



# DrHouse: An LLM-empowered Diagnostic Reasoning System through Harnessing Outcomes from Sensor Data and Expert Knowledge

BUFANG YANG\*, The Chinese University of Hong Kong, China

SIYANG JIANG\*, The Chinese University of Hong Kong, China

LILIN XU, The Chinese University of Hong Kong, China

KAIWEI LIU, The Chinese University of Hong Kong, China

HAI LI, The Chinese University of Hong Kong, China

GUOLIANG XING, The Chinese University of Hong Kong, China

HONGKAI CHEN, The Chinese University of Hong Kong, China

XIAOFAN JIANG, Columbia University, United States

ZHENYU YAN<sup>†</sup>, The Chinese University of Hong Kong, China

Large language models (LLMs) have the potential to transform digital healthcare, as evidenced by recent advances in LLM-based virtual doctors. However, current approaches rely on patient's subjective descriptions of symptoms, causing increased misdiagnosis. Recognizing the value of daily data from smart devices, we introduce a novel LLM-based multi-turn consultation virtual doctor system, DrHouse, which incorporates three significant contributions: 1) It utilizes sensor data from smart devices in the diagnosis process, enhancing accuracy and reliability. 2) DrHouse leverages continuously updating medical knowledge bases to ensure its model remains at diagnostic standard's forefront. 3) DrHouse introduces a novel diagnostic algorithm that concurrently evaluates potential diseases and their likelihood, facilitating more nuanced and informed medical assessments. Through multi-turn interactions, DrHouse determines the next steps, such as accessing daily data from smart devices or requesting in-lab tests, and progressively refines its diagnoses. Evaluations on three public datasets and our self-collected datasets show that DrHouse can achieve up to an 31.5% increase in diagnosis accuracy over the state-of-the-art baselines. The results of a 32-participant user study show that 75% medical experts and 91.7% test subjects are willing to use DrHouse.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**.

Additional Key Words and Phrases: LLMs, Sensor Data, Diagnostic Reasoning Systems, Knowledge Retrieval, Proactive Conversational Systems, Up-to-Date, Internet of Things

## ACM Reference Format:

Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. 2024. DrHouse: An LLM-empowered Diagnostic Reasoning System through Harnessing Outcomes from Sensor Data and

\*Both authors contributed equally to this research.

<sup>†</sup>Corresponding author.

Authors' Contact Information: [Bufang Yang](#), The Chinese University of Hong Kong, China, [bfiyang@link.cuhk.edu.hk](mailto:bfiyang@link.cuhk.edu.hk); [Siyang Jiang](#), The Chinese University of Hong Kong, China, [syjiang@ie.cuhk.edu.hk](mailto:syjiang@ie.cuhk.edu.hk); [Lilin Xu](#), The Chinese University of Hong Kong, China, [lilinxu@cuhk.edu.hk](mailto:lilinxu@cuhk.edu.hk); [Kaiwei Liu](#), The Chinese University of Hong Kong, China, [1155189693@link.cuhk.edu.hk](mailto:1155189693@link.cuhk.edu.hk); [Hai Li](#), The Chinese University of Hong Kong, China, [lh123@ie.cuhk.edu.hk](mailto:lh123@ie.cuhk.edu.hk); [Guoliang Xing](#), The Chinese University of Hong Kong, China, [glxing@cuhk.edu.hk](mailto:glxing@cuhk.edu.hk); [Hongkai Chen](#), The Chinese University of Hong Kong, China, [hkchen@ie.cuhk.edu.hk](mailto:hkchen@ie.cuhk.edu.hk); [Xiaofan Jiang](#), Columbia University, United States, [jiang@ee.columbia.edu](mailto:jiang@ee.columbia.edu); [Zhenyu Yan](#), The Chinese University of Hong Kong, China, [zyyan@ie.cuhk.edu.hk](mailto:zyyan@ie.cuhk.edu.hk).



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2024 Copyright held by the owner/author(s).

ACM 2474-9567/2024/12-ART153

<https://doi.org/10.1145/3699765>

## 1 Introduction

According to the views of the fictional character Dr. Gregory House in the popular TV show *House*,<sup>1</sup> instead of relying on patient's subjective descriptions of symptoms, doctors should take advantage of more objective data inputs such as smart devices, wearable sensors, and in-lab tests for diagnosis. Within the realm of large language models (LLMs) [2, 43] and other advanced AI technologies [52, 59], many variations of LLMs have been developed for medical tasks, such as understanding reports [13] and diagnostic reasoning systems [51, 69] that can provide diagnosis suggestions for doctors. We refer to them as *virtual doctors* throughout this paper. However, existing LLM-based virtual doctors still depend on the patient's subjective description of symptoms. In addition, many of these models do not incorporate the most recent medical corpus and focus on the medical question-answering tasks [41]. They only provide general medical summaries or suggestions instead of proactively inquiring about additional symptoms or lab tests from patients, thus missing out on potential illnesses. The inability to quantify the likelihood of each diagnosis also hampers the confidence of both doctors and patients in the diagnostic results provided by these LLM-based virtual doctors.

However, simply incorporating medical conversations into LLMs is not feasible as patients' descriptions can be susceptible to subjective perceptions or memory biases, thus leading to the ambiguity and unreliability of the reported symptoms [34, 36]. Moreover, it can be more challenging for patients to provide precise answers about objective metrics such as respiratory rate and blood oxygen levels. Consequently, the subjective perception of patients and uncertainty about their symptoms will hinder timely and accurate diagnosis and treatments [36]. Additionally, after examining the medical diagnosis specifications for various common diseases on Up-to-Date,<sup>2</sup> we found that many physiological indicators crucial for diagnosing numerous diseases can be obtained from smart devices used in daily life. Motivated by these observations, this paper aims to incorporate the knowledge from patients' daily sensor data to assist LLM-based virtual doctors with multi-turn diagnosis.

According to our observations which highlight the potential benefits of incorporating daily sensor data into the diagnostic process, existing studies on LLMs for sensor data primarily focus on utilizing LLMs to interpret sensor signals for understanding the physical world [16, 22, 54] or make health predictions [24, 38, 39]. None of the existing works use the sensor data for multi-turn diagnosis. To address this research gap, we first explore the integration of objective sensor data into the multi-turn diagnosis of LLM-based virtual doctors. We summarize several challenges we encountered as follows. First, existing LLM-based virtual doctors [5, 41, 69] are only fine-tuned with medical corpus. How to enable the LLM-based virtual doctor to follow continuously updated medical diagnostic standards and proactively inquire about disease-related physiological indicators that can be obtained from the patient's smart devices is challenging. Second, the retrieval of required knowledge from the patient's extensive daily sensor databases remains challenging due to the vast amount of data collected and the complex and diverse questions posed by the LLM-based virtual doctor. Third, sensor data can be influenced by various factors such as environment and calibration, making it challenging to select and integrate the patients' symptom descriptions and the sensor data to determine the next-step actions during multi-turn medical consultations.

In this paper, we propose *DrHouse*, the first LLM-empowered Diagnostic reasoning system through Harnessing outcomes from sensor data and expert knowledge. *DrHouse* first retrieves the latest expert knowledge including potential disease-related diagnosis guidelines and then proactively inquires about patients' symptoms through multi-turn medical consultations. Through multi-turn interactions, *DrHouse* integrates patients' symptom descriptions, sensor data from smart devices (e.g., wearable devices), and the latest medical expert knowledge,

<sup>1</sup>House, also known as "House, M.D.", is an American medical drama television series.

<sup>2</sup>Up-to-Date [47] is a medical knowledge database that contains the latest knowledge and constantly evolving techniques in healthcare.

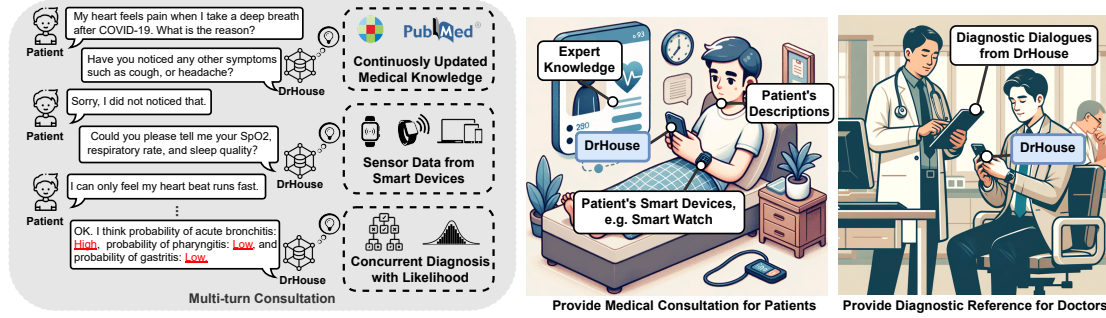


Fig. 1. Overview of DrHouse. DrHouse incorporates patients' sensor data from smart devices into the multi-turn diagnosis process to enhance accuracy and reliability. DrHouse can provide medical consultations for patients at their homes, or offer diagnostic references to doctors to reduce their workload.

to determine the next steps such as accessing daily sensor data or requesting in-lab tests. Simultaneously, DrHouse concurrently evaluates potential diseases, progressively refines its diagnoses, and finally generates an explainable diagnosis with a likelihood for each specific disease. We summarize the contributions of this paper as follows:

- We propose DrHouse, the first LLM-empowered virtual doctor system that can both initiate multi-turn medical consultations about disease-related physiological indicators and integrate the knowledge from patients' sensor data into the multi-turn diagnosis process to enhance accuracy and reliability.
- We develop three modules to enable DrHouse to conduct multi-turn conversations with the patients while incorporating both the latest medical expert knowledge and sensor data from smart devices. First, we construct the knowledge base in DrHouse, and develop a multi-source knowledge retrieval approach to iteratively retrieve the required sensor data knowledge and medical expert knowledge during each round of conversations. Then, we develop a knowledge integration approach to fuse patient's descriptions, sensor data knowledge, and medical knowledge. Finally, we design a diagnostic decision-making strategy through knowledge selection and concurrent checking of candidate diseases with likelihood, allowing DrHouse to consider the quality of sensor data and provide more nuanced and informed medical assessments.
- We validate DrHouse on both synthetic and real-world datasets. We also conduct two types of user studies involving both volunteers (N=12) and medical experts (N=20) as participants. Evaluations on both synthetic and real-world datasets show that DrHouse outperforms state-of-the-art baselines by up to 31.5% in diagnostic accuracy, and user studies indicate that DrHouse exhibits promising market potential.

## 2 Related work

### 2.1 LLM-based Virtual Doctors

**Fine-tuning LLMs for Clinical Diagnosis.** Several LLMs have been proposed for medical purposes, such as Med-PaLM 2 [41], DISC-MedLLM [5] and HuatuoGPT [69]. They collect extensive medical datasets comprising question-answering (QA) and diagnosis dialogues, employing the supervised fine-tuning (SFT) technique to LLMs. These specialized medical LLMs can outperform GPT-4 on several benchmarks in the medical domain. Google [44] employs a self-play-based environment to generate simulated doctor-patient dialogues. They fine-tune the LLM based on public medical datasets, real doctor-patient dialogues in hospitals, and simulated dialogues generated by LLMs. LLaVA-Med [27] leverages comprehensive datasets of biomedical figure captions to fine-tune LLaVA, a visual large language model. LLaVA-Med can perform medical visual question answering (VQA), such as interpreting CT scans. Other works like AMSC [48] propose to use multiple LLMs working in collaboration

Table 1. A summary of the recent LLM-based medical applications (● means included).

Approach	Base LLM	Proactive Multi-turn Consultation	Sensor Data Knowledge	Diagnosis Guidelines	Latest Knowledge	Probabilistic Diagnosis
MedicalGPT [56]	LLaMA, Baichuan	○	○	○	○	○
LLM-AMT [51]	GPT-3.5, GPT-4, LLaMA-2	○	○	○	○	○
HuatuoGPT-II [69]	Baichuan2-Base	●	○	○	○	○
DISC-MedLLM [5]	Baichuan2-Base (7B,13B)	●	○	○	○	○
Med-PaLM 2 [41]	PaLM 2	○	○	○	○	○
MedDM [26]	GPT-3.5	●	○	●	○	○
Health-LLM [24]	Alpaca	○	●	○	○	○
CaiTI [38]	GPT-3.5, GPT-4, LLaMA-2	●	●	●	○	○
<b>DrHouse</b>	GPT-3.5, GPT-4, LLaMA-3	●	●	●	●	●

for automated diagnosis, with each LLM functioning as a specialized doctor focused on a single subject. Previous works for LLM-based virtual doctors primarily concentrate on adapting LLMs to the medical domain and overlook the influence of subjects' perceptions. DrHouse employs a different approach to address this challenge by incorporating the objective sensor data from smart devices into the multi-turn medical consultations.

**Knowledge Retrieval with LLMs.** Due to the significant computational resources and costs for LLM training, several recent studies [21, 51] have explored the utilization of RAG to enhance LLMs with medical knowledge, eliminating the need for additional training. LLM-AMT [51] and Health-LLM [21] develop a database of medical textbooks for retrieval by LLMs. JMLR [49] employs a joint training approach for LLM and retrieval simultaneously to improve LLM's ability of medical knowledge reasoning. MedDM [26] proposes a framework for preprocessing diagnostic guidelines into a format comprehensible by LLMs. However, MedDM retrieves the diagnosis guidelines based on the patient's descriptions only. The discrepancy between patients' subjective expressions and the specialized medical terminology in diagnostic guidelines can result in reduced retrieval performance. In addition, previous works on retrieval-based virtual doctors primarily focus on retrieving medical knowledge to assist LLM diagnosis, with none exploring the retrieval of sensor data knowledge during multi-turn medical consultations.

