Extent-Med: A Dynamic Decision-Making Approach under Incomplete Information for Community Healthcare

Anonymous ACL submission

Abstract

Large language model (LLM)-based agents have shown promise in supporting healthcare consultations, yet their effectiveness remains limited in community healthcare contexts where patients often provide incomplete, ambiguous, or non-technical descriptions of symptoms. These conditions introduce significant incomplete information and uncertainty during multi-turn consultations. To address these challenges, we introduce Extent-Med, an uncertainty-guided dynamic decision-making framework designed for multi-turn medical consultation under incomplete information in community healthcare. We evaluate Extent-Med on both public medical dialogue datasets and real-world scenarios. Experimental results demonstrate that our system significantly outperforms baseline LLM-based agents and knowledge-augmented variants. This research contributes to the emerging literature on LLM-based agents in healthcare by proposing a novel approach for uncertainty management in conversational healthcare settings. The findings provide practical insights for developing reliable AI systems for medical consultation.

1 Introduction

The emergence of Large Language Models (LLMs) has substantially transformed the design and deployment of autonomous conversational agents in professional domains (Wang et al., 2025; Han et al., 2023). In particular, LLM-based agents are increasingly viewed as promising tools for community healthcare, a setting characterized by rapidly growing service demand, limited medical resources, and substantial variation in physicians' expertise and experience (Abada and Altalafha, 2023). By enabling increasingly sophisticated decision-support capabilities, LLM-based systems have evolved from simple information retrieval tools into conversational agents capable of assisting complex professional tasks.

Despite these advances, the application of LLM-based agents to medical consultation in community healthcare faces significant challenges. First, many community healthcare consultations involve patients providing incomplete, imprecise, or non-technical descriptions of symptoms, while comprehensive diagnostic tests or medical histories are often unavailable at the time of consultation (Shortliffe and Sepúlveda, 2018). Such input introduces substantial uncertainty and information incompleteness into community healthcare decision-making. Traditional chatbots are primarily reactive question-answering systems that do not proactively guide information acquisition (Li et al., 2023; Wang et al., 2025). These systems are limited in their capacity to support accurate, efficient, and decision-relevant clinical reasoning under incomplete information.

Second, even when LLMs are augmented with reasoning capabilities or external knowledge sources, they remain prone to generating fluent yet factually unsupported outputs–a phenomenon often referred to as hallucination (Pal et al., 2023; Ji et al., 2023a). These outputs can include clinically incorrect inferences or fabricated medical statements, undermining patient safety and the interpretability of AI-assisted decisions. Existing mitigation strategies, such as prompt engineering (Ji et al., 2023b), retrieval-augmented generation (RAG) (Li et al., 2024; Lewis et al., 2020), or output guardrails (Dong et al., 2024), show inconsistent effectiveness, particularly in multi-turn, complex diagnostic dialogues.

Third, across multi-turn interactions, dialogue trajectories may progressively accumulate irrelevant or low-value content, leading to diagnostic trajectory drift (Guan et al., 2025). This deviation undermines the systematic acquisition of high-value information, disrupts task alignment, and reduces the reliability of sequential decision-making. Drift can manifest as task drift (Abdelnabi et al., 2025), identity drift (Choi et al., 2024), or knowledge drift

(Fastowski and Kasneci, 2024), highlighting the need for mechanisms that dynamically guide the consultation toward the most informative and clinically relevant information.

These challenges remain unsolved and raise an important research question: **how to design AI systems that can dynamically acquire informative clinical signals, manage uncertainty across multi-turn interactions, and support reliable decision-making under incomplete information in community healthcare settings.** Addressing this question requires AI agents capable of strategic information acquisition, uncertainty reduction, and interpretable decision support throughout the consultation process.

In this study, we introduce Extent-Med, an uncertainty-guided dynamic decision-making framework designed for multi-turn medical consultation under incomplete information in community healthcare. Unlike conventional LLM-based chatbots that passively respond to user queries or rely on heuristic prompting strategies, Extent-Med treats medical consultation as a sequential information acquisition problem. The system explicitly models diagnostic uncertainty and dynamically selects the next conversational action that maximally reduces this uncertainty. Extent-Med consists of two tightly integrated modules: a Symptom Acquisition Engine (SAE), and a Structured Clinical Reporter (SCR). Central to the framework is a Diagnostic Uncertainty Reduction (DUR) strategy that evaluates candidate symptoms according to their expected contribution to reducing diagnostic uncertainty. This strategy effectively converts open-ended conversational space into a structured decision process, enabling the agent to prioritize high-value clinical signals rather than generating ad-hoc follow-up questions.

Beyond strategic information acquisition, Extent-Med addresses three limitations of LLM-based medical consultation systems aforementioned. First, to handle incomplete and ambiguous patient input, the framework translates unstructured patient narratives into structured clinical signals through knowledge-grounded symptom extraction. Second, to mitigate stochastic hallucination, each conversational step is anchored in a symptom-disease knowledge graph that constrains the agent's reasoning space and ensures that generated inquiries remain clinically grounded. Third, to prevent diagnostic trajectory drift, Extent-Med introduces a sequential uncertainty-reduction objective that continuously evaluates the informational value of candidate questions, guiding the dialogue toward diagnostically meaningful trajectories rather than allowing topic diffusion over multiple turns.

We evaluate Extent-Med using both public medical dialogue datasets and real-world clinical consultation scenarios. Across multiple backbone LLMs, the proposed framework consistently outperforms baseline models and knowledge-augmented variants in Top-k diagnostic accuracy, triage performance, and consultation efficiency. Importantly, the results demonstrate that explicitly modeling uncertainty and information value in conversational decision-making significantly improves the reliability of LLM-based medical agents. These findings highlight the potential of uncertainty-guided dialogue control as a generalizable design paradigm for building interpretable, robust, and decision-oriented LLM agents in high-stakes professional environments.

This study makes several contributions. First, we contribute to the emerging literature on LLM-based agents in healthcare by reconceptualizing multi-turn AI consultation as a sequential information acquisition problem under incomplete information. While information-theoretic approaches such as decision trees based on information gain have a long history in machine learning (Quinlan, 1986) they have largely focused on improving response generation through prompt engineering, retrieval augmentation, or knowledge injection. They treat conversational interaction primarily as a response-generation task. In contrast, we conceptualize AI-assisted consultation as a dynamic decision process in which the agent must strategically acquire information over time to reduce uncertainty and improve decision quality. By framing multi-turn dialogue as an information decision problem, our work highlights the value of designing LLM-based agents that actively guide interactions rather than passively respond to user inputs.

Second, we introduce an diagnostic uncertainty reduction strategy that operationalizes uncertainty reduction in LLM-based interactions. Specifically, our proposed framework, Extent-Med, integrates a diagnostic uncertainty reduction strategy to dynamically identify the most diagnostically informative questions during consultation. This strategy transforms open-ended conversational interactions into a structured decision process that prioritizes high-value information acquisition. By explicitly modeling diagnostic uncertainty and optimizing

question selection, our approach addresses two persistent challenges in LLM-based systems, i.e., hallucination and diagnostic drift, through principled dialogue guidance rather than post-hoc output filtering.

2 Literature Review

2.1 LLM-based Agents in Healthcare

LLM-based agents are increasingly adopted in healthcare, including clinical decision support, medical documentation, training simulations, and service optimization (Wang et al., 2025). Representative systems include ChatDoctor (Li et al., 2023), MedAlpaca (Han et al., 2023), and Elixr (Xu et al., 2023). Despite their growing popularity, real-world deployment remains challenging. LLM responses may be stochastic and affected by hallucination (Pal et al., 2023). In multi-turn diagnostic dialogues, drift can occur where the agent's reasoning deviates from the diagnostic objective (Guan et al., 2025). Overall, ensuring factual accuracy, process stability in LLM-based agent design remains a key challenge in healthcare.

2.2 Hallucination and Drift in LLM-based Agents

Alongside the advancement of LLM-based agents, attention to hallucination has also increased. Hallucination refers to factually incorrect or unsupported content generated by LLMs (Ji et al., 2023a). Previous research, like (Bang et al., 2025), divided it into intrinsic hallucination and extrinsic hallucination. In healthcare, hallucination affects documentation generation (Vishwanath et al., 2024). Therefore, healthcare applications require LLM agents that are trustworthy, transparent, and explainable (Ahmad et al., 2023). In this context, hallucination is not merely a generation error but safety issue.

Besides hallucinations, drift also affects agent performance. We consider drift as the deviation from the intended context or task objective, leading to uncertain outcomes. Recent LLM research further distinguishes task drift (execute user instructions) (Abdelnabi et al., 2025), identity drift (interaction patterns or styles change) (Choi et al., 2024), and knowledge drift (susceptibility to misinformation) (Fastowski and Kasneci, 2024). In agent practice, multi-turn interactions may result in contextual deviations over time (Dongre et al., 2025). Besides, Ambiguous or misleading user inputs can further induce behavioral shifts (Bao et al., 2026). In medical AI, such drifts are typically observed as system degradation (Guan et al., 2025).

2.3 Evaluation via LLM-as-a-Judge

LLM-as-a-Judge, where LLM serves as evaluators for complex tasks, is increasingly adopted. This paradigm is supported by multiple studies. Surveys (Gu et al., 2024; Li et al., 2025) highlight its broad applicability and also demonstrate some strategies for constructing reliable LLM judges; Empirical benchmarks (Zheng et al., 2023) show strong agreement with human preferences, achieving over 80%. In healthcare, LLM-as-a-Judge has been successfully applied to evaluating real-world Electronic Health Record (EHR) multi-document summaries, demonstrating strong reliability compared with human evaluators (Croxford et al., 2025); It has also been examined in a clinical trial design (Curran et al., 2024). Nevertheless, limitations remain in some domain-specific or highly specialized tasks, indicating that human expert assessments remain important (Szymanski et al., 2025).

3 Methodology

3.1 Problem Definition and Design Framework

We study LLM-based medical agents that interact with patients over multiple turns, gather clinical evidence, and produce structured medical reports. In a typical medical consultation scenario, the system is expected to (1) interactively inquire informative symptoms, (2) provide diagnostic results, and (3) produce clinically meaningful summaries for clinicians.

Formally, let $\{D_1, \cdots, D_n\}$ denote the disease space and $\{S_1, \cdots, S_{n_S}\}$ denote the symptom space derived from the disease set. At each dialogue turn t, the agent generate question Q and acquire the patient's response A. Then it obtain an observation set of symptom in turn t: $\mathcal{O}^{(t)}$. Based on accumulated observations, the agent maintains a belief distribution over candidate diseases $\mathcal{D} = \{\mathbb{P}(D_i) : i = 1, \cdots, n\}$ and decides which symptom S^* to inquire next, with the objective of efficiently reducing diagnostic uncertainty and converging to the correct disease.

Table 1 summarizes the design framework that considers key task complexities in community healthcare consultations and design requirements for the proposed AI consultation agent.

Task Complexity	Design Requirement
Ambiguous and Partial Patient Input. Patients often describe symptoms using incomplete, imprecise, or non-technical language, creating uncertainty in symptom identification and prioritization.	Structure and disambiguate patient narratives to convert unstructured descriptions into diagnostically meaningful symptom representations.
Epistemic Uncertainty in Model Output. LLM agents may produce fluent but unsupported medical statements due to reliance on linguistic plausibility rather than verified knowledge.	Ensure evidence-grounded reasoning by anchoring model outputs in external clinical knowledge.
Sequential Task Deviation. In multi-turn consultations, the agent may prioritize low-value information, causing diagnostic trajectory drift.	Prioritize high-information symptoms to sequentially reduce diagnostic uncertainty and maintain task alignment.

Table 1: Design framework. The framework identifies three core challenges: ambiguous patient input, epistemic uncertainty in model outputs, and sequential task deviation during multi-turn interactions. For each challenge, we derive a corresponding design requirement that guides the development of mechanisms to structure patient narratives, ensure evidence-grounded reasoning, and maintain diagnostically efficient information acquisition throughout the consultation process.

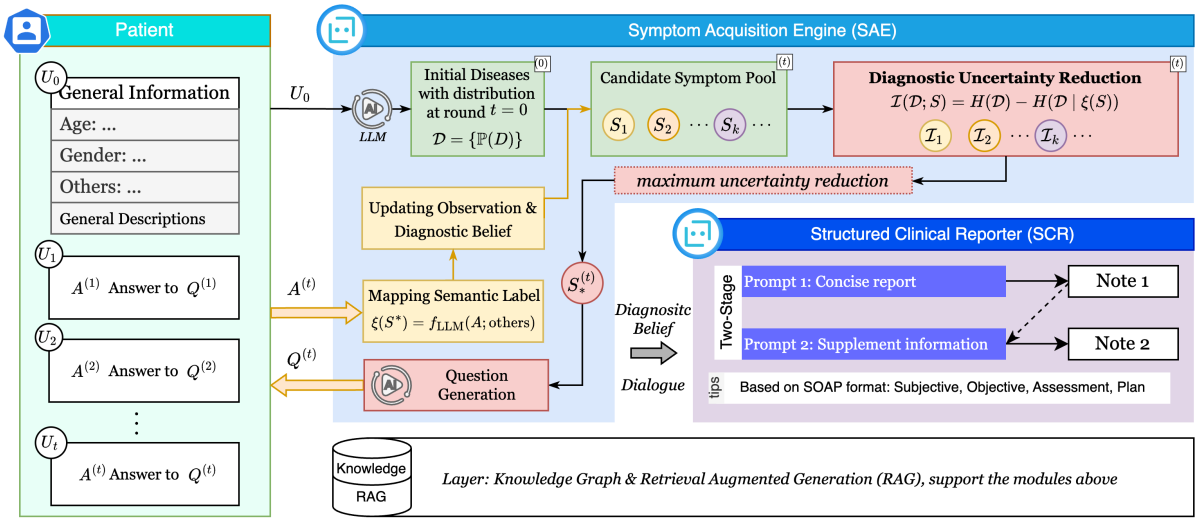

Figure 1: Extent-Med framework. It consists of two main components: (1) Symptom Acquisition Engine (SAE) integrates RAG, a medical knowledge graph, and the Diagnostic Uncertainty Reduction (DUR) strategy to actively identify and elicit the most diagnostically informative symptoms, and (2) Structured Clinical Reporter (SCR) produces professional, structured SOAP reports that support physician decision-making and maintain comprehensive clinical records.

3.2 Extent-Med

We propose an uncertainty-guided LLM-based agent, Extent-Med, as illustrated in Figure 1.

3.3 Symptom Acquisition Engine (SAE)

The Symptom Acquisition Engine (SAE) is responsible for guiding patients to answer required questions and collecting information from them. SAE communicates with patients, collects their descriptions, and retrieves relevant information from the Knowledge Graph. Based on LLMs with RAG, SAE generates questions using guidance information, including probable diseases and symptoms.

The core of SAE is the Diagnostic Uncertainty Reduction (DUR) strategy, which aims to identify the most important symptom in the current situation. By selecting the symptom that can best reduce diagnostic uncertainty, SAE guides the subsequent question generation and improves the efficiency of information acquisition.

3.3.1 Knowledge Graph Construction

As for Knowledge Graph (KG), we construct four domain-specific datasets: (1) a disease-symptom relation dataset, (2) a disease prior probability dataset, (3) a symptom prior score dataset, and (4) a supplement knowledge dataset. It is constructed from authoritative medical resources, including MSD Manual, UMLS (Bodenreider, 2004), SNOMED-CT (Bos and Donnelly, 2006), public health statistics from World Health Organization (WHO) and Centers for Disease Control and Prevention (CDC), and additional clinical references and medical textbooks. The resulting graph covers 8,807 diseases

and associated clinical symptoms. For each disease D, we define the disease-symptom relation as:

$$\mathbb{I}(D, S) = \begin{cases} 1, & \text{if } S \text{ is associated with } D, \\ 0, & \text{otherwise.} \end{cases}$$

The graph also incorporates disease prior probability $\mathbb{P}(D)$ from public health statistics. Based on disease prevalence and disease-symptom associations, we define a prevalence-weighted symptom prior score for 5,885 unique symptoms. Using Bayes theorem and the above relations, the prior score of symptom S is

$$\begin{aligned} \text{Score}(S) &\propto \sum\nolimits_D \mathbb{P}(D) \cdot \mathbb{P}(S \mid D) \\ &= \sum\nolimits_D \mathbb{P}(D) \cdot \mathbb{E}[\mathbb{I}(D, S)] \\ &\approx \sum\nolimits_D \mathbb{P}(D) \cdot \mathbb{I}(D, S) \end{aligned}$$

where $\text{Score}(S)$ is the prior score rather than a normalized marginal probability, and is proportional to $\mathbb{P}(S)$ under the same modeling assumptions (Appendix A).

We further construct a medical supplement knowledge dataset that provides clinical reference information of each disease, like etiological factors, disease categories, relevant departments, recommended treatments, and commonly prescribed medications. It is worth noting that the proposed framework does not strongly depend on prior distributions. In practice, the framework remains robust under weakly informative or uniform priors.

3.3.2 Diagnostic Uncertainty Reduction Strategy

By means of carefully designed prompts, the LLM is instructed to assume a professional physician and to produce a structured list of potential diseases according to the patient's first descriptions. Formally, this process can be expressed as

$$\mathcal{D} = \{D_1, D_2, \cdots, D_n\} = f_{\text{LMM}}(\text{patient input})$$

where the LLM is further augmented with domain-specific medical knowledge through RAG. Searching the disease-symptom relation in KG, we get the symptom set of disease D_i. Combined the symptoms of $D_1, \cdots, D_n$, we form the symptom set: $\{S_1, \cdots, S_{n_S}\}$, where n_S representing the number of distinct symptoms. From the KG, we establish the disease distribution $\mathcal{D} = \{\mathbb{P}(D_i) : i = 1, \cdots, n\}$, and the symptom distribution $\mathcal{S} = \{\mathbb{P}(S_k) : k = 1, \cdots, n_S\}$. Based on the disease distribution, the initial diagnostic uncertainty is computed by entropy as

$$H(\mathcal{D}) = -\sum\nolimits_{i=1}^{n} \mathbb{P}(D_i) \log \mathbb{P}(D_i)$$

For each symptom S, we define a random variable $\xi(S)$ to represent the semantic label derived from the patient response to the symptom S. For instance, replies like "yes", "that's right", etc, denotes presence. While responses such as "I do not know" are excluded, since they do not provide evidence for updating disease probability, i.e., $H(\mathcal{D} \mid \xi \in \{\text{uncertainty reply}\}) = H(\mathcal{D})$. While the responses are discretized, the underlying symptom-disease relations are not treated as uniform or purely 0/1. With the KG, an affirmative or negative answer does not update all diseases equally; instead, it changes the posterior probabilities in terms of these prior scores, thereby capturing richer information even from seemingly simple answers. After receiving the response for S, this symptom is removed from the set $\mathcal{S}$.

To quantify how informative a symptom S is for disease inference, we define the diagnostic uncertainty reduction as

$$\mathcal{I}\big(\mathcal{D}; \xi(S)\big) = H(\mathcal{D}) - H(\mathcal{D} \mid \xi(S))$$

Since the patient response $\xi(S)$ is discrete, seen as a conditional distribution $\xi(S) \sim \mathbb{P}(\xi \mid S)$, we estimate the expected diagnostic uncertainty reduction over possible responses

$$\begin{aligned} &\mathbb{E}_\xi\Big[\mathcal{I}\big(\mathcal{D}; \xi(S)\big)\Big] = H(\mathcal{D}) - \mathbb{E}_\xi\big[H(\mathcal{D} \mid \xi(S))\big] \\ &= H(\mathcal{D}) - \sum\nolimits_\xi H(\mathcal{D} \mid \xi(S)) \cdot \mathbb{P}(\xi \mid S) \end{aligned}$$

However, accurately estimating $P(\xi \mid S)$, i.e., the probability that a patient reports the symptom, is challenging in practice due to individual variability. Therefore, we evaluate the infimum of the diagnostic uncertainty reduction:

$$\begin{aligned} &\mathbb{E}_\xi\Big[\mathcal{I}\big(\mathcal{D}; \xi(S)\big)\Big] \\ &\geq H(\mathcal{D}) - \max\nolimits_\xi \Big\{H(\mathcal{D} \mid \xi(S))\Big\} \\ &:= \inf\nolimits_\xi \mathbb{E}_\xi\Big[\mathcal{I}\big(\mathcal{D}; \xi(S)\big)\Big] \end{aligned}$$

This worst-case strategy selects the symptom with the largest guaranteed uncertainty reduction under the least favorable informative response. The conditional uncertainty $H(\mathcal{D} \mid \xi(S))$ and relevant distribution, like $\mathbb{P}(D \mid S)$, can be estimated from the

Algorithm 1: Procedure of SAE Module

Input : LLM, Knowledge Graph (KG), patient input x, max turns T, threshold ϵ
Initialize diseases $\mathcal{D} \leftarrow \mathrm{LLM}(x)$, symptom $\mathcal{S} \leftarrow \mathrm{KG}(\mathcal{D})$, history observation $\mathcal{O} \leftarrow \phi$
while *not StopCriterion*$(\mathcal{D}, \mathcal{S}, T, \epsilon)$ **do**
  Select most informative symptom S^* [Eq. 1]
  Question generate $Q \leftarrow \mathrm{LLM}(S^*; \text{history})$ with KG and RAG; Get patient's answer A
  Add $\mathrm{LLM}(A)$ to $\mathcal{O}$ and $\mathcal{D} \leftarrow \mathrm{Update}(\mathcal{D}; \mathcal{O})$
return $\mathcal{D}$ and other information

KG using Bayes' rule based on disease-symptom relations and prior probabilities. All detailed formulas are shown in Appendix B.

Finally, the next symptom is selected as

$$S^* = \arg\max_{S \in \mathcal{S}} \left\{ \inf_{\xi} \mathbb{E}_{\xi} \left[\mathcal{I}(\mathcal{D}; \xi(S)) \right] \right\} \quad (1)$$

which guarantees a robust reduction in diagnostic uncertainty, leading to a stable questioning process.

3.3.3 Consultation Process and Multi-turn Information Acquisition

At round t, given the optimal symptom S^* selected by the DUR strategy algorithm and the current disease distribution $\mathcal{D}^{(t)} = \{\mathbb{P}^{(t)}(D)\}$, we employ a large language model (LLM) to generate a natural-language question that explicitly queries the presence of this symptom: $Q^{(t)} = f_{\mathrm{LLM}}(S^*; \mathcal{D}^{(t)}, \text{history})$.

After receiving the patient's textual response $A^{(t)}$, the LLM interprets the answer and maps it to the category label $\xi(S^*) = f_{\mathrm{LLM}}(S^*; Q^{(t)}, A^{(t)})$. Once the observation is obtained, the symptom is added to the observed set $\mathcal{O}^{(t)} = \mathcal{O}^{(t-1)} \cup \{(S^*, \xi(S^*))\}$ and removed from the candidate symptom space. The disease distribution is then updated accordingly. This process terminates when either (1) the fluctuation of the maximal diagnostic uncertainty reduction becomes smaller than a predefined threshold ϵ, or (2) a maximum number of rounds is reached. A simplified procedure is outlined in the pseudocode in 1. More details are shown in Appendx C.

3.4 Structured Clinical Reporter (SCR)

Directly presenting raw information from SAE (e.g., candidate diseases, symptom status, and dialogue content) may be inefficient for clinical use (Asgari et al., 2024). We therefore introduce a Structured Clinical Reporter (SCR) to convert it into professional, structured medical documents. SCR uses a Two-Stage Generation Paradigm, widely adopted (Han and Lyu, 2025; Yin et al., 2024), consisting of an Initial Diagnostic Report and a SOAP-based Clinical Report.

Stage-I: Initial Diagnostic Report with Reasoning Process. SCR first generates a concise diagnostic report focused on reasoning (Goh et al., 2024; Savage et al., 2024). Given the candidate disease set, symptom dictionary, and dialogue history, SCR generates reasoning traces with supporting, opposing, and uncertain evidence. Diagnoses remain consistent with knowledge graph information.

Stage-II: SOAP-Based Clinical Documentation Generation. SCR then expands the preliminary report into a structured clinical document following the SOAP format (Weed et al., 1968; Pearce et al., 2016). The Subjective (S) section summarizes patient self-report in standardized medical language. The Objective (O) section contains observable and measurable findings. The Assessment (A) section provides diagnostic analysis with a Disease-Symptom Comparison Table, categorizing the symptoms into "occurred", "absent", and "unobserved". The Plan (P) section provides supportive medical recommendations. SCR also generates a brief supplementary section with relevant medical knowledge from KG associated with the candidate diagnoses.

4 Experiments

4.1 Public Dataset

We evaluate our system on a public multi-turn medical dialogue dataset: MTS-Dialog (Abacha et al., 2023), which contains 1,700 doctor-patient conversations and diagnosis. Due to the computational cost, we randomly sample one-third of the data for experiments.

Experimental Setup. We formulate a multi-turn medical dialogue between a virtual doctor and a virtual patient. At each turn, the doctor asks a symptom-related question, and the patient responds based on its hidden clinical state. Eventually, the doctor outputs the candidate diseases, and we evaluate diagnostic accuracy and efficiency.

Virtual Patient. Following prior studies (Cook et al., 2025; Zhu et al., 2025; Grévisse, 2024; Wei et al., 2018), we construct a Virtual Patient (VP) agent based on structured clinical cases from the MTS-Dialog using Qwen-Plus, built upon the Qwen3 (Yang et al., 2025). The final diagnosis is treated as the hidden ground-truth label and is not disclosed during interaction, preventing infor-

Benchmark (%)	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8
GLM-based Agent	33.51	45.97	51.17	54.03	55.58	57.14	57.14	58.18
DeepSeek-based Agent	32.54	43.54	49.28	52.39	56.46	58.13	60.05	61.96
Qwen-based Agent	32.72	42.63	48.16	52.30	55.30	58.06	58.99	60.14
GLM-based Agent + RAG	41.19	51.80	56.48	60.30	62.63	63.91	64.97	65.82
DeepSeek-based Agent + RAG	39.48	48.60	54.38	58.85	61.64	64.62	65.74	66.85
Qwen-based Agent + RAG	43.31	51.41	54.93	58.63	60.74	63.20	64.61	65.85
GLM-based Agent + KG	32.91	45.86	52.87	55.84	58.17	59.66	60.51	61.57
DeepSeek-based Agent + KG	33.52	46.74	56.42	62.38	66.11	67.60	69.27	70.58
Qwen-based Agent + KG	34.33	48.59	55.28	58.45	61.80	64.08	65.32	66.02
Our[GLM] - DUR	34.18	49.26	53.72	58.17	59.66	60.93	62.63	63.06
Our[DeepSeek] - DUR	36.50	50.65	56.61	61.82	65.92	67.60	69.46	71.32
Our[Qwen] - DUR	34.15	48.06	55.81	59.86	62.32	63.91	64.96	65.49
Our[GLM]	54.78	66.24	72.40	75.80	77.71	79.62	79.83	81.32
Our[DeepSeek]	59.59	69.83	75.61	77.84	79.70	79.89	81.38	82.50
Our[Qwen]	**67.78**	**74.82**	**76.58**	**79.58**	**81.51**	**82.92**	**82.92**	**82.92**

Table 2: Top-k accuracy across benchmarks. Our full model achieves the best accuracy for all $k = 1, \ldots, 8$. RAG and KG ground diagnoses in domain knowledge, reducing unsupported speculation. Integrating the Knowledge Graph (KG) and RAG consistently enhances performance over Agent alone.

mation leakage. The VP responds faithfully and consistently based on its clinical case.

Virtual Doctor. The virtual doctor agent is implemented using large language models (LLMs) to conduct iterative diagnostic questioning and output a ranked list of candidate diseases. We evaluate representative LLMs, including Qwen-Plus, DeepSeek-V3.2 (Liu et al., 2025), and GLM-4.7 (Zeng et al., 2025) under identical configurations. The virtual doctor is prompted with structured instructions, and some agents are equipped with RAG and KG within specific benchmarks.

Our System & Other Benchmark Implementations. We implement multiple virtual doctor variants, differing in prompting strategies and component integration (RAG and KG), across different backbone LLMs. (1) Agent: multi-turn diagnostic dialogue without external knowledge augmentation. (2) Agent + RAG: augmented with retrieval-based external medical knowledge. (3) Agent + KG: guided by structured knowledge graph, such as disease-symptom relations. (4) Our System without DUR (Agent + RAG +KG): our full system without uncertainty-based symptom selection. (5) Our System: integrates RAG, KG reasoning, and DUR for symptom selection.

Evaluation Metrics. We adopt Top-k accuracy (Diagnosis ACU@k) as the primary evaluation metric. For each consultation sample i, let the ground-truth diagnosis be denoted as K_i $(i = 1, \cdots, N)$, and $L_i(k)$ be the ranked list of top-k predicted diseases. The metric is defined as

$$\text{Diagnosis ACU@}k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(K_i \in L_i(k))$$
$$\approx \frac{1}{N}\sum_{i=1}^{N} f_{\text{LMM}}(\text{whether } K_i \in L_i(k))$$

We use LLM-as-a-judge to build an evaluator. The evaluator is given the diagnostic report, dialogue history, and the predicted disease list, and outputs a judgment. In addition, we measure output token cost each sample to evaluate the efficiency of the dialogue process.

Benchmark (%)	Qwen	DeepSeek	GLM
Agent	339.05	288.10	255.44
Agent + RAG	350.89	347.46	250.05
Agent + KG	310.23	285.15	234.85
Our[·] - DUR	313.19	317.61	235.24
Our[·]	**268.23**	**230.85**	**148.13**

Table 3: Output token costs across benchmarks. Our full system uses the fewest tokens. RAG increases token usage, whereas KG does not. Removing DUR increases tokens, showing the efficiency of DUR strategy.

Results. As shown in Table 2, our system achieves the highest accuracy, indicating that our agent with DUR guides the dialogue toward informative symptoms and helps mitigate hallucinations. Table 3 shows that our full system also has the lowest token cost, implying that our agent also improves efficiency.

