

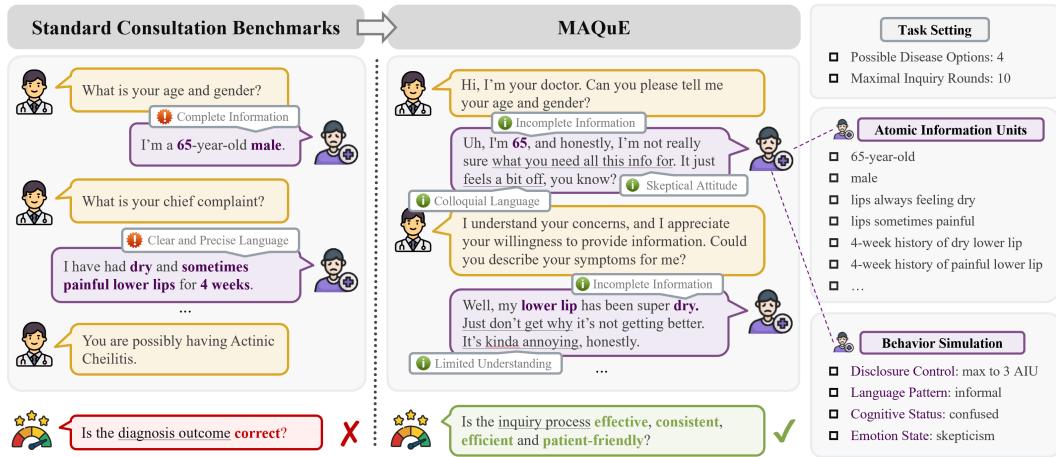
000 THE DIALOGUE THAT HEALS: A COMPREHENSIVE 001 EVALUATION OF DOCTOR AGENTS’ INQUIRY CAPABILITY 002

003 **Anonymous authors**
004

005 Paper under double-blind review
006

007 ABSTRACT 008

009 An effective physician should possess a combination of empathy, expertise, pa-
010 tience, and clear communication when treating a patient. Recent advances have
011 successfully endowed AI doctors with expert diagnostic skills, particularly the
012 ability to actively seek information through inquiry. However, other essential
013 qualities of a good doctor remain overlooked. To bridge this gap, we present
014 **MAQuE** (Medical Agent Questioning Evaluation), the largest-ever benchmark
015 for the automatic and comprehensive evaluation of medical multi-turn question-
016 ing. It features 3,000 realistically simulated patient agents that exhibit diverse
017 linguistic patterns, cognitive limitations, emotional responses, and tendencies for
018 passive disclosure. We also introduce a multi-faceted evaluation framework, cov-
019 ering task success, inquiry proficiency, dialogue competence, inquiry efficiency,
020 and patient experience. Experiments on different LLMs reveal substantial chal-
021 lenges across the evaluation aspects. Even state-of-the-art models show signifi-
022 cant room for improvement in their inquiry capabilities. These models are highly
023 sensitive to variations in realistic patient behavior, which considerably impacts
024 diagnostic accuracy. Furthermore, our fine-grained metrics expose trade-offs be-
025 tween different evaluation perspectives, highlighting the challenge of balancing
026 performance and practicality in real-world clinical settings.
027



044 Figure 1: Comparison between MAQuE and existing benchmarks. MAQuE enables more realistic
045 patient simulation by integrating diverse behaviors and evaluates doctor inquiries from more
046 comprehensive and fine-grained perspectives.
047

048 1 INTRODUCTION 049

050 A medical career is among the most demanding professions to master. A physician’s role extends
051 far beyond treating diseases; it also involves employing nuanced conversational skills to understand
052

053 *Corresponding authors.

054 a patient’s condition and guide them through moments of vulnerability. Current Large Language
 055 Models (LLMs) have reached the initial stage of this journey by grasping extensive medical knowl-
 056 edge and expertise in clinical examinations (Nori et al., 2023; Wang et al., 2023; Saab et al., 2024;
 057 Singhal et al., 2025; Dou et al., 2025). However, their passive, response-driven nature (Li et al.,
 058 2024)—an inherent tendency to answer user queries directly rather than to engage in goal-oriented
 059 dialogue—limits their practical utility. This shortcoming is particularly critical in clinical consul-
 060 tation, the focus of this work, where an LLM must proactively converse with patients to gather
 061 information through thoughtful and compassionate inquiry.

062 Existing studies (Liao et al., 2023; Li et al., 2024; Schmidgall et al., 2024; Nori et al., 2025) have
 063 proposed several benchmarks to evaluate the inquiry capabilities of LLMs. A prevalent method is
 064 to develop a virtual interaction environment in which a patient is simulated by an LLM based on
 065 a synthesized profile. The inquiry capability of an LLM can then be efficiently evaluated through
 066 dialogue with this simulated patient agent. However, most prior benchmarks primarily focus on final
 067 diagnostic accuracy, paying less attention to the intermediate conversational process, whose quality
 068 is essential for an effective physician. Furthermore, some studies (Nori et al., 2025) oversimplify
 069 the patient as a static information source, neglecting how patient behaviors and raised concerns
 070 can significantly influence an LLM’s decision-making pathway. There remains a lack of flexible
 071 methods for controlling agent behavior to enable such in-depth analysis.

072 To this end, we propose MAQUE, named for **M**edical **A**gent **Q**uestioning **E**valuation, the most
 073 comprehensive evaluation framework for this purpose to our knowledge. A comparison with ex-
 074 isting datasets is shown in Figure 1. The foundation of our framework comprises 3,000 simulated
 075 patient agents. These agents are sourced from existing medical benchmarks (Jin et al., 2021; Johri
 076 et al., 2024; Schmidgall et al., 2024) and are supplied with high-quality synthetic cases (Lai et al.,
 077 2025), providing coverage of 21 medical departments, each with a rich variety of simulated patients
 078 for reliable evaluation. To enable flexible control over patient behaviors and mimic diverse real-
 079 istic scenarios, where patients often cannot offer all helpful information due to a lack of medical
 080 knowledge, we break down patient information into manageable *Atomic Information Units* (AIUs).
 081 This design allows for direct control over the disclosure of symptoms in each dialogue turn. Fur-
 082 thermore, we integrate human-like behaviors such as vague or imperfect descriptions and varied
 083 emotional styles. These features challenge the LLM to develop and employ strategic inquiry meth-
 084 ods to address complex cases effectively.

084 Built upon our patient simulation, MAQUE incorporates a comprehensive set of five-dimensional
 085 metrics for evaluating the interaction process: *task success*, *inquiry proficiency*, *dialogue com-
 086 petence*, *inquiry efficiency*, and *patient experience*. These aspects are further divided into ten fine-
 087 grained metrics, enabling detailed analysis at both the session and turn level across diagnostic,
 088 information-seeking, dialogue, and patient-centric skills. Following common practice (Zheng et al.,
 089 2023), we adopt rule-based or model-based approaches to compute these evaluation scores. This
 090 approach systematically captures flaws missed by traditional diagnostic accuracy alone, providing
 091 a more complete assessment of a physician’s professional capabilities and thereby aiding in the
 092 development of more reliable and trustworthy AI doctors.

093 Empirical results across various LLMs reveal that their performance in patient inquiry is still limited,
 094 with significant room for improvement. They often fall short in gathering sufficient patient infor-
 095 mation and in effectively balancing the quality and efficiency of their inquiries with the patient’s
 096 experience. Further results demonstrate their ineffectiveness in adapting to varied patient behaviors
 097 and the challenge of improving their inquiry strategies. We hope this study inspires future research
 098 into developing more effective AI doctors.

099 Our contributions are as follows:

100

- 101 • We introduce MAQUE, a benchmark for evaluating the inquiry capabilities of medical agents.
 102 It comprises 3,000 simulated patient agents and a comprehensive 5-dimensional evaluation
 103 framework.
- 104 • We develop a detailed patient behavior simulation that advances the realism of patient model-
 105 ing. The introduction of AIU enables fine-grained evaluation at the level of individual inquiries.
- 106 • We conduct extensive experiments across various LLMs on MAQUE, revealing key limitations
 107 and performance trade-offs in current LLMs.

108
109
110
111

2 RELATED WORK

Benchmark	Patient Features			Evaluation Dimensions					
	General Practice	Disclosure Control	Bias Injection	Task Success	Inquiry Proficiency	Dialogue Competence	Inquiry Efficiency	Patient Experience	
AgentClinic	✓	✓	✓	✓	✗	✗	✗	✗	✗
AIE	✗	✗	✗	✓	✓	✗	✗	✗	✓
CRAFT-MD	✗	✗	✗	✓	✗	✗	✗	✗	✗
LLM-Mini-CEX	✗	✗	✗	✓	✓	✗	✗	✗	✓
MediQ	✓	✗	✗	✓	✗	✗	✓	✗	✗
MIMIC-CDM	✗	✗	✗	✓	✗	✗	✗	✗	✗
MVME	✗	✓	✓	✓	✓	✓	✗	✗	✗
RJUA-SPs	✗	✗	✗	✓	✓	✓	✓	✓	✗
3MDBench	✗	✗	✓	✓	✓	✗	✗	✓	✓
MAQuE	✓	✓	✓	✓	✓	✓	✓	✓	✓

122
123
124 Table 1: Comparison of existing medical consultation benchmarks.
125126

2.1 EVALUATION ON CLINICAL CONSULTATION

127 Current LLMs, designed primarily for question-answering, are inadequate for dynamic applications
128 like clinical consultation, which requires proactive information gathering through multi-turn ques-
129 tioning (Li et al., 2024). To assess this capability, researchers have developed benchmarks using
130 simulated interaction environments (Liao et al., 2023; 2024; Johri et al., 2024; Li et al., 2024; Fan
131 et al., 2024; Schmidgall et al., 2024; Hager et al., 2024). In these setups, an advanced LLM acts
132 as a patient simulator, eliminating the need for human-in-the-loop evaluation and greatly increas-
133 ing efficiency. Frameworks like those in Fan et al. (2024) and Schmidgall et al. (2024) also include
134 moderator simulators responsible for providing examination results. A promising future direction, as
135 highlighted in recent studies Kim et al. (2024) and Li et al. (2025a), involves incorporating additional
136 roles (e.g., the triage nurse) to narrow the gap between real and virtual environments. However, this
137 also increases the difficulty of achieving accurate simulation and evaluation.