## 2.2 LLM Understanding Sensor Data

Smart devices such as smartphones and wearable devices have become pervasive in everyday life [8, 55, 60], serving as passive sensors that effortlessly collect a multitude of data [17, 18, 61, 62]. Recent works explore using sensor data to assist LLM in understanding the physical world [16, 19, 54, 63]. Penetrative AI [54] shows that with suitable prompt design and expert knowledge, LLMs can interpret diverse sensor data, like WiFi and ECG signals. HARGPT [19] directly utilizes raw sensor data as input and chain-of-thought prompts to perform human activity recognition in the physical world. Additionally, many studies investigate LLM reasoning with sensor data for healthcare applications [10, 22, 24, 38]. Kim *et al* [24] integrate contextual information, such as user demographics and health knowledge, with physiological data, including resting heart rate and sleep duration, to enhance the comprehensive understanding of LLMs. The work in [10] leverages large language models to synthesize clinically useful insights from multi-sensor data, generating reasons about how trends in data relate to mental conditions. CaiTI [38] analyzes the user's daily functioning through several fixed and predetermined dimensions and employs LLM for psychological therapy. However, CaiTI is specialized in mental health and is not designed for multi-turn medical consultations. Previous work primarily focuses on utilizing LLMs to interpret sensor signals for understanding the physical world or predicting health outcomes. DrHouse takes a further step to incorporate the knowledge from sensor data into multi-turn medical consultations.

In summary, as shown in Table 1, most existing works either focus on adapting LLMs to the medical domain or utilizing LLMs to interpret diverse sensor signals for understanding the physical world. How to incorporate



Diagnosis: Hyperthyroidism	Diagnosis: Gastritis
<p><b>Patient:</b> When I'm hungry, my stomach feels uncomfortable. After eating, it's as if I experience indigestion. I must eat when hunger strikes; otherwise, I become dizzy, lack energy, experience stomach discomfort, often feel sleepy, and tire easily.</p> <p><b>GPT-4:</b> Do you have symptoms of acid reflux and heartburn?</p> <p><b>Patient:</b> No, I don't have.</p> <p><b>GPT-4:</b> That should be caused by an upset stomach. It could be gastritis. Eat good digestive food recently, such as porridge and soup.</p> <p><b>Patient:</b> Okay, I make soup every day now.</p> <p><b>GPT-4:</b> Are you flustered? What is your respiratory rate, sleep score, heart rate, and SpO2 levels? If your heart rate is high, consider testing your thyroid function.</p> <p><b>Patient:</b> <u>I am not sure about my heartbeat. But I feel my heart runs fast, maybe 130 bpm.</u></p> <p><b>Physiological Indicators in the Abnormal Range</b></p> <p><b>GPT-4:</b> Based on your heart rate of 130 beats per minute, which is above the normal range, it's important to investigate further as this could indicate <u>hyperthyroidism. Hyperthyroidism can cause symptoms like feeling dizzy, tiredness, and indigestion issues due to increased metabolism.</u></p>	<p><b>Patient:</b> When I'm hungry, my stomach feels uncomfortable. After eating, it's as if I experience indigestion. I must eat when hunger strikes; otherwise, I become dizzy, lack energy, experience stomach discomfort, often feel sleepy, and tire easily.</p> <p><b>GPT-4:</b> Do you have symptoms of acid reflux and heartburn?</p> <p><b>Patient:</b> No, I don't have.</p> <p><b>GPT-4:</b> That should be caused by an upset stomach. It could be gastritis. Eat good digestive food recently, such as porridge and soup.</p> <p><b>Patient:</b> Okay, I make soup every day now.</p> <p><b>GPT-4:</b> Are you flustered? What is your respiratory rate, sleep score, heart rate, and SpO2 levels? If your heart rate is high, consider testing your thyroid function.</p> <p><b>Patient:</b> <u>My heartbeat is 69 bpm.</u></p> <p><b>Physiological Indicators in the Normal Range</b></p> <p><b>GPT-4:</b> Given your heart rate is around 69 beats per minute, which falls within the normal range of 60 to 100 beats per minute, <u>it doesn't seem like your symptoms are related to an issue with your thyroid function. The patient has reported symptoms that could be indicative of gastritis.</u></p>

Fig. 2. An example showing that responding to the LLM-based virtual doctor with varying physiological indicators can lead to different diagnostic outcomes, such as hyperthyroidism and gastritis. We use GPT-4 as a naive LLM in the examples.

the knowledge of sensor data from patients' smart devices to enhance the capability of LLMs for diagnostic decision-making and multi-turn medical consultations remains unexplored.

### 3 Background and Motivation

In this section, we first show the limitations of existing LLM-based virtual doctors for patients' subjective perception. Next, we show the diagnostic guidelines for various diseases on the Up-to-Date database [47]. The key insights from these observations motivate the design of DrHouse.

**Potential Misdiagnosis Risks Due to Patient's Inaccurate Subjective Perception.** Previous studies in medicine have revealed that the subjective perception of patients and uncertainty about their symptoms can impede timely and accurate diagnosis and treatments [34, 36]. In clinical practice, physicians typically inquire about various objective physiological metrics from patients, including respiratory rate, sleep quality, blood oxygen levels, and heart rate [32]. However, it can be challenging for patients to provide precise answers about these objective metrics, particularly in the context of LLM-based virtual doctors. This is because patients typically use virtual doctors from their homes without the presence of in-lab tests or healthcare professionals like nurses who could aid them in obtaining these measurements.

We first present an example showing how, in the absence of objective sensor data, existing LLM-based virtual doctors may misdiagnose due to inaccurate subjective symptom descriptions. Specifically, we utilize the public medical dialogue dataset, i.e., DIALMED [15], for experiments. We conduct multi-turn consultations with the virtual doctor, utilizing symptoms from the dataset. When virtual doctors inquire about objective physiological indicators, we test responses using the physiological indicators in the normal and abnormal ranges, respectively. Figure 2 shows that responding to the virtual doctor with varying physiological indicators can lead to different diagnostic outcomes, such as hyperthyroidism and gastritis. However, in the real world, patients' responses may be inaccurate or vague, increasing the risk of misdiagnosis. This motivates us to incorporate objective sensor data into LLM-based medical consultations, which is crucial for ensuring their reliability.

**Sensor Data Indicators in Medical Diagnostic Guidelines.** Medical diagnosis guidelines [26] serve as the gold standard that doctors directly consult and utilize in clinical practice, enabling LLM to learn with minimal data and execute actions and decisions that closely align with those of real doctors. These diagnostic guidelines typically employ a tree-like structure. Doctors will inquire about the patient's symptoms and subsequently make informed decisions and diagnoses by referencing the diagnosis guidelines. Furthermore, our observations indicate that numerous latest medical diagnosis guidelines incorporate the assessment of diverse physiological indicators,

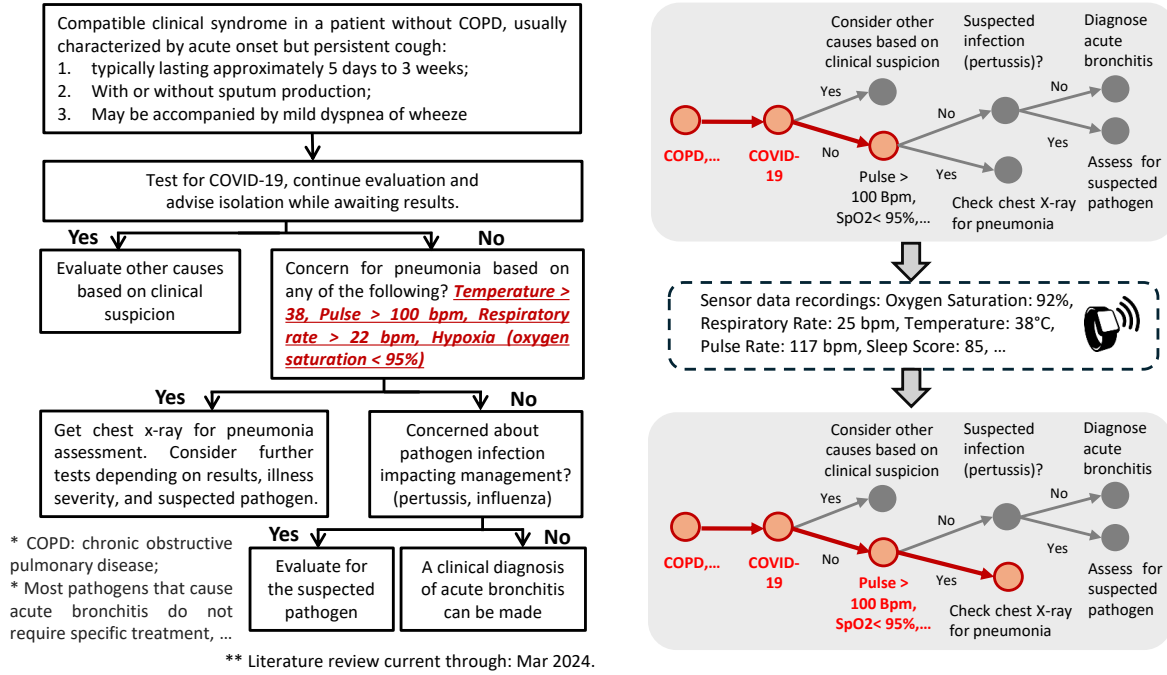


Fig. 3. An example of sensor data indicators in medical diagnosis guidelines. The left figure shows the latest medical diagnostic guidelines for acute bronchitis on the Up-to-Date database [47]. The right figure shows that accessing patients' daily sensor data can take the diagnostic process one step further and assist virtual doctors in decision-making.

which can be conveniently obtained through smart devices such as FitBit [14], worn by patients. We examined the diagnostic guidelines for various common diseases from the Up-to-Date [47] database and found that many of these guidelines include physiological indicators that can be obtained from wearable devices. For instance, 71.4% of the diagnostic guidelines for common respiratory conditions such as acute bronchitis, asthma, and chronic obstructive pulmonary disease (COPD) involve physiological indicators that can be obtained using smart wearable devices. Figure 3 shows an example of the diagnostic guideline for acute bronchitis from the Up-to-Date [47], where the red highlight shows the sensor data request in the diagnostic guideline, such as respiratory rate > 22bpm and oxygen saturation < 95%. Accessing patients' daily sensor data directly from smart devices can assist the virtual doctor in decision-making and enhance the probability of accurate diagnosis.

**Summary.** We summarize the motivations as follows.

- First, the patient's subjective perception, particularly regarding objective physiological indicators, can increase the potential risk of misdiagnosis by LLM-based virtual doctors.
- Second, we found that the diagnostic guidelines for many common diseases in the latest medical databases [47] include indicators that can be obtained from sensor data of smart devices. Integrating sensor data from smart devices into the diagnostic process of LLM-based virtual doctors can enhance diagnostic reliability.
- Third, the action space of existing virtual doctors [5, 69] is limited to inquiring about patients' symptoms and requesting in-lab tests. Accessing patients' sensor data from smart devices offers virtual doctors an additional option, expanding their action space. This expansion not only assists in cutting down on unnecessary in-lab tests but also facilitates patients in providing subjective answers, thereby reducing the risk of misdiagnoses.

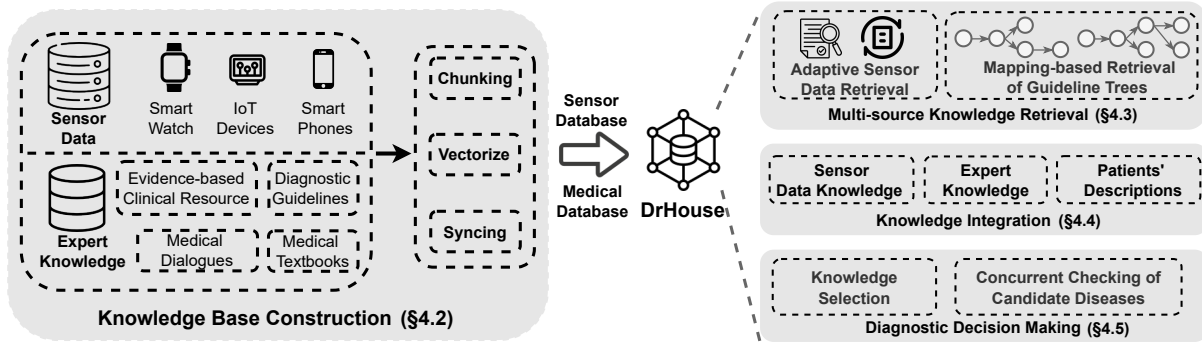


Fig. 4. System overview of DrHouse.

## 4 System Design

### 4.1 System Overview

DrHouse is an LLM-empowered multi-turn consultation virtual doctor system capable of leveraging both the patient’s descriptions of their symptoms, patients’ sensor data from smart devices, and medical knowledge for diagnosis. Figure 4 overviews the design of the DrHouse. The left part of Figure 4 shows DrHouse’s construction of two knowledge bases. In particular, DrHouse constructs knowledge bases (§ 4.2) using daily sensor data from patients’ smart devices and continuously updated medical expert knowledge, respectively. After construction, DrHouse processes the data into databases for LLM’s usage. The right part of Figure 4 illustrates the three runtime technologies to support multi-turn medical consultations. First, DrHouse employs two approaches to iteratively retrieve the required sensor data and the latest medical knowledge during each round of conversations, respectively (§ 4.3). Then, to make LLM understand the retrieved knowledge, DrHouse employs a knowledge integration approach to fuse patients’ descriptions, sensor data knowledge, and medical knowledge (§ 4.4). Finally, DrHouse oversees the integrated knowledge from sensors, medical knowledge, and in-lab tests for estimating the probabilities of multiple candidate diseases (§ 4.5).