4.2 Real-World Scenarios

Experimental Design. In collaboration with a local community hospital, we conducted an experiment using 100 real-world patients' records. For each one, it was used to simulate a patient-agent

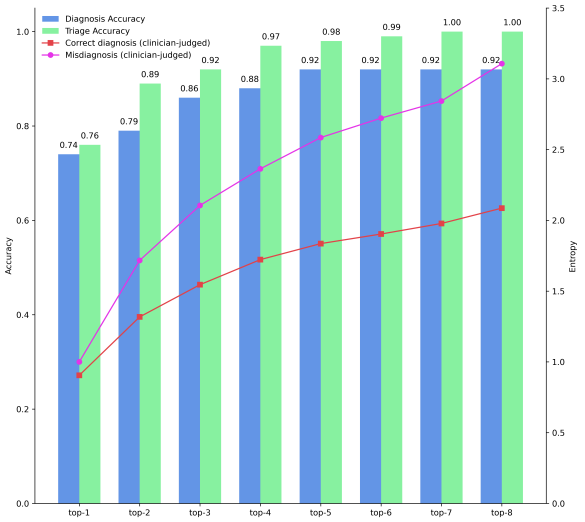

Figure 2: Top-k diagnostic and triage performance. Triage accuracy exceeds diagnostic accuracy and increases as k increases. Top-8 diagnostic accuracy is 92%, approaching physician's evaluation (95%). Lower department entropy in correct cases indicates more concentrated predictions and better triage.

interaction: the system asked questions, and responses were provided according to the medical record. After the dialogue, the system generated two outputs: (1) a preliminary diagnostic result and (2) a SOAP clinical document. The diagnosis in the electronic health record was seen as the reference. All case records and generated texts were stored on hospital servers and anonymized.

Expert Review Protocol. For each case, licensed physicians independently reviewed (i) the anonymized clinical record and (ii) the system-generated outputs. Using the original medical record as references, they assessed whether the system's candidate diagnoses matched the reference diagnosis. Diagnostic accuracy was computed based on these expert annotations. Disagreements were resolved through discussion after the initial independent evaluations.

Diagnosis Accuracy & Triage Performance judged by LLM. We also compute Top-k diagnostic accuracy (Diagnosis ACU@k) using the LLM-as-a-judge as previously introduced. Meanwhile, we further examined the triage performance by mapping all predicted diseases into clinical departments using KG. The metric measures are similar to that of Diagnosis ACU@k.

Results. The physicians independently judged whether the system provided a clinically appropriate diagnosis. Among the 100 cases, 95 predictions correctly included the ground-truth diagnosis, resulting in a 95% hit rate. As shown in Figure 2, our system demonstrates strong performance in both diagnostic prediction and clinical triage across different Top-k settings. See appendix D for more details.

5 Conclusion

We design an LLM-based agent architecture that structures patient narratives, grounds reasoning in external medical knowledge, and prioritizes high-information symptoms during sequential interactions. By integrating knowledge-grounded reasoning and uncertainty-reduction mechanisms, the proposed approach improves the performance and efficiency of AI-assisted symptom consultation. Meanwhile, we acknowledge the potential risks associated with deploying such agents in community healthcare settings, including possible overreliance on automated recommendations. These considerations underscore the importance of human oversight and responsible deployment practices.

Our work contributes to research on AI-enabled decision support by demonstrating how large language model–based agents can be systematically designed to address task uncertainty and information acquisition challenges in healthcare consultations. This study also has practical implications for the deployment of AI consultation systems in community healthcare settings. By guiding patients through structured, evidence-grounded dialogues, AI agents can help improve symptom information quality before clinical encounters, potentially supporting more efficient triage and decision-making. At the same time, this work highlights the importance of grounding AI-generated medical interactions in reliable knowledge sources and uncertainty-aware dialogue strategies to mitigate risks associated with purely generative models.

6 Limitations

This study has several limitations that also suggest directions for future work. First, the disease-symptom relationships in our current setting are simplified for deployment; future work should establish a more complex and explainable model. and future studies will incorporate broader benchmarking. Second, our real-world experiments are still limited in both scale and design sophistication, indicating the need for larger and more well-designed evaluations. Third, our current assessment relies on an evaluation agent that is related to the experi-

mental backbone LLM, which may introduce bias; subsequent work will explore independent judges.

References

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302.

Amir Abada and Samir Ibrahim Altalafha. 2023. Chronic disease management in primary care setting. *Sch Acad J Biosci*, 2:63–8.

Sahar Abdelnabi, Aideen Fay, Giovanni Cherubin, Ahmed Salem, Mario Fritz, and Andrew Paverd. 2025. Get my drift? catching llm task drift with activation deltas. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 43–67. IEEE.

Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. 2023. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv preprint arXiv:2311.01463*.

Elham Asgari, Japsimar Kaur, Gani Nuredini, Jasmine Balloch, Andrew M Taylor, Neil Sebire, Robert Robinson, Catherine Peters, Shankar Sridharan, and Dominic Pimenta. 2024. Impact of electronic health record use on cognitive load and burnout among clinicians: narrative review. *JMIR Medical Informatics*, 12:e55499.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156.

Han Bao, Zheyuan Zhang, Pengcheng Jing, Zhengqing Yuan, Kaiwen Shi, and Yanfang Ye. 2026. Driftbench: Diagnosing cooperative breakdowns in llm agents under input faults via multi-turn interaction. *arXiv preprint arXiv:2602.02455*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

L Bos and K Donnelly. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Stud Health Technol Inform*, 121:279–290.

Junhyuk Choi, Yeseon Hong, Minju Kim, and Bugeun Kim. 2024. Examining identity drift in conversations of llm agents. *arXiv preprint arXiv:2412.00804*.

David A Cook, Joshua Overgaard, V Shane Pankratz, Guilherme Del Fiol, and Chris A Aakre. 2025. Virtual patients using large language models: Scalable, contextualized simulation of clinician-patient dialogue with feedback. *Journal of Medical Internet Research*, 27:e68486.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, and 1 others. 2025. Automating evaluation of ai text generation in healthcare with a large language model (llm)-as-a-judge. *medRxiv*.

Corey Curran, Nafis Neehal, Keerthiram Murugesan, and Kristin P Bennett. 2024. Examining trustworthiness of llm-as-a-judge systems in a clinical trial design benchmark. In *2024 IEEE International Conference on Big Data (BigData)*, pages 4627–4631. IEEE.

Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. 2024. Building guardrails for large language models. *arXiv preprint arXiv:2402.01822*.

Vardhan Dongre, Ryan A Rossi, Viet Dac Lai, David Seunghyun Yoon, Dilek Hakkani-Tür, and Trung Bui. 2025. Drift no more? context equilibria in multi-turn llm interactions. *arXiv preprint arXiv:2510.07777*.

Alina Fastowski and Gjergji Kasneci. 2024. Understanding knowledge drift in llms through misinformation. In *International Workshop on Discovering Drift Phenomena in Evolving Landscapes*, pages 74–85. Springer.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, and 1 others. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open*, 7(10):e2440969.

Christian Grévisse. 2024. Raspatient pi: a low-cost customizable llm-based virtual standardized patient simulator. In *International Conference on Applied Informatics*, pages 125–137. Springer.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Hao Guan, David Bates, and Li Zhou. 2025. Keeping medical ai healthy and trustworthy: A review of detection and correction methods for system degradation. *IEEE Transactions on Biomedical Engineering*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Yewei Han and Chen Lyu. 2025. Multi-stage guided code generation for large language models. *Engineering Applications of Artificial Intelligence*, 139:109491.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

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

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Patricia F Pearce, Laurie Anne Ferguson, Gwen S George, and Cynthia A Langford. 2016. The essential soap note in an ehr age. *The Nurse Practitioner*, 41(2):29–36.

J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning*, 1(1):81–106.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20.

Edward H Shortliffe and Martin J Sepúlveda. 2018. Clinical decision support in the era of artificial intelligence. *Jama*, 320(21):2199–2200.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th international conference on intelligent user interfaces*, pages 952–966.

Prathiksha Rumale Vishwanath, Simran Tiwari, Tejas Ganesh Naik, Sahil Gupta, Dung Ngoc Thai, Wenlong Zhao, Sunjae Kwon, Victor Ardulov, Karim Tarabishy, Andrew McCallum, and 1 others. 2024. Faithfulness hallucination detection in healthcare ai. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025. A survey of llm-based agents in medicine: How far are we from baymax? *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10345–10359.

Lawrence L Weed and 1 others. 1968. Medical records that guide and teach. *N Engl J Med*, 278(11):593–600.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.

Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Atilla Kiraly, Sahar Kazemzadeh, Zakkai Melamed, and 1 others. 2023. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv preprint arXiv:2308.01317*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yuan-Jhe Yin, Bo-Yu Chen, and Berlin Chen. 2024. A novel llm-based two-stage summarization approach for long dialogues. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–6. IEEE.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

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

A Some normalization methods

For completeness, we summarize several normalization operations omitted from the main text.

Symptom prior probabilities Given the unnormalized symptom prior scores $\mathrm{Score}(S)$, we obtain the normalized marginal probabilities for a target symptom set $\mathcal{S} = \{S_1, \ldots, S_{n_S}\}$ via

$$\mathbb{P}(S) = \frac{\mathrm{Score}(S)}{\sum_{S \in \mathcal{S}} \mathrm{Score}(S)}$$

Log/Exp-space normalization Whenever probabilities are obtained by products of many small factors, we compute in log-space to avoid numerical underflow. Given unnormalized p, we compute normalized probabilities via

$$\hat{p} = \frac{\exp(\log p - \max_p \log p)}{\sum_p \exp(\log p - \max_p \log p)}$$

B Details of some conditional probability estimation

Condition on symptom Given symptom marginals $\mathbb{P}(S)$ (Appendix A), we construct a relevance score

$$r(S \mid D) = \mathbb{P}(S) \cdot \mathbb{I}(D, S)$$

which is non-zero only when S is associated with D. We then normalize over all symptoms associated with D:

$$\hat{\mathbb{P}}(S \mid D) = \frac{r(S \mid D)}{\sum_S r(S \mid D)} = \frac{\mathbb{P}(S) \cdot \mathbb{I}(D, S)}{\sum_S \mathbb{P}(S) \cdot \mathbb{I}(D, S)}$$

This can be interpreted as a disease-specific distribution over its possible symptoms.

Condition on response As for certain symptom S, we obtain its response $\xi(S)$. Based on Bayes, the distribution over diseases under $\xi(S) = a$ can be estimated as

$$\mathbb{P}(D \mid \xi(S) = a) = \frac{\mathbb{P}(\xi(S) = a \mid D) \cdot \mathbb{P}(D)}{\mathbb{P}(\xi(S) = a)}$$

For instance, if we consider $\{\mathrm{yes}, \mathrm{no}\}$ as the useful label space for patients' responses, a well estimation of $\mathbb{P}(\xi(S) = a \mid D)$ is the conditional probability $\hat{\mathbb{P}}(S \mid D)$ and $\hat{\mathbb{P}}(\neg S \mid D) = 1 - \hat{\mathbb{P}}(S \mid D)$. In this case, we obtain

$$\hat{\mathbb{P}}(D \mid \xi(S) = \mathrm{yes}) = \frac{\hat{\mathbb{P}}(S \mid D) \cdot \mathbb{P}(D)}{\mathbb{P}(S)}$$

$$\hat{\mathbb{P}}(D \mid \xi(S) = \mathrm{no}) = \frac{\hat{\mathbb{P}}(\neg S \mid D) \cdot \mathbb{P}(D)}{\mathbb{P}(\neg S)}$$

C Disease probabilities updating

By the end of round t, we need to update the current disease distribution for the next round $t+1$, with current observation $\mathcal{O}^{(t)}$. By Bayes theorem, we have

$$\mathbb{P}^{(t+1)}(D) = \mathbb{P}(D \mid \mathcal{O}^{(t)}) = \frac{\mathbb{P}(\mathcal{O}^{(t)} \mid D)\mathbb{P}^{(t)}(D)}{\mathbb{P}(\mathcal{O}^{(t)})}$$

