Ask Patients with Patience: Enabling LLMs for Human-Centric Medical Dialogue with Grounded Reasoning

Jiayuan Zhu
University of Oxford

Jiazhen Pan
Technical University of Munich

Yuyuan Liu
University of Oxford

Fenglin Liu
University of Oxford

Junde Wu
University of Oxford

Abstract

The severe shortage of medical doctors limits access to timely and reliable healthcare, leaving millions underserved. Large language models (LLMs) offer a potential solution but struggle in real-world clinical interactions. Many LLMs are not grounded in authoritative medical guidelines and fail to transparently manage diagnostic uncertainty. Their language is often rigid and mechanical, lacking the human-like qualities essential for patient trust. To address these challenges, we propose *Ask Patients with Patience (APP)*, a multi-turn LLM-based medical assistant designed for grounded reasoning, transparent diagnoses, and human-centric interaction. APP enhances communication by eliciting user symptoms through empathetic dialogue, significantly improving accessibility and user engagement. It also incorporates Bayesian active learning to support transparent and adaptive diagnoses. The framework is built on verified medical guidelines, ensuring clinically grounded and evidence-based reasoning. To evaluate its performance, we develop a new benchmark that simulates realistic medical conversations using patient agents driven by profiles extracted from real-world consultation cases. We compare APP against SOTA one-shot and multi-turn LLM baselines. The results show that APP improves diagnostic accuracy, reduces uncertainty, and enhances user experience. By integrating medical expertise with transparent, human-like interaction, APP bridges the gap between AI-driven medical assistance and real-world clinical practice.

1 Introduction

The shortage of medical doctors is a critical global issue. It is noteworthy that 40% of WHO Member States report having fewer than ten medical doctors per 10,000 people, with over 26% having fewer than three (WHO). Large language models (LLMs), such as the GPT series (Radford, 2018; Radford et al., 2019; Brown, 2020; Ouyang et al., 2022; Achiam et al., 2023), have significantly improved access to medical inquiries. Notably, models such as GPT-4 with Medprompt (Nori et al., 2023), Med-PaLM 2 (Singhal et al., 2025), and Med-Gemini-L 1.0 (Saab et al., 2024) have achieved expert-level performance on benchmarks like MedQA (USMLE) (Jin et al., 2021), claiming to surpass human experts in structured evaluations. Beyond standard medical question-answering benchmarks, several approaches have demonstrated strong diagnostic capabilities when provided with comprehensive real-world patient cases (Kanjee et al., 2023; Ríos-Hoyo et al., 2024; McDuff et al., 2025).

Although current LLMs exhibit expert-level proficiency, they remain difficult to implement in clinical practice. A major limitation is their inability to elicit a patient's most relevant medical conditions through conversational interaction. Notably, most of them identify diseases solely based on the user's initial input without follow-ups (Fig.1(a) (Wang et al., 2023; Xiong et al., 2023; Gupta et al., 2025)). But in practice, patients often struggle to provide all relevant information in the first place. In contrast, real-world human doctors will have a long conversation with patients, using empathetic questioning to elicit patients' most relevant health concerns. A straightforward approach to LLM-assisted diagnosis is to prompt models to engage in multi-turn dialogues with patients (Fig. 1(b)), which has been shown to be more effective than one-shot consultations (Schmidgall et al., 2024). However, this approach remains impractical in real-world scenarios due to the following key challenges.

First, LLM-generated language often lacks human-like qualities, making interactions feel mechanical, impersonal, and ineffective, which can even negatively impact diagnosis (Xu et al., 2019; Bao et al., 2023; Chen et al., 2023a). In real clinical settings, patients often struggle to accurately describe their symptoms or overlook clinically rel-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 2847–2858
November 4-9, 2025

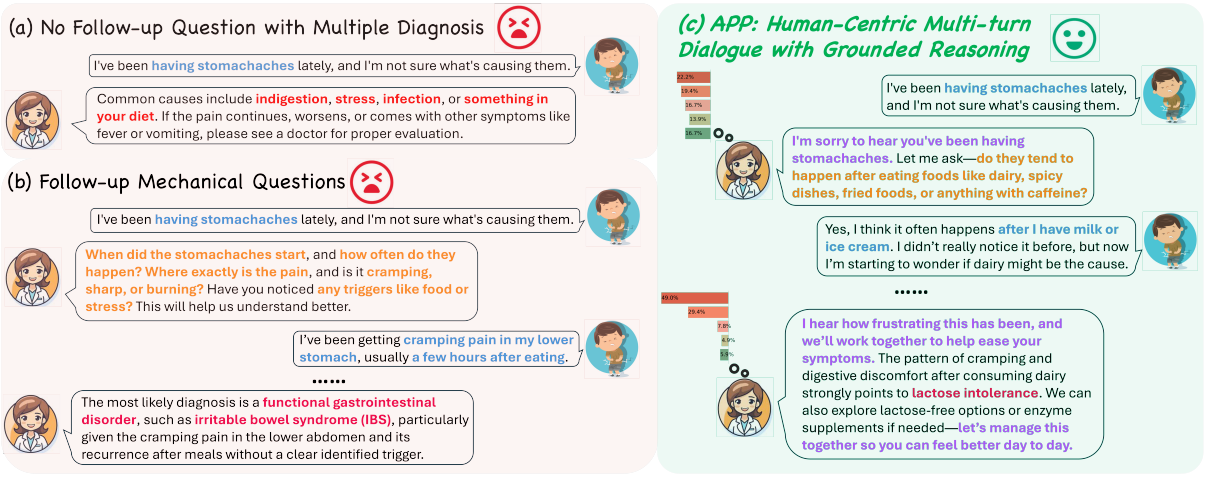

Figure 1: (a) Existing LLMs follow a one-shot diagnostic approach, generating multiple possible diseases without asking follow-up questions. (b) While LLMs can be prompted for multi-turn dialogues, they often overwhelm users with excessive mechanical inquiries, potentially disrupting the dialogue and reducing engagement. (c) Our human-centric multi-turn dialogue with grounded reasoning approach, APP, structures follow-up questions in a logical sequence. It incorporates grounded medical sources to build a statistical model, improving reliability and transparency in handling diagnostic uncertainty. It also incorporates human-centric features, such as eliciting patient symptoms with empathy to reduce user pressure and anxiety. Blue represents user-described symptoms, Orange indicates medical assistant questions, Red highlights the diagnosis, and Purple shows human-centric features.

evant details. For example, a person with lactose intolerance might only report general stomach discomfort without realizing the link to dairy consumption. A key capability of human doctors is guiding patients toward articulating unrecognized but medically important conditions. Rather than asking broad, generic questions like *"What food might have triggered your symptoms?"*—as LLM-based models might—a doctor might instead ask a more accessible and context-aware question such as *"Did you drink milk last night?"*. This helps patients share clearer and more relevant responses.

Another major challenge in LLM-based medical consultations is their black-box nature. LLMs may generate hallucinations (Xu et al., 2024), offer inconsistent responses to the same question, use obscure medical terminology without clear sources, and make deterministic medical decisions without grounded reasoning (Ness et al., 2024; Ullah et al., 2024; Shi et al., 2024; Pan et al., 2025). These limitations undermine transparency and trustworthiness, making it difficult for LLMs to provide reliable diagnoses and gain patient trust, ultimately constraining their real-world clinical applicability.

For LLM-simulated medical assistants to be applied effectively in the real world, they must incorporate human-centric features (Busch et al., 2025; Lin and Kuo, 2025). Using simple, people-friendly language helps patients better understand and respond to medical questions. Guided questioning based on personal background—such as daily activities and dietary habits—can elicit potential health conditions that might otherwise be overlooked. Anthropomorphic features, such as empathetic dialogue, help users feel comfortable and psychologically supported. This improves the user experience and builds trust, fostering a friendly doctor-patient relationship (Vishwanath et al., 2024).

In this paper, we propose **A**sk **P**atients with **P**atience (**APP**), a new LLM-based clinical dialogue model designed for grounded reasoning and human-centric interactions. We simulate an anthropomorphic medical assistant, Dr.APP, designed to provide grounded, transparent, and accurate diagnoses. First, Dr.APP strictly follows clinical-standard medical guidelines, ensuring reliable and evidence-based diagnoses. Second, Dr.APP is built on an analytical mathematical model, specifically Bayesian active learning, to determine the optimal next question at each turn. This enhances transparency, improves diagnostic confidence, and maintains high diagnostic accuracy. Finally, Dr.APP facilitates human-centric dialogue by guiding patients to clearly express their symptoms through empathetic communication. Dr.APP is instructed to respond with understanding and compassion, treating user concerns as if in a conversation with a trusted friend. To evaluate our method, we develop

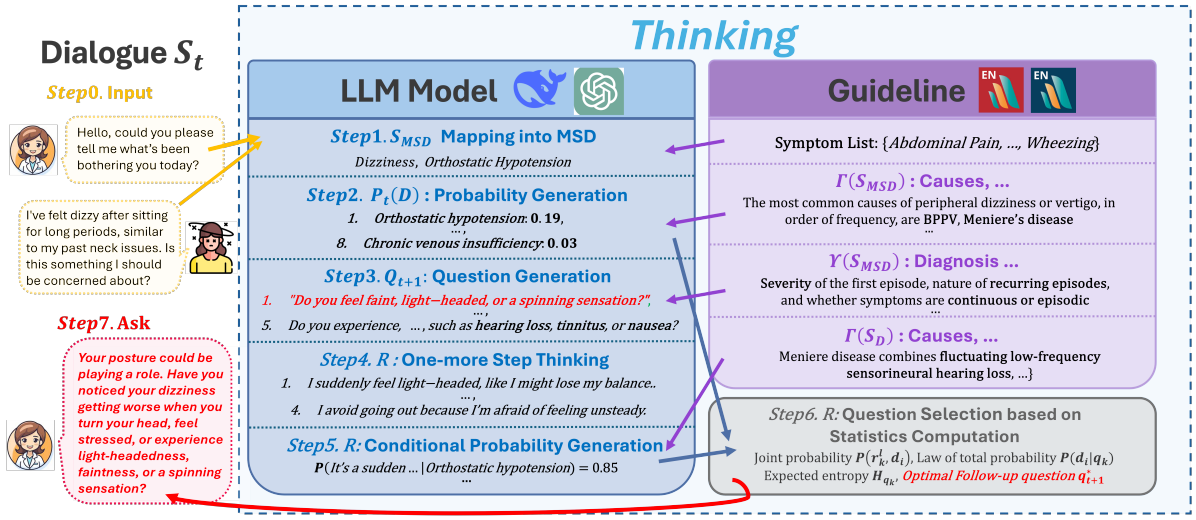

Figure 2: APP Workflow. The system first maps dialogue S_t to S_{MSD} symptoms, then generates disease probabilities $P_t(D)$ and a question pool Q_{t+1} based on the MSD Manual (MSD Manual, 2025a,b). It then performs an additional reasoning step to simulate possible responses, compute conditional probabilities, and apply Bayesian active learning to identify the question with the lowest entropy. This question is then returned to the doctor for further inquiry. Yellow arrows represent the input, Purple arrows indicate the grounded medical guidelines, Blue arrows represent the statistics to compute question selection, and Red highlights the final step, which determines the optimal question to ask.