### 4.2 Knowledge Base Construction

In this section, we introduce the knowledge base in DrHouse, including the two types of knowledge sources, the latest clinical resources, and the vectorization and synchronization of the knowledge base.

**4.2.1 Medical Expert Knowledge.** The medical knowledge sources in DrHouse consist of multi-turn medical dialogues, medical textbooks, and diagnostic guidelines.

**Medical Dialogues.** To enable LLMs to emulate the role of a doctor and facilitate multi-turn medical consultations, we incorporate real medical dialogue datasets [30, 67] into DrHouse’s medical knowledge base. These dialogue datasets can teach DrHouse the doctor’s language to raise multi-turn conversations.

**Medical Textbooks.** Despite being trained extensively on massive datasets, LLMs such as GPT-4 still lack the professional knowledge required in the medical field. This prevents LLMs from giving professional diagnoses and advice. Therefore, we teach general medical knowledge including the definition of medical terms and normative range of medical indicators, e.g., “Upper endoscopy is not required in the presence of typical Gastroesophageal reflux disease (GERD) symptoms of heartburn or regurgitation”. A large number of medical textbooks, including MedQA [67] and PubMedQA [23], are stored in the medical knowledge base of DrHouse.

**Diagnostic Guidelines.** The objective of clinical decision-making is to minimize risks and potential harm while maximizing effectiveness [3]. However, solely relying on a doctor’s subjective intuition can lead to unpredictable

and non-reproducible diagnoses. Therefore, experts have proposed medical diagnostic trees to reduce uncertainty in decision-making [4]. As one of the evidence-based clinical decision methods, diagnostic guideline trees serve as the gold standard for doctors during diagnosis [46]. Clinicians should meticulously refer to the guidelines and utilize the most effective evidence gathered during consultations to make the most beneficial decisions for their patients [33]. The characteristics of ease of interpretation of diagnostic guideline trees are also suitable for LLMs, which can improve the predictability and stability of LLM reasoning.

To this end, DrHouse also integrates diagnostic guidelines from the public medical knowledge search engine Up-to-Date [47] into its knowledge sources. The diagnostic guidelines are structured as a decision tree, which is suitable for LLMs to ask relevant follow-up questions and determine the next steps, as shown in Figure 3. Currently, DrHouse includes diagnostic guidelines for thirty common diseases. Since the original diagnostic guidelines are in figure format, we first use the optical content recognition model PaddleOCR [1] to convert them into a text-based format, facilitating better comprehension by LLMs. After that, we follow the method in MedDM [26] to transform the text-based diagnostic guidelines into if-else formatted **diagnosis guideline trees**. Subsequently, a manual inception is performed to ensure the correctness.

**4.2.2 Sensor Data.** DrHouse will access the patient's daily sensor data from smart devices during the diagnosis. This subsection details the sensor data knowledge sources in DrHouse.

**Knowledge Source of Sensor Data.** The sensor data knowledge source contains daily data collected from patients' smart devices, such as the Fitbit Sense smartwatch [14]. The recorded data includes various physiological indicators of the patient over each time interval, such as oxygen saturation, sleep scores, and heart rate. These physiological indicators provide LLM-based virtual doctors with reliable evidence for diagnosis. Patient's sensor data are saved in a tabular format and stored on their local device or personal cloud storage, like iCloud Drive.

**Check of Sensor Data's Uncertainty.** Although sensor data are generally more objective than patient statements, they can still suffer from quality issues due to environmental factors or motion artifacts [42]. To address this, we first check the sensor data's uncertainty, assigning an uncertainty score for each sample. If the uncertainty is high, the virtual doctor should request in-lab tests instead of relying solely on the sensor data. Specifically, we profile the uncertainty level for each sensor data record using a window-based probability density function (PDF) approach [66] and store this information in the sensor database. This method computes the mean and variance within a sliding window, thereby avoiding the misinterpretation of sensor values during exercise. For example, a higher heart rate during exercise will not be considered an outlier if the overall average remains stable. Additionally, our approach can also identify erroneous sensor readings, such as anomalous zero values.

**4.2.3 Latest Clinical Resources.** We leverage evidence-based clinical resources to enable continuous learning from the latest clinical research output. DrHouse will embed the lifelong learning capability like the doctors by using evidence-based clinical resources as a knowledge resource. Up-to-date [47] and PubMed [40] are two popular such data resources that contain evidence-based clinical cases, the latest research outputs, disease management, and treatment. To ensure our model remains at the forefront of medical diagnosis standards, DrHouse continuously updates these evidence-based clinical resources from Up-to-date and PubMed.

**4.2.4 Knowledge Base Vectorization and Synchronization.** The medical and sensor data knowledge sources mentioned above are stored in the format of text. To facilitate the knowledge querying by DrHouse during the diagnosis process, we convert these text knowledge sources into vector databases in advance, which is called **knowledge base** in this paper. Specifically, we first split the raw document into multiple chunks. These chunks are converted to embedding vectors using the OpenAI embedding model text-embedding-ada-002 [37]. The knowledge databases are subsequently utilized for knowledge retrieval in response to queries made by patients or virtual doctors. DrHouse employs an event-driven synchronization strategy to dynamically embed knowledge, including both the latest medical knowledge and continual sensor data, into the knowledge database.

The synchronization process for the sensor data is triggered every hour. The medical knowledge is synchronized when newer medical guidelines are online.

### 4.3 Multi-source Knowledge Retrieval

Existing LLM-based virtual doctors [5, 41, 69] only retrieve medical knowledge based on patients descriptions. DrHouse takes a unique approach by multi-source knowledge retrieval, enabling it to leverage sensor data from smart devices and medical expert knowledge simultaneously.

**4.3.1 Retrieval of Guidelines Trees.** Accurate retrieving relevant diagnosis guidelines based on patients' self-reported symptoms in the initial round of conversation is challenging. MedDM [26] directly retrieves the guideline based on patients' descriptions. However, the **discrepancy** between the non-specialized expressions of patients and the specialized medical statements in the guideline tree results in poor retrieval performance. Our results show the accuracy of directly retrieving guidelines based on the patient's descriptions is only about 37%, suggesting a huge gap between patients' expressions and the statements in guidelines trees (see § 5.3.2).

**Mapping-based Retrieval of Guidelines Trees.** To address this challenge, we design a mapping-based approach for the retrieval of multiple guideline trees, as shown in Figure 9. Specifically, we first retrieve the top-k similar patients' symptoms from a pre-collected symptom-disease dataset [50], i.e.,  $d_{sym} = \text{Retriever}(sym, \mathcal{D}_{sym})$ , where  $sym$  is the patient's descriptions about their symptoms, and  $\text{Retriever}$  is the retrieval model.  $\mathcal{D}_{sym}$  denotes the dataset [50], which contains common clinical symptoms, each associated with potential disease labels. The  $\text{Retriever}$  calculates the cosine similarity score between the query vector  $q$  and stored embeddings. A retrieval is deemed successful if the similarity score surpasses a predefined threshold [10], after which the top-k stored documents are returned, denoted as  $d = \text{TopK}_{s \in \mathcal{D}}(\text{sim}(q, s))$ , where  $q$  denotes the question and  $s$  denotes the sample in the dataset  $\mathcal{D}$ . In addition, we use the embedding model `text-embedding-ada-002` [37] from OpenAI as  $\text{Retriever}$  throughout this study. Once we acquire the top-k potential diseases of the patients, we can precisely obtain the diagnosis guideline trees associated with these diseases. Finally, we incorporate these top-k guideline trees into the prompt of DrHouse, and input to the LLM (see § 4.4).

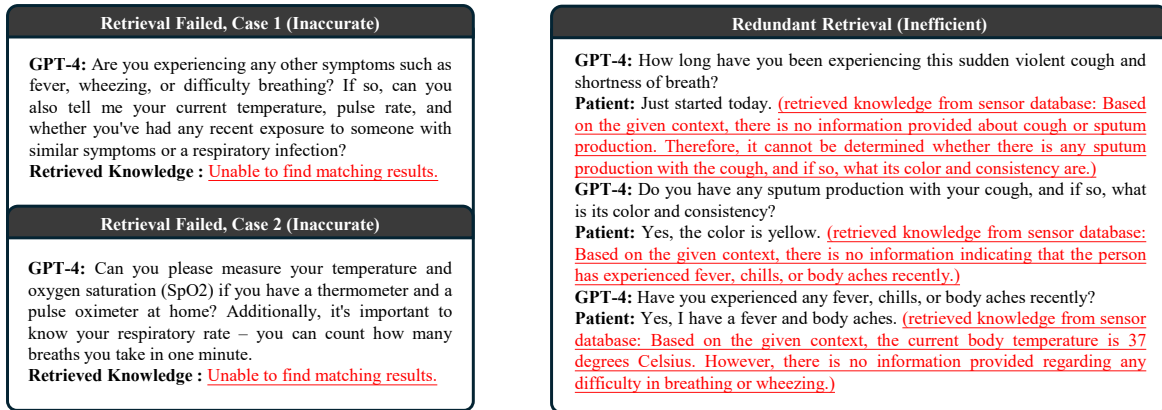
**Dynamic Guideline Tree Update.** As multiple rounds of medical consultations progress, DrHouse accumulates an increasing number of patient symptoms, resulting in the change of potential diseases. Consequently, after each round of consultation, DrHouse dynamically updates the top-k diagnosis guideline trees to adapt to the changes in symptoms. On the other hand, the increased number of patient symptoms also helps DrHouse to narrow down the potential diseases.

**4.3.2 Adaptive Sensor Data Retrieval.** In this subsection, we introduce the basic pipeline of sensor data knowledge retrieval in DrHouse, followed by our motivations and the design of semantic-based retrieval filtering approach.

**Retrieval Pipeline.** The pipeline of sensor data retrieval in DrHouse contains two processes: raw sensor data retrieval and summarizing. Specifically, we first retrieve the raw context of the sensor data recordings based on the query posed by the virtual doctor, which is presented as  $c_{sensor} = \text{Retriever}(q_{doctor}, \mathcal{D}_{sensor})$ , where  $q_{doctor}$  is the query (question) of the virtual doctor, and  $\mathcal{D}_{sensor}$  is the sensor data knowledge base. Subsequently, we employ a summary LLM called  $\text{LLM}_{sum}(\cdot)$  to interpret and summarize the raw context  $c_{sensor}$ , which can be expressed as  $d_{sensor} = \text{LLM}_{sum}(q_{doctor}, c_{sensor})$ . The summary output  $d_{sensor}$  is more readable and better suited for DrHouse than the raw context. Given the costs, we use GPT-3.5 as  $\text{LLM}_{sum}$  in this study. Next, we will introduce the motivation of semantic-based retrieval filtering and how it works in the pipeline.

**Observations and Motivation.** Figure 5a shows that the outputs of LLM doctors can contain very complex and diverse actions, such as: "Are you experiencing any other symptoms such as fever, wheezing, or difficulty breathing? If so, can you also tell me your current temperature, pulse rate, and whether you've had any recent exposure to someone with similar symptoms or a respiratory infection". Keywords related to sensor data indicators





(a) An example of inaccurate sensor data retrieval.

(b) An example of inefficient sensor data retrieval.

Fig. 5. Examples of inaccurate and inefficient sensor data knowledge retrieval. The words highlighted in red represent the results of sensor data knowledge retrieval. Note that we use GPT-4 as a naive LLM for example.

can be overwhelmed by a large number of words, making sensor data retrieval challenging. As mentioned above, Retriever requires a predefined threshold to determine whether the retrieved contexts should be used. A higher threshold can obscure or overwhelm the keywords for retrieving the sensor database, leading to retrieval failure. On the other hand, shown in Figure 5b, a lower threshold could lead to many redundant retrievals, thereby increasing system overhead. Figure 12 shows the overhead of sensor data retrieval. Since the process of sensor data knowledge retrieval contains both data retrieval and LLM summarizing, retrieving sensor knowledge in each round of diagnosis will bring a large overhead, including both latency and cost.

**Semantic-Based Retrieval Filtering.** The above observations show that not all queries necessitate the retrieval of sensor data. To this end, we develop a semantic-based retrieval filtering mechanism to determine whether to initiate sensor data retrieval. One straightforward way is to employ LLM itself for this determination. However, the prompt for DrHouse already contains a long context, which can make it challenging for LLMs to follow instructions accurately [68]. In addition, existing studies [7] show that fine-tuned small language models with application-specific training data consistently achieve superior performance compared to zero-shot LLMs in all tested classification tasks. Our experimental results in § 5.3.3 show the same phenomenon.

To address these challenges, we develop a semantic-based retrieval filtering mechanism that employs a small language model as a semantic filter. The goal of the semantic-based retrieval filtering is to categorize inquiries posed by LLMs and determine whether the retrieval of sensor data is necessary. Specifically, we first collect queries from our virtual doctor platform to construct a dataset. This dataset contains diverse questions posed by our developed LLM virtual doctor, including those that should initiate the retrieval of sensor data and those that should not. We manually annotated each query with labels (i.e., whether to start retrieval of sensor data or not). Recent studies [51] have demonstrated the effectiveness of LLMs for data augmentation. To mitigate the annotation cost, we use LLMs to rewrite queries and expand the diversity of descriptions in our collected dataset, while keeping the label of the query unchanged. Finally, we use the augmented dataset and cross-entropy loss to train the semantic filter.

**4.3.3 Medical Knowledge Retrieval.** DrHouse's retrieval of medical expert knowledge contains two stages: preceding retrieval and runtime retrieval. The preceding stage enables LLMs with the capabilities of multi-turn medical conversations, while the runtime stage empowers LLMs with diverse medical knowledge.

