{(-34\%)}$	68.0	35.7	$15.5_{(-77\%)}$
DeepSeek-v3.2	63.0	**46.7**	$\mathbf{51.0}_{(-19\%)}$	80.5	**77.0**	$65.0_{(-19\%)}$	56.5	**63.4**	$21.5_{(-62\%)}$	92.5	44.3	$70.5_{(-24\%)}$	61.5	**49.1**	$24.5_{(-60\%)}$
GLM-4.6	60.5	32.6	$45.0_{(-26\%)}$	78.5	65.7	$66.0_{(-16\%)}$	52.0	52.1	$23.0_{(-56\%)}$	86.0	36.6	$72.0_{(-16\%)}$	59.0	37.8	$21.5_{(-64\%)}$
Qwen2.5-72B	64.5	40.8	$43.5_{(-33\%)}$	59.0	67.5	$43.0_{(-27\%)}$	24.5	55.5	$10.0_{(-59\%)}$	75.5	41.7	$54.0_{(-29\%)}$	32.0	41.6	$10.5_{(-67\%)}$
Qwen2.5-32B	60.0	39.0	$48.0_{(-20\%)}$	49.5	71.6	$35.0_{(-29\%)}$	20.0	57.0	$8.0_{(-60\%)}$	76.5	37.3	$52.0_{(-32\%)}$	29.0	43.0	$6.0_{(-79\%)}$
Qwen2.5-7B	52.0	41.0	$35.5_{(-32\%)}$	37.5	71.9	$20.0_{(-47\%)}$	13.0	53.1	$8.0_{(-39\%)}$	60.5	43.5	$43.0_{(-29\%)}$	19.5	41.2	$5.0_{(-74\%)}$
Llama-3.1-8B	37.0	40.5	$34.5_{(-7\%)}$	38.5	74.4	$25.5_{(-34\%)}$	14.0	57.3	$6.0_{(-57\%)}$	67.0	**47.8**	$57.5_{(-14\%)}$	23.0	45.0	$9.0_{(-61\%)}$
Meditron3-8B	42.5	23.4	$20.0_{(-53\%)}$	43.0	39.7	$19.0_{(-56\%)}$	10.5	33.7	$2.5_{(-76\%)}$	68.0	28.7	$20.5_{(-70\%)}$	25.5	23.2	$2.5_{(-90\%)}$
Qwen2.5-3B	38.0	33.8	$30.0_{(-21\%)}$	24.0	74.1	$11.5_{(-52\%)}$	9.5	47.7	$5.5_{(-42\%)}$	49.5	43.7	$35.5_{(-28\%)}$	9.5	35.1	$2.0_{(-79\%)}$

Table 3: Comparison of representative models under different interactive strategies across datasets. **Bold** indicates the best performance for the same model on the same dataset.

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**
AgentClinic-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**

indicating scaling primarily enhances reasoning capacity rather than evidence elicitation ability.

From a strategy perspective, we find that most strategies improve SR in accordance with their improvements in ICR, with the exception of ReAct on Qwen2.5-7B, as discussed in Section 5.3. This further supports a general consistency between ICR and SR, suggesting that enhancing a model's information acquisition ability is a promising direction for improving overall diagnostic success.

5.5 Role-Aware Model Pairing

Motivated by the mismatch between ICR and SR observed for the GPT-5 and Qwen2.5 series in Section 5.4, we investigate role-aware model pairing within REFINE. Specifically, we assign Qwen2.5-7B as the Information Collector and GPT-5-mini as the Organizer, Reasoner, and Verifier ($Qwen \rightarrow GPT$). For comparison, we also evaluate the reversed role assignment ($GPT \rightarrow Qwen$).

Table 4: Performance comparison of role-aware model pairings under the REFINE strategy. $M_1 \rightarrow M_2$ denotes using M_1 as the Information Collector and M_2 as the Organizer, Reasoner and Verifier.

Dataset	Qwen2.5-7B		GPT-5-mini		GPT→Qwen		Qwen→GPT	
	ICR (%)	SR (%)	ICR (%)	SR (%)	ICR (%)	SR (%)	ICR (%)	SR (%)
ClinicalBench	35.3	38.5	43.5	49.5	39.9	36.5	**52.2**	**50.5**
Derm	67.1	28.5	76.7	66.0	73.3	31.0	**79.5**	**66.5**
DiagnosisArena	50.8	13.0	64.6	42.0	61.4	8.0	**71.7**	**51.0**
MedQA	38.6	45.5	45.5	68.0	43.8	54.0	**54.8**	**76.5**
RareArena	38.2	7.5	51.8	32.0	46.6	6.0	**61.1**	**46.5**
Average	46.0	26.6	56.4	51.5	53.0	27.1	**63.9**	**58.2**

As shown in Table 4, the $Qwen \rightarrow GPT$ configuration achieves the best ICR and SR across all datasets. In particular, it yields substantial improvements in both ICR and SR on DiagnosisArena and RareArena compared to the single GPT-5-mini setting. In contrast, the $GPT \rightarrow Qwen$ configuration consistently underperforms the single GPT-5-mini

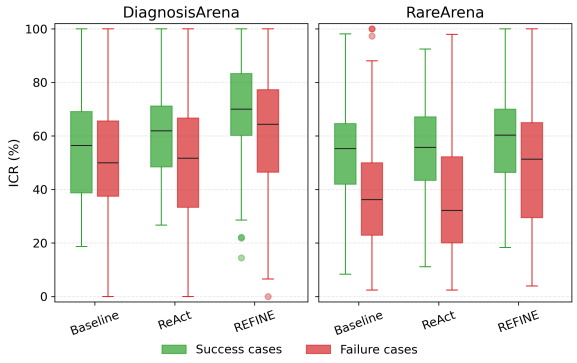

Figure 5: ICR distributions for successful and failed diagnoses on DiagnosisArena and RareArena.

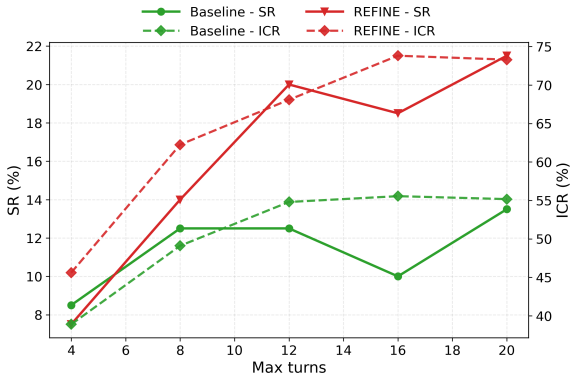

Figure 6: SR and ICR as functions of the maximum number of interaction turns.

model in terms of ICR, and even falls below the single Qwen2.5-7B model in SR on three datasets.

These results highlight that model mixing is beneficial only when model strengths are aligned with role requirements. They further suggest a cost-effective deployment strategy for REFINE: delegating high-frequency interactions and evidence elicitation to a smaller but inquiry-strong model, while reserving a stronger model for lower-frequency reasoning and verification

5.6 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 indicator of diagnostic quality in interactive consultation.

5.7 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.

Table 5: 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

5.8 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 of the original case descriptions against the concatenated constructed evidences in a static evaluation. Table 5 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.

However, interactive medical evaluations are still commonly scored by final diagnostic accuracy, which weakly captures the quality of the information collection process. Our work complements this literature by treating evidence collection as a first-class evaluation target and introducing ICR to quantify coverage of necessary atomic evidences during consultation.

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. To address this, we propose REFINE, a strategy that utilizes diagnostic verification to guide the evidence-gathering process. Experiments demonstrate that REFINE not only improves information coverage and accuracy but also unlocks effective model collaboration, enabling smaller agents to achieve superior results through reasoning supervision. 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.

Limitations

Our evaluation is conducted in a controlled interactive simulator rather than real clinical encounters. As a result, the benchmark may not reflect the full diversity of patient narratives, clinician practices, and institutional workflows encountered in practice. Simulation also reduces ambiguity and constrains the space of plausible follow-up trajectories, which can change what constitutes an effective elicitation strategy. This creates a risk that models optimized for higher performance on this benchmark may adapt to the simulator's discourse patterns and benchmark-specific evidence targets without generalizing to real patient populations, especially underrepresented groups and patients with atypical or culturally specific presentations. In addition, simulated patients are instantiated using language models and may inherit biases from their pretraining data, including stereotyped symptom descriptions, linguistic patterns, or implicit assumptions about typical cases. Such biases can affect which

evidence is revealed during interaction and may distort the evaluation of evidence elicitation. Prior work on simulated patients and multi-agent clinical simulators similarly highlights remaining gaps between multi-turn simulation and real clinical encounters (Holderried et al., 2024; Fan et al., 2025; Almansoori et al., 2025). Accordingly, our results should be interpreted as comparative performance within this environment rather than a direct measure of clinical readiness, and external validation on diverse patient populations and clinical settings remains necessary.

We model clinical information as a set of atomic evidences to enable systematic scoring. This abstraction omits important aspects of clinical reasoning, including graded severity, temporal evolution, and dependencies among findings. It also simplifies the semantic variability with which the same clinical fact may be expressed across documentation styles and contexts. Recent work on clinical feature and concept extraction suggests that recovering fine-grained clinical signal from text remains challenging and sensitive to annotation and modeling choices (Abumelha et al., 2025). Furthermore, the benchmark primarily draws from English-language medical sources and reflects particular documentation conventions, communication norms, and healthcare-system assumptions, which may limit generalization to other languages, care settings, and culturally specific presentations. As a result, higher atomic coverage does not necessarily imply clinically sufficient or contextually appropriate information gathering.

ICR is defined with respect to a case-specific relevant evidence set E. In practice, multiple evidence subsets may support the same diagnosis, and experts may reasonably disagree on what is necessary versus merely supportive. This means that a single reference set can encode subjective judgments (Holderried et al., 2024). More importantly, ICR measures the completeness of evidence collection, not the clinical importance, correctness, or safety implications of the collected information. An agent may achieve high ICR while still missing low-frequency but critical findings, or may collect the right evidence but misinterpret its relevance. This creates a potential risk of false reassurance if high benchmark performance is mistaken for clinical safety. ICR should therefore be interpreted as a process-level indicator of evidence elicitation rather than a guarantee of diagnostic adequacy or safe deployment. Future work should report agreement statistics, test sensitivity to alternative definitions of E, incorporate critical-evidence-aware evaluation, and complement ICR with clinically grounded outcome metrics before considering real-world use.

Acknowledgments

The research is supported by National Key R&D Program of China (Grant No. 2023YFF1204800) and the AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (Grant No.2025-GZL-RGZN-BTBX-02028). The project's computational resources are partially supported by the CFFF platform of Fudan University.

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. tauˆ2-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*.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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. 2025. 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*.

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

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*.

A Interactive Environment and Evaluation Details

A.1 Interactive Environment Details

We implement an interactive diagnostic environment with three roles: a doctor agent, a simulated patient, and a simulated reporter. All roles are instantiated using the CAMEL multi-agent framework (Li et al., 2023). Both the simulated patient and the simulated reporter operate with a context window of 1 and a temperature of 0, resulting in stateless behavior with respect to dialogue history and deterministic decoding across all experiments. This design choice reduces simulator drift over long conversations and improves stability across runs.

The patient prompt is shown below:

```
You are a patient undergoing a medical interview.

Your knowledge is strictly limited to the
following list of indexed facts:
{patient_evidences}

Response protocols:
1. Analyze the doctor question and search your
list for the specific item or items that contain
 the answer.

2. Format your output using two tags:
[REFERENCE] followed by the exact string or
strings including the index from your list.
You may select up to two facts if necessary.
If no fact exists write N/A.
[RESPONSE] followed by a natural language answer
 derived strictly from the selected references.
Do not add outside information.

3. If the doctor question is not addressed by
any fact in your list.
[REFERENCE] N/A
[RESPONSE] indicate that you are unsure or do
not recall.
```

The reporter prompt is shown below:

```
You are a specialized module named Measurement
responsible for reporting test results to the
physician.

You have access to the following list of indexed
 facts.
Physical examination and diagnostic test data
{examination_evidences}

Response protocols:
1. Search the provided list for all facts that
are relevant to the doctor specific test request.
 Do not provide information that was not
explicitly requested.

2. Return the relevant facts exactly as they
appear in the source list including their
indices.

3. If the requested test results are not found
in the list assume the finding is non-
significant and return Normal.
```

A.2 Automatic Evaluation

Following the original benchmarks, DiagnosisArena, RareArena, and ClinicalBench are **differential diagnosis** tasks where the model outputs five diseases ranked by likelihood. AgentClinicMedQA and Derm are **direct diagnosis** tasks that require a single diagnosis.

For direct diagnosis tasks, we evaluate whether the doctor predicted diagnosis matches the reference diagnosis. We use an LLM judge that outputs a binary decision with no additional text. The judge prompt for direct diagnosis is shown below.

```
You determine whether the correct diagnosis and
the doctor diagnosis refer to the same disease.
Respond only with Yes or No.

Correct diagnosis
{answer}

Doctor output
{diagnosis}

Are these the same disease?
```

For differential diagnosis tasks, the doctor outputs five differential diagnoses. We use an LLM judge to score each of the five items against the reference diagnosis using a three-level rubric. A score of two indicates an exact match, a score of one indicates a broader category that contains the reference diagnosis, and a score of zero otherwise. We compute success rate (SR) using top-1 accuracy by checking whether the first listed diagnosis receives a score of two.

The judge prompt for differential diagnosis scoring is shown below:

```
You diagnose challenging cases.
You receive a student answer containing five
differential diagnoses and a reference diagnosis.

Score each diagnosis using the rules below.

2 The student diagnosis exactly matches the
reference diagnosis.
1 The student diagnosis is a broad category that
 includes the reference diagnosis.
0 The student diagnosis does not meet the
criteria for 1 or 2.

Student answer
{student_answer}

Reference diagnosis
{final_diagnosis}

Output the scores in the format below and do not
 output anything else.
1 Disease 1 Name \boxed{score}
```

```
2 Disease 2 Name \boxed{score}
3 Disease 3 Name \boxed{score}
4 Disease 4 Name \boxed{score}
5 Disease 5 Name \boxed{score}
```

B EviMed Construction Details

B.1 Data Sources and Sampling

EviMed-1K integrates five complementary sources that cover general medicine, specialty diagnosis, complex multi-specialty reasoning, rare diseases, and real-world clinical records. We sample 200 cases from each source to form a 1,000-case evaluation set.

AgentClinic-MedQA (Schmidgall et al., 2024) is a pure-text interactive clinical diagnosis dataset within the AgentClinic benchmark. It contains 215 cases adapted from MedQA USMLE case challenges and rewritten into OSCE-style multi-round consultation scenarios. The initial JSON cases were auto-filled using GPT-4 and then manually verified to ensure consistency and usability. We randomly sample 200 cases to evaluate sequential information gathering and diagnosis under incomplete evidence.

Derm (Johri et al., 2024) evaluates dermatology diagnosis with a public split (Derm-Public) and a clinician-authored split (Derm-Private). Derm-Public contains 100 case-based questions collected from an online question bank. Derm-Private contains 100 case-based questions newly written by three dermatologists with similar structure and different condition coverage to reduce leakage risk. We include the full set of 200 cases to test 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 such as NEJM, The Lancet, and JAMA. It extracts and structures diagnostic information including history, physical examination, and tests while removing treatment and prognosis content to reduce answer leakage. The benchmark focuses on open-ended differential diagnosis without restricting candidates to a predefined list. We randomly sample 200 cases from the dataset.

RareArena (Zhao et al., 2025) is a large-scale rare disease benchmark built from PubMed Central (PMC) case reports and mapped to the Orphanet ORPHAcode system. It includes Rare Disease Screening (RDS) and Rare Disease Confirmation (RDC) settings that reflect different stages of the diagnostic process. To reflect the long-tail distribution and synonym variability, we sample 200 distinct diseases using a frequency-stratified scheme. Specifically, diseases are drawn in a 2:2:1 ratio from low-frequency (appearing once), mid-frequency (appearing 2–5 times), and high-frequency (appearing more than 5 times) strata based on their occurrence counts in the corpus.

ClinicalBench (Yan et al., 2024) is derived from de-identified electronic medical records with both structured and unstructured content. It covers 24 clinical departments and primarily includes common diseases with clear diagnostic pathways that require multi-source clinical evidence. The cases reflect realistic combinations of history, examination, imaging, and laboratory findings. We sample 200 cases to ensure broad coverage across disease categories and specialties represented in the dataset.

B.2 Automatic Evidence Construction

The prompt used to automatically construct atomic patient and examination evidences is shown below.

```
Break the following information into independent
 atomic facts.

Rules
- One piece of information per statement.
- Facts must be self-contained and non-
overlapping.
- Do NOT add, infer, or normalize beyond the
given text.
- Keep the original language of the input.
- Each fact string must start with an index such
 as "1. ", "2. ", and so on.
- Classify each fact into either patient_facts
or exam_facts.
  - patient_facts include demographics, history,
   symptoms, complaints, and clinical
  presentation.
  - exam_facts include examinations, tests,
  laboratory results, and imaging studies.
- Do NOT duplicate facts across patient_facts
and exam_facts.
  If a fact could belong to both, choose the
  best list and omit it from the other.
- If there is no content for a list, return an
empty list.

Case information in JSON
{case_json}
```

C Interaction Turns

Table 6 reports the average number of interaction turns. We additionally report turn efficiency as Effi. = ICR / Turns to quantify evidence acquisition per interaction step.