Since the evidence set $\mathcal{O}^{(t)}$ is the same fixed observation when comparing different diseases, the denominator $\mathbb{P}(\mathcal{O}^{(t)})$ does not depend on D and serves only as a constant. Equivalently,

$$\mathbb{P}^{(t+1)}(D) \propto \mathbb{P}(\mathcal{O}^{(t)} \mid D) \cdot \mathbb{P}^{(t)}(D)$$

Assuming conditional independence among symptoms given the disease, that is, $(S_{k_1} \perp\!\!\!\perp S_{k_2}) \mid D$ for $k_1 \neq k_2$, we obtain

$$\begin{aligned}\mathbb{P}(\mathcal{O}^{(t)} \mid D) &= \mathbb{P}\Big(\bigcup_\xi \{S : S \text{ in } \mathcal{O}_\xi^{(t)}\} \,\Big|\, D\Big) \\ &= \prod_\xi \prod_{S \text{ in } \mathcal{O}_\xi^{(t)}} \hat{\mathbb{P}}(S \mid D)\end{aligned}$$

where $\hat{\mathbb{P}}(S \mid D)$ is the estimated probability of D given symptom S (Appendix B). $\mathcal{O}_\xi^{(t)}$ represents the observation $(S, \xi(S))$ at round t with $\xi(S) = \xi$. Thus, the unnormalized posterior at round t can be written as

$$\mathbb{P}_*^{(t+1)}(D) = \left[\prod_\xi \prod_{S \text{ in } \mathcal{O}_\xi^{(t)}} \hat{\mathbb{P}}(S \mid D)\right] \mathbb{P}^{(t)}(D)$$

and then normalized it.

Naive Bayesian updates can become overconfident. To improve probability calibration, we apply a temperature scaling based on the Softmax transformation (Guo et al., 2017; Jang et al., 2016). Let $\hat{\mathbb{P}}^{(t+1)}(D)$ be the normalized posterior above, and $\mathrm{logit}(p) = \log[p/(1-p)]$. We define the calibrated posterior

$$\mathbb{P}^{(t+1)}(D) = \frac{\exp\left\{\mathrm{logit}[\hat{\mathbb{P}}^{(t+1)}(D)]/T\right\}}{\sum_D \exp\left\{\mathrm{logit}[\hat{\mathbb{P}}^{(t+1)}(D)]/T\right\}}$$

where $T > 0$ is a temperature parameter: larger T produces a softer distribution, while smaller T yields sharper predictions.

D Evaluation for triage performance

To quantify the dispersion of the triage behavior, we computed the Shannon entropy of the predicted department distribution across cases:

• As for medical consultation i, its predicted diseases are $D_1^i, \cdots D_m^i$.

• Find the relevant clinical departments are $C_1^i, \cdots, C_l^i$ where $l > m$ under normal circumstances.

• Obtain the distribution of these departments at this sample:

$$\mathbb{P}(C_j^i) = \frac{\sum_u \mathbb{I}(D_u^i \text{ in } C_j^i)}{\sum_j \sum_u \mathbb{I}(D_u^i \text{ in } C_j^i)}$$

where D_u^i in C_j^i means disease D_u^i belongs to department C_j^i.

• Calculate the Shannon entropy of departments distribution

$$H(\{C_1^i, \cdots, C_l^i\}) = -\sum \mathbb{P}(C_j^i) \log \mathbb{P}(C_j^i)$$

To compare the entropy among different distribution, we propose to calculate the entropy of correct samples (selected by the physicians) and wrong samples.

E Prompt Templates

E.1 Virtual Patient Prompt

Virtual Patient Prompt

[Role]
You are a virtual patient participating in a medical consultation. You must simulate a real patient based strictly on the provided clinical information.

[Requirements]
- You are NOT aware of the final diagnosis.
- You must NEVER reveal, imply, or guess the diagnosis.
- Respond only from the perspective of a patient describing personal experiences.

[Rules]
- Keep your answers as brief as possible.
- Prefer short responses such as: "Yes." / "No." / "I don't know." / "Not sure."
- Do NOT exceed 10 words per reply.
- Do NOT provide extra explanations unless explicitly asked.
- If the symptom exists in your record, answer truthfully.
- If the symptom does not exist, clearly deny it.
- If the information is not mentioned in the record, reply: "I don't know." or "I'm not sure."

[Patient's Real Clinical Information]
{INFORMATION}

E.2 Virtual Doctor Prompt

Virtual Doctor Prompt

[Role]
You are a professional doctor conducting a multi-turn diagnostic consultation. Your objective is to infer the patient's underlying condition through iterative symptom-based questioning.

[Requirements]
- Do NOT directly ask about any disease name.
- Ask exactly ONE symptom or clinical feature per turn.
- Describe the symptom in clear, patient-friendly language.
- Determine whether the patient has experienced the described symptom.
- Keep each question concise and specific. The question must not exceed 45 words.
- Do NOT provide reasoning process, diagnosis, or medical advice.

- Focus only on symptom inquiry. Do not output anything other than the question.

IF RAG is acquired {
[External Medical Knowledge]
The following retrieved medical information may assist your reasoning: ${DOCUMENTS}.
}

IF Knowledge Graph is acquired {
[Knowledge Base]
Candidate diseases and their related symptoms define the current hypothesis space: {KG}.
}

E.3 Diagnosis Evaluator Prompt

Virtual Diagnosis Evaluator Prompt

[Role]
You are a professional physician. Your task is to read the patient's diagnostic report and evaluate whether the predicted disease list is accurate. If the prediction is accurate, identify the indices of the correctly predicted diseases in the list.

[Requirements]
- 'Accurate' means the predicted disease list contains the disease(s) mentioned in the diagnostic report.
- If the prediction is accurate, determine the indices of correct diseases in the predicted list. Indexing starts from 0. Multiple selections are allowed. If multiple diseases are correct, include all corresponding indices.

[Output Example]

```
{
    "judge": true,
    "accuracy_list": [0, 2, 3]
}
```

[Important Constraints]
- Strictly output valid JSON exactly like the example above. Do NOT include any additional text.
- If the prediction list is completely incorrect, return:

```
{
    "judge": false,
    "accuracy_list": []
}
```