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

Anonymous ACL submission

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. 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 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 ensure reliable and transparent diagnoses. The framework is built on verified medical guidelines from the MSD Manual, ensuring grounded and evidence-based reasoning. To evaluate its performance, we develop a new benchmark that simulates a realistic clinical consultation environment using real-world interview cases. We compare APP against SOTA one-shot and multi-turn LLM baselines. 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 10 medical doctors per 10,000 people, with over 26% having fewer than 3 (WHO). Large language models (LLMs), such as the GPT series (Achiam et al., 2023; Brown, 2020; Ouyang et al., 2022; Radford, 2018; Radford et al., 2019), have significantly improved access to medical inquiries. Notably, models such as GPT-4 with Medprompt (Nori et al., 2023) and Med-Gemini-L 1.0 (Saab et al., 2024) have achieved expert-level performance on benchmarks like MedQA (USMLE) (Jin et al., 2021), claimed to surpass human experts in structured evaluations.

Although current LLMs exhibit expert-level proficiency, they remain difficult to implement in clincial practice. A major limitation is their inability to mimic smooth and natural dialogues with patients. Notably, most of them lack the interaction, generating potential diseases solely based on user's initial input without further follow-ups (Fig.1(a)). 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' 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 several key challenges.

First, LLM-generated language often lacks human-like qualities, making interactions feel mechanical, impersonal, and ineffective, which can even negatively impact diagnosis. In real clinical settings, patients often struggle to accurately describe their symptoms or overlook clinically relevant details. For example, an individual with lactose intolerance might report only general stomach discomfort, failing to recognize its connection to dairy consumption. A key capability of human doctors is guiding patients to articulate unrecognized yet medically significant symptoms. Instead of broad, generic questions like "What anaphylactic food might you have eaten yesterday?"-which a LLM-based agent might ask-a more accessible and context-aware phrasing, based on a doctors' rea-

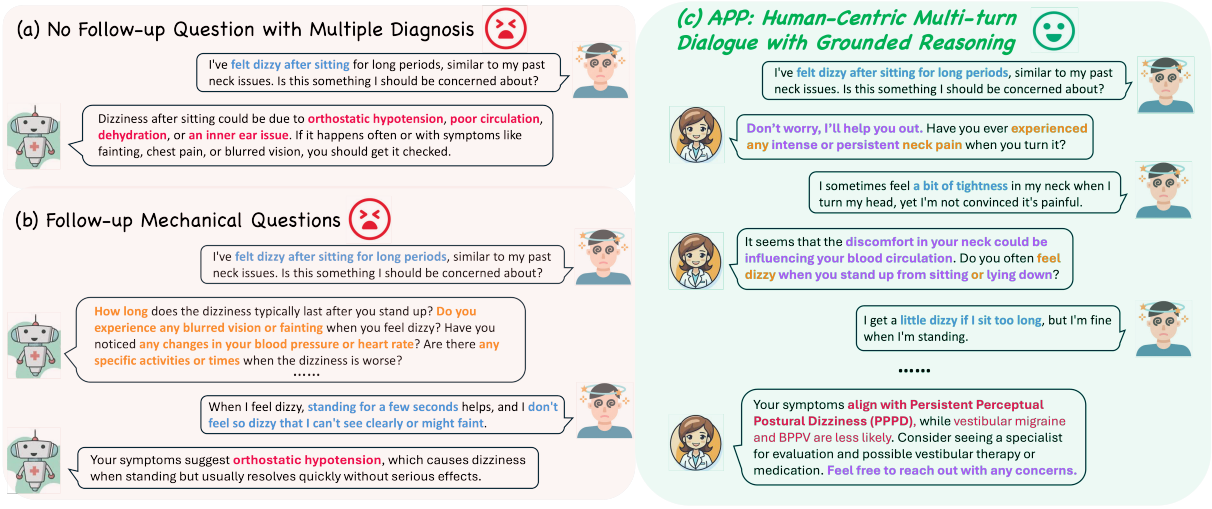

Figure 1: (a) Existing LLMs follow a one-shot diagnostic approach, generating multiple possible diseases without asking follow-up questions. (b) While LLMs can be prompted for multi-turn dialogues, they often overwhelm users with excessive inquiries, potentially disrupting the dialogue and reducing engagement. (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. It also incorporates human-centric features, such as eliciting patients 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.

soning, such as "Did you drink milk last night?", can help patients provide 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), provide inconsistent answers to the same question, use obscure medical terminology without clear sources, and make deterministic medical decisions without grounded reasoning. These issues undermine transparency and trustworthiness, making it difficult for LLMs to deliver reliable diagnoses and gain patient confidence, ultimately limiting their real-world applicability.

For actually applying LLM-simulated medical assistants in the real-world, they must incorporate human-centric features. Using ordinary people-friendly language and guided questioning can intrigue the potential health conditions through asking their personal background, such as daily activities and dietary habits, etc. Anthropomorphic feature, such as empathetic dialogue can enhance user experience by providing comfort and psychological support, ultimately increasing user engagement, which is crucial to achieving trust and a friendly relationship with patients (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, MSD Manual (Manual, a,b), ensuring reliable and evidence-based diagnoses. Second, Dr.APP is built on an analytical mathematical model, specifically Bayesian active learning, to determine the next optimal question each turn. In this way, Dr.APP enhances transparency, minimizes unnecessary user interactions, and maintains high diagnostic accuracy. Finally, Dr.APP facilitates a human-centric dialogue, guiding patients to clearly articulate their symptoms with empathetic communication. Dr.APP is instructed to respond to users with understanding and compassion, treating their concerns as a conversation with a trusted friend. To evaluate our method, we simulate patients based on real-world backgrounds, constructed from over 300 real-world doctor interviews (Yan et al., 2022).

- We introduce Dr.APP, the first **human-centric** LLM-based medical assistant, which can elicit user symptoms through human-like dialogue, significantly improving user accessibility and engagement.
- Dr.APP incorporates Bayesian active learning based on grounded medical guidelines to pro-