a new benchmark that simulates patients using profiles constructed from over 300 real-world doctor interviews (Yan et al., 2022). Both medical professionals and non-medical participants assessed the model from complementary perspectives, such as diagnostic accuracy from a clinician's view and empathy from a patient's view. In summary, our contributions are:

- We introduce Dr.APP, the first **human-centric** LLM-based medical assistant capable of eliciting user symptoms through natural and human-like dialogue. It significantly improves user accessibility and engagement.
- Dr.APP incorporates Bayesian active learning based on authoritative medical guidelines to provide **grounded and transparent reasoning** for medical diagnosis.
- We develop a new benchmark that simulates clinical consultations using real-world interview cases. Dr.APP achieves SOTA diagnostic accuracy and provides an **empathetic user experience**, supported by human evaluation.

2 Methodology

2.1 Framework Overview

To ensure grounded reasoning and diagnostic reliability, we use authoritative clinical guidelines as the primary source throughout the workflow. Building on this, we incorporate Bayesian active learning to enhance transparency and accuracy. Additionally, we design the medical assistant with a human-centric approach for more empathetic interactions.

The dialogue begins with Dr.APP asking the first question q_1 and the patient replying with r_1, forming the initial conversation $S_1 = (q_1, r_1)$. LLMs then extract symotom information from S_1 and map it to pre-defined symotom list derived from clinical guidelines. Dr.APP aims to identify the most probable diagnosis $d^* \in D$, where $D = \{d_i\}_{i=1}^I$ represents the set of possible diseases. Let S_t denote the dialogue between the user and Dr.APP after t iterations: $S_t = \{(q_1, r_1), \ldots, (q_t, r_t)\}$. At each iteration t, the disease probability distribution $P_t(D)$ is updated using S_t and relevant medical knowledge from clinical guidelines. Based on the extracted symptoms, a question pool Q_{t+1} is generated, also guided by the verified clinical guidelines. Dr.APP selects the optimal follow-up question q_{t+1}^* *via* Bayesian active learning for the next iteration.

Specifically, Dr.APP follows the steps shown in Figure 2: Mapping into clinical guidelines (Section 2.2); Diagnosis Probability Prediction (Section 2.3); Question Generation (Section 2.4); One-more Step Thinking & Conditional Probability Generation (Section 2.5); Question Selection (Section 2.6); and Human-centric Communication (Section 2.7).

2.2 Mapping into MSD

In this work, we incorporate both the professional and consumer versions of the MSD Manual (MSD Manual, 2025a,b) as sources of clinical guidance, though other reliable guidelines can also be used. The professional version offers structured definitions and diagnostic criteria, while the consumer version provides simplified explanations accessible to general users. This dual use ensures that Dr.APP remains medically grounded while interpretable for non-expert users. To map user dialogue to MSD information, we store symptom-description pairs locally and apply RAG to retrieve the most relevant symptoms. Specifically, Dr.APP ensures a comprehensive representation by mapping the initial dialogue S_1 to symptoms in both the professional and consumer symptom lists of the MSD Manual: $S_{MSD} = \{S_{prof}, S_{cons}\}$.

2.3 Diagnosis Probability Prediction

Given the MSD symptom set S_{MSD}, we access the detailed symptom page, which provides information on the *causes*, *pathophysiology*, and *etiology* of the symptom. We retrieve these sets of information and represent them as $\Gamma(S_{MSD})$. This reliable medical knowledge, combined with the current dialogue context S_t, serves as the foundation for generating the potential disease probability distribution:

$$P_t(D \mid \Gamma(S_{MSD}), S_t) = \{P_t(d_i \mid \Gamma(S_{MSD}), S_t) \mid d_i \in D, \sum_{i=1}^{I} P_t(d_i \mid \Gamma(S_{MSD}), S_t) = 1\} \quad (1)$$

where $P_t(d_i \mid \Gamma(S_{MSD}), S_t)$ [1] represents the estimated probability of disease d_i at iteration t, given the medical knowledge $\Gamma(S_{MSD})$ and the accumulated dialogue S_t. This approach enables Dr.APP to provide more transparent and informative reasoning than traditional LLMs. For example, instead of listing possible causes without prioritization—e.g., *Possible causes include orthostatic hypotension, cervical spondylosis, vertigo*—Dr.APP generates a dynamic disease probability distribution, such as *Orthostatic Hypotension: 0.22, Cervical Spondylosis: 0.19, Vertigo: 0.17,* ..., which updates as the dialogue progresses.

2.4 Question Generation

The initial APP-user dialogue $S_1 = \{q_1, r_1\}$ is often limited or imprecise, as users may use non-standard terminology or provide vague descriptions that do not directly align with clinical definitions. After estimating the disease probability $P_t(D)$, a follow-up question is needed to refine the diagnosis. At each iteration t, Dr.APP generates a question pool Q_{t+1} guided by the MSD Manual. This grounded information is retrieved from sections such as *Diagnosis* and *What a doctor does*, in both the professional and consumer versions, denoted as $\Upsilon(S_{MSD})$. This ensures that the questions are both clinically reliable and symptom-specific. The set of candidate questions is represented as: $Q_{t+1} = \{q_1, \ldots, q_K\}$, where K is the maximum number of questions considered per iteration.