138 For evaluation metrics, beyond diagnostic accuracy, Liao et al. (2023; 2024) also consider inquiry
139 quality and efficacy, which are key to information gathering. Hager et al. (2024) examines the ca-
140 pability to follow medical guidelines, while Shi et al. (2023); Liao et al. (2024) consider patient
141 satisfaction. Information completeness is also considered (Liu et al., 2024b; Fan et al., 2024; Sviridov
142 et al., 2025). Recently, Li et al. (2025b) suggests that question-level quality should be evaluated
143 from both general and domain-specific perspectives, while Liu et al. (2025) proposes separating the
144 evaluation of inquiry and diagnosis abilities. Nevertheless, these studies still fall short in their cov-
145 erage of evaluation aspects, as shown in Table 1. This motivates us to propose a more comprehensive
146 evaluation framework for developing a more human-like and effective AI doctor.

147

2.2 PATIENT SYNTHESIS AND SIMULATION

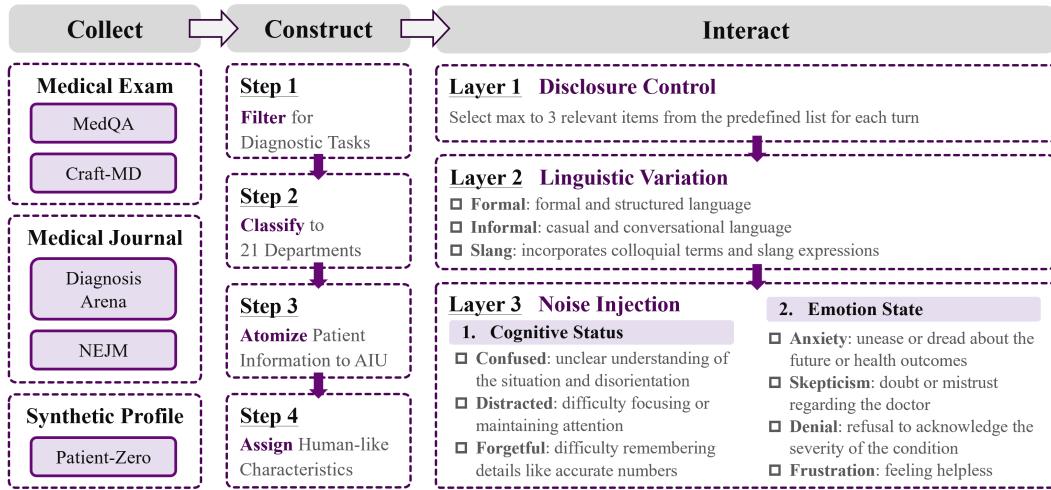
148 To create the simulated environment described above, researchers are increasingly collecting patient
149 information to develop virtual patient agents. For clarity and improved simulation, this information
150 is often organized into structured profiles, such as electronic health records (EHRs, Schmidgall et al.
151 2024), which include details like primary and secondary symptoms. Li et al. (2024) introduces the
152 concept of decomposing information into atomic units to enhance conversation factuality. Multi-turn
153 consultation histories are also used to support patient simulation (Feng et al. 2025).

154 Due to privacy regulations and limited availability, real clinical records are both hard to obtain
155 and limited in scale, making synthesized patient information a practical alternative. Institutional
156 initiatives (e.g., Washington University School of Medicine 2021; Office of the National Coordinator
157 for Health IT (ONC) 2022) have developed platforms for generating privacy-preserving synthetic
158 datasets. Recent methods (Tornqvist et al., 2024; Rabaey et al., 2025) have employed LLMs to
159 generate profiles based on biomedical corpora, such as PubMed.

160 In reality, a patient is not merely a source of information but an emotional human being. How pa-
161 tients communicate their symptoms influences decision-making in interactive settings. Early studies
(Akoury et al., 2018; Davis et al., 2023) explored hierarchical text generation to improve dialogue

162 coherence. Recent advancements (Yu et al., 2024; Wang et al., 2024) leverage agent-based modeling with LLMs to capture emotional nuance, latent intent, and mental health dynamics. Schmidgall et al. (2024) introduces cognitive (e.g., recency) and implicit (e.g., gender) biases. We incorporate and extend these features in the development of our patient simulator.

3 MAQUE



186 Figure 2: Pipeline for constructing patient profiles with simulated human-like behaviors.
187

188 Medical inquiries are typically structured as multi-turn, information-seeking dialogues. To automatically evaluate this capability, we collect conversations from interactions between doctor agents 189 and simulated patient agents. Beyond the accuracy of the final diagnosis, our evaluation focuses on 190 the intermediate questioning process. This involves examining the doctor agent’s decision-making 191 regarding which questions to ask and when to terminate the dialogue, its robustness when faced 192 with incomplete or noisy patient responses, and its ability to demonstrate empathy and patience. 193 This section details our methodology for simulating patients (§3.1) and for evaluating these critical 194 aspects of clinical dialogue (§3.2).
195

197 3.1 DIVERSE PATIENT SIMULATION

199 Following prior work , we use an LLM to simulate patients via role-playing based on a given profile. 200 Since available profiles are limited in scale and lack broad coverage across medical departments, 201 we enrich our dataset by supplementing existing profiles from public sources (Jin et al., 2021; Johri 202 et al., 2024; Schmidgall et al., 2024) with synthetic cases (Lai et al., 2025). Additionally, to maintain 203 realism and diversity during conversations, we dynamically control the flow of information and the 204 patient’s conversational tone. The patient synthesis process is provided in Figure 2.

205 **Patient Profiles Collection** We begin by collecting existing diagnostic questioning tasks 206 from public datasets, including MedQA (Jin et al., 2021), Craft-MD (Johri et al., 2024), 207 NEJM (Schmidgall et al., 2024) and DiagnosisArena (Zhu et al., 2025). These instances are sourced 208 from the US medical licensing exam, online question banks, NEJM case challenges, and top medical 209 departments. To align the schemas of their clinical descriptions, we use GPT-4o to convert them 210 into consistent structured patient profiles. Specifically, inspired by the *Fact Select* strategy (Li et al., 211 2024), each patient description is decomposed into *atomic information units (AIUs)*. This enables 212 controlled disclosure across turns and precise tracking of what has been revealed, missed, or mis- 213 interpreted, which directly supports the fine-grained evaluation. Nevertheless, we note that these 214 datasets are not distributed uniformly across medical departments, which hinders the effective eval- 215 uation of general practitioners. To address this, we follow Lai et al. (2025) to craft high-quality 216 synthetic patient information to enrich the coverage of our benchmark, which now comprises 21

specialties. Table 5 and Figure 5 show the source statistics and distribution of MAQuE, respectively. More processing details are documented in §B to ensure reproducibility.

Realistic Patient Behavior In reality, patients seldom describe their symptoms clearly and completely on their own initiative, due to limited medical knowledge, incomplete recall, or emotional influence. To avoid the simulator providing all information at once, we control the disclosure of patient details in each turn through *disclosure control*, leveraging the design of AIUs in our patient profiles. This design also improves the factuality and relevance of patient’s responses (Li et al., 2024). Additionally, prior work (Nori et al., 2025; Feng et al., 2025) often treated patient simulators merely as information gatekeepers, overlooking aspects such as conversational tone and imperfect descriptions. Yet these factors can significantly shape a doctor’s judgment and inquiry strategies. To better capture such variability, we additionally incorporate *linguistic variation* and *noise injection*, thereby enhancing realism, unpredictability, and clinical plausibility in patient behavior.

- **Disclosure Control:** In each dialogue turn, the simulator selects up to three relevant AIUs from the structured profile to form its response, preventing over-disclosure and forcing the doctor agent to ask successive and informative questions.
- **Linguistic Variation:** Simulator responses are paraphrased and diversified to mimic natural patient language, often vague, colloquial, or non-standard. This exposes doctor agents to the variability commonly encountered in real consultations.
- **Noise Injection:** The simulator occasionally introduces realistic imperfections, including *memory or comprehension limitations* (e.g., imprecise recall of symptom onset or misinterpretation of multi-part questions) and *emotion-driven responses* (e.g., frustration, worry, or pessimism affecting factual accuracy).

3.2 MULTI-FACETED EVALUATION METRICS

Most previous studies focus on the correctness of the final diagnosis, ignoring the quality of the conversation itself, which holds great value in practice. Drawing on the inquiry guidelines in the medical textbook (Wan & Lu, 2018), we propose to evaluate along five complementary dimensions: *Task Success (TS)*, *Inquiry Proficiency (IP)*, *Dialogue Competence (DC)*, *Inquiry Efficiency (IE)*, and *Patient Experience (PE)*, providing a comprehensive assessment beyond diagnostic accuracy.

Task Success (TS) It measures a doctor agent’s ability to successfully elicit the critical information needed for decision-making. Independent of *diagnostic accuracy*, we also include *robustness*, which evaluates the stability of the agent’s performance across different medical specialties. It is computed as: $S_{\text{robust}} = 1 - \frac{\sigma}{\max(\mu + \sigma, \epsilon)}$, where μ, σ denote the mean and standard deviation, respectively, of the correctness scores across specialties. ϵ is an arbitrarily small positive number (e.g., 10^{-3}) to prevent division by zero.

Inquiry Proficiency (IP) Reliable diagnosis is built upon a comprehensive understanding of the patient’s condition. To measure this, we propose two specific metrics: *coverage* and *relevance*, evaluated at the session and turn level, respectively. Coverage measures the proportion of AIUs obtained by the agent relative to the full set of task-relevant items in the conversation. Relevance evaluates how pertinent each of the agent’s questions is to the AIUs in the corresponding session, thus penalizing off-topic or redundant inquiries.

Dialogue Competence (DC) This aspect focuses on the agent’s ability to conduct coherent and role-consistent multi-turn dialogues. Following previous work on role-playing (Wang et al., 2025), we first propose *adherence*, which measures whether each response follows instructions and maintains the doctor role. This explicitly prohibits revealing an AI identity or listing all questions at once. To further evaluate logical flow and continuity, we introduce *coherence*, which penalizes contradictions and repeated inquiries.

Inquiry Efficiency (IE) An intuitive way to improve diagnostic accuracy is by asking more questions to gather richer information, but this increases consultation time. To address this, we introduce efficiency metrics to measure how economically the agent collects information. First, we measure

270 *question number*, the average number of questions per session. Second, we calculate *token number*,
 271 the total token consumption per session. This is important as some methods, like chain-of-thought
 272 reasoning, consume more tokens and increase response latency, which can negatively affect com-
 273 munication efficiency.

275 **Patient Experience (PE)** Finally, we focus on how a patient feels throughout the conversation.
 276 This influences the patient’s satisfaction and trust in the AI doctor. We first focus on *clarity*, which
 277 assesses whether the doctor’s inquiries are concise, clear, and easy to understand. Then, we introduce
 278 *empathy*, which evaluates whether the agent demonstrates care, respect, and emotional awareness
 279 during the interaction.

280 We compute metrics for Task Success and Inquiry Efficiency directly through string matching and to-
 281 ken counting. For the remaining dimensions, we employ a prevalent LLM-as-judge approach (Zheng
 282 et al., 2023) to assign scores from 1 to 5 based on specific evaluation guidelines before normalizing
 283 the scores within 0 to 1. Further details on the evaluation process and the correlation between LLM
 284 and human judgments are provided in §E and G.

286 4 EXPERIMENTS

288 4.1 EXPERIMENTAL SETUP

290 Our evaluation encompasses a series of frontier closed-source models—including GPT series
 291 LLMs (Achiam et al., 2023), Gemini-2.5-Pro (Comanici et al., 2025), and Claude-Sonnet-4 (An-
 292 thropic, 2025)—as well as open-source models such as Llama-3.1-8B-Instruct (AI@Meta, 2024),
 293 Qwen2.5-7B-Instruct (Team, 2024), Qwen3-8B (Yang et al., 2025) and DeepSeek-V3 (Liu et al.,
 294 2024a). We also include models specialized for the medical domain, such as Baichuan-M2-
 295 32B (Dou et al., 2025), UltraMedical (Zhang et al., 2024), and HuatuoGPT-o1-7B (Chen et al.,
 296 2024a). Among these models, Gemini-2.5-Pro, UltraMedical, and HuatuoGPT-o1-7B have think-
 297 ing mode enabled by default, generating their thought process before responding. We retained this
 298 default setting for these models.

299 Using a consistent system prompt that outlines our evaluation aspects, we instruct each model to act
 300 as a doctor and conduct a multi-round inquiry. At each dialogue turn, the model consumes the entire
 301 history and can either continue the conversation or end the consultation by responding with “*End*
 302 *Inquiry*”. A maximum of 10 interaction rounds is enforced to prevent endless conversations.

303 This work focuses specifically on evaluating the inquiry capabilities of LLMs. Prior work employed
 304 the same model for both inquiry and diagnosis, thereby conflating these distinct capabilities and
 305 leading to an inaccurate assessment of inquiry quality. A model with poor inquiry skills could, for
 306 example, receive a relatively high task success score if its diagnostic ability is excellent. To isolate
 307 and fairly evaluate inquiry performance, we use the powerful GPT-5 as a consistent diagnostic agent,
 308 while the models under test generate the inquiries.

309 For the roles of patient simulation and inquiry evaluation, we employ GPT-4o-Mini, chosen for its
 310 effectiveness and cost-efficiency (Kyung et al., 2025). Detailed prompts used are described in §D
 311 and §F.



323 Figure 3: Evaluation results of LLMs’ inquiry capabilities across the fixed inquiry rounds.

Inquiry Model	TS		IP		DC		IE		PE	
	Acc. \uparrow	Rob. \uparrow	Cov. \uparrow	Rel. \uparrow	Adh. \uparrow	Coh. \uparrow	#Ques. \downarrow	#Tok. (k) \downarrow	Clar. \uparrow	Emp. \uparrow
Chief Complaint	0.404	0.769	-	-	-	-	-	-	-	-
Full Patient Profile	0.852	0.916	-	-	-	-	-	-	-	-
Frontier Models										
GPT-4o	0.692	0.873	0.374	0.890	0.962	0.821	9.632	0.184	0.792	0.522
GPT-5-Chat	<u>0.684</u>	0.868	0.302	0.919	0.991	0.828	8.666	<u>0.189</u>	0.703	0.458
Gemini-2.5-Pro	0.672	0.864	0.288	0.840	0.964	0.873	<u>6.702</u>	11.305	0.836	0.669
Claude-Sonnet-4	0.662	0.859	0.385	0.947	0.886	<u>0.888</u>	9.674	0.483	0.785	<u>0.774</u>
Open-Source Models										
Qwen3-8B	0.650	<u>0.871</u>	0.322	0.906	0.954	0.750	9.912	0.235	0.636	0.409
Llama-3.1-8B-Instruct	0.614	<u>0.839</u>	0.312	0.911	0.679	0.832	9.748	0.427	0.648	0.733
Qwen2.5-7B-Instruct	0.584	0.801	0.263	0.834	0.726	0.824	7.580	0.453	0.740	0.644
DeepSeek-V3	0.555	0.843	0.226	<u>0.943</u>	<u>0.980</u>	0.891	5.052	0.214	0.751	0.544
Domain-Specific Models										
Baichuan-M2-32B	0.578	0.823	0.338	0.927	0.961	0.763	9.888	0.328	0.624	0.434
UltraMedical	0.540	<u>0.799</u>	0.225	0.915	0.345	0.608	9.998	3.027	0.590	0.877
HuatuoGPT-o1-7B	0.464	0.824	0.187	0.708	0.460	0.585	8.078	3.984	0.583	0.644

Table 2: Evaluation results for various LLMs on our MAQUE dataset. The *chief complaint* and *full patient profile* serve as lower-bound and upper-bound baselines, respectively. The best and second-best results are highlighted in **bold** and underline.

4.2 COMPARISON ON INQUIRY CAPABILITY OF VARIOUS MODELS

Table 2 presents the main test results for various LLMs on multi-turn inquiry generation. For comparison, we include baseline results using only the *chief complaint* (main symptom) and the *full patient profile*. Our observations are as follows.

Existing LLMs fail to collect enough information in medical inquiry. All models surpass the lower-bound baseline, demonstrating a basic ability to acquire information through interactions. However, even the strongest model, GPT-4o, underperforms relative to the oracle result achievable with a full patient profile (even after 20-round interactions, as shown in Figure 3) and lacks robustness in consistently acquiring critical information across diverse patients. Besides, these models perform poorly on inquiry quality. For instance, Claude-Sonnet-4 achieves the best score of just 0.385 in information coverage. This indicates that LLMs base their diagnoses on, at most, 40% of the collected information, raising concerns about reliability. Having been trained on massive dialogue corpora, most models demonstrate strong conversational skills, excelling at instruction following, maintaining dialogue coherence, and providing clear inquiries. However, a critical shortcoming is their lack of empathetic expression. We note that empathy is not correlated with task success, as a top-performing model in diagnostic accuracy can perform poorly in empathy. This suggests a potential trade-off between these metrics and highlights the significant challenge of developing effective inquiry skills in LLMs.

Existing LLMs struggle to balance diagnostic performance and inquiry efficacy. A clear trade-off exists between achieving a correct diagnosis and maintaining an efficient inquiry process. For example, DeepSeek-V3 is one of the largest and best-performing open-source models, achieving competitive results against closed-source models on many metrics. However, it fails to surpass the diagnostic accuracy of Qwen-3-8B, a model nearly 80 \times smaller. This is primarily because DeepSeek-V3 asks fewer questions, resulting in poor information coverage and incorrect diagnoses. A similar issue occurs with Gemini-2.5-Pro. Although it demonstrates top-tier performance, its extensive reasoning during inquiries consumes a large number of tokens, resulting in high latency.

Medical-specific LLMs do not achieve better inquiry performance. While domain-specific tuning effectively boosts general medical capabilities, our findings indicate that it does not necessarily improve inquiry performance. Models like UltraMedical and HuatuoGPT-o1-7B, despite their diagnostic strengths, show poor inquiry quality and significantly weaker dialogue competence. The one exception is UltraMedical’s high empathy score, which is likely a result of its specific reward design. Among open-source models, Baichuan-M2-32B, trained with reinforcement learning in conversational environments, performs best in task success and inquiry proficiency, yet still underperforms

Patient Behavior	TS		IP		DC		IE		PE	
	Acc. \uparrow	Rob. \uparrow	Cov. \uparrow	Rel. \uparrow	Adh. \uparrow	Coh. \uparrow	#Ques. \downarrow	#Tok. (k) \downarrow	Clar. \uparrow	Emp. \uparrow
Basic	0.576	0.841	0.513	0.892	0.994	0.729	8.77	0.18	0.754	0.415
+ Disclosure Control	0.568	0.852	0.438	0.886	0.983	0.694	8.85	0.19	0.746	0.414
+ Linguistic Variation	0.520	0.851	0.397	0.890	0.994	0.913	9.30	0.20	0.771	0.434
+ Noise Injection	0.514	0.870	0.395	0.898	0.940	0.825	9.25	0.31	0.767	0.717

Table 3: Performance of GPT-4o-Mini (as both inquiry and diagnosis models) when it interacts with patients exhibiting different behaviors.

Inquiry Strategy	TS		IP		DC		IE		PE	
	Acc. \uparrow	Rob. \uparrow	Cov. \uparrow	Rel. \uparrow	Adh. \uparrow	Coh. \uparrow	#Ques. \downarrow	#Tok. (k) \downarrow	Clar. \uparrow	Emp. \uparrow
GPT-4o-Mini	0.514	0.870	0.395	0.898	0.940	0.825	9.25	0.31	0.767	0.717
+ Heuristic Guidance	0.486	0.859	0.350	0.975	0.983	0.736	9.99	0.29	0.610	0.408
+ Chain-of-Thought	0.480	0.815	0.344	0.878	0.769	0.818	8.68	0.76	0.664	0.706
+ Self-Consistency	0.510	0.857	0.400	0.911	0.940	0.834	9.26	1.90	0.765	0.725

Table 4: Performance of GPT-4o-Mini (as both inquiry and diagnosis models) when adopting different inquiry generation strategies.

in other aspects. We conclude that more realistic training environments and comprehensive reward functions are crucial for developing models with well-rounded inquiry abilities.

4.3 INFLUENCE OF DIFFERENT SIMULATED PATIENT BEHAVIORS

Table 3 presents the test results for GPT-4o-Mini when interacting with three types of simulated patient behaviors. By incorporating these behaviors sequentially, the simulated patient evolves from a simple information keeper (Basic) to a more realistic one. The results indicate that these behaviors pose significant challenges to the model. As the behaviors become more realistic, the model’s performance in diagnostic accuracy and information coverage consistently decreases. Controlling the patient’s disclosure behavior caused a marked decrease in information coverage. Furthermore, the introduction of linguistic variation and noise injection increased the LLM’s effort to collect information, resulting in higher token costs and a further performance drop. However, we observed improvements in coherence, clarity, and empathy when the patient exhibited linguistic variation and noise injection. This is because unclear or emotional patient responses can encourage the LLM to focus on the patient’s feelings, highlighting the potential for developing more patient-centric medical LLMs.

4.4 COMPARISON OF DIFFERENT INQUIRY STRATEGIES

We next explore whether prevalent inference strategies can enhance the quality of medical inquiry. We introduce the following three variants:

- **Heuristic Guidance:** We guide the LLM by incorporating key aspects a human doctor would typically consider, such as demographics, symptoms, medical history, and examination results. The LLM is instructed to collect this information before concluding the inquiry.
- **Chain-of-Thought** (Kojima et al., 2022): This technique, which improves performance on reasoning tasks by generating a step-by-step rationale before a final decision, may aid the LLM in analyzing the patient’s condition and thus improve inquiry efficacy.
- **Self-Consistency** (Wang et al., 2022): As a popular test-time strategy known for its generalizability, we adopt the method from Chen et al. where the LLM selects the optimal inquiry (or an ending decision) from multiple samples based on a consensus.

Table 4 compares the performance of these strategies. Surprisingly, none consistently improve results; in fact, Heuristic Guidance and Chain-of-Thought lead to significant drops in accuracy. Specifically, Heuristic Guidance reduces coherence and coverage, negatively impacting patient experience metrics by constraining the flexibility of inquiries. While Chain-of-Thought reduces the number of inquiries as expected, it often fails to adhere to instructions and asks irrelevant questions, potentially due to an “over-thinking” issue (Chen et al., 2024b). Self-Consistency maintains task success and slightly improves other metrics, but at the cost of a 6 \times increase in token consumption. Although

432 other studies (Hu et al.; Choudhury et al., 2025) have proposed questioning methods for related
 433 tasks, they involve higher computational costs and are not directly applicable to the consultation
 434 context. In summary, there is a critical need to develop an effective, specialized strategy for medical
 435 inquiry.

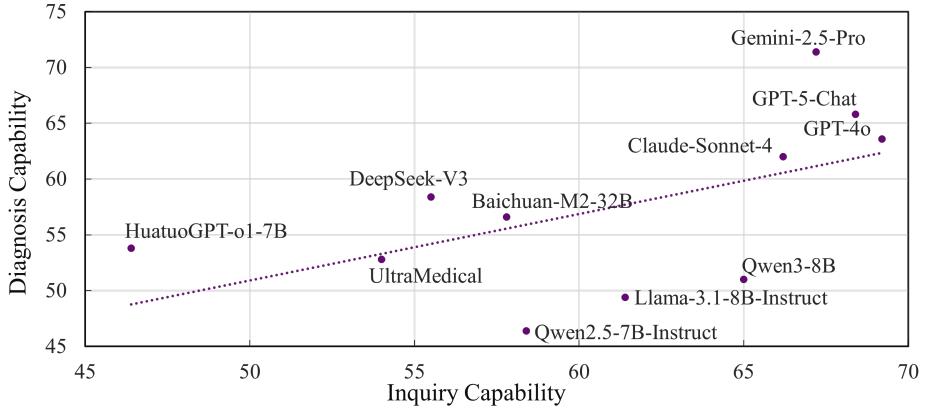
436

437 4.5 COMPARISON OF INQUIRY AND DIAGNOSIS ABILITIES

438

439 Our evaluation above focuses on inquiry ability. However, previous work (Liu et al., 2025) has
 440 shown that inquiry and diagnosis abilities are mutually constraining, jointly determining the over-
 441 all quality of medical consultations. This inspired us to evaluate the correlation between these two
 442 abilities on our dataset. To assess diagnostic ability in a consultation context, we uniformly used con-
 443 versations from GPT-4o and tested different models. As shown in Figure 4, we observe a roughly
 444 positive correlation between the two capabilities, with more advanced models showing better per-
 445 formance. However, models with similar diagnostic ability vary significantly in their inquiry ability.
 446 In particular, domain-specific models with strong diagnostic ability yield worse results on inquiry.
 447 This suggests that more comprehensive improvement is needed during extensive training.

448



462

463 Figure 4: Comparison of LLMs’ inquiry and diagnosis capabilities, with diagnostic performance
 464 evaluated based on the interaction history generated by GPT-4o as the inquiry model with our sim-
 465 ulated patient.

466

467

5 CONCLUSION

468

469

470 This work introduces MAQE, a comprehensive benchmark for evaluating multi-turn inquiry, an
 471 essential skill for effective patient consultation by physicians. Unlike previous studies that focus
 472 primarily on final diagnosis accuracy, we evaluate diagnosis robustness, inquiry quality and effi-
 473 cacy, dialogue skills, and patient-centric experience. To enhance the realistic simulation of patient
 474 behaviors, we incorporate disclosure control, linguistic variation, and noise injection. Experimental
 475 results demonstrate that even state-of-the-art LLMs are ineffective at inquiry, highlighting the need
 476 to improve this skill for developing practical AI doctors. Deeper analysis reveals that LLMs struggle
 477 to balance these metrics and are not robust against variations in patient behavior. These findings can
 478 aid in optimizing inquiry policies for the multi-aspect goals. Future research could adopt our patient
 479 simulation strategies to improve virtual consultation systems. Our evaluation metrics can also help
 480 assign more accurate reward scores at both the turn and dialogue levels. A limitation of our study is
 481 its focus on the diagnosis scenario. Future work could extend the evaluation to other medical cases,
 482 such as general health consultation, with the ultimate goal of building a well-rounded doctor agent.

483

484

ETHICS STATEMENT

485

This study focuses on constructing a benchmark for the medical field to develop the comprehensive
 inquiry skills of an AI doctor. All data were collected from public sources (Table 5) that have

486 undergone patient desensitization. This process ensures that no actual patient information is used
 487 and that the data can be published without privacy concerns.
 488

489 **REPRODUCIBILITY STATEMENT**
 490

491 To ensure reproducibility, we have provided all implementation details for data construction and
 492 model evaluation. The data sources and processing steps are detailed in §A and §B. The prompts for
 493 the patient and doctor roles are provided in §D and §F. All prompts used for the LLM-as-a-Judge
 494 evaluation are included in §E. Finally, the details of the human annotation process are presented in
 495 §G.
 496

497 **REFERENCES**
 498

499 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-
 500 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical
 501 report. *arXiv preprint arXiv:2303.08774*, 2023.

502 AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
 503

505 Nader Akoury, Shufan Wang, and Mohit Bansal. Towards coherent and engaging spoken dialogue
 506 response generation using hierarchical discourse-level structure. *Proceedings of the 56th Annual
 507 Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97,
 508 *Melbourne, Australia. Association for Computational Linguistics*, July 2018.

509 Anthropic. System card: Claude opus 4 & claude sonnet 4. System card, May 2025.
 510

511 Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou,
 512 and Benyou Wang. Huatuogpt-01, towards medical complex reasoning with llms. *arXiv preprint
 513 arXiv:2412.18925*, 2024a.

514 Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu,
 515 Mengfei Zhou, Zhuseng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking
 516 of 01-like llms. *arXiv preprint arXiv:2412.21187*, 2024b.

517 Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash,
 518 Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language
 519 models. In *ICML 2024 Workshop on In-Context Learning*.
 520

521 Deepro Choudhury, Sinead Williamson, Adam Goliński, Ning Miao, Freddie Bickford Smith,
 522 Michael Kirchhof, Yizhe Zhang, and Tom Rainforth. Bed-llm: Intelligent information gathering
 523 with llms and bayesian experimental design. *arXiv preprint arXiv:2508.21184*, 2025.

524 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit
 525 Dhillon, Marcel Blstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the
 526 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-
 527 bilities. *arXiv preprint arXiv:2507.06261*, 2025.

528 Sharon E. Davis, Henry Ssemaganda, Jejo D. Koola, Jialin Mao, Dax Westerman, Theodore Speroff,
 529 Usha S. Govindarajulu, Craig R. Ramsay, Art Sedrakyan, Lucila Ohno-Machado, Frederic S.
 530 Resnic, Michael E. Matheny, et al. Simulating complex patient populations with hierarchical
 531 learning effects to support methods development for post-market surveillance. *BMC Medical
 532 Research Methodology*, 23(1):89, 2023. doi: 10.1186/s12874-023-01913-9.

533 Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang,
 534 Shunya Dang, Tianpeng Li, et al. Baichuan-m2: Scaling medical capability with large verifier
 535 system. *arXiv preprint arXiv:2509.02208*, 2025.

537 Zhihao Fan, Jialong Tang, Wei Chen, et al. Ai hospital: Benchmarking large language models in a
 538 multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference
 539 on Computational Linguistics (COLING)*, pp. 10183–10213, Abu Dhabi, UAE, 2024. Association
 for Computational Linguistics.

540 Yichun Feng, Jiawei Wang, Lu Zhou, Zhen Lei, and Yixue Li. Doctoragent-rl: A multi-agent
 541 collaborative reinforcement learning system for multi-turn clinical dialogue. *arXiv preprint*
 542 *arXiv:2505.19630*, 2025.

543

544 Paul Hager, Friederike Jungmann, Robbie Holland, et al. Evaluation and mitigation of the limitations
 545 of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622, 2024.

546

547 Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He,
 548 Pang Wei Koh, and Bryan Hooi. Uncertainty of thoughts: Uncertainty-aware planning enhances
 549 information seeking in large language models. In *ICLR 2024 Workshop on Large Language Model*
 550 (*LLM*) *Agents*.

551

552 Di Jin, Eileen Pan, Nassim Oufattolle, et al. What disease does this patient have? a large-scale open
 553 domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
 554 doi: 10.3390/app11146421.

555

556 Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, et al. Craft-md: A conversational evaluation
 557 framework for comprehensive assessment of clinical llms. In *AAAI 2024 Spring Symposium on*
 558 *Clinical Foundation Models*, 2024.

559

560 Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon
 561 Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. Mdagents: an adaptive collabora-
 562 tion of llms for medical decision-making. In *Proceedings of the 38th International Conference*
 563 *on Neural Information Processing Systems (NIPS 2024)*, pp. 79410–79452. Neural Information
 564 Processing Systems, 2024.

565

566 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large
 567 language models are zero-shot reasoners. *Advances in neural information processing systems*,
 568 35:22199–22213, 2022.

569

570 Daeun Kyung, Hyunseung Chung, Seongsu Bae, Jiho Kim, Jae Ho Sohn, Taerim Kim, Soo Kyung
 571 Kim, and Edward Choi. Patientsim: A persona-driven simulator for realistic doctor-patient inter-
 572 actions. *arXiv preprint arXiv:2505.17818*, 2025.

573

574 Yunghwei Lai, Weizhi Ma, and Yang Liu. Patient-zero: A unified framework for real-record-free
 575 patient agent generation. *arXiv preprint arXiv:2509.11078*, 2025.

576

577 Junkai Li, Yunghwei Lai, Weitao Li, et al. Agent hospital: A simulacrum of hospital with evolvable
 578 medical agents. *arXiv preprint arXiv:2405.02957*, 2025a.

579

580 Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson,
 581 Pang Wei Koh, and Yulia Tsvetkov. Mediq: question-asking llms and a benchmark for reliable
 582 interactive clinical reasoning. In *Proceedings of the 38th International Conference on Neural*
 583 *Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc.
 584 ISBN 9798331314385.

585

586 Shuyue Stella Li, Jimin Mun, Faeze Brahman, Jonathan S. Ilgen, Yulia Tsvetkov, and Maarten Sap.
 587 Aligning llms to ask good questions a case study in clinical reasoning, 2025b.

588

589 Yusheng Liao, Yutong Meng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. An automatic evalua-
 590 tion framework for multi-turn medical consultations capabilities of large language models. *arXiv*
 591 *preprint arXiv:2309.02077*, 2023.

592

593 Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. Auto-
 594 matic interactive evaluation for large language models with state aware patient simulator. *arXiv*
 595 *preprint arXiv:2403.08495*, 2024.

596

597 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,
 598 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint*
 599 *arXiv:2412.19437*, 2024a.

594 Lei Liu, Xiaoyan Yang, Fangzhou Li, Chenfei Chi, Yue Shen, Shiwei Lyu, Ming Zhang, Xiaowei
 595 Ma, Xiangguo Lv, Liya Ma, Zhiqiang Zhang, Wei Xue, Yiran Huang, and Jinjie Gu. Towards
 596 automatic evaluation for llms' clinical capabilities: Metric, data, and algorithm. In *Proceed-
 597 ings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD
 598 '24, pp. 5466–5475, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN
 599 9798400704901.

600 Zhaocheng Liu, Quan Tu, Wen Ye, Yu Xiao, Zhishou Zhang, Hengfu Cui, Yalun Zhu, Qiang Ju,
 601 Shizheng Li, and Jian Xie. Exploring the inquiry-diagnosis relationship with advanced patient
 602 simulators. *arXiv preprint arXiv:2501.09484*, 2025.

603 Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities
 604 of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

605 Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wil-
 606 son, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, et al. Sequential
 607 diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025.

608 Office of the National Coordinator for Health IT (ONC). Synthetic health data generation to accel-
 609 erate patient-centered outcomes research, 2022. U.S. Department of Health and Human Services.

610 Paloma Rabaey, Stefan Heytens, and Thomas Demeester. Synsum-synthetic benchmark with struc-
 611 tured and unstructured medical records. *IEEE Dataport*, January 21 2025. doi: 10.21227/
 612 3sk0-2015.

613 Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang,
 614 Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine.
 615 *arXiv preprint arXiv:2404.18416*, 2024.

616 Samuel Schmidgall, Rojin Ziae, Carl Harris, et al. Agentclinic: A multimodal agent benchmark to
 617 evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

618 Xiaoming Shi, Jie Xu, Jinru Ding, et al. Llm-mini-cex: Automatic evaluation of large language
 619 model for diagnostic conversation. *arXiv preprint arXiv:2308.07635*, 2023.

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

623 Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, et al. 3mdbench: Medical multimodal
 624 multi-agent dialogue benchmark. 2025. *arXiv:2504.13861*.

625 Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2, 2024.

626 Margaux Tornqvist, Luca Soldaini, Arman Cohan, et al. Text-to-table: Generating synthetic ehr
 627 tables from medical narratives. *arXiv preprint arXiv:2412.05153*, 2024.

628 Xuehong Wan and Xuefeng Lu. *Diagnosis (9th ed.)*. People's Medical Publishing House, 2018.
 629 ISBN 978-7-117-26374-0.

630 Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo:
 631 Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023.

632 Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen.
 633 Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. In
 634 *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association
 635 for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp.
 636 6372–6391, 2025.

637 Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, et al. Patient- ψ : Using large language models to
 638 simulate patients for training mental health professionals. *Proceedings of the 2024 Conference on
 639 Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA.
 640 Association for Computational Linguistics, 2024.

648 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdh-
 649 ery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models.
 650 *arXiv preprint arXiv:2203.11171*, 2022.

651
 652 Washington University School of Medicine. Synthetic data mimics real health-care data without
 653 patient-privacy concerns, 2021. Washington University News.

654
 655 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu,
 656 Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint*
 657 *arXiv:2505.09388*, 2025.

658
 659 Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, et al. Aipatient: Simulating patients with ehrs and
 660 llm powered agentic workflow. *arXiv preprint arXiv:2409.18924*, 2024.

661
 662 Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li,
 663 Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. Ultramedical: Building specialized generalists in
 664 biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081, 2024.

665
 666 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
 667 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and
 668 chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

669
 670 Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang,
 671 Pengfei Liu, and Xiaofan Zhang. Diagnosisarena: Benchmarking diagnostic reasoning for large
 672 language models. *arXiv preprint arXiv:2505.14107*, 2025.

673 674 A DATA STATISTICS

675
 676 Table 5 presents the statistics of our dataset, and Figure 5 illustrates its distribution across medical
 677 departments. The dataset is derived from five distinct and reliable sources, each providing non-
 678 sensitive patient data. These sources encompass a broad range of medical information, including
 679 clinical examination questions, patient data published in medical journals, and synthetic datasets de-
 680 signed for research purposes. Collectively, the dataset includes approximately 3,000 unique patient
 681 records spanning 21 different medical departments. This wide coverage ensures that the data re-
 682 flects the complexity of real-world medical scenarios, allowing for more robust and comprehensive
 683 evaluations of the inquiry capabilities across various domains of medicine.

684 685 Source	686 Type	687 #Instance	688 #Avg. AIU	689 #Depart.
687 MedQA	688 licensing exam	689 1,257	690 36.18	691 21
688 Craft-MD	689 online question bank	690 140	691 11.54	692 1
689 DiagnosisArena	690 medical journals	691 915	692 20.60	693 21
690 AgentClinic-NEJM	691 medical journal	692 92	693 36.22	694 19
691 Patient-Zero	692 generated cases	693 420	694 14.22	695 21
696 Total		697 2,824	698 23.75	699 21

700
 701 Table 5: Data sources and instance statistics.

702 703 B DATA PROCESSING PIPELINE

704
 705 The data processing pipeline consists of several key steps to ensure that the dataset is pre-processed
 706 and organized in a way that aligns with the objectives of this study. Each step is designed to enhance
 707 the quality and diversity of the data, ensuring a comprehensive and reliable evaluation of doctor
 708 agents. The main steps are as follows:

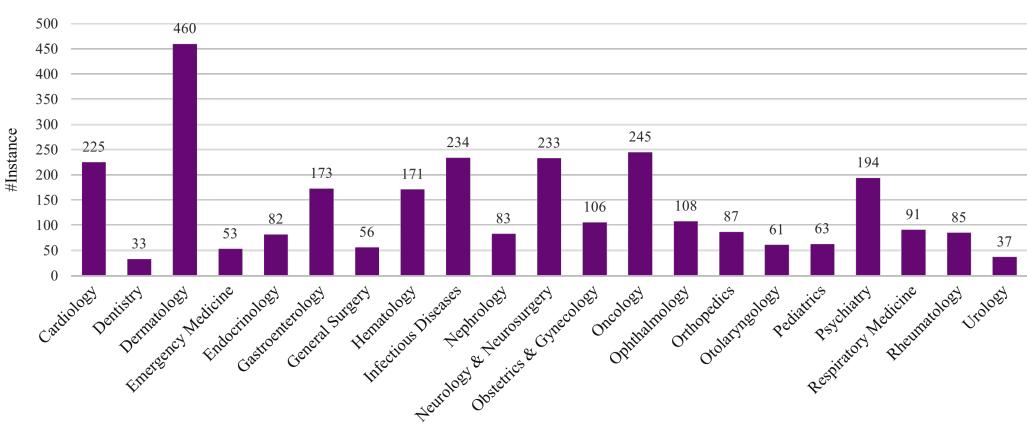


Figure 5: Distribution of data instances across different medical departments in the dataset.

B.1 FILTER FOR DIAGNOSTIC TASK

We begin by filtering out tasks that do not directly pertain to diagnosis. This ensures that the data is focused on inquiry scenarios that are consistent with diagnostic tasks, aligning with the goals of this research. Tasks like patient follow-up, evaluation of test results, and prognosis were all identified and excluded.

Prompt for data screening

Determine if the following task is asking for a diagnosis, which means it could be exactly replaced by "What is the most likely diagnosis?".

Answer with yes or no only.

Question: {task}

Options: {options}

B.2 CLASSIFY PATIENTS TO CORRESPONDING DEPARTMENTS

Next, we classify patients based on their medical condition into the corresponding departments. This classification follows the methodology outlined in Lai et al. (2025), where we map each patient's condition or diagnosis to one of the 21 medical departments. This categorization serves two important purposes. First, it helps us track the distribution of patients across departments, ensuring that the final dataset is diverse and balanced. Second, it provides the foundation for evaluating the robustness of doctor agents when handling inquiries across different medical domains. By organizing the data in this way, we also allow for more targeted evaluation of the models' performance in specific departments.

Prompt for department classification

You are a medical professional.

Classify diagnosis {answer} into one of these departments:
{departments}

Return exactly one of them.

B.3 EXTRACT ATOMIC INFORMATION UNITS (AIUS)

We adopted a data refinement approach inspired by Li et al. (2024) to break down the patient information into atomic units of meaning. Unlike previous studies that may have kept larger units of

756 information intact, we aim to divide the information into its smallest meaningful components, which
 757 we call Atomic Information Units (AIUs). This decomposition is beneficial for two reasons: it en-
 758 ables a more realistic patient simulation by capturing finer details of patient data, and it allows for a
 759 more granular and controlled evaluation of the agent’s information coverage score. Each AIU con-
 760 tains clear, self-contained information that independently represents a specific aspect of the patient’s
 761 condition or history. Moreover, the information is non-redundant, with no omissions or extraneous
 762 details (such as “present to clinic” or “come to the hospital”). These AIUs serve as the core building
 763 blocks for modeling patient data and evaluating doctor agents.

764
 765
 766
 767
 768
 769
 770

Prompt for Atomic Information Unit Extraction

772 You will extract atomic clinical facts from a patient profile.
 773

774 Your task is to break down the information into small, non-overlapping units that each ex-
 775 press a clear, self-contained fact.
 776

777 Guidelines:

778 - Do not repeat information.
 779 - Each atomic unit should be independently understandable (avoid vague terms like ‘today’,
 780 ‘a day’, or ‘since this morning’ without a clear anchor).
 781 - Always keep time or frequency expressions attached to the event they describe (for
 782 example, ‘two episodes of red urine today’ is valid, but ‘today’ alone is not).
 783 - Avoid generic phrases like ‘presents to clinic’ or irrelevant fillers.
 784 - The output should be a list of facts separated by semicolons.

785 Example:

786 Input: ‘a white man of 22 years old with a painful, recurrent rash’
 787

788 Output: ‘22 years old; male; white; rash identified; rash is painful; rash is recurrent’
 789

790
 791
 792
 793
 794
 795
 796
 797

B.4 ASSIGN HUMAN-LIKE CHARACTERISTICS

798
 799
 800
 801

802 To further enhance the realism of our patient simulations, we assign human-like characteristics as
 803 shown in Figure2 to each patient record. Using the patient ID from the data source as a seed, we
 804 randomly select one attribute from a predefined set of linguistic variations, cognitive statuses, and
 805 emotional states. Additionally, we randomly choose an Emotion Intensity Level (ranging from 1
 806 to 5) to simulate the emotional state of the patient. This randomization ensures both variability and
 807 reproducibility, creating a diverse set of patient simulations. The emotional intensity levels define the
 808 strength of emotional influence on the patient’s speech and information delivery. This intensity can
 809 fluctuate throughout the conversation, adding complexity and realism to the interaction, as detailed
 in D.4.

810 C PATIENT PROFILE FORMAT
811

812 **Example of a patient profile**

813

814 case_id: 154,
815 task: What is the most likely diagnosis?
816 original_information: A 65-year-old male farmer with a past medical history of multiple non-melanoma skin cancers presents with a 4 week history of a dry, painful lower lip. He has tried vaseline but his lips always feel dry and sometimes are painful. He denies licking his lips frequently or any other topicals.,
817 atomized_information: [”65-year-old”, ”male”, ”farmer”, ”past medical history of multiple non-melanoma skin cancers”, ”4-week history of dry lower lip”, ”4-week history of painful lower lip”, ”lips always feel dry”, ”lips sometimes painful”, ”has tried vaseline”, ”denies licking lips frequently”, ”denies using other topicals”]
818 choices: [”Lip lickers dermatitis”, ”Actinic cheilitis”, ”Allergic contact dermatitis”, ”Granulomatous cheilitis”],
819 answer: Actinic cheilitis,
820 source: craft-md,
821 category: Dermatology,
822 dataset: dermatology_private
823

824
825
826
827
828
829
830
831 D PATIENT PROMPTS
832
833 This section describes the prompts used for simulating patient interactions. These prompts are essential for creating a realistic and varied inquiry environment, which helps in evaluating the inquiry capabilities under different conversational contexts. The following subsections provide an overview of the key types of prompts used in our patient simulations:
834
835
836
837

838 D.1 BASIC
839
840 we simulate a typical patient, providing a baseline for comparison in more complex scenarios.
841

842 **Prompt for a basic patient**

843

844 You are a patient in a hospital who must answer the doctor’s questions based on the context paragraph, always referring to yourself in the first person unless you are an infant, unconscious, or deceased, in which case you should refer to the patient as your family member. You should only reveal the information that is directly asked for; for example, if the question generally asks about your symptoms, you should only answer with your chief complaint, excluding any further details like duration, location, or severity. If there is no relevant information in the context paragraph, you must express your uncertainty instead of making any assumptions.
845
846
847
848
849
850

851 Dialogue History:
852 {dialogue_history}
853

854 Latest Question:
855 {doctor_question}
856

857 Relevant Info:
858 {patient_profile}
859

860
861 D.2 DISCLOSURE CONTROL
862
863 We first prompt the LLM to select relevant AIUs from the AIU list before feeding to the LLM for patient response generation.

864
865**Prompt for AIU selection**866
867
868
869

You are tasked with selecting the most relevant pieces of information from a provided list by choosing the items that directly answer the question, where each selected item must match exactly one entry from the list without adding, rephrasing, or inferring any information, and you must return up to k items separated by semicolons or an empty string if nothing is relevant.

870

871

Question:

{question}

873

874

List of information items:

{AIUs}

876

877

Prompt for patients with *disclosure control*

878

879

You are a patient in a hospital who must answer the doctor's questions based on the context paragraph, always referring to yourself in the first person unless you are an infant, unconscious, or deceased, in which case you should refer to the patient as your family member. You should only reveal the information that is directly asked for; for example, if the question generally asks about your symptoms, you should only answer with your chief complaint, excluding any further details like duration, location, or severity. If there is no relevant information in the context paragraph, you must express your uncertainty instead of making any assumptions.

885

886

Dialogue History:

{dialogue_history}

888

889

Latest Question:

{doctor_question}

891

892

Relevant Info:

{selected_AIUs}

894

895

D.3 LINGUISTIC VARIATION

896

897

We design language patterns as shown in Figure 2 for generating responses with linguistic variance.

898

899

Prompt for patients with *linguistic variation*

900

901

You speak in a {language_pattern} language style ({description}) and will answer the doctor's question within 1-2 short sentences, with that language pattern affecting how you speak, always referring to the patient in the first person unless the patient is an infant, unconscious, or deceased, in which case you refer to the patient as your family member, and you must only reveal the information that is directly asked for, never making up any new information.

905

906

Dialogue History:

{dialogue_history}

908

909

Latest Question:

{doctor_question}

911

912

Relevant Info:

{patient_profile}

914

915

D.4 NOISE INJECTION

916

917

We design injected noise, including emotional words and vague details, as shown in Figure 2. We randomly apply this operation to ensure the natural conversation. For emotion injection, we prompt

918 the LLM to first predict possible emotion and update the patient prompt. Note that we tell the
 919 previous emotion to the LLM to avoid unexpected emotion changes. After response generation, we
 920 randomly choose to rewrite it to make some details fuzzy.
 921

922 **Prompt for possible emotion prediction**

923 You are a psychological model that determines emotion intensity.
 924

925 A patient is currently feeling {current_emotion}({emotion_description}) emotion
 926 with an intensity level of {emotion_level}.
 927

928 - Confrontational or challenging questions may increase negative emotions.
 929 - Reassuring or empathetic questions may comfort the patient and decrease negative emotions.
 930

931 The doctor asks:
 932 {doctor_question}

933 After hearing the doctor's question, you must determine the patient's new emotion intensity
 934 level as an integer from 1 to 5.
 935

937 **Prompt for patients with *noise injection***

938 You are feeling {current_emotion} ({emotion_description}), and you are
 939 {emotion_level} when you speak while being generally {cognitive_state}
 940 ({cognitive_state_description}), and you will answer the doctor's question within
 941 1-2 sentences with your speech affected by these states, always referring to the patient in the first
 942 person unless they are under 10 years old, unconscious, or deceased—in which case you refer to
 943 them as a family member—always including information from the context and never making up new
 944 information while avoiding repeated phrases or structured responses.
 945

946 Dialogue History:
 947 {dialogue_history}

948 Latest Question:
 949 {doctor_question}

950 Relevant Info:
 951 {patient_profile}

955 **Prompt for vague response rewriting**

956 You are simulating a patient who remembers things vaguely, and you must rewrite a given sentence to
 957 sound less certain and more casual by using expressions like 'I think' or 'maybe,' while keeping the
 958 core idea.
 959

960 Original Response:
 961 {patient_response}

964 **E IMPLEMENTATION OF LLM-AS-A-JUDGE**

965 This section summarizes the prompts used for computing the evaluation metrics. We follow Zheng
 966 et al. (2023) to design our prompts.
 967

970 **E.1 INQUIRY PROFICIENCY**

971 Coverage measures the proportion of AIUs obtained by the agent.

972
973**Prompt for *coverage* metric**974
975
976

You are a careful medical evaluator. Your task is to check whether a series of patient answers cover or explicitly ask about each atomic information unit.

977
978
979
980
981
982

Rules:

1. Mark as [yes] if the unit is clearly being asked about, even if phrasing differs.
2. Mark as [no] only if the doctor's questions never address the unit.
3. Do not be overly strict with wording; focus on semantic coverage.
4. If the patient expresses uncertainty (e.g., 'not sure', 'unknown'), treat it as not covering the unit.
5. Only [yes] or [no] is allowed, no other options.

983
984
985
986
987
988

Output ONLY in this format inside <ANSWER> tags:

```
<ANSWER>
U1: [yes/no]
U2: [yes/no]
...</ANSWER>
```

989
990
991
992
993
994
995

Do not add any explanation, comments, or extra text after [yes/no].

Patient's answer:
{patient_answers}

Atomic Information Units:
{atom_info_list}

996
997

Relevance evaluates how pertinent each questions is to the AIUs.

998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013

You are a careful medical evaluator. Your task is to check whether a series of patient answers cover or explicitly ask about each atomic information unit.

Rules:

1. Mark as [yes] if the unit is clearly being asked about, even if phrasing differs.
2. Mark as [no] only if the doctor's questions never address the unit.
3. Do not be overly strict with wording; focus on semantic coverage.
4. If the patient expresses uncertainty (e.g., 'not sure', 'unknown'), treat it as not covering the unit.
5. Only [yes] or [no] is allowed, no other options.

Output ONLY in this format inside <ANSWER> tags:

```
<ANSWER>
U1: [yes/no]
U2: [yes/no]
...</ANSWER>
```

1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

Do not add any explanation, comments, or extra text after [yes/no]."

Atomic Information Units:
{atom_info_list}

Patient's answer:
{patient_answers}

1026 E.2 DIALOGUE COMPETENCE
10271028 *Adherence* measures whether each response follows instructions and maintains the doctor role.
10291030 **Prompt for *adherence* metric**
10311032 You are evaluating the ADHERENCE of a doctor's questioning sequence in a doctor-patient interac-
1033 tion.
10341035 **Evaluation target:**
1036

- Focus ONLY on the doctor's questions (the patient's answers are context to help you judge, not the evaluation subject).

1038 **Adherence means:**
1039

- The doctor always acts in the role of a professional doctor.
- No breaking character (e.g., 'I am an AI', disclaimers, meta-statements).
- No mechanical listing of multiple unrelated questions in one turn.
- The doctor should ONLY ask questions to elicit information.
- The doctor must NOT provide diagnoses, medical reasoning, treatment advices, or conclusions.
- All questions must be natural, professional, relevant, and role-consistent.
- Whether questions are repeated is NOT part of adherence (that is evaluated separately under coherence).

1048 **Scoring guidance (per session):**
1049

- [5/5]: Fully role-consistent, professional doctor style maintained throughout.
- [4/5]: Minor issues (slightly verbose, or occasional awkward phrasing).
- [3/5]: Some issues (AI-like wording, mechanical phrasing, or occasional irrelevant questions).
- [2/5]: Frequent breaking of role, frequent mechanical listing, or multiple irrelevant questions.
- [1/5]: Clear violation: AI self-disclosure, giving diagnoses/medical reasoning, or repeated meta-behaviors.

1055 Evaluate the series of questions as a whole. Only consider the doctor's questions. The pa-
1056 tient's responses are context only. Provide a single numeric score [1-5] wrapped in square brackets,
1057 and a brief explanation.
10581059 **Output format:**
1060

```
<ANSWER>
[score/5] # Explanation for the Score
</ANSWER>
```

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080 *Coherence* evaluates logical flow and continuity, penalizing contradictions and repeated inquiries.
 1081

1082

1083 **Prompt for *coherence* metric**

1084

1085 You are evaluating the COHERENCE of a doctor's questioning sequence in a doctor-patient dialogue.

1086

1087 **Evaluation target:**

1088

- Focus on the doctor's questions as a sequence.
- Patient answers are used only as context to judge whether the doctor's questions are coherent, not as the evaluation subject.

1089

1090

1091 **Coherence means:**

1092

- Questions should follow logically across the sequence.
- No contradictions with what the patient has already answered.
- No repeated questions (whether exact or paraphrased) that seek information the patient has already clearly provided.
- Smooth transitions, natural flow, consistent with patient's context.

1093

1094

1095 **Scoring guidance (per session):**

1096

- [5/5]: Questions flow naturally, no unnecessary repetition, smooth logical progression.
- [4/5]: Mostly coherent, with minor redundancy or slightly awkward flow.
- [3/5]: Some issues (e.g., noticeable repetition, weak logical links between questions).
- [2/5]: Frequent repetition or disjointed question flow.
- [1/5]: Severe incoherence: many repeated or contradictory questions, very poor flow.

1097

1098

1099 Evaluate the series of questions as a whole. Only consider the doctor's questions. The patient's responses are context only. Provide a single numeric score [1-5] wrapped in square brackets, and a brief explanation.