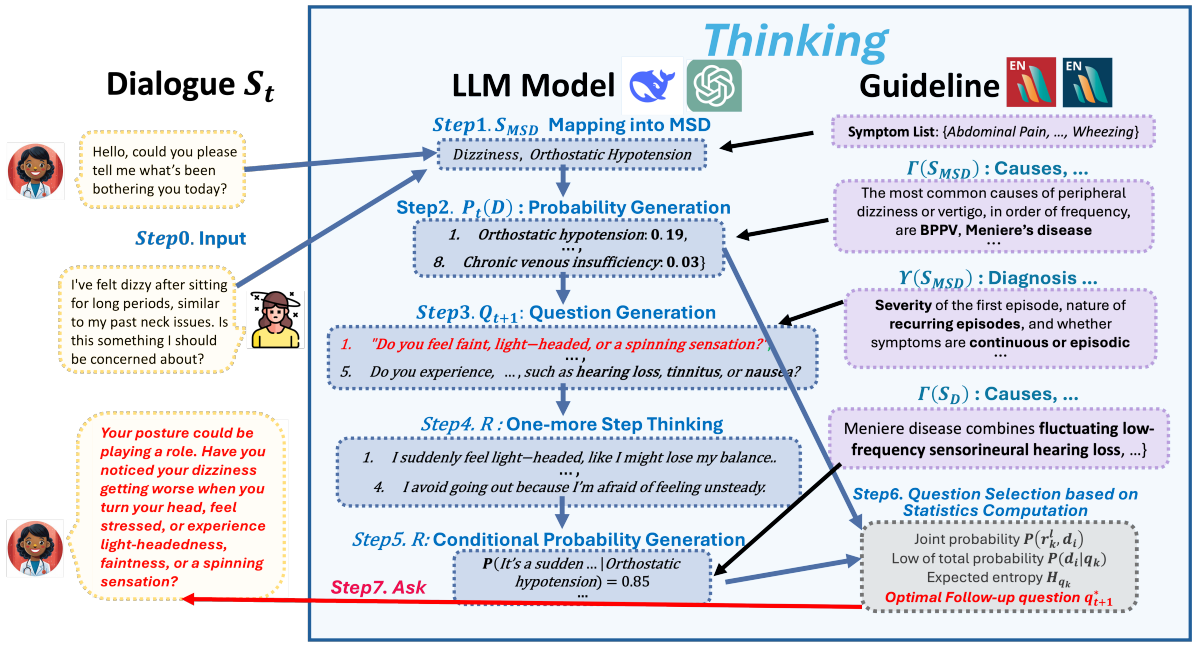

Figure 2: APP Workflow. The system first maps dialogue S_t to S_{MSD} symptoms, then generates disease probabilities and a question pool based on the MSD Manual (Manual, a,b). It then performs an additional reasoning step to simulate possible responses, compute conditional probabilities, and apply Bayesian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

vide **grounded and transparent reasoning** for medical diagnosis.

- We develop a new benchmark that simulates a more realistic clinical consultation environment from real-world interview cases. Dr.APP achieves SOTA diagnostic accuracy with a empathetic user experience, supported by human evaluation.

2 Methodology

2.1 Framework Overview

To ensure grounded reasoning, we use the MSD Manual (Manual, a,b) as a primary source through the whole workflow. Building on this, we incorporate Bayesian active learning to enhance transparency and accuracy. Additionally, we design the doctor agent with a human-centric approach for more effective and engaging interactions.

Our dialogue starts from Dr.APP asking the first question q_1 and the patient responding with r_1, forming the initial conversation $S_1 = (q_1, r_1)$. Then, LLMs extract symotom information from S_1 and map it to pre-defined MSD symotom list. The objective of Dr.APP is to determine the most probable diagnosis $d^* \in D$, through medical dialogue, where $D = \{d_i\}_{i=1}^I$ represents the set of possible diseases. Let S_t denotes the dialogue between the user and Dr.APP after t iterations: $S_t = \{(q_1, r_1), (q_2, r_2), \ldots, (q_t, r_t)\}$. At each iteration t, the probability distribution $P_t(D)$ is updated by using LLMs based on S_t and disease information retrieved from MSD Manual. Based on the symotom, a question pool Q_{t+1} is generated, also guided by MSD Manual. Dr.APP selects the optimal follow-up question q_{t+1}^* based on Bayesian active learning for the next iteration.

To ensure the diagnostic reliability, Dr.APP leverages the MSD Manual (Manual, a,b) as its primary knowledge source, incorporating both professional and consumer versions. The professional version offers structured clinical definitions, diagnostic criteria, and treatment guidelines, ensuring medical precision. Meanwhile, the consumer version presents simplified medical concepts, improving accessibility for general users. By retrieving the information from both, Dr.APP remains grounded in authoritative medical knowledge while ensuring interpretability for non-expert users.

Specifically, Dr.APP involves following steps (Figure 2): Mapping into MSD (Section 2.2); Diagnosis Probability Prediction (Section 2.3); Question

Generation (Section 2.4); One-more Step Thinking 2.5 & Conditional Probability Generation; Question Selection (Section 2.6); and Human-centric Communication Incorporation (Section 2.7).

2.2 Mapping into MSD

In order to map user dialogue to MSD information, we first store pairs wiki-like introduction of each symptom and its name in MSD manual to the local storage. Then we use RAG to retrieve the most relevant MSD symptoms from the user dialogue. Dr.APP ensures a comprehensive representation by mapping S_1 to one symptom in the professional and one in the consumer symptom list of the MSD Manual: $S_{MSD} = \{S_{prof}, S_{cons}\}$.

2.3 Diagnosis Probability Prediction

Given MSD symptiom S_{MSD}, we access the detailed symptom page, which provides information on *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 available dialogue 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 from $\Gamma(S_{MSD})$ and the cumulative dialogue S_t.

2.4 Question Generation

The initial meaningful APP-user dialogue $S_1 = \{q_1, r_1\}$, is often limited and imprecise, as users may use non-standard terminology or provide vague descriptions that do not directly align with clinical definitions. So after generating the disease probability $P_t(D)$, we need a follow-up question to further determine the mostly likely disease. At each iteration t, Dr.APP generates a question pool Q_{t+1} based on guidance from the MSD Manual. These grounded information is retrieved from sections such as *Diagnosis* and *What a doctor does*, from MSD professional and consumer version, respectively. They are represented as $\Upsilon(S_{MSD})$. It ensures that the generated question is both clinically reliable and symptom-specific. The set of candidate diagnostic questions are represented as: $Q_{t+1} = \{q_1, q_2, \ldots, q_K\}$, where K is the maximum number of questions considered per iteration.

2.5 One-more Step Thinking & Conditional Probability Generation

For each candidate question $q_k \in Q_{t+1}$, we would think one more step further to anticipate the patients' response. In this case, we can select the most effective question to ask in the next turn. Specifically, a set of plausible responses are generated by LLM for each candidate question, given the current dialogue S_t. The set of responses for question q_k is denoted as $R_k = \{r_k^1, r_k^2, \ldots, r_k^L\}$, where R_k represents the possible responses for question q_k, and L is the number of generated responses. 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 MSD Manual.

2.6 Question Selection

Then we use Bayesian active learning to select the optimal question from Q_{t+1}. Once the responses for candidate question are generated, Dr.APP calculates the virtual next step disease probability distribution $P(d_i|q_k)$ using Bayesian inference. Using Bayes' 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 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 next iteration, Dr.APP evaluates the expected entropy for 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 information gain:

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

[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 $P_t(D)$ and $P_t(d_i)$, respectively.

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 repeatedly predicting the potential diagnosis probability $P_{t+1}(D)$ and determining the optimal follow-up question, Dr.APP reaches the final diagnosis d^*.

2.7 Human-Centric Communication

To make the diagnostic process more accessible for individuals without medical background, Dr.APP simplifies complex medical terminology and symptom descriptions. When asking each optimal question q^*_t, Dr.APP is instructed to use clear, easy-to-understand language, such as *"Simplify medical terminology and jargon into everyday language."* It ensures effective communication and minimizes misunderstandings.

Individuals may not always recognize or articulate abnormal behaviors or symptoms from a clinical perspective. To address this, Dr.APP guides users with contextual hints that help them to recall relevant information they might otherwise overlook. Dr.APP is explicitly prompted with "the question should be answerable with a simple yes/no or a straightforward multiple choice response." For example, instead of asking a broad question like *"Have you eaten anything unusual?"*, the system offers specific cues such as *"Have you consumed foods like milk or beverages like soda (e.g., Coke)?"* This approach helps users to recall information that could be otherwise overlooked.

Even with simplified yes/no questions, users may struggle with medical terminology or subtle differences in symptom descriptions. To mitigate this, Dr.APP formulates "specific, descriptive questions with explanations or examples". For instance, rather than asking *"Do you feel dizzy?"*, Dr.APP refines the inquiry to: *"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 ensures users can accurately identify and describe their symptoms, leading to more precise and efficient communication.

To reduce patient anxiety and encourage engagement, Dr.APP is designed to exhibit anthropomorphic qualities by responding with reassurance and empathy, helping to reduce anxiety and provide comfort. Prompts such as "Use a warm and empathetic tone to ensure the patient feels comfortable" is provided to Dr.APP. For example, it may explicitly say, *"Let's not worry excessively for the moment,"* or implicitly convey understanding with phrases like, *"I understand your concerns."* These responses help create a more supportive and calming interaction.

3 Experiment

3.1 Dataset

To evaluate the performance of our proposed approach, APP, we use a subset of the ReMeDi (Yan et al., 2022) dataset, which consists of real-world multi-turn conversations between doctors and patients. This ensures that the dialogues reflect realistic, natural interactions, capturing the inherent variability and complexity of user-provided information. We use ReMeDi-base, which originally contained 1,557 labeled dialogues, as the foundation of our dataset. In this dataset, doctors' responses are annotated with seven different action labels: "Informing", "Inquiring", "Chitchat", "QA", "Recommendation", "Diagnosis", and "Others". For our study, we extracted dialogues that exclusively contain the "Diagnosis" label, resulting in 329 real-world, multi-turn diagnostic conversations between doctors and patients. In this paper, we randomly selected 70 dialogues, covering 58 distinct diseases across 15 specialties, such as Orthopedics (e.g. osteoarthritis), Gynecology (e.g. polycystic ovary syndrome), Dermatology (e.g. androgenetic alopecia).

3.2 Experimental Setup

To evaluate Dr.APP's performance, we simulate patients using the real-world dataset mentioned above. DeepSeek-v3 first summarizes the patient's condition, background, personality, and intent etc. based on real-world data, then role-plays as the patient in the dialogue. To further mimic real-world patient interactions, patient agents are prompted with instructions such as: *"Reasonably incorporate daily life details that align with the patient's personality and background."*

In our experimental setup, we set $K = 5$, meaning a maximum of five candidate questions are generated at each iteration. For each question, at least two and at most of 5 candidate responses($2 \leq L \leq 5$) are generated.

3.3 Evaluation Matrix

3.3.1 Accuracy

We first use OpenAIEmbeddings (OpenAI) to generate numerical embeddings of the predicted and

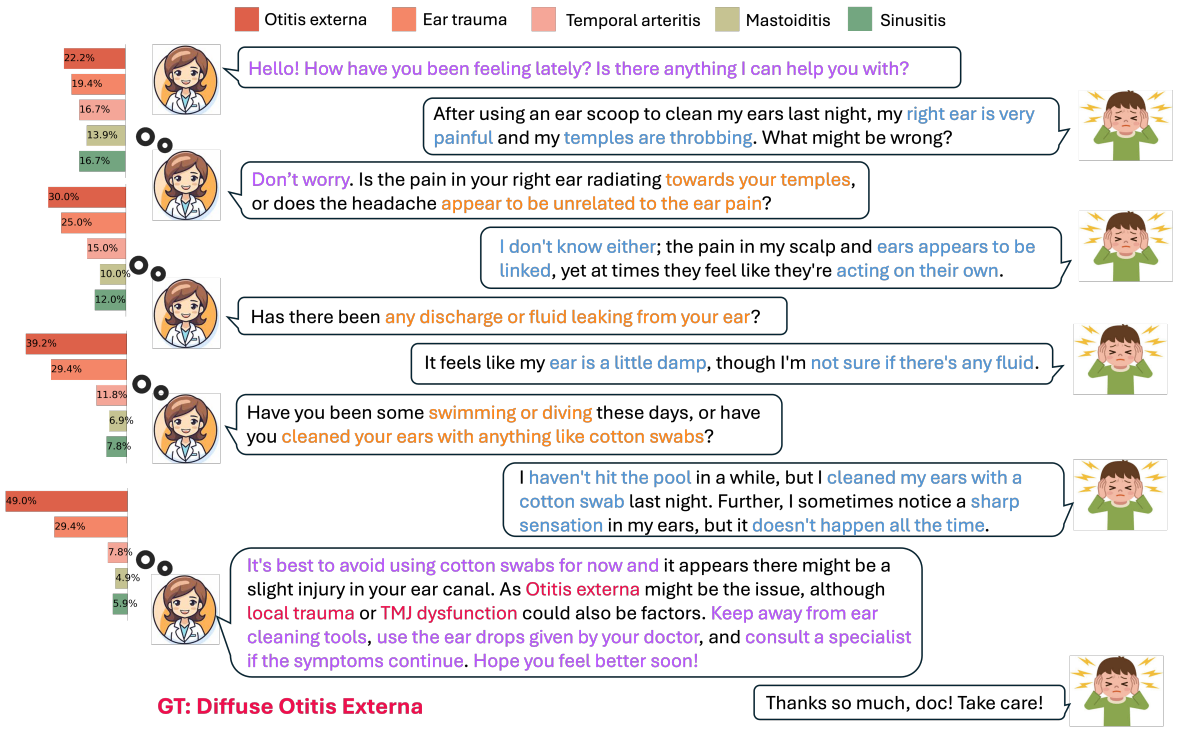

Figure 3: An APP case study of human-centric multi-turn dialogue based on medical guidelines. The estimated disease distribution is updated with the progression of the conversation. Disease items from top to bottom: Otitis externa, ear trauma, TMJ dysfunction and the others. The ground truth is Diffuse Otitis Externa, where our diagnosis is Otitis Externa. Blue represents user-described symptoms, Orange indicates questions raised by APP, Red highlights the diagnosis, and Purple shows human-centric features.

ground truth diagnosis, capturing the semantic relationships between them. Cosine similarity is then computed between each pair of the prediction and ground truth to measure their alignment. Diagnosis accuracy is determined based on this similarity score. A prediction is considered correct if its similarity score exceeds a predefined threshold Θ. To ensure a robust evaluation, we define a threshold range $\Theta \in \{0.5, 0.51, ..., 1.0\}$ with an interval of 0.01. The final accuracy is computed as the average accuracy across all threshold values: $acc = \frac{1}{|\Theta|} \sum_{\theta \in \Theta} acc(\theta)$ where $acc(\theta)$ represents the proportion of correct predictions at threshold θ.

3.3.2 Entropy

Given the current probability distribution of potential diseases $P_t(D)$, we aim for the system to increase confidence in certain diagnoses and rule out less likely 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.

3.3.3 Human-Centric

Accessibility Score To assess whether the questions posed by Dr.APP are easy for users without medical background to understand, we evaluate the language accessibility using GPT-4o. The model rate the clarity and simplicity of the doctor's language on a scale from 0 to 1.

Empathy Score This score reflects the level of empathy demonstrated by the Dr.APP during the conversation with the user. The degree of empathy is rated on a scale from 0 to 1 using GPT-4o, with higher values indicating more empathetic communication.

Relevant Response Rate In some cases, users may ask the doctor follow-up questions. Ideally, the doctor should address these concerns before proceeding with the next question. This metric evaluates whether the doctor's response directly answers the user's question, with GPT-4o assigning a score

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

Table 1: **Diagnosis Accuracy (%) Comparison with SOTA Methods**: APP-DeepSeek-v3 achieves the highest overall accuracy in both one-round and multi-round evaluations, demonstrating the effectiveness of multi-turn interactions driven by statistical modeling and grounded medical guidelines.

Model	One Round				Multiple Rounds			
	Cardiology	Allergy	General med.	Overall	Cardiology	Allergy	General Med.	Overall
QWen-72B	66.92	83.07	63.46	70.93	68.46	**83.07**	64.42	70.43
Claude-3	62.31	72.31	58.65	67.16	60.00	76.92	59.61	67.27
GPT-4o	70.00	83.07	**66.34**	70.10	68.46	**83.07**	67.30	70.66
LLaMA-70B	**69.23**	74.61	61.53	71.59	70.00	78.46	60.58	71.42
APP-LLaMA-70B	66.15	76.92	58.65	67.96	64.61	**83.07**	60.58	69.67
DeepSeek-v3	60.00	78.46	58.65	67.96	66.15	80.00	62.50	67.91
APP-DeepSeek-v3	64.62	**83.85**	**66.34**	**72.14**	**75.38**	82.31	**70.19**	**73.02**

of 0 or 1.

Reliability To assess whether APP's predicted disease aligns with the medical information extracted from the MSD Manual, we conduct a human evaluation of reliability. Reliability are rated on the rating scale from 1 (completely incorrect) to 5 (fully accurate). This assessment ensures that APP's predictions are clinically grounded, trustworthy, and aligned with established medical knowledge.

3.4 Accuracy Analysis versus Baselines

To evaluate the diagnostic accuracy of Dr.APP, we compared it against SOTA LLMs across multiple medical domains, including cardiology, allergy, and general medicine (Table 1). The 'Overall' represents the performance across all 15 specialties. We assess performance in both single-round and multi-round diagnostic settings.

In the one-round evaluation, models were given only the initial user input and required to generate a diagnosis without follow-up interaction. Our APP-DeepSeek-v3 achieved the highest overall accuracy (72.14%), outperforming other models, including GPT-4o (70.10%) and LLaMA-70B (71.59%). In the multi-round evaluation, where models engaged in iterative questioning to refine their diagnoses, APP-DeepSeek-v3 again outperformed other methods, reaching an overall accuracy of 73.02%, with particularly strong results in cardiology (75.38%) and general medicine (70.19%).

Table 2 demonstrates that multi-turn interactions significantly enhance diagnostic accuracy, with APP-DeepSeek-v3 consistently outperforming baseline models across all iterations. Its adaptive questioning strategy, based on a statistical model and verified medical sources, enables more effective diagnosis refinement, achieving the highest overall accuracy of 72.37%.

Table 2: **Diagnosis Accuracy (%) Comparison across Iterations.** APP-DeepSeek-v3 consistently outperforms baseline models across all iterations

Methods	Per-Iteration Results					Overall
	1	2	3	4	5	
Calude-3	67.16	66.94	67.61	68.72	67.27	67.53
GPT-4o	70.10	71.25	70.16	70.21	70.65	70.47
DeepSeek-v3	67.96	67.36	67.85	67.74	67.91	67.76
APP-DeepSeek-v3	**72.14**	**71.98**	**72.03**	**72.69**	**73.02**	**72.37**

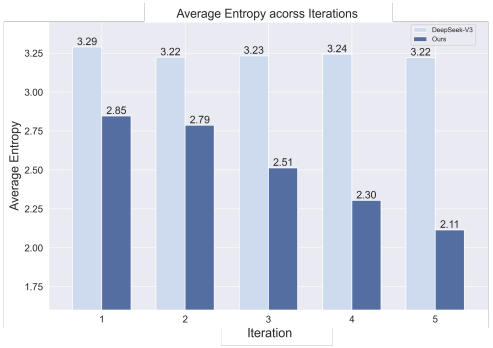

Figure 4: **Entropy Comparison across Iterations.** APP consistently shows a sharper decrease in entropy (the lower, the better), indicating increased diagnostic confidence and reduced uncertainty through iterative dialogues.

3.5 Confidence Analysis across Iterations

Figure 4 illustrates the evolution of diagnostic confidence across iterations by comparing the entropy values of APP-DeepSeek-v3 and the DeepSeek-v3 baseline. In the initial iteration, Dr.APP exhibits lower diagnostic uncertainty, with an entropy of 2.85, compared to 3.29 for DeepSeek-v3. This suggests that APP generates more confident predictions even before follow-up interactions, likely due to its reliance on verified medical sources for initial reasoning.

As iterations progress, Dr.APP shows a sharper and more consistent decline in entropy, refining

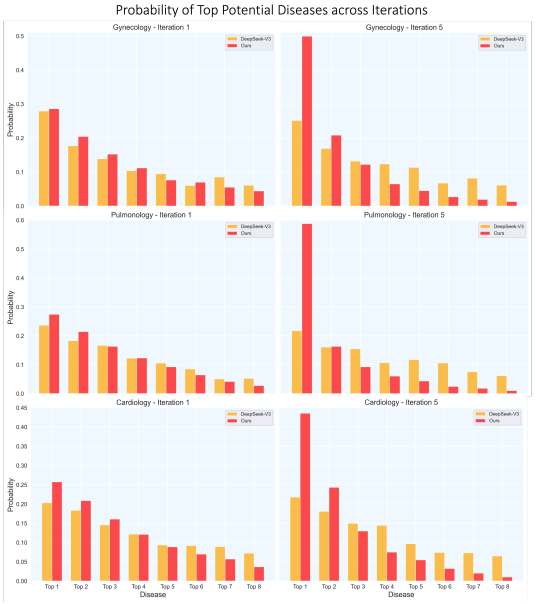

Figure 5: **Confidence Analysis across Iterations.** APP-DeepSeek-v3 shows increased confidence in the top predicted disease while reducing confidence in less likely conditions over multiple iterations, demonstrating improved diagnostic confidence with interpretability.

diagnoses more effectively. After six iterations, Dr.APP reaches it entropy to 1.95, indicating high certainty, whereas DeepSeek-v3 retains 3.18, suggesting persistent uncertainty. This reduction highlights Dr.APP's superiority in managing diagnostic uncertainty and improving prediction confidence.

Figure 5 further illustrates how the distribution of top potential diseases evolves across iterations for different specialties, including gynecology, pulmonology, and cardiology. The results indicate that APP consistently assigns higher confidence to the most probable disease while reducing confidence in less likely conditions, leading to a clearer separation in probability rankings. This 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 over iterations, Dr.APP improves model transparency and diagnostic certainty. The increase in confidence reduces ambiguity, leading to more reliable and trustworthy medical guidance. These enhancements ultimately foster greater user trust in AI-assisted diagnosis while improving clinical reliability and usability.

3.6 Human-Centric Analysis with Real-world Dialogue

Our human-centric system, Dr.APP, shows notable performance in user accessibility, question empathy and relevance compared to the original dialogues collected from real-world online platform (ReMeDi-base). In terms of accessibility, Dr.APP achieved an average score of 0.91, outperforming the original dialogues, which scored 0.85. This highlights the system's ability to present medical information in a way that is easier for users to understand. For empathy, Dr.APP scored 0.66, compared to 0.50 in the original dialogues. This indicates that our system encourages more compassionate and human-centric dialogues, helping to reduce user anxiety and create a better overall experience. Regarding relevance, Dr.APP maintained a high score of 0.79, closely aligning with the original dialogues' score of 0.82. Additionally, we invited four medical professionals with over five years of graduate-level expertise to evaluate whether Dr.APP's diagnoses align with grounded medical resources. The assessment yielded an average reliability score of 4.5/5, further confirming Dr.APP's enhanced diagnostic reliability. Overall, these results demonstrate that Dr.APP enhances human-friendly communication, leading to better user understanding and engagement.

4 Conclusion

In this study, we introduce Dr.APP, the first human-centric LLM-based medical assistant built upon grounded medical resources and Bayesian active learning. Dr.APP enhances diagnostic accuracy and reliability by integrating verified medical guidelines and leveraging Bayesian active learning to optimize follow-up questioning. Through entropy minimization, Dr.APP effectively refines diagnoses and improves efficiency via iterative user interactions. Our experiments demonstrate that Dr.APP significantly enhances both diagnostic accuracy and efficiency compared to one-shot and current multi-turn LLM baselines. Entropy analysis confirms that Dr.APP rapidly reduces diagnostic uncertainty over successive iterations, leading to greater confidence in its predictions. Additionally, Dr.APP prioritizes user accessibility and empathetic dialogue, eliciting users to express medically relevant information more effectively. By bridging the gap between clinical expertise and patient communication, Dr.APP fosters greater user engagement and trust.

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 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 effectively integrates medical guidelines to improve diagnostic accuracy, its reliability is still constrained by the quality and coverage of these guidelines. The MSD Manual provides grounded medical knowledge, but there are many additional real-world medical sources. Expanding the system to integrate additional medical knowledge bases could enhance its clinical applicability.

Third, APP's statistical framework optimizes follow-up question selection, but it assumes an idealized patient interaction where users provide consistent and accurate responses. In reality, patients may misinterpret questions, provide inaccurate answers, or experience cognitive biases that affect their descriptions. Further human-in-the-loop refinements and adaptive questioning strategies are needed to account for user variability and uncertainty.

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

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*.

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

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.

MSD Manual. a. MSD Manual Consumer Version.

MSD Manual. b. MSD Manual Professional Edition.

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*.

OpenAI. OpenAI Platform.

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.

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.

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*.

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.

WHO. Medical doctors (number).

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.

A Appendix

A.1 Survey Question

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. **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?
5. **Interpretability Score (Int.)**
    - The model provides disease probabilities at each stage of the diagnosis. How clear and helpful is this information in understanding the reasoning behind the diagnosis?
    - **1:** Completely unclear - probabilities are confusing or not useful. **2:** Mostly unclear - difficult to interpret without additional explanation. **3:** Somewhat clear - but could be more intuitive. **4:** Mostly clear - probabilities help in understanding the diagnosis. **5:** Completely clear - easy to interpret and useful for assessing the diagnosis.
    - **Optional Comment**: Does the probability information improve your understanding of the diagnosis? If not, what could be improved?