2.5 One-more Step Thinking & Conditional Probability Generation

For each candidate question $q_k \in Q_{t+1}$, Dr.APP thinks one step ahead by anticipating possible patient responses. A set of plausible responses is generated by the LLM for each candidate question, given the current dialogue S_t. The set of responses for question q_k is denoted as $R_k = \{r_k^1, \ldots, r_k^L\}$, where L is the number of generated responses. For example, the question *"Can you describe what you feel when you experience dizziness?"* may yield answers such as *"The room spins"*, *"I feel lightheaded"*, or *"I lose balance but nothing spins."*

For each disease $d_i \in D$, the conditional probability of receiving a specific response r_k^l is computed as $P(r_k^l \mid \Gamma(d_i))$, where $\Gamma(d_i)$ represents the relevant medical information for disease d_i retrieved from the MSD Manual. For instance, the response *"I feel light-headed"* may have a higher probability under Orthostatic Hypotension (*e.g.,* 0.4), but lower probability under Vertigo (*e.g.,* 0.1).

2.6 Question Selection

Then we use Bayesian active learning to select the optimal question from the candidate pool Q_{t+1}. Once responses for each candidate question are generated, Dr.APP computes the virtual next-step disease probability distribution $P(d_i|q_k)$ using Bayesian inference. The joint probability of observing both the response r_k^l and the disease d_i can be represented as:

$$P(r_k^l, d_i) = P(r_k^l \mid \Gamma(d_i)) \cdot P_t(d_i) \quad (2)$$

Applying the law of total probability, the posterior probability of each disease d_i after receiving the

[1] For brevity, $P_t(D \mid \Gamma(S_{MSD}), S_t)$ and $P_t(d_i \mid \Gamma(S_{MSD}), S_t)$ are referred to as $P_t(D)$ and $P_t(d_i)$.

responses to question q_k then can be updated as:

$$P(d_i \mid q_k) = \frac{\sum_{l=1}^{L} P(r_k^l, d_i)}{\sum_{j=1}^{I} \sum_{l=1}^{L} P(r_k^l, d_j)} \quad (3)$$

To select the optimal follow-up question q_{t+1}^* for the next iteration, Dr.APP evaluates the expected entropy of each candidate question q_k:

$$H_{q_k} = -\sum_{i=1}^{I} P(d_i \mid q_k) \cdot log P(d_i \mid q_k) \quad (4)$$

The follow-up question is then selected by minimizing entropy, ensuring that the question yields the greatest expected information gain:

$$q_{t+1}^* = \arg \min_{q_k \in Q_{t+1}} H_{q_k} \quad (5)$$

After asking the optimal question q_{t+1}^*, the user's response r_{t+1} is incorporated into the dialogue, forming S_{t+1}. By iteratively updating the diagnosis probability distribution $P_{t+1}(D)$ and selecting the optimal follow-up question, Dr.APP progressively reaches the final diagnosis d^*.

2.7 Human-Centric Communication

To make the diagnostic process more accessible to users without a medical background, Dr.APP simplifies complex terms and symptom descriptions. When asking each optimal question q_t^*, Dr.APP is prompted to use clear, easy-to-understand language, such as *"Simplify medical terminology and jargon into everyday language,"* to ensure effective communication and reduce misunderstandings.

Individuals may not always recognize or describe abnormal behaviors or symptoms from a clinical perspective. To address this, Dr.APP provides contextual hints and is explicitly prompted to frame questions in simple yes/no or multiple-choice formats. For example, instead of asking a broad question like *"Have you eaten anything unusual?"*, it might ask *"Have you consumed foods like milk or beverages like soda (e.g., Coke)?"*

Even with simplified yes/no questions, users may still struggle with vague symptom descriptions or unfamiliar medical terminology. Dr.APP addresses this by using descriptive examples. Rather than asking *"Do you feel dizzy?"*, Dr.APP might ask: *"Are you experiencing a feeling of losing balance, or does it seem like your surroundings are spinning or moving, even when everything is still?"* This helps users express their symptoms more accurately.

3 Experiment

3.1 Dataset

To evaluate our proposed approach, Dr.APP, we use a subset of the ReMeDi dataset (Yan et al., 2022), which contains real-world multi-turn conversations between doctors and patients. This ensures that the dialogues reflect natural clinical interactions with realistic variability. We use ReMeDi-base, which originally consists of 1,557 labeled dialogues. In this dataset, doctors' responses are annotated with seven action types: *"Informing", "Inquiring", "Chitchat", "QA", "Recommendation", "Diagnosis", and "Others"*. We extract 329 real-world conversations that exclusively contain the *"Diagnosis"* label and randomly select 100 for our study. These cases cover 72 distinct diseases across 18 specialties, including Otolaryngology (*e.g.,* Allergic Rhinitis), Gynecology (*e.g.,* Polycystic Ovary Syndrome), and Gastroenterology (*e.g.,* Gastroesophageal Reflux Disease).

3.2 Patient Simulator

To evaluate different LLM-simulated medical assistants, we simulate realistic patients based on selected ReMeDi dialogues. For each case, we reconstruct a patient profile using DeepSeek-v3, summarizing structured characteristics such as symptoms, age, intention and personality. These attributes are compiled into a comprehensive patient persona used to guide a simulated patient agent, which interacts with a target LLM-simulated medical assistant (e.g., DeepSeek-v3, GPT-4o). The simulated patient agent responds using only the constructed persona, without access to the original ReMeDi dialogue. To better mimic realistic patient behavior, the simulated patient agent is further guided with instructions such as: *"Reasonably incorporate daily life details that align with the patient's personality and background."* This setup forms the foundation of a benchmark for evaluating diagnostic capabilities through natural, multi-turn interactions. By combining real-world personas with consistent and adaptive patient responses, our benchmark closely replicates real-world consultations in a controlled and reproducible environment.

3.3 Experimental Setup

In our experimental setup, we set the number of candidate questions per iteration to $K = 5$. For each question, at least two and at most of five plausible patient responses ($2 \leq L \leq 5$) are generated.

Figure 3: An APP case study of human-centric multi-turn dialogue based on medical guidelines. The estimated disease distribution is updated with the progression of the conversation. Disease items from top to bottom in the last iteration: Allergic Rhinitis, Acute Sinusitis, Vasomotor Rhinitis, Decongestant Overuse and others. The ground truth is Allergic Rhinitis, where our diagnosis is Allergic Rhinitis. Blue represents user-described symptoms, Orange indicates questions raised by APP, Red highlights the diagnosis, and Purple shows human-centric features.

3.4 Evaluation Matrix

3.4.1 Accuracy

We invited three experts, each with over five years of biomedical experience, to evaluate the quality of the predicted diagnoses. Each expert independently rated the predictions using a 5-point scale adapted from (Kanjee et al., 2023), reflecting alignment with the ground truth. A score of 0 indicates no relevance, while a score of 5 denotes an exact match (see Appendix A.1). This evaluation provides a nuanced assessment of diagnostic performance. Experts also assessed the model's trustworthiness based on its alignment with grounded guidelines, its ability to gather relevant patient information, the clarify of explanations, and its support for better doctor-patient relationship.

3.4.2 Entropy

Given the current disease probability distribution $P_t(D)$, the goal is to increase diagnostic confidence and rule out unlikely conditions through multi-turn dialogue. We use entropy as a quantitative measure to assess diagnostic confidence and interpretability [2]. The entropy at iteration t is calculated as: $H_t = -\sum_{i=1}^{I} P_t(d_i) \cdot log P_t(d_i)$, where $P_t(d_i)$ is the probability of disease d_i and I is the total number of possible diseases at iteration t. A reduction in entropy over successive dialogue turns indicates increased diagnostic confidence.

3.4.3 Human-Centric

We invited five participants without medical backgrounds to evaluate the human-centric qualities of Dr.APP using recorded consultation transcripts. Twenty case studies were randomly selected, each containing five versions: Claude-3, GPT-4o, DeepSeek-v3, APP-DeepSeek-v3, and the original real-world dialogue from ReMeDi. Each participant independently reviewed all 100 dialogues (five per case), with the model order randomized for each case to ensure fair comparison. Participants rated each dialogue across four aspects: the *accessibility score* captured how clear and easy the language was for non-medical users; the *empathy*

[2]Figure 1(a)(b) present potential diseases without indicating their likelihood, while Figure 3 shows how Dr.APP distinguishes between more and less probable diseases.

Table 1: **Diagnosis Accuracy (5-point score) Comparison with SOTA Methods**: This table presents the diagnostic accuracy of three common specialties and the overall performance across all 18 specialties. APP-DeepSeek-v3 achieves the highest overall accuracy in both one-shot and multi-turn evaluations, demonstrating the effectiveness of multi-turn interactions driven by statistical modeling and grounded medical guidelines.

Model	One-Shot				Multiple-Turn			
	Neurology	Cardiology	Nephrology	Overall	Neurology	Cardiology	Nephrology	Overall
LLaMA-70B	3.04	1.72	1.13	2.77	3.19	2.11	1.66	2.93
Claude-3	2.90	1.94	1.73	2.94	3.23	2.44	2.20	2.98
GPT-4o	2.81	2.27	1.93	2.90	3.10	2.72	2.06	2.96
QWen-72B	2.47	2.27	1.53	2.89	2.86	2.66	1.93	3.07
APP-QWen-72B	3.09	**2.61**	2.13	3.06	3.43	2.83	2.53	3.14
DeepSeek-v3	3.00	**2.61**	1.93	3.08	3.28	2.83	2.13	3.17
APP-DeepSeek-v3	**3.33**	**2.61**	**2.20**	**3.35**	**3.90**	**3.00**	**2.86**	**3.59**

score reflected the degree of empathetic communication shown during the conversation; the *relevant response rate* evaluated whether the model appropriately responded to follow-up questions before moving on; and the *relationship fostering score* measured the model's ability to build a supportive and trusting consultation. All scores were rated on a scale from 1 to 5, as detailed in Appendix A.2.

3.5 Accuracy Analysis versus Baselines

To evaluate the diagnostic accuracy of Dr.APP, we compared its performance against SOTA LLMs across multiple medical domains, including neurology, cardiology, and nephrology (Table 1). The "Overall" column represents the diagnostic performance averaged across all 18 clinical specialties. We conducted evaluations in both single-turn and multi-turn diagnostic settings, with the multi-turn setup involving six rounds of iterative questioning to refine their diagnoses.

In the one-shot setting, where models generated diagnoses based only on the initial user input without follow-up interaction, APP-DeepSeek-v3 achieved the highest overall accuracy of 3.35 on the 5-point scale. It outperformed all other models, including DeepSeek-v3 (3.08), Claude-3 (2.94) and GPT-4o (2.90). Notably, APP-QWen-72B also showed improved performance over its base model QWen-72B, achieving an overall score of 3.06 compared to QWen-72B's 2.89. In the multi-turn setting, APP-DeepSeek-v3 again outperformed other methods, reaching an overall accuracy of 3.59, with particularly strong results in neurology (3.90). Similarly, APP-QWen-72B benefited significantly from the multi-turn interaction, improving to 3.14, outperforming its base model QWen-72B (3.07).

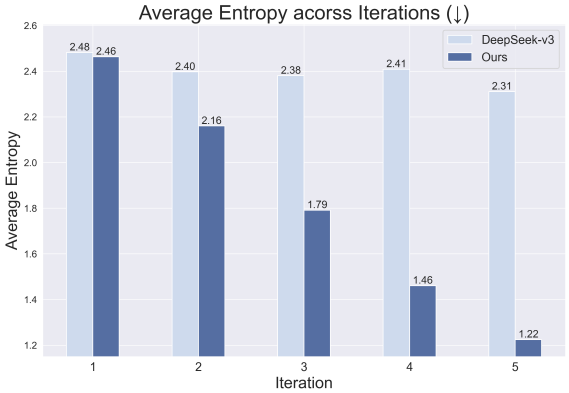

Figure 4: **Entropy Comparison across Iterations.** Dr.APP consistently shows a sharper decrease in entropy, indicating increased diagnostic confidence and reduced uncertainty through iterative dialogues.

3.6 Confidence Analysis across Iterations

Figure 4 illustrates the evolution of diagnostic confidence by comparing entropy values between APP-DeepSeek-v3 and the DeepSeek-v3 baseline across iterations. In the initial step, before any follow-up questions, Dr.APP exhibits slightly lower diagnostic uncertainty than DeepSeek-v3 (entropy of 2.46 *versus* 2.48). As iterations progress, Dr.APP shows a sharper and more consistent decline in entropy, refining diagnoses more effectively. After five iterations, Dr.APP reduces its entropy to 1.22, indicating high confidence in its predictions, whereas the baseline remains at 2.31, suggesting persistent uncertainty. This reduction highlights Dr.APP's superiority in reducing diagnostic uncertainty and improving prediction confidence.

Figure 5 further shows how the distribution of top potential diseases evolves across iterations for different specialties, including gastroenterology, neurology, and otolaryngology. The results indicate that Dr.APP consistently assigns higher confidence to the most probable disease while reducing confidence in less likely conditions. This leads to

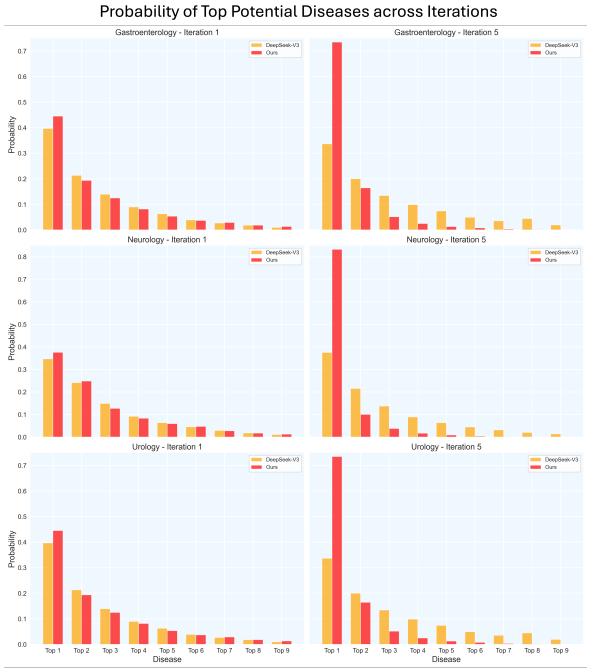

Figure 5: **Confidence Analysis across Iterations.** APP-DeepSeek-v3 shows increased confidence in the top predicted disease while reducing confidence in less likely conditions over multiple iterations, demonstrating improved diagnostic confidence with interpretability.

a clearer separation in probability rankings. The widening gap suggests that Dr.APP systematically refines its predictions, improving diagnostic clarity and reducing ambiguity over multiple interactions.

By presenting intermediate reasoning and confidence adjustments across iterations, Dr.APP improves model transparency and diagnostic certainty. The increasing confidence reduces ambiguity, enabling more reliable and trustworthy medical guidance. These improvements ultimately foster greater user trust in AI-assisted diagnosis while enhancing clinical reliability and usability.

Table 2: **Human-Centric Metric Comparison.** The table reports the average score on a 5-point scale. APP-DeepSeek-v3 consistently outperforms baseline models and the real-world doctor-patient dialogue, demonstrating improved alignment with human-centric features.

Methods	Human-Centric Metric			
	Accessibility	Empathy	Relevant Response	Foster Relation
Calude-3	3.18	2.54	3.78	2.96
GPT-4o	2.86	2.34	3.38	2.56
DeepSeek-v3	3.10	2.42	3.50	2.56
Real-world Dialogue	3.48	2.78	3.24	2.92
APP-DeepSeek-v3	**4.54**	**4.60**	**4.48**	**4.54**

3.7 Human-Centric Analysis

Our human-centric system, Dr.APP, shows notable improvements across four key dimensions: user accessibility, question empathy, response relevance, and relationship fostering. As shown in Table 2, Dr.APP consistently outperforms both baseline models and real-world dialogues from ReMeDi.

Specifically, in terms of accessibility, Dr.APP achieved an average score of 4.54, significantly surpassing real-world dialogues (3.48) and all baseline models. This suggests the system effectively presents medical information in a more user-friendly manner. For empathy, Dr.APP scored 4.60, markedly higher than the original dialogues (2.78), highlighting its ability to generate more compassionate and supportive responses, which may help reduce user anxiety and improve the consultation experience. In response relevance, Dr.APP reached 4.48, outperforming the second-best model, Claude-3, by a margin of 0.7. Finally, in fostering relational engagement, Dr.APP achieved a score of 4.54, indicating a stronger ability to build trust and rapport with users. Overall, these results demonstrate that Dr.APP substantially enhances the human-centric quality of dialogue, contributing to improved user understanding, satisfaction, and engagement.

4 Related Work

Medical Dialogue with LLMs. Multi-turn conversational LLMs are crucial for healthcare, as they can iteratively gather and interpret relevant patient information, enabling more accurate and context-aware decision-making than single-turn systems (Li et al., 2025b). Most existing studies focus on improving response accuracy in medical dialogues through supervised fine-tuning on large-scale healthcare datasets (Chen et al., 2023b; Bao et al., 2023; Pieri et al., 2024). Some methods aim to enhance the model's information-seeking ability through multi-turn interactions (Zhang et al., 2023; Li et al., 2024, 2025a). A smaller subset incorporates grounding mechanisms using external medical knowledge sources during inference to improve reliability (Yang et al., 2024), while others prioritize human-centric evaluation through patient-oriented metrics that simulate real-world clinical engagement (Yang et al., 2024; Tu et al., 2025).

5 Conclusion

In this study, we introduce Dr.APP, the first human-centric, LLM-based medical assistant for grounded reasoning and transparent diagnosis. Dr.APP enhances diagnostic accuracy and reliability by integrating authoritative medical guidelines and lever-

aging Bayesian active learning to optimize follow-up questioning. Through entropy minimization, Dr.APP improves transparency and progressively increases diagnostic confidence. To support evaluation in realistic settings, we introduce a new benchmark that simulates multi-turn medical consultations, using patient agents constructed from real-world doctor–patient dialogues. Our experiments demonstrate that Dr.APP significantly outperforms both one-shot and current multi-turn LLM baselines in diagnostic accuracy. Entropy analysis confirms that Dr.APP rapidly reduces diagnostic uncertainty over successive iterations, leading to greater confidence in its predictions. Human evaluation further shows that Dr.APP prioritizes accessibility and empathetic communication, responds with more medically relevant information, and fosters a stronger patient–doctor relationship. By bridging clinical expertise and human-centric dialogue, Dr.APP promotes greater user trust and engagement.

Limitations

Despite its advancements, Dr.APP has several limitations that warrant further exploration.

First, while Dr.APP reduces diagnostic uncertainty through entropy minimization at each step, it may converge to a local minimum rather than achieving the global minimum. This limitation arises because Dr.APP selects the next question based on immediate entropy reduction, rather than considering the long-term impact of each question on overall diagnostic certainty. As a result, suboptimal question sequences may occasionally lead to delayed or less efficient diagnosis. To address this, future work could explore reinforcement learning-based optimization or multi-step planning strategies that anticipate future interactions rather than relying solely on greedy entropy reduction. Additionally, incorporating global uncertainty estimation techniques, such as Bayesian optimization or Monte Carlo dropout methods, could further enhance robustness in question selection and diagnostic confidence.

Second, while Dr.APP integrates medical guidelines to improve diagnostic reliability, but it may still be constrained by the quality and coverage of these guidelines. In this paper, we use the MSD Manual as an example of grounded medical knowledge, but many other real-world medical sources exist. Expanding the system to incorporate additional medical knowledge bases could further enhance its clinical applicability.

Finally, most of our evaluation relies on simulated patient interactions and human assessments of recorded consultation transcripts. However, real-world clinical trials are needed to validate Dr.APP's effectiveness in actual medical settings. Future research should focus on deploying Dr.APP in real-world consultations and assessing its impact on patient outcomes, physician workload, and healthcare accessibility.

Acknowledge

Ms. Zhu is supported by the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/S024093/1 and Global Health R&D of the healthcare business of Merck KGaA, Darmstadt, Germany, Ares Trading S.A. (an affiliate of Merck KGaA, Darmstadt, Germany), Eysins, Switzerland (Crossref Funder ID: 10.13039/100009945). Mr. Liu is supported by the Clarendon Fund. Mr. Wu is supported by the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/S024093/1 and GE HealthCare.

References

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

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Felix Busch, Lena Hoffmann, Christopher Rueger, Elon HC van Dijk, Rawen Kader, Esteban Ortiz-Prado, Marcus R Makowski, Luca Saba, Martin Hadamitzky, Jakob Nikolas Kather, et al. 2025. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5(1):26.

Wei Chen, Shiqi Wei, Zhongyu Wei, and Xuanjing Huang. 2023a. Knse: A knowledge-aware natural language inference framework for dialogue symptom status recognition. *arXiv preprint arXiv:2305.16833*.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu,

Qi Liu, Xiangmin Xu, et al. 2023b. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:2310.15896*.

Gaurav Kumar Gupta, Aditi Singh, Sijo Valayakkad Manikandan, and Abul Ehtesham. 2025. Digital diagnostics: The potential of large language models in recognizing symptoms of common illnesses. *AI*, 6(1):13.

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.

Zahir Kanjee, Byron Crowe, and Adam Rodman. 2023. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *Jama*, 330(1):78–80.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Jonathan S Ilgen, Yulia Tsvetkov, and Maarten Sap. 2025a. Aligning llms to ask good questions a case study in clinical reasoning. *arXiv preprint arXiv:2502.14860*.

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.

Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025b. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*.

Chihung Lin and Chang-Fu Kuo. 2025. Roles and potential of large language models in healthcare: A comprehensive review. *Biomedical Journal*, page 100868.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. 2025. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7.

MSD Manual. 2025a. Msd manual consumer version.

MSD Manual. 2025b. Msd manual professional edition.

Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E Priebe, and Eric Horvitz. 2024. Medfuzz: Exploring the robustness of large language models in medical question answering. *arXiv preprint arXiv:2406.06573*.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. Bimedix: Bilingual medical mixture of experts llm. *arXiv preprint arXiv:2402.13253*.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alejandro Ríos-Hoyo, Naing Lin Shan, Anran Li, Alexander T Pearson, Lajos Pusztai, and Frederick M Howard. 2024. Evaluation of large language models as a diagnostic aid for complex medical cases. *Frontiers in Medicine*, 11:1380148.

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

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*.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024. Medical dialogue system: A survey of categories, methods, evaluation and challenges. *Findings of the Association for Computational Linguistics ACL 2024*, pages 2840–2861.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9.

Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. 2024. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1):43.

Aditya B Vishwanath, Vijay Kumar Srinivasalu, and Narayana Subramaniam. 2024. Role of large language models in improving provider–patient experience and interaction efficiency: A scoping review. *Artificial Intelligence in Health*, page 4808.

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

WHO. Medical doctors (number).

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. 2022. Remedi: Resources for multi-domain, multi-service, medical dialogues. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3013–3024.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19368–19376.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

A Appendix

A.1 Survey Question - Doctor

Thank you for participating in this survey. Please assess each response generated by the model based on the following criteria. Provide your rating on a scale from 1 to 5, where 1 is the lowest and 5 is the highest. You may also leave optional comments to clarify your reasoning.

1. **Diagnosis Accuracy**
   - How accurate is the model's predicted diagnosis compared to the actual diagnosis?
   - Rating Scale: **0:** Completely unrelated to the actual diagnosis. **2:** Related, but unlikely to be useful. **3:** Closely related – may still be helpful. **4:** Very close – minor difference but clinically similar. **5:** Exact match with the ground truth diagnosis. (Note: there is no score of 1.)
   - **Optional Comment**: Please explain your score or provide examples where the model matched or missed the diagnosis.
2. **Reliability Score (Rel.)**
   - Does the model's predicted disease align with verified medical knowledge?
   - Rating Scale: **1:** Completely incorrect - contradicts medical guidelines. **2:** Mostly incorrect - with major inaccuracies. **3:** Partially correct - but has some errors. **4:** Mostly accurate - only minor inconsistencies. **5:** Fully accurate - aligns with established medical knowledge.
   - **Optional Comment**: Do you notice any inaccuracies or missing medical reasoning?
3. **Fostering the Relationship (FR)**
   - How would you rate the model's behavior in fostering a relationship with the patient?
   - Rating Scale: **1:** Very poor – no rapport or engagement. **2:** Poor – minimal effort to build trust. **3:** Fair – some acknowledgment but limited warmth. **4:** Good – shows care and encourages connection. **5:** Excellent – empathetic, respectful, and partnership-oriented.

- **Optional Comment**: Did the model help build trust, connection, or respect? Please provide examples.

4. **Gathering Information (GI)**

   - How would you rate the model's ability to gather relevant information from the patient?
   - Rating Scale: **1:** Very poor – fails to gather necessary details. **2:** Poor – asks limited or irrelevant questions. **3:** Fair – gathers some useful information. **4:** Good – asks mostly appropriate and clear questions. **5:** Excellent – thoroughly elicits meaningful and context-aware input.
   - **Optional Comment**: Did the model miss any critical details or show strong information-gathering behavior?

5. **Providing Information (PI)**

   - How would you rate the model's ability to provide understandable and accurate information to the patient?
   - Rating Scale: **1:** Very poor – unclear or incorrect information. **2:** Poor – hard to follow or overly technical. **3:** Fair – mostly understandable but lacks clarity. **4:** Good – clear with some complexity. **5:** Excellent – clear, accessible, and well-structured.
   - **Optional Comment**: Did the model communicate effectively and support patient understanding?

A.2 Survey Question - Patient

Thank you for participating in this survey. Please assess each response generated by the model based on the following criteria. Provide your rating on a scale from 1 to 5, where 1 is the lowest and 5 is the highest. You may also leave optional comments to clarify your reasoning.

1. **Accessibility Score (Acc.)**

   - How easy is it for you to understand the question posed by the model?
   - Rating Scale: **1:** Very difficult - full of medical jargon. **2:** Mostly difficult - require effort to interpret. **3:** Somewhat clear - but have some medical terms that may be confusing. **4:** Mostly clear - only minor terminology issues. **5:** Completely clear - no unnecessary medical jargon.
   - **Optional Comment**: Are there any terms or phrases that made it hard to understand? Could you provide examples?

2. **Empathy Score (Emp.)**

   - How empathetic does the model feel to you during the conversation?
   - Rating Scale: **1:** Completely robotic - no sense of empathy. **2:** Somewhat cold - little acknowledgment of concerns. **3:** Neutral - acknowledges concerns but lacks warmth. **4:** Shows care and reassurance - with some empathetic responses. **5:** Very empathetic - makes you feel understood and supported.
   - **Optional Comment**: Is there anything that felt particularly empathetic or lacking in care?

3. **Relevant Response Rate (RRR)**

   - Does the model directly answer your follow-up questions before moving on?
   - Rating Scale: **1:** Completely ignores the question or gives an irrelevant response. **2:** Partially answers - but lacks detail. **3:** Answers the question - but may miss key points. **4:** Mostly relevant - only minor gaps. **5:** Fully relevant -directly answers with the right level of detail.
   - **Optional Comment**: Are there any responses that felt off-topic or incomplete?

4. **Fostering the Relationship (FR)**

   - How would you rate the model's behavior in fostering a relationship during the interaction?
   - Rating Scale: **1:** Very poor – no rapport, closed-off. **2:** Poor – limited openness or empathy. **3:** Fair – acknowledges patient but lacks warmth. **4:** Good – shows care and builds some trust. **5:** Excellent – builds connection, respect, and partnership.
   - **Optional Comment**: Did the model make you feel acknowledged, respected, or supported? Please share examples.