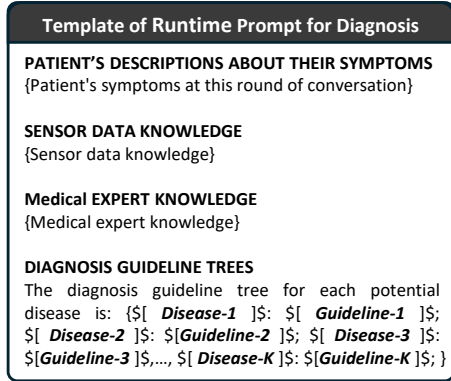


Fig. 6. Template of the runtime prompt.

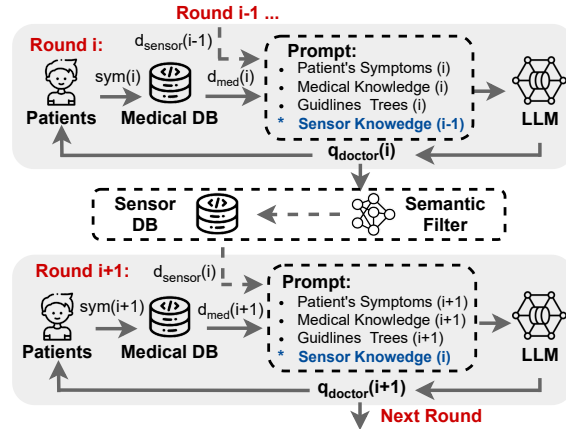


Fig. 7. Pipeline of DrHouse for knowledge update at runtime.

**Preceding Retrieval.** DrHouse retrieves relevant dialogue demonstrations based on the symptoms described by patients during the first round of conversation:  $d_{\text{dlg}} = \text{Retriever}(\text{sym}, \mathcal{D}_{\text{dlg}})$ , where  $\mathcal{D}_{\text{dlg}}$  is the dialogue dataset [30, 67]. These demonstrations cover multi-turn dialogues of diverse diseases. By incorporating this information into the prompt, DrHouse can have the capability to perform multi-turn medical consultations. The performance of LLMs will be enhanced with an increase in the number of demonstrations and their relevance to the patient's descriptions. However, due to the limited processing context of LLMs [29], DrHouse retrieves only the top- $k$  similar dialogues from the vector database and incorporates them into the prompt. Considering the context length and cost,  $k$  is set to 3 in our experiments.

**Runtime Retrieval.** At each round of conversation, DrHouse retrieves the medical knowledge based on the patient's reported symptoms. This process enables DrHouse to provide a more professional reply and mitigate LLM hallucinations [21, 51]. The retrieval of medical knowledge can be expressed as  $d_{\text{med}} = \text{Retriever}(\text{sym}, \mathcal{D}_{\text{med}})$ , where  $\mathcal{D}_{\text{med}}$  is the medical textbooks database. Finally,  $d_{\text{med}}$  is integrated with the patient's symptoms  $\text{sym}$  and the knowledge from sensor data  $d_{\text{sensor}}$ , and inputted to the LLM for diagnostic decision-making.

#### 4.4 Knowledge Integration in DrHouse

Existing studies either adapt LLMs to the medical domain [5, 41, 69], or utilize LLMs to interpret sensor signals for understanding the physical world [16, 22, 54]. How to fuse the knowledge of sensor data from patients' smart devices to assist LLM-based virtual doctors in multi-turn diagnosis is challenging. In this section, we introduce the prompts used in DrHouse, which empower LLMs with the capacity to initiate multi-turn medical consultations and utilize the sensor data from patients' smart devices. The prompt utilized in DrHouse consists of two parts:  $\text{Prompt} = \text{Prompt}_{\text{preceding}} + \text{Prompt}_{\text{runtime}}$ , where  $\text{Prompt}_{\text{preceding}}$  and  $\text{Prompt}_{\text{runtime}}$  are the preceding prompt and the runtime prompt. The former remains fixed once the multi-turn diagnosis commences, while the content of the latter evolves throughout the multi-turn conversation.

**4.4.1 Preceding Prompt.** The preceding prompt aims to invoke the LLM's capability of multi-turn consultations and provide initial diagnostic guidelines. Figure 8 shows the template of the preceding prompt. It contains four parts: overall instruction, task instruction, retrieved diagnosis guidelines, and retrieved dialogue demonstrations. Overall instruction prompts the LLM to play the role of a virtual doctor for a multi-turn diagnosis. Task instruction emphasizes critical points for the diagnosis, including following diagnosis guidelines, meticulous evaluation of sensor data indicators, and providing instructions for multi-disease diagnosis. Diagnosis guidelines and dialogue

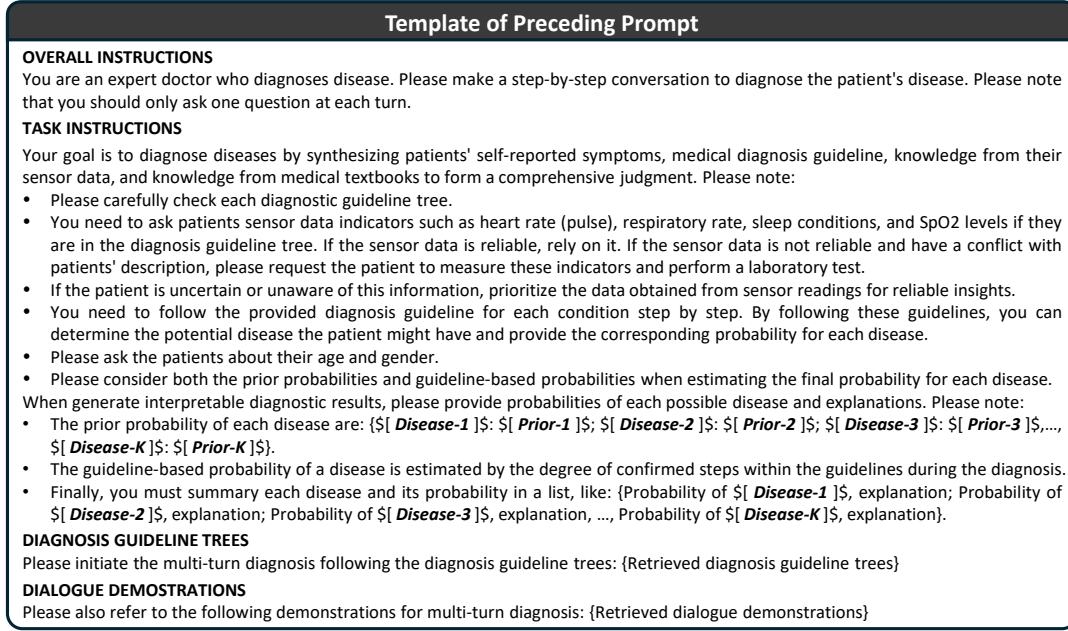


Fig. 8. Template of the preceding prompt in DrHouse.

demonstrations are retrieved from our constructed medical knowledge based on the patient's reported symptoms (see § 4.3.1). Once the patient initially reports their symptoms, the content within the  $Prompt_{preceding}$  remains fixed and does not change throughout the subsequent conversation rounds.

**4.4.2 Runtime Prompt for Diagnosis.** Figure 6 shows the template of the runtime prompt in DrHouse. The content within the  $Prompt_{runtime}$  changes with each round of conversation. Figure 7 shows the pipeline of DrHouse for updating its knowledge at runtime. At each round of consultation, the input to the LLM contains four parts:  $Prompt_{runtime} = sym(i) + d_{med}(i) + d_{tree}(i) + d_{sensor}(i - 1)$ , where  $sym(i)$  is the patient's description of their symptoms at the  $i$ th round of the conversation.  $d_{tree}(i)$  and  $d_{med}(i)$  are the retrieved diagnostic guidelines and medical knowledge based on  $sym(i)$ .  $d_{sensor}(i - 1)$  refers to the retrieved sensor data knowledge based on DrHouse's query (question) posed by the LLMs in the previous round of conversation. Note that sensor data is available only when the semantic filter initiates data retrieval; otherwise, it is set to empty (see § 4.3.2).

## 4.5 Diagnostic Decision Making

Patients' descriptions may be influenced by subjective perceptions or memory biases [34, 36], while sensor data can be affected by environmental factors [25], posing challenges for LLM to make diagnostic decisions. To address this challenge, we develop a diagnostic decision-making strategy for DrHouse that performs knowledge selection and provides explainable diagnosis results.

**4.5.1 Knowledge Selection.** The decision space of DrHouse encompasses four types, including further inquiring about symptoms, requesting in-lab tests, accessing patients' sensor knowledge base, and summarizing diagnosis results. In cases where diseases lack sensor data indicators, DrHouse relies on patients' symptoms for decision-making. Conversely, for diseases where guidelines consider sensor data indicators crucial for diagnosis, DrHouse relies on the knowledge retrieved from sensor data. To mitigate the influence of sensor data quality on

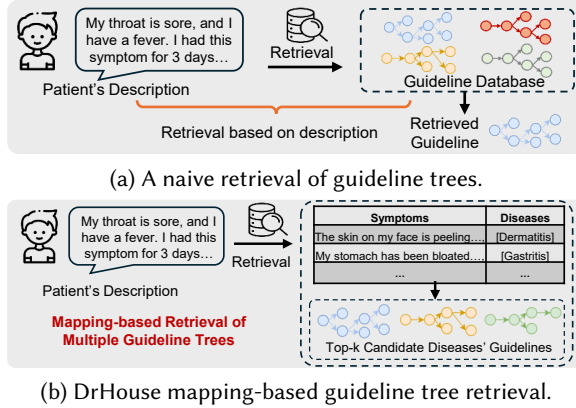


Fig. 9. Comparing guideline tree retrieval of DrHouse and MedDM.

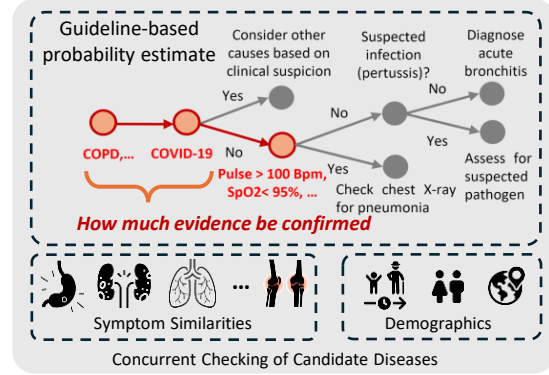


Fig. 10. Key idea of concurrent checking of candidate diseases in DrHouse.

the diagnosis, we include instructions in the prompt for DrHouse to consider the uncertainty level of the retrieved sensor data during decision-making. Specifically, we incorporate the instruction “If the sensor data is reliable, rely on it, else please request the patient to measure these indicators and perform in-lab tests.” into the prompt, which can be seen in Figure 8. This instruction guides DrHouse to trust the retrieved sensor data knowledge only when it possesses a low uncertainty level. Otherwise, DrHouse will request the patient to perform in-lab tests to ensure a reliable diagnosis.

**4.5.2 Concurrent Checking of Candidate Diseases.** Despite the substantial advancements in AI technology, doctors continue to prioritize relying on their own judgment when making final decisions due to concerns related to safety and liability [36]. To this end, we enable DrHouse to concurrently check each candidate disease and generate interpretable diagnostic results, including probability estimates and explanations for each candidate disease, which physicians can use as a reference.

Figure 10 shows the key idea of DrHouse’s concurrent checking of candidate diseases. The final probability for each disease is determined by the LLM leveraging the information of prior probability and the guideline-based probability. The prior probability is determined after the initial round of the diagnostic conversation (preceding stage), while the guideline-based probability dynamically evolves as the conversation progresses.

**Prior probability.** Based on the patient’s self-reported symptoms, we first retrieve the top-k most relevant diseases from a pre-collected symptom-disease dataset (see § 4.3.1). For each disease, the retrieval similarity is referred to as symptom similarity. Given that treatment guidelines vary by age and physicians recommend considering age and gender in virtual doctor dialogues [47], we incorporate disease demographics into our system. Demographic probabilities can be retrieved from disease distribution statistics and integrated into the prompts.

**Guideline-based probability.** As the diagnosis advances and additional symptoms and medical test results are obtained, the probability of the patient having a particular disease changes. This probability is denoted as the guideline-based probability in this study. Confirmation of disease requires ensuring that the patient’s symptoms meet the criteria at all branching points of the diagnostic tree until reaching the leaf node at the end of the diagnostic guide tree [3]. Figure 10 illustrates an example of guideline-based probability estimation for acute bronchitis, where red points indicate confirmed steps based on the patient’s symptoms, and grey points indicate a lack of evidence, i.e., symptoms not yet confirmed. The key idea of estimating guideline-based probability is to calculate the ratio of confirmed steps in the diagnostic guideline tree to the total depth of the tree. Since we have incorporated the medical diagnosis guidelines in DrHouse, we prompt LLM to inquire about the patient’s

condition following the diagnosis guidelines of each disease. The prompt also includes instructions that enable the LLM to consider the degree of confirmed evidence outlined in the guidelines during the diagnosis, thereby obtaining a guideline-based probability.

Incorporating both the prior probability and guideline-based probability of each disease, DrHouse can progressively narrow down diseases with low likelihood during each round of the diagnostic conversation. Additionally, we prompt the LLM to provide the final probability estimation and explanations for each disease based on prior and guideline-based probabilities. Details of the prompt setting can be seen in Figure 8.

## 5 Evaluation

In this section, we first introduce the experiment settings and datasets used for evaluation. Next, we introduce the performance metrics and the baseline methods used for comparison. Subsequently, we conduct both simulation and real-world experiments to validate DrHouse. Finally, we conduct a user study involving both test subjects and medical experts, followed by a comprehensive analysis through their feedback.

### 5.1 Evaluation Setup

To the best of our knowledge, no public dataset contains both diagnostic conversations and sensor data from smart devices. To validate our system, we conduct three types of experiments, including simulation experiments, real-world experiments, and user studies.

**5.1.1 Evaluation on Synthetic Datasets.** Our synthetic datasets are based on public medical dialogue datasets [15, 28, 30]. We will first introduce the three public dialogue datasets, followed by an explanation of how the synthetic dataset is constructed and the experimental settings.

**Medical Dialogue Datasets.** Many studies on medical LLMs employ medical knowledge question answering (QA) tasks [20, 49] or multiple choice tasks [41] for evaluation, which is inconsistent in real medical consultation scenarios. In this work, we use three public medical dialogue datasets for evaluation, which contain real patient descriptions of their symptoms and doctor-patient conversations.

- **DIALMED** [15] dataset comprises medical dialogues between patients and doctors from three departments, including respiratory, dermatology, and gastroenterology. Each dialogue includes a ground truth label indicating the disease diagnosed by the doctor.
- **MedDG** [30] dataset comprises 17,864 multi-round medical dialogues. It encompasses twelve prevalent gastrointestinal disorders. These dialogues are gathered from the gastroenterology section of a web-based medical consultation platform.
- **KaMed** [28] is a large-scale multi-round medical dialogue dataset encompassing more than 63K dialogues from over one hundred hospital departments.

**Synthetic Dataset Construction.** We construct a synthetic dataset based on each dialogue dataset. Each sample in the dialogue dataset contains descriptions of patients' symptoms and corresponding disease labels. Recent LLM-based methods have been proven to be able to produce reliable simulations of sensor data [35] and tabular data [6]. Since sensor data captured by smart devices is also formatted as tabular data (e.g., the public wearable sensor dataset, LifeSnaps [65]), we refer to [6, 35] for sensor data simulation. Specifically, for each sample in the dialogue dataset, we generate simulated sensor data according to the disease associated with the sample. We consider **two factors** to ensure the validity of the simulated sensor data. First, we ensure the **validity of the sensor data ranges**. Specifically, for each sample, we use LLMs to generate simulated sensor data based on the disease type and the reasonable sensor data range for this disease as documented in medical literature [12] (e.g., the  $SpO_2 < 95\%$ , respiratory rate  $> 22$  bpm, and pulse  $> 100$  bpm for pneumonia). We then incorporate this medical knowledge into the prompt to guide the LLM in generating sensor data within a reasonable range for that disease. Second, we ensure the **validity of the sensor data formats**. We refer to the public wearable sensor



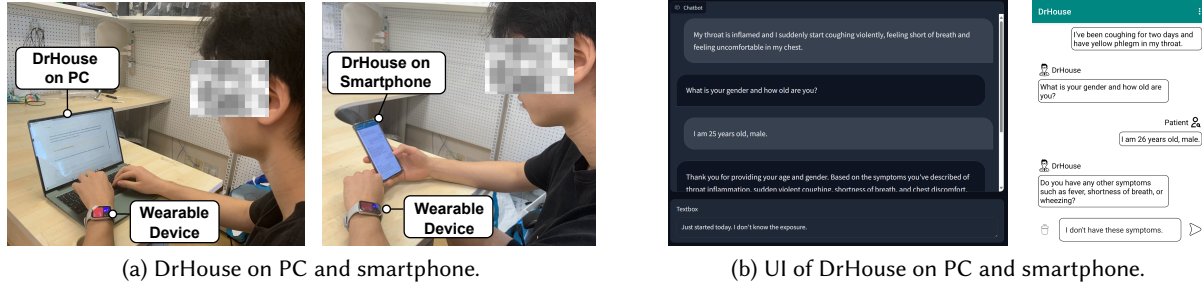


Fig. 11. Real-world evaluation settings. Participants engage in multi-turn diagnosis interactions using DrHouse on their mobile phones or personal computers.

dataset, LifeSnap [65], which provides data formats from off-the-shelf wearable devices. The dataset includes a wide range of daily physiological indicators such as SpO<sub>2</sub>, sleep scores, heart rate, and stress levels, all formatted as tabular data. By incorporating several samples from LifeSnap into the LLM prompts as examples, we ensure the validity of the simulated sensor data formats.

**5.1.2 Evaluation on Real-world Profiles.** To further validate our proposed system, we conducted real-world experiments and user studies.

**Real-world Profiles.** We recruited 12 test subjects who have experienced illnesses in the past 12 months and consistently wear Apple Watches in their daily lives. Their symptoms and wearable sensor data are strictly paired and originate from the same individual. Apple Watch continuously records data via the Health app on iPhone devices. With their permission, we extracted their historical sensor data during their illness periods as they continued to wear the smart devices. Disease types of test subjects include acute bronchitis, chondromalacia, pneumonia, influenza, and dermatitis. Each participant engages in multi-turn diagnosis interactions using DrHouse and other baseline approaches. All participants have consent to the data collection and study of this project, which has been approved by the authors' IRB.

**User Study.** To evaluate the user experience, we conducted a user study involving two types of DrHouse users, including 12 test subjects and 20 medical experts. Please refer to § 5.4 for details.

**5.1.3 Performance Metrics.** It is crucial to assess the virtual doctor's performance throughout the entire multi-turn conversations during the diagnosis, rather than solely focusing on the performance of single-turn QA. However, most existing medical LLMs are evaluated in a non-interactive manner such as single-turn QA and summary tasks [11]. In this study, we propose a comprehensive evaluation criteria that encompasses three dimensions:

- **Compliance.** The virtual doctor's adherence to diagnostic guidelines during diagnosis. We use *compliance* to validate the LLM's instruction-following performance for diagnostic guidelines.
- **Sensor Data Utilization.** The degree to which the knowledge from sensor data contributes to the virtual doctor's diagnostic process. We use *sensor data utilization* to validate the LLM's instruction-following performance for integration of sensor data.
- **Accuracy.** Consistency between virtual doctor's diagnosis results and the ground truth disease label.

Guided by this criterion, we conduct evaluations using LLMs and expert manual assessment, respectively.

**GPT Evaluation.** Recent studies have proven the feasibility of using LLMs for evaluation [11]. Figure 13 shows the prompt used for GPT evaluation. We incorporate the ground truth disease label, the corresponding disease guideline, and the dialogue into the prompt and input it into GPT-4. Subsequently, GPT-4 generates scores for individual dimensions and an overall score, which we refer to as **GPT-score** in this study. We also prompt GPT-4 to generate scoring explanations. Figure 16 shows an example of an explanation from GPT-4 scoring a dialogue.

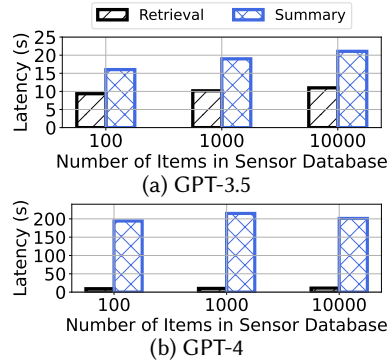


Fig. 12. Overhead of sensor data retrieval during multi-turn consultations.

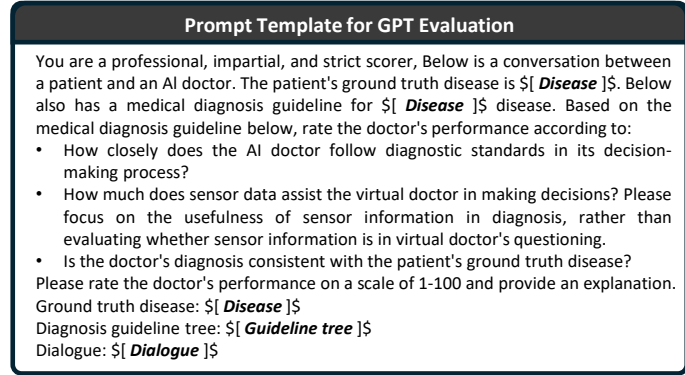


Fig. 13. Prompt template used for GPT evaluation.

**Expert's Manual Evaluation.** We recruit 20 medical experts to manually assess the results using the same evaluation criteria. We provide the scorer with the dialogue, the ground truth disease label, and the corresponding disease guideline. Scorers are required to evaluate the dialogue based on the above criteria, completing the questionnaire. For details about the results of the manual evaluation, please refer to the user study in §5.4.

**5.1.4 Baseline.** In our experiments, we compare the diagnosis performance of DrHouse with the following baselines, including the state-of-the-art knowledge retrieval approaches and SFT approaches for LLM-based virtual doctors. Since many works on LLM-based virtual doctors do not open-source their code or model weights, such as Med-PaLM 2 [41], they are not used in our experiments.

- **GPT-3.5** [64]. We utilize the GPT-3.5-Turbo version in our experiment, which is one of the most widely used models in GPT-3.5 family. We utilize the API provided by OpenAI.
- **GPT-4** [2]. This is the largest LLM provided by OpenAI. We utilize the GPT-4-1106-preview version of GPT-4.
- **Llama-3-70B** [43]. This is the latest and largest version of Llama series LLMs provided by Meta. We employ the Llama-3-70-Instruct version.
- **HuatuoGPT-II** [69]. This is one of the state-of-the-art SFT approaches for LLM-based virtual doctors. HuatuoGPT-II is uses Baichuan-Base [57] as the base LLM. It is fine-tuned by SFT and reinforced learning from AI feedback. We utilize the HuatuoGPT-II-34B version in the experiments.
- **MedDM** [26]. This is one of the state-of-the-art knowledge retrieval approaches for LLM-based virtual doctors. MedDM retrieves the most relevant diagnostic guideline trees based on the patient's descriptions during multi-turn diagnoses. The original MedDM employs GPT-3.5 as the base LLM. To provide a fair comparison, we adopt the more powerful GPT-4 as the base LLM in the experiments.

**5.1.5 Implement Details.** The prompts for the baseline approach GPT-3.5, GPT-4, and Llama-3-70B are identical, each containing overall instructions and retrieved medical dialogues. To ensure a fair comparison, these two sections are consistent with DrHouse. MedDM adopts the same two sections in its prompt and also incorporates the retrieved diagnosis guideline trees. We directly test HuatuoGPT-II without adding additional prompts.

For the task of evaluating adaptive sensor data retrieval, the ground truth is sourced from our virtual doctor platform, with each query manually annotated to indicate whether sensor data retrieval should start. We split the collected data into 60%, 20%, and 20% for training, validation, and testing, respectively. We use BERT [9] and its variants [31] as the semantic filter, and use cross-entropy loss for training. We use an hour-wise sliding window in sensor data uncertainty checking. We use the Adam optimizer with a learning rate of  $1 \times 10^{-5}$  and a weight

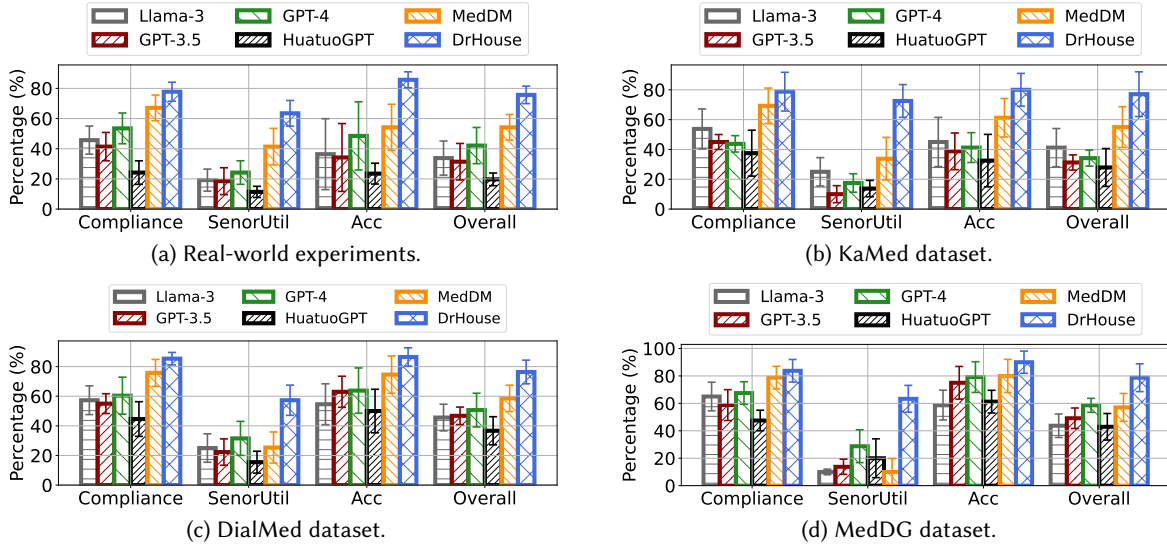


Fig. 14. Overall diagnostic performance. (a) shows the performance on real-world test subjects and (b)~(c) represents the performance on synthetic datasets. X-axis: Compliance means adherence to diagnosis guidelines. SensorUtil is the sensor data utilization score. Acc is the diagnostic accuracy of disease. Overall is the average score of the three criteria.

decay of  $1 \times 10^{-7}$ . For the tasks of evaluating diagnostic performance and guideline retrieval, the ground truth is derived from the disease labels in the dialogue datasets.

For simulation experiments, we conduct multi-turn consultations with different virtual doctors based on the symptoms provided in the dialogue datasets. During interactions, we strictly adhere to the symptoms described in the datasets to answer the questions posed by the virtual doctors, avoiding any subjective preferences or descriptions. If the dataset lacks information relevant to the questions posed, we respond with “I don’t know”. We randomly select 50 samples from the dialogue datasets for experiments. We obtain multi-turn dialogues through consultations between humans and virtual doctors and evaluate them using the GPT-score.

## 5.2 Overall Performance

This section shows the main experimental results of DrHouse from both simulation and real-world experiments.

**5.2.1 Quantitative Results.** We first compare DrHouse with baselines using quantitative diagnosis metrics, including *compliance*, *sensor data utilization*, and *accuracy*. Subsequently, we conduct an in-depth analysis of the diagnostic performance for different disease categories.