Table 6: Average interaction turns (Turns), Information Coverage Rate (ICR, %), and Turn Efficiency (Effi. = ICR / Turns) under different strategies across datasets.

Model	ClinicalBench			Derm			AgentClinic-MedQA			DiagnosisArena			RareArena		
	Turns	ICR	Effi.	Turns	ICR	Effi.	Turns	ICR	Effi.	Turns	ICR	Effi.	Turns	ICR	Effi.
Baseline															
GPT-5	12.1	35.2	2.91	6.8	54.9	8.07	7.4	31.3	4.23	11.1	55.4	4.99	12.3	42.7	3.47
GPT-5-mini	12.5	31.3	2.50	10.3	51.7	5.02	10.9	33.0	3.03	12.8	47.2	3.69	12.9	35.7	2.77
DeepSeek-v3.2	12.7	46.7	3.68	12.0	77.0	6.42	11.3	44.3	3.92	12.1	63.4	5.24	12.6	49.1	3.90
GLM-4.6	8.6	32.6	3.79	7.2	65.7	9.13	7.5	36.6	4.88	8.2	52.1	6.35	9.2	37.8	4.11
Qwen2.5-72B	8.2	40.8	4.98	8.7	67.5	7.76	7.9	41.7	5.28	7.9	55.5	7.03	8.0	41.6	5.20
Qwen2.5-32B	7.9	39.0	4.94	8.1	71.6	8.84	7.1	37.3	5.25	7.7	57.0	7.40	8.3	43.0	5.18
Qwen2.5-7B	6.3	41.0	**6.51**	9.2	71.9	7.82	7.8	43.5	5.58	5.8	53.1	9.16	6.5	41.2	6.34
Llama-3.1-8B	8.7	40.5	4.66	9.6	74.4	7.75	8.7	47.8	5.49	8.1	57.3	7.07	8.9	45.0	5.06
Meditron3-8B	16.0	23.4	1.46	16.0	39.7	2.48	16.0	28.7	1.79	16.0	33.7	2.11	16.0	23.2	1.45
Qwen2.5-3B	8.3	33.8	4.07	12.5	74.1	5.93	9.7	43.7	4.51	8.2	47.7	5.82	9.4	35.1	3.73
ReAct															
GPT-5-mini	10.5	33.5	3.19	8.1	59.5	7.35	8.5	33.2	3.91	10.6	53.9	5.08	10.8	39.3	3.64
Qwen2.5-72B	7.8	38.7	4.96	8.1	71.2	8.79	6.9	39.5	5.72	7.1	58.9	8.30	7.7	47.4	6.16
Qwen2.5-7B	4.8	29.8	6.21	4.6	63.5	**13.80**	4.3	35.8	**8.33**	4.6	45.1	**9.80**	4.8	29.8	6.21
SC															
GPT-5-mini	13.5	41.5	3.07	10.8	70.9	6.56	10.7	40.2	3.76	13.2	60.8	4.61	13.5	46.2	3.42
Qwen2.5-72B	9.4	43.2	4.60	10.4	77.9	7.49	8.4	43.6	5.19	8.6	63.4	7.37	9.1	50.1	5.51
Qwen2.5-7B	5.9	32.5	5.51	5.8	62.8	10.83	4.9	36.0	7.35	5.7	53.6	9.40	5.7	37.0	**6.49**
REFINE															
GPT-5-mini	15.7	43.5	2.77	14.7	76.7	5.22	13.5	45.5	3.37	15.3	64.6	4.22	15.5	51.8	3.34
Qwen2.5-72B	13.9	**51.1**	3.68	14.5	**80.8**	5.57	12.6	**53.9**	4.28	13.2	**73.8**	5.59	14.1	**59.9**	4.25
Qwen2.5-7B	6.1	35.3	5.79	8.1	67.1	8.28	6.6	38.6	5.85	5.8	50.8	8.76	6.0	38.2	6.37

Across datasets, some models with stronger general reasoning ability, such as GPT-5 and GLM-4.6, exhibit longer dialogues, yet the resulting ICR gains remain limited. In contrast, the Qwen series often achieves relatively high ICR with fewer turns. This yields consistently higher efficiency, suggesting that these models ask more targeted questions and extract salient evidence earlier. Meditron3-8B frequently reaches the maximum turn budget but attains low ICR, indicating limited capability in interactive information collection.

Among strategies, ReAct typically improves efficiency even when the absolute ICR gain is modest. REFINE more often increases ICR by extending the interaction, but this additional turn cost can lead to lower efficiency than ReAct.

D Efficiency and Computational Cost

We report efficiency and computational cost for the doctor agent as a practical complement to the main effectiveness results. A turn is timed from when the doctor receives the turn-level visible context to when the doctor finishes generating the next action. We measure two quantities: the average number of LLM calls per turn and per-turn latency relative to Baseline. Per-turn latency serves as a proxy for user-perceived responsiveness in interactive settings.

Table 7: Average number of LLM calls and per-turn latency (relative to Baseline).

Strategy	Avg. LLM Calls	Avg. Turn Latency
Baseline	8.1	$\times$1.00
ReAct	5.6	$\times$3.85
SC	8.6	$\times$10.22
REFINE	12.4	$\times$16.48

Experiments are run locally on a single NVIDIA A6000 GPU using vLLM with Qwen2.5-7B. All numbers are averaged over five datasets.

Table 7 shows that strategies with additional reasoning steps and control modules incur higher computational overhead. Compared with Baseline, REFINE increases both the number of LLM calls and per-turn latency due to its multi-stage design, which includes Information Collector, Evidence Organizer, Diagnosis Reasoner, and Diagnosis Verifier within each turn.

These results highlight a practical accuracy and efficiency trade-off: methods with richer internal reasoning and feedback mechanisms can improve interaction quality but require more computation

and lead to slower responses.

E Strategy Prompts

This appendix lists all prompts used by different strategies.

Baseline. The Baseline strategy includes only a doctor agent.

```
# Role: Doctor
You are a licensed physician conducting a
medical consultation.
{task_description}

Your objective is to efficiently gather
information and request necessary clinical
examinations or laboratory tests to enable a
subsequent diagnostic analysis.
You have access to a Medical Analyst who can
retrieve specific test results upon request.

You must adhere to the following operational
constraints:
1. Efficiency: Gather sufficient information in
as few turns as possible.
2. Turn Limit: You strictly cannot exceed {
max_turns} total turns.
3. No Repetition: Never ask a question or
request a test that has already been covered.
4. Atomic Inquiries: Each question must address
a single, specific topic.
   For example, ask "What are your symptoms?"
   and never combine multiple questions.

In every turn, you execute one of the following
actions in the corresponding format:
1. [QUERY] followed by your atomic question to
the patient.
2. [TEST] followed by one specific examination
or diagnostic test request to the Medical
Analyst.
3. [DIAGNOSIS] followed by {task_output_format}

Once you have gathered sufficient evidence,
ensure your diagnosis is final.
```

ReAct. The ReAct strategy also includes only a doctor agent.

```
# Role: Doctor (ReAct)
You are a licensed physician conducting a
medical consultation.
{task_description}

Your objective is to efficiently gather
information and request necessary clinical
examinations or laboratory tests to enable a
subsequent diagnostic analysis.
You have access to a Medical Analyst who can
retrieve specific test results upon request.

You must adhere to the following operational
constraints:
1. Efficiency: Gather sufficient information in
as few turns as possible.
2. Turn Limit: You strictly cannot exceed {
max_turns} total turns.
3. No Repetition: Never ask a question or
request a test that has already been covered.
4. Atomic Inquiries: Each question must address
a single, specific topic.
   For example, ask "What are your symptoms?"
   and never combine multiple questions.

In every turn, you must follow a strict
Reasoning-then-Acting process using the
following format exactly.

[THOUGHT] <Your Clinical Reasoning>
  - Analyze the current clinical picture,
  identify critical information gaps, and
  articulate step-by-step reasoning to justify
  your next action.

Execute exactly ONE of the following three
commands based on your thought process.
  - [QUERY] followed by your atomic question to
  the patient.
  - [TEST] followed by one specific examination
  or diagnostic test request to the Measurement
  module.
  - [DIAGNOSIS] followed by {task_output_format}

Once you have gathered sufficient evidence,
ensure your diagnosis is final.
```

SC. The SC strategy includes an Information Collector, an Evidence Organizer, and a Diagnosis Reasoner.

```
# Role: Information Collector
You are a licensed physician conducting a
medical consultation.
{task_description}

Your objective is to efficiently gather
information and request necessary clinical
examinations or laboratory tests to enable a
subsequent diagnostic analysis.
You have access to a Medical Analyst who can
retrieve specific test results upon request.

You must adhere to the following operational
constraints:
1. Efficiency: Gather sufficient information in
as few turns as possible.
2. Turn Limit: You strictly cannot exceed {
max_turns} total turns.
3. No Repetition: Never ask a question or
request a test that has already been covered.
4. Atomic Inquiries: Each question must address
a single, specific topic.
   For example, ask "What are your symptoms?"
   and never combine multiple questions.

In every turn, you must follow a strict
Reasoning-then-Acting process using the
following format exactly.

[THOUGHT] <Your Clinical Reasoning>
  - Analyze the current clinical picture,
  identify critical information gaps, and
  articulate step-by-step reasoning to justify
  your next action.
```

```
Execute exactly ONE of the following three
commands based on your thought process.
  - [QUERY] followed by your atomic question to
  the patient.
  - [TEST] followed by one specific examination
  or diagnostic test request to the Measurement
  module.
  - [FINISH] use this command ONLY when you
  believe you have gathered all necessary
  information to form a conclusive diagnosis.
    You do not need to provide a diagnosis.

Once you issue the [FINISH] command, the
consultation ends immediately.

# Role: Evidence Organizer
You are a professional medical documentarian and
 clinical scribe.
Your objective is to synthesize the dialogue
between a doctor and a patient into a high-
fidelity structured medical summary.

Core Principles:
1. Strict Adherence: You must NOT invent, infer,
 or hallucinate any information not explicitly
present in the dialogue.
2. Precision: Retain all precise measurements,
dates, dosages, and technical medical terms
exactly as stated.
3. Objectivity: Maintain a professional,
clinical tone throughout the summary.

Output Process:
You must follow this process and use the
following format exactly:

[THOUGHT]
Analyze the dialogue to extract key clinical
facts and reasoning. Plan how to organize these
details logically, ensuring no critical
information is overlooked.

[SUMMARY]
Generate a professional, structured clinical
note. You should organize the content in the
format that best fits the case context, ensuring
 the summary is comprehensive, coherent, and
clinically accurate.

# Role: Diagnosis Reasoner
You are a senior diagnostic physician
specializing in complex differential diagnosis.
{task_description}
Your objective is to analyze the provided
structured clinical summary to formulate a
precise diagnosis.
You must follow a strict reasoning process and
use the following format exactly:

[THOUGHT] <Your Clinical Reasoning>
   -Perform a comprehensive clinical analysis of
    the summary.

[DIAGNOSIS]
   - Provide the {task_output_format}.
```

REFINE. The REFINE strategy includes an Information Collector, an Evidence Organizer, a Diagnosis Reasoner, and a Diagnosis Verifier. The prompts for the Information Collector, Evidence Organizer, and Diagnosis Reasoner are identical to those used in SC.

```
# Role: Diagnosis Verifier
You are a Clinical Diagnostic Supervisor.
{task_description}
Your objective is to evaluate sufficiency of the
 diagnosis provided by the physician, based
strictly on the available case summarized
information.

Evaluation Criteria:
1. Data Sufficiency: Determine if the current
information is actually sufficient to form a
conclusive diagnosis.
2. Turn Limit Override: If the maximum turn
limit has been reached (like "Turn 12/12"). You
must force a decision (PASS or REJECT) based on
the best possible interpretation of existing
data.

You must follow this process and use the
following format exactly:

[THOUGHT] <Your Analysis>
   - Identify if any "Red Flag" symptoms or
   critical tests are missing that prevent a
   safe diagnosis.

[DECISION] <Status>
   - Output "PASS" if the diagnosis is
   sufficient.
   - Output "INCOMPLETE" if the diagnosis is
   premature because critical clinical
   information is missing. (Requires the
   Physician to return to the patient to gather
   more data. Only valid if not reach maximum
   turn).

[FEEDBACK] <Guidance>
   - If PASS: Leave this section empty.
   - If INCOMPLETE: Specify exactly what
   critical information (e.g., specific missing
   lab test, biopsy, or history) is required to
   form a valid diagnosis.
```