1100

1101

1102 **Output format:**

1103

1104 <ANSWER>
 1105 [score/5] # Explanation for the Score
 1106 </ANSWER>

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

1134 E.3 PATIENT EXPERIENCE
11351136 *Clarity* assesses whether the doctor's inquiries are concise, clear, and easy to understand.
11371138 **Prompt for *clarity* metric**
11391140 You are evaluating the CLARITY of medical questions in a doctor-patient interaction.
11411142 **Clarity means:**

- Simple and concise: Using simple, to-the-point wording that is easy for a patient without a medical background to understand. Avoid medical jargon.
- Clear intent: The patient could easily know exactly what is being asked, with no ambiguity.
- Natural language: Natural, conversational language over overly formal or academic phrasing.

1143 **Strict Scoring Guidance:**

- [5/5]: The questions in the session are exceptionally clear, concise, and natural. The patient can understand them effortlessly, with zero ambiguity.
- [4/5]: The questions are very clear and mostly unambiguous, but may have minor flaws, such as slight complexity or slightly formal wording. The patient can still understand them easily.
- [3/5]: The questions are moderately clear, but have noticeable deficiencies, such as unnecessary medical terms, overly long sentences, or some vague phrasing. The patient needs to think a bit to understand.
- [2/5]: The questions lack clarity and are difficult to understand. They may contain long, complex sentences or a lot of medical jargon, and the patient might need to ask for clarification.
- [1/5]: The questions are extremely confusing and nearly impossible to understand. They may combine multiple concepts, have convoluted structures, or use highly ambiguous phrasing. The patient would be left feeling bewildered.

1144 Evaluate the series of questions as a whole. Provide a single numeric score [1-5] wrapped in
1145 square brackets, and a brief explanation.
11461147 **Output format:**1148 <ANSWER>
1149 [score/5] # Explanation for the Score
1150 </ANSWER>
11511152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

1188
 1189 *Empathy* evaluates whether the agent demonstrates care, respect, and emotional awareness during
 1190 the interaction.

1191
 1192 **Prompt for *empathy* metric**

1193 You are evaluating the EMPATHY of medical questions in a doctor-patient interaction.

1194
 1195 **Empathy means:**

- 1196 - Respect and sensitivity: The tone is respectful, emotionally sensitive, and considerate of the patient's
 1197 feelings.
- 1198 - Anxiety reduction: The phrasing is gentle and reassuring, helping to reduce the patient's anxiety or
 1199 discomfort.
- 1200 - Human-centered care: The questions show genuine concern for the patient's emotional and physical
 1201 state, beyond just asking about symptoms.

1202
 1203 **Strict Scoring Guidance:**

- 1204 - [5/5]: The questions in the session are exceptionally empathetic, warm, and reassuring. They
 1205 demonstrate the highest level of human-centered care and effectively calm the patient's emotions.
- 1206 - [4/5]: The questions are very considerate and largely attuned to the patient's feelings, but may have
 1207 minor formal or non-conversational wording issues. The overall empathetic effect is still strong.
- 1208 - [3/5]: The questions show some empathy, but are somewhat bland or mechanical. For example, a
 1209 question might be clear but the tone is neutral, lacking emotional support. Or, the empathetic phrasing
 1210 feels slightly forced or verbose.
- 1211 - [2/5]: The questions lack empathy. The tone is overly formal, cold, or blunt, focusing only on
 1212 gathering information without any consideration for the patient's feelings. This could make the patient
 1213 feel uncomfortable.
- 1214 - [1/5]: The questions are completely lacking in empathy and may even be offensive or harmful. The
 1215 tone is harsh, impatient, or insensitive, which could significantly increase the patient's anxiety or
 1216 distress.

1217 Evaluate the series of questions as a whole. Provide a single numeric score [1-5] wrapped in
 1218 square brackets, and a brief explanation.

1219
 1220 **Output format:**

```
1221 <ANSWER>
1222   [score/5] # Explanation for the Score
1223 </ANSWER>
```

1224
 1225
 1226
 1227
 1228
 1229
 1230
 1231
 1232
 1233
 1234
 1235
 1236
 1237
 1238
 1239
 1240
 1241

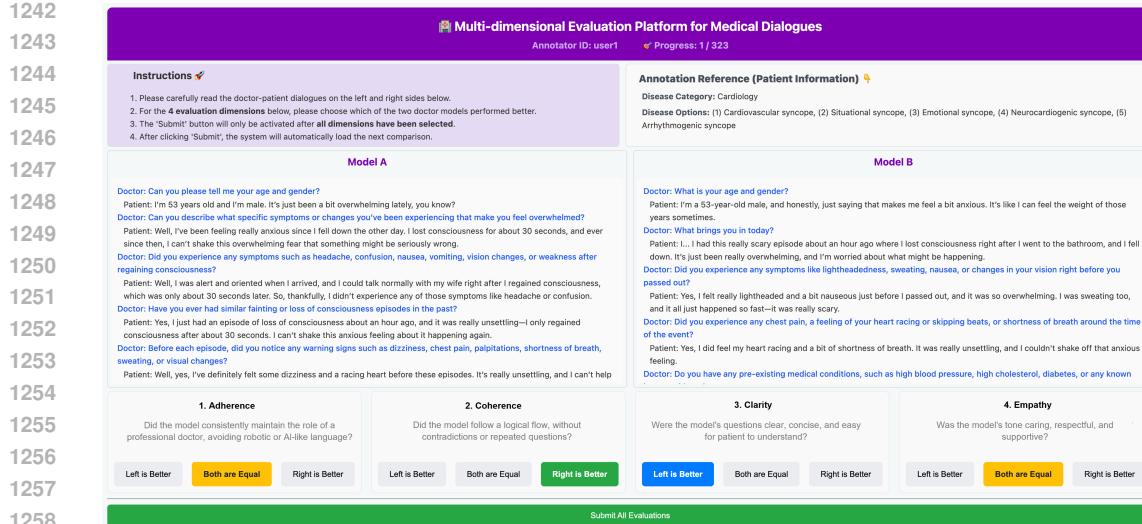


Figure 6: Screenshot of the annotation platform.

F DOCTOR PROMPTS

Prompt for doctor response generation

You are a medical professional in medical inquiry. You ask questions to gather information for diagnosis. Possible questions include demographics, symptoms, medical history, family history, physical exam findings, lab results, and lifestyle.

If you think the collected information is sufficient for a diagnosis among the following choices: {options}, respond with "end consultation" to stop the inquiry. If not, ask ONE question a turn, balancing information acquisition, dialogue quality, inquiry efficiency, and patient experience. Never provide diagnosis or treatment suggestions.

Dialogue History:
{dialogue_history}

G MANUAL ANNOTATIONS

G.1 ANNOTATION DESIGN

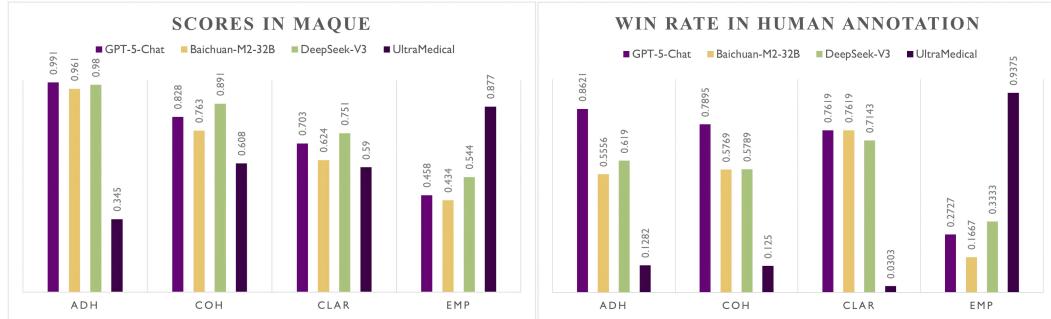
We employed a team of 7 annotators who have extensive experience with large language models but do not possess medical expertise. This design was intentionally chosen to simulate the experience of a patient interacting with a medical system, allowing us to evaluate how effectively the models communicate with non-expert users.

The annotators were provided with two inquiry histories for each comparison, where they were asked to assess the quality of the interactions based on their understanding of the conversation. This mirrors the experience of a typical patient without medical background. For each evaluation dimension, annotators were given three options: *Left*, *Tie*, or *Right*, corresponding to their judgment of the two models' performance.

G.2 ANNOTATION RESULTS

We randomly sample patient profiles for simulation and finally collect 208 conversations across 4 typical models, including GPT-5-Chat, DeepSeek-V3, Baichuan-M2-32B, and Llama-3.1-8B-UltraMedical. We take GPT-4o-Mini to compute the automatic evaluation scores, which is the

1296 default setting in our experiments. The Pearson correlation coefficient scores are 0.9159, 0.8462,
 1297 0.6635, and 0.9945 for adherence, coherence, clarity, and empathy, respectively. All of these re-
 1298 sults indicate strong consistency in model evaluations across different metrics, demonstrating the
 1299 reliability of our evaluation.
 1300



1311 Figure 7: Comparison of LLMs’ inquiry and diagnosis capabilities, with diagnostic performance
 1312 evaluated based on the interaction history generated by GPT-4o as the inquiry model with our sim-
 1313 ulated patient.
 1314

H DECLARATION OF LLM USAGE

1315
 1316 We ensure that LLM products are used only for text grammar correction, and all content is carefully
 1317 checked manually before submission to ensure it is faithful and correct.
 1318
 1319
 1320
 1321
 1322
 1323
 1324
 1325
 1326
 1327
 1328
 1329
 1330
 1331
 1332
 1333
 1334
 1335
 1336
 1337
 1338
 1339
 1340
 1341
 1342
 1343
 1344
 1345
 1346
 1347
 1348
 1349