**Overall Performance of Diagnosis.** Figure 14 shows the quantitative diagnosis performance of DrHouse and baselines. DrHouse achieves state-of-the-art results across all datasets, validating its effectiveness in integrating sensor data and up-to-date medical knowledge into diagnosis. Results show that DrHouse can achieve up to 31.5% *accuracy* improvement than the top-performing baselines in the real-world experiment, and also achieve up to 18.7% higher *accuracy* than the best baseline approaches in the simulation experiments. Additionally, results show that DrHouse achieves up to 38.8% higher *sensor data utilization* and 10.7% higher *compliance* than the best baseline approaches, validating its effectiveness in integrating sensor data into medical diagnoses and following diagnostic guidelines, respectively. Note that while baselines do not incorporate sensor data, their *sensor data utilization* scores are not zero. This is because GPT-4’s scoring mechanism assigns a positive score if the virtual doctor inquires about physiological indicators from sensors. However, since baselines cannot access the sensor

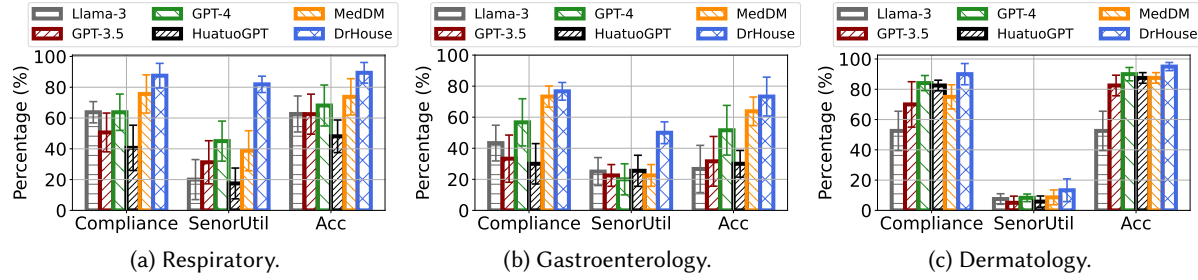


Fig. 15. Diagnostic performance across different types of diseases on DialMed dataset.

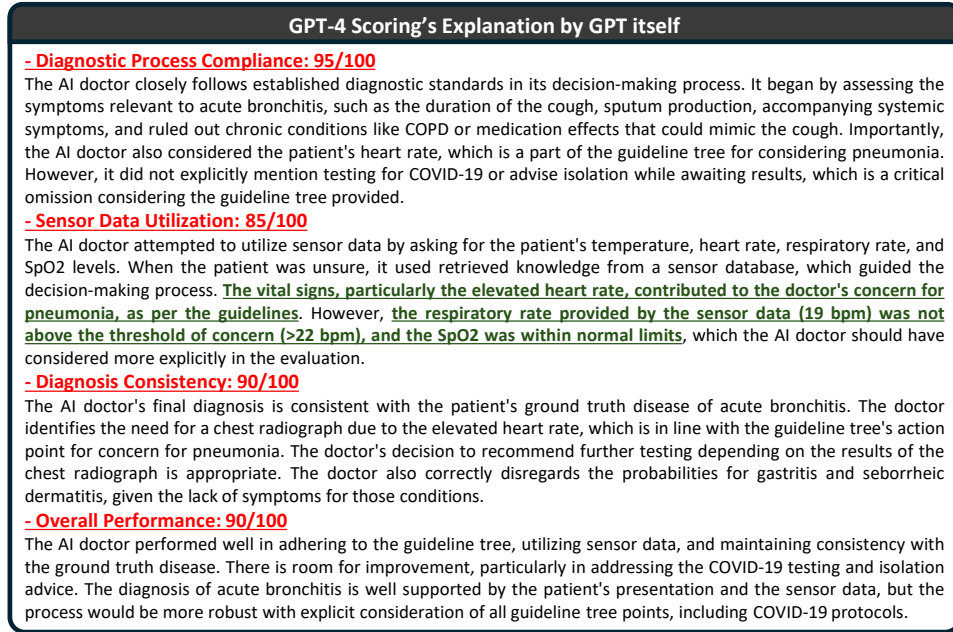


Fig. 16. An example of GPT-4 scoring's explanation by GPT itself. The words highlighted in red and green represent the GPT-score and the explanation for the sensor data utilization metric, respectively.

database, sensor data does not contribute to the diagnosis, resulting in significantly lower scores compared to DrHouse.

**Diagnosis Performance Across Different Diseases.** We further evaluate DrHouse and baseline methods with a finer granularity: assessing diagnostic performance across various disease categories. Figure 15 shows the diagnosis performance on the DialMed dataset, including respiratory, dermatology, and gastroenterology diseases. As shown in Figure 15, DrHouse achieves 36.8% and 24.5% higher *sensor data utilization* than baselines on respiratory and gastroenterology diseases, respectively, thereby enabling DrHouse to achieve the highest accuracy compared to baseline approaches. However, when diagnosing dermatology diseases, nearly all virtual doctors exhibit poor performance in terms of *sensor data utilization* scores. This is because the diagnosis of dermatology diseases is less relevant to the physiological indicators that can be collected from smart devices, while respiratory and gastroenterology diseases are more associated with these indicators.

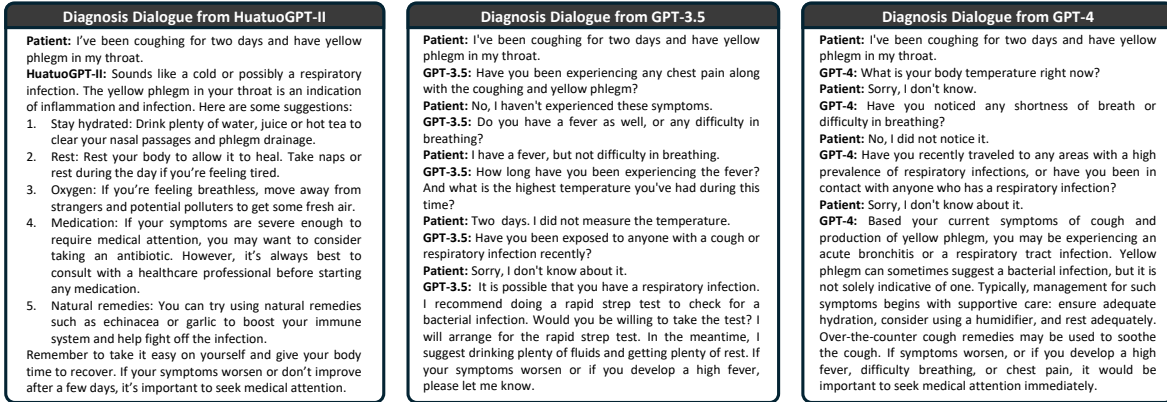


Fig. 17. Example diagnosis dialogues of baseline LLM-based virtual doctors. The dialogues from left to right are from HuatuoGPT-II, GPT-3.5, and GPT-4, respectively.

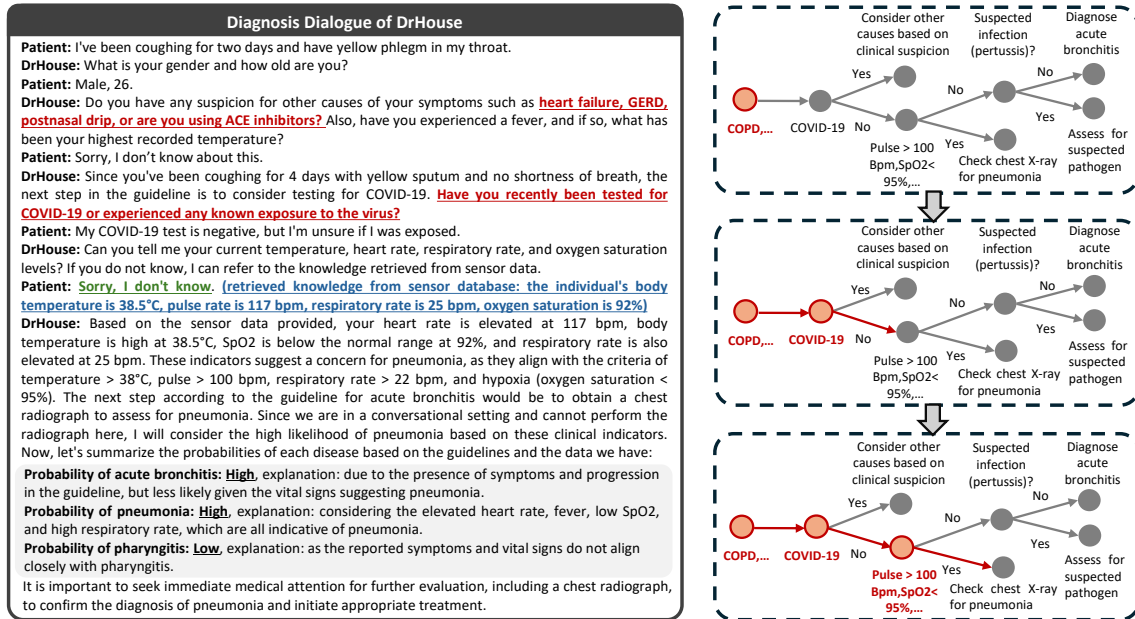


Fig. 18. An example diagnosis dialogue of DrHouse. Words highlighted in blue are the retrieved knowledge from the test subject's sensor data. The right part shows the simplified diagnostic guideline trees.

**GPT-score Explanation.** To demonstrate the rationale behind the GPT-score, we provide an example to showcase the GPT-4 scoring's explanation by GPT itself. Figure 16 shows an example of calculating the GPT-score for diagnostic dialogue using GPT-4. When provided with the dialogue, ground-truth disease label, and corresponding diagnosis guideline, GPT-4 is capable of generating reasonable scores and explanations. Notably, it can determine the extent to which sensor data contributes to the diagnosis of the disease according to the guideline, thus providing a reliable score of *sensor data' utilization*. For example, as shown in Figure 16, the respiration and the elevated heart rate contribute to the doctor's concern for pneumonia, thus GPT-4 gives a higher score.



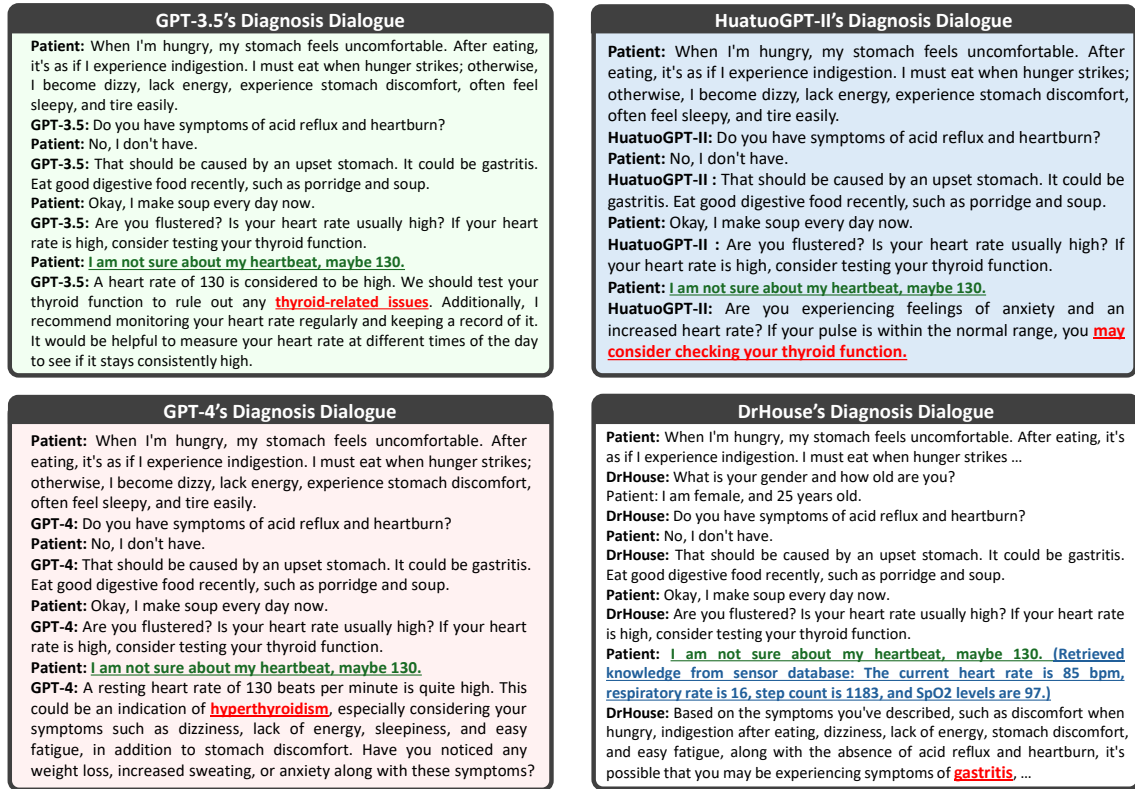


Fig. 19. An example of showing the effectiveness of DrHouse in avoiding misdiagnosis by incorporating sensor data.

**5.2.2 Qualitative Results.** To better demonstrate the performance of DrHouse, we present examples of diagnosis dialogues. The key observations from these dialogues are summarized as follows:

**Observation 1: DrHouse can follow up-to-date medical guidelines for diagnosis.** Figure 17 shows that existing LLM-based virtual doctors often have too autonomy in their questioning process during diagnosis, leading to a failure to adhere to the latest. HuatuoGPT-II made diagnostic conclusions and recommendations directly based on the patient's self-reported symptoms, without asking the patient for more symptoms or performing laboratory tests. The left and right parts of Figure 18 show the diagnostic dialogue from DrHouse and the simplified guideline trees, respectively. Results validate that DrHouse exhibits reliable instruction-following performance and faithfully adheres to the guideline trees for diagnosis.

**Observation 2: DrHouse can successfully leverage sensor data knowledge to assist in diagnostic decision-making and advance the diagnostic process further.** Figure 18 shows that DrHouse can proactively initiate questions about disease-related physiological indicators that can be obtained from the patient's smart devices. Additionally, results show that DrHouse can retrieve sensor data and effectively leverage this information for decision-making and advance the diagnostic process to another step forward, validating the LLM's instruction-following performance in integrating sensor data.

**Observation 3: Incorporating knowledge of sensor data from subjects' smart devices can reduce the risks of misdiagnosis.** Existing medical studies suggest numerous diseases exhibit shared symptoms [70]. For example, gastritis and hyperthyroidism have shared symptoms like increased hunger and weight loss. The physiological indicator of heart rate plays a vital role in distinguishing between these two diseases. To compare the performance

Decision-making on Patient's Reported Symptoms	Decision-making on Low Confidence Sensor Data	Decision-making on High Confidence Sensor Data
<p><b>Patient:</b> I have severe hair loss, my hair roots are wrapped with white stuff, and my scalp is itchy.</p> <p><b>DrHouse:</b> What is your gender and how old are you?</p> <p><b>Patient:</b> Male, 23.</p> <p><b>DrHouse:</b> Do you have any scaly, red skin patches on your scalp or other hairy areas of your body?</p> <p><b>Patient:</b> Yes, I have.</p> <p><b>DrHouse:</b> It is possible that you may have seborrheic dermatitis. How extensive these patches are, have you noticed any inflammation or lesions beyond the scalp area?</p> <p><b>Patient:</b> I don't have dandruff, but I have small white particles on my scalp.</p> <p><b>DrHouse:</b> It seems you have a moderate to severe seborrheic dermatitis. The next step in the treatment guideline would be to use an antifungal shampoo 2 to 3 times per week, along with a high-potency topical corticosteroid. You should assess the improvement after 4 weeks. If there is improvement, you would continue with the antifungal shampoo once or twice weekly to prevent relapse. Here is the current probability assessment for your condition:</p> <ul style="list-style-type: none"> <li>- Probability of seborrheic dermatitis: High, explanation: Based on your symptoms and the lack of improvement with over-the-counter treatments, which indicates a moderate to severe form of the condition.</li> <li>- Probability of esophagitis: Low, explanation: You have not mentioned any symptoms related to gastroesophageal reflux disease, such as heartburn or difficulty swallowing.</li> </ul>	<p><b>Patient:</b> I had a fever for four days, and the fever lasted for about four hours a day, and the rest of the time was normal.</p> <p><b>DrHouse:</b> What is your gender and how old are you?</p> <p><b>Patient:</b> Female, 24.</p> <p><b>DrHouse:</b> Do you have any other symptoms such as cough, sputum production, dyspnea, or wheeze?</p> <p><b>Patient:</b> I have yellow phlegm.</p> <p><b>DrHouse:</b> Have you experienced any shortness of breath or wheezing along with your cough and fever?</p> <p><b>Patient:</b> No, I don't have these symptoms.</p> <p><b>DrHouse:</b> Can you tell me your current temperature, heart rate, respiratory rate, and SpO2 levels?</p> <p><b>Patient:</b> Sorry, I don't know. (retrieved knowledge from sensor database: the current temperature is 37 degrees Celsius with a confidence score of 31%. The heart rate is 108 bpm with a confidence score of 23%. The respiratory rate is 19 breaths per minute with a confidence score of 26%. The SpO2 level is 98% with a confidence of 18%.)</p> <p><b>DrHouse:</b> Given that you are unsure of your heart rate, respiratory rate, and SpO2 levels, but the sensor data has a <b>low confidence score</b>, we will <b>request you to perform in-lab tests</b>. I recommend obtaining a <b>chest radiograph</b> to assess for pneumonia and pursue additional testing based on the results, severity of illness, and whether a specific pathogen is suspected. <b>Please proceed with getting a chest radiograph as the next step in your evaluation.</b></p>	<p><b>Patient:</b> I've been coughing for three or four days.</p> <p><b>DrHouse:</b> What is your gender and how old are you?</p> <p><b>Patient:</b> Male, 28.</p> <p><b>DrHouse:</b> Can you describe if you've experienced other symptoms such as fever, shortness of breath, or wheezing?</p> <p><b>Patient:</b> I have a fever. No shortness of breath.</p> <p><b>DrHouse:</b> Could you tell me your current pulse rate, respiratory rate, and if you've been experiencing any shortness of breath or a drop in oxygen saturation?</p> <p><b>Patient:</b> Sorry, I don't know. (retrieved knowledge from sensor database: Based on the given context, the person's heart rate is 90 bpm with a confidence score of 90%, and their respiratory rate is 20 breaths per minute with a confidence score of 88%. The SpO2 level is 97% with a confidence score of 95%.)</p> <p><b>DrHouse:</b> I would consider a clinical diagnosis of acute bronchitis. Your symptoms are consistent with this diagnosis: a persistent cough with sputum production and fever, without severe symptoms like high fever, rapid heart rate, or increased respiratory rate that would suggest pneumonia or other serious conditions.</p> <ul style="list-style-type: none"> <li>- Probability of acute bronchitis: High. Explanation: You have a persistent cough with yellow phlegm and a fever, which are symptoms of acute bronchitis. No shortness of breath or wheezing, and your COVID-19 test is negative.</li> <li>- Probability of gastritis: Low. Explanation: You have not reported any symptoms related to gastritis, such as loss of appetite, stomach pain, nausea, vomiting, or bleeding.</li> </ul>

Fig. 20. An example of DrHouse's decision-making considering the sensor data uncertainty. The words highlighted in red show that DrHouse considers the sensor data uncertainty when making decisions.

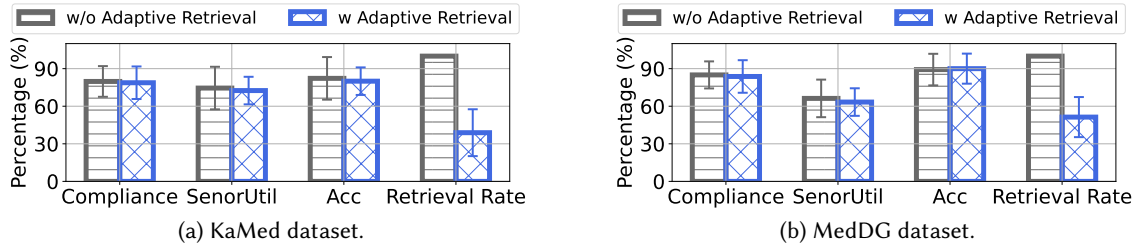


Fig. 21. Effectiveness of adaptive sensor data retrieval in DrHouse. “w” and “w/o” denote the use and non-use of adaptive sensor data retrieval, respectively. The retrieval rate is the ratio of retrievals to total diagnostic dialogue rounds.

of different approaches on patients' subjective descriptions, we use the same initial three rounds of dialogue, which contain queries about the patient's physiological indicators, and send these contexts to the virtual doctor simultaneously. Figure 19 shows that when the patient's subjective description is input, baseline approaches misdiagnose gastritis as hyperthyroidism. However, DrHouse can effectively avoid misdiagnosis by incorporating objective sensor data (highlighted in blue in Figure 19) into the diagnosis process.

**Observation 4: DrHouse incorporates sensor data uncertainty into its decision-making process.** Figure 20 shows an example of DrHouse's decision-making considering the sensor data uncertainty. The left figure illustrates how DrHouse leverages patients' descriptions for diagnosis when sensor data retrieval is unnecessary. The middle and right figures present two examples where DrHouse relies on sensor data knowledge for diagnostic decision-making. When sensor data uncertainty is low, DrHouse relies more on the knowledge from the sensor data for decision-making. Conversely, if the uncertainty level of the retrieved sensor data is high, DrHouse requests the subject to **perform an in-lab test** to ensure the accuracy of the diagnosis.

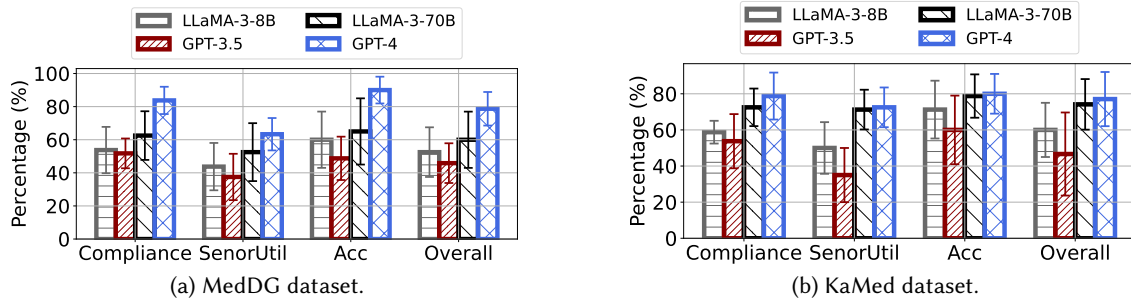


Fig. 22. Overall performance of DrHouse on MedDG and KaMed datasets when using different base LLMs.

### 5.3 Evaluation of System Modules

In this section, we first perform an ablation study of DrHouse's individual modules including the effectiveness of adaptive sensor data retrieval filtering and mapping-based guideline tree retrieval approach. Next, we perform experiments of DrHouse under different parameter settings.

**5.3.1 Effectiveness of Adaptive Sensor Data Retrieval.** We first evaluate the effectiveness of adaptive sensor data retrieval in DrHouse. The retrieval rate is denoted as the ratio of retrievals to total diagnostic dialogue rounds. Figure 21 shows that the utilization of adaptive retrieval in DrHouse achieves 2.2x retrieval efficiency improvement on average while exhibiting negligible impact on the diagnosis performance metrics including *compliance*, *sensor data utilization*, and *accuracy*.

**5.3.2 Effectiveness of Mapping-based Guideline Tree Retrieval.** We further evaluate the effectiveness of mapping-based guideline retrieval. Figure 25 shows the diagnostic guideline retrieval accuracy of DrHouse and baseline method MedDM [26]. To our knowledge, MedDM is the only work that retrieves diagnostic guidelines in LLM-based virtual doctors for multiple-turn diagnosis. MedDM saves the guideline tree in the vector database and directly retrieves it based on patient's symptoms. We use MedDM with varying parameters as baselines. Specifically, the vector database of MedDM utilizes chunk sizes of 100, 200, 400, 800, 1000, and 2000, with a fixed overlapping length of 100. Results show that mapping-based guideline retrieval achieves up to 10.8% higher retrieval accuracy than the best baseline setting.

**5.3.3 Impact of Hyper-parameters.** In this subsection, we perform ablation study and analyze the performance of DrHouse under different parameter settings.

**Impact of Different Base LLMs.** We first evaluate the diagnostic performance of DrHouse using different base LLMs, including Llama-3-8b-Instruct, Llama-3-70b-Instruct, GPT-3.5-Turbo, and GPT-4-1106. Figure 22 shows that using the GPT-4-1106 as the base LLM in DrHouse yields the best diagnostic performance, achieving 13.7%, 8.5%, 13.8%, 9.5% higher *compliance*, *sensor data utilization*, *accuracy*, and average score than the Llama-3-70b-Instruct, respectively. It should be noted that Llama-3-8b-Instruct and Llama-3-70b-Instruct are open-source LLMs. They exhibit faster inference speed and can be deployed on the edge devices of patients, making them a promising choice as the base LLM in DrHouse.

**Effectiveness of Semantic Filter.** In this subsection, we compare the retrieval filtering accuracy of our semantic filter with two baselines: the LLM-based approach and the hard matching approach [53]. We utilize the RoBERTa-base model, trained on our augmented dataset, as the semantic filter. The LLM-based approach uses LLM itself to determine whether to initiate the retrieval of sensor data. We test the performance of GPT-3.5 and GPT-4 both under the zero-shot and few-shot settings. In the zero-shot setting, we include the task instructions in the prompt of the LLMs, following the approach used in the existing study [7], and adapt prompt instructions to

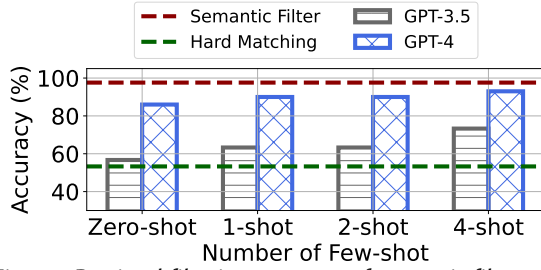


Fig. 23. Retrieval filtering accuracy of semantic filter, and two baselines, LLM-based approach and hard matching.

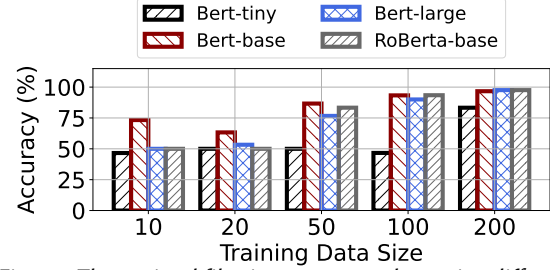


Fig. 24. The retrieval filtering accuracy when using different models as the semantic filter.

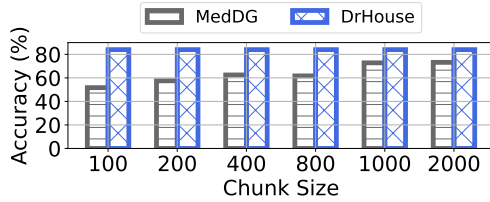


Fig. 25. Diagnosis guideline retrieval accuracy of DrHouse and baseline method MedDM. The vector database of MedDM utilizes chunk ranges from 100 to 2000.

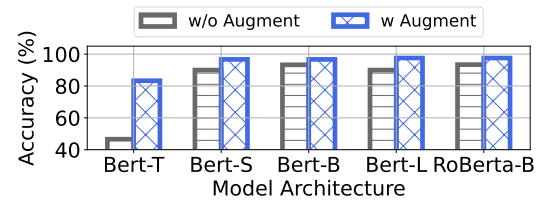


Fig. 26. Adaptive retrieval performance in multiple models. w and w/o denote using and not using GPT-4 for data augmentation, respectively.

suit our classification task. In the few-shot setting, we randomly select few-shot examples from the training set and incorporate them into the prompt of LLMs. Figure 23 shows that directly incorporating task instructions of sensor data retrieval determination into the prompt results in only 57% accuracy for GPT-3.5 and 86% accuracy for GPT-4, both of which are lower than the accuracy of the semantic filter. In addition, our approach can even achieve higher accuracy than 4-shot LLMs. Although adding more examples to the prompt of LLMs can improve their performance, an overly long prompt will increase the cost of LLM calls and reduce instruction-following performance [68]. Additionally, results show that the semantic filter achieves 43.3% higher accuracy compared to the hard matching approach.

**Effect of Different Models as Semantic Filter.** Next, we evaluate retrieval filtering accuracy using various BERT and RoBERTa models [31] as semantic filters. Figure 24 shows that Bert-base achieves 46.7% higher accuracy compared to Bert-tiny while it only demonstrates a 3.3% lower accuracy compared to Bert-large. However, the model size of Bert-base (110M) is 3 times smaller than Bert-large (340M). Therefore, Bert-base is used as the semantic filter, balancing accuracy and overhead. Figure 24 illustrates that with a training dataset of 200 samples, the Bert-base model achieves a 96.7% filtering accuracy, validating DrHouse's effectiveness in determining whether to retrieve sensor data.

**Effect of Data Augmentation.** We further evaluate the effectiveness of data augmentation in semantic-based retrieval filtering. Figure 26 shows that by employing GPT-4 to rewrite the queries of the LLM-based virtual doctor, there is a significant enhancement in the accuracy of retrieval filtering. The accuracy of Bert-tiny, Bert-small, and Bert-base increases by 36.7%, 6.7%, and 3.4%, respectively. Results show that our semantic filter can well understand DrHouse's questions and query sensor data on demand.

## 5.4 User Study

This section shows our user study of DrHouse, including real-world test subjects (using DrHouse for disease diagnosis) and medical experts (using DrHouse for diagnosis reference).

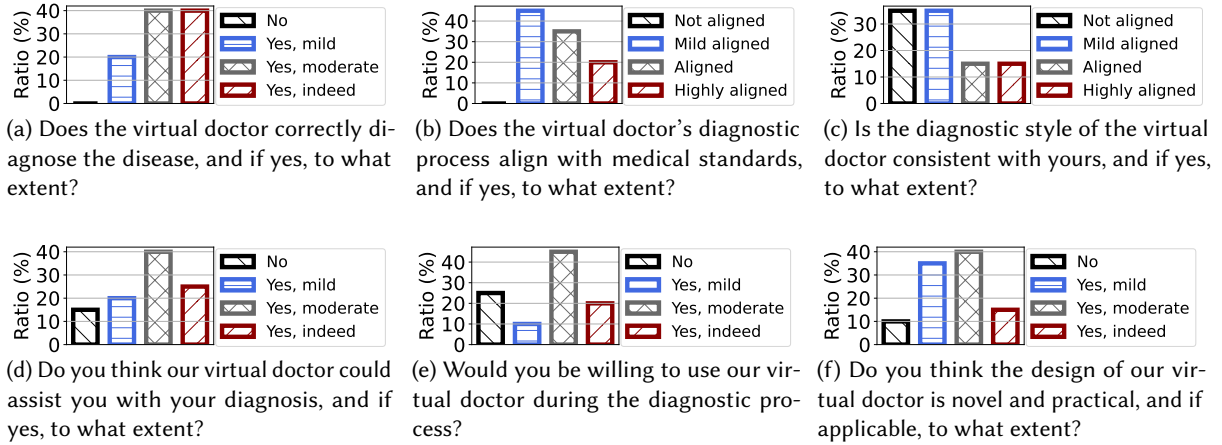


Fig. 27. DrHouse's user study for medical experts.

**5.4.1 User Study for Medical Experts.** We first conduct a user study for medical experts. We recruit 20 medical experts including clinicians and PhDs in medicine. The diagnosis dialogues between DrHouse and test subjects are presented to the medical experts for evaluation. We design a questionnaire comprising six questions and gather feedback from these medical experts on their experience using DrHouse. The questions are as follows:

- **Q1:** Does the virtual doctor correctly diagnose the diseases, and if yes, to what extent?
- **Q2:** Does the virtual doctor's diagnostic process align with medical standards, and if yes, to what extent?
- **Q3:** Is the diagnostic style of the virtual doctor consistent with yours, and if applicable, to what extent?
- **Q4:** Do you think our virtual doctor could assist you with your diagnosis, and if applicable, to what extent?
- **Q5:** Would you be willing to use our virtual doctor during the diagnostic process?
- **Q6:** Do you think the design of our virtual doctor is novel and practical?

Figure 27 shows the study results of medical experts. The feedback shows that 80% of medical experts find the diagnosis of DrHouse aligned or mild aligned with the standard diagnostic procedures, and 20% of medical experts think the diagnosis is highly aligned. In addition, 80% of medical experts think DrHouse correctly diagnoses the diseases. However, 35% of medical experts think that the diagnostic style of DrHouse is inconsistent with their own, which could be due to the varying diagnostic habits and styles among different clinicians. In addition, 85% of medical experts think DrHouse is beneficial for their diagnosis, and 75% of medical experts would like to use DrHouse during their diagnostic process. These experts agree that the diagnostic consultations between patients and DrHouse can provide valuable references and help reduce their workload. Besides, they believe the demographic information in the diagnostic dialogue of DrHouse is very useful as a reference for their diagnosis. In addition, 90% of the medical experts believe the design of DrHouse is novel and practical.

Overall, feedback from medical experts suggests that DrHouse's diagnoses are universally considered reliable. The medical experts express their willingness to utilize DrHouse as a pre-screening tool that offers valuable references, enhances diagnostic efficiency, and reduces overhead, showing its promising market potential.

**5.4.2 User Study for Real-world Test Subjects.** Twelve real-world test subjects who consistently wear Apple Watches in their daily lives participated in the user study. Each participant engages in multi-turn diagnosis interactions using DrHouse on their mobile phones or personal computers, as shown in Figure 11. This is followed by filling out a questionnaire about their user experience. The questions are as follows:

- **Q1:** Have you ever had an experience with virtual doctors for consultations before?
- **Q2:** Are you satisfied with the diagnosis provided by the virtual doctor, if yes, to what extent?



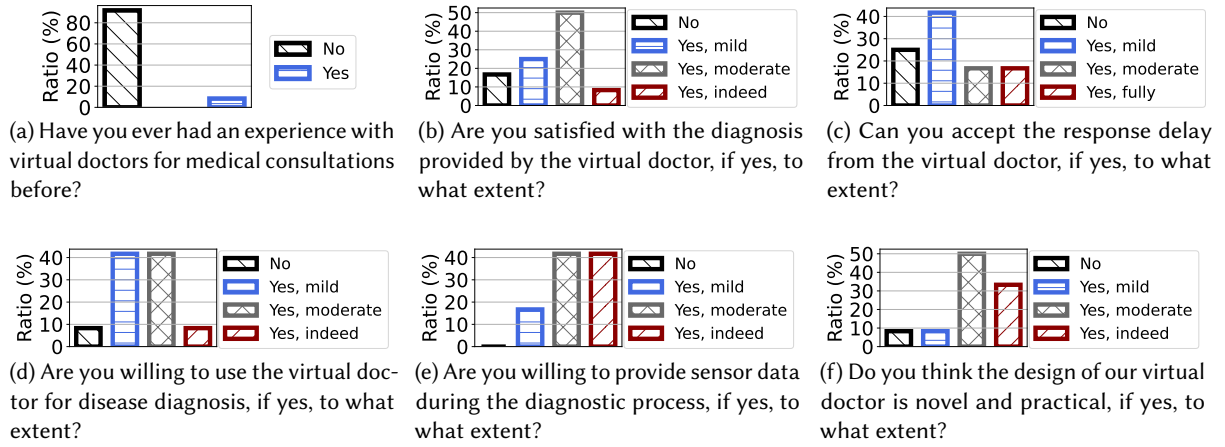


Fig. 28. DrHouse's user study for real-world test subjects.

- **Q3:** Can you accept the response delay of the virtual doctor, if yes, to what extent?
- **Q4:** Are you willing to use the virtual doctor for disease diagnosis, if yes, to what extent?
- **Q5:** Are you willing to provide sensor data, such as respiratory and heart rate, during the diagnostic process, if yes, to what extent?
- **Q6:** Do you think the design of our virtual doctor is novel and practical?

The results in Figure 28 show that only 8.3% of the participants have previously consulted virtual doctors. In addition, 83.4% of participants report satisfaction with the diagnoses provided by DrHouse. Regarding the response delay of DrHouse, 75% of the participants find it acceptable. Moreover, despite these delays, most participants find using DrHouse is much more convenient than visiting a hospital in person. Furthermore, 91.7% of the participants express willingness to use DrHouse for future diagnoses. When asked for access to personal sensor data, 83.4% of the participants are strongly or moderately willing to provide such data to DrHouse for diagnostic purposes. A key reason for the high willingness to share data is that most participants in the user study have already authorized smartphone apps to access this data. Consequently, they are generally comfortable providing this data to another application, particularly one that offers health consultation services. Additionally, 91.7% of the participants appreciate the innovative and practical design of DrHouse. Overall, feedback from participants in the user study indicates that DrHouse can provide reliable and satisfactory medical diagnoses from the patient's perspective. The substantial willingness of participants to share personal sensor data and to utilize DrHouse for future medical consultations demonstrates the promising practical applications of DrHouse.

## 6 Discussion and Limitations

In this section, we discuss the limitations and future directions of DrHouse.

**Integration of Additional Expert Knowledge.** DrHouse requires the retrieval of guideline trees to initiate multi-turn medical consultations. This process is impacted by the number of diagnosis guidelines incorporated in DrHouse and the number of cases included in the symptom-disease dataset [50]. Currently, DrHouse incorporates the diagnostic guidelines from the Up-to-Date database [47] for thirty common diseases. In the future, we plan to include additional diagnostic guidelines to support a wider range of diseases.

**Sensor Data Integration.** DrHouse focuses on integrating sensor data from wearable devices like smartwatches. The data is preprocessed by the algorithms within these devices and saved as text and tabular data. Enabling DrHouse to integrate sensor data from other modalities, such as images and IMU data, into diagnostic decision-making is a promising future direction. In addition, DrHouse incorporates the uncertainty of patients' sensor

data into its decision-making process. Integrating the latest approaches in anomaly detection [45] to enhance the system's ability to verify sensor data reliability remains our future work.

**Response Delay.** The base LLM in DrHouse is deployed on a cloud server, utilizing the API of the LLM service provider for inference. The latency of these API calls is unstable, as server congestion and instability can prolong the inference delay of the LLM. However, according to user study feedback (Figure 28), the majority of test subjects think the diagnostic delay of DrHouse is acceptable as it is much shorter than traveling to a hospital. In the future, we will consider deploying DrHouse's base LLM on mobile devices or employing edge-cloud collaboration [58] to reduce delays.

**Subjective Factors in Diagnosis.** Figure 27 shows that there are 35% of medical experts think the diagnostic style of DrHouse is not consistent with their own. This is because each doctor possesses a unique diagnostic style. Some doctors lean towards a conservative approach and adhere strictly to diagnostic guidelines, while others rely heavily on their own experience. Therefore, ensuring the consistency of DrHouse's diagnostic style with every doctor is challenging. On the other hand, diagnostic experiences tend to be highly subjective. Currently, DrHouse only relies on objective medical guidelines for diagnosis. Incorporating different diagnostic styles into DrHouse is a promising future direction.

**Other Medical LLMs as Base Models.** In this study, we perform experiments using different base LLMs in DrHouse, including Llama-3, GPT-3.5, and GPT-4. Leveraging existing medical LLMs as the base model can also be considered. Existing open-sourced medical LLMs like HuatuoGPT [69] and DISC-MedLLM [5] adapt the small LLMs to the medical domain through fine-tuning, compromising their generalization capability. We leave exploring other existing medical LLMs as the base LLM in DrHouse as our future work.

## 7 Conclusion

This paper proposes DrHouse, the first LLM-empowered virtual doctor system that incorporates patients's daily sensor data and expert knowledge for multi-turn diagnosis. DrHouse leverages the latest diagnostic guidelines to proactively initiate multi-turn diagnosis. During the multi-turn diagnosis process, DrHouse employs a multi-source knowledge retrieval approach to retrieve required sensor data and expert knowledge. In addition, DrHouse integrates the patient's descriptions and the two types of knowledge for diagnostic decision-making. We evaluate DrHouse on both synthetic and real-world datasets. Results show that DrHouse outperforms existing LLM-based virtual doctors by 31.5% diagnostic accuracy.

## Acknowledgments

This paper was supported in part by the Research Grants Council (RGC) of Hong Kong under GRF 14207123 and GRF 14212323, and the Alzheimer's Drug Discovery Foundation under Grant RDADB-201906-2019049. This research was also partially supported by the National Science Foundation under Grant Number CNS-1943396. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of The Chinese University of Hong Kong, Columbia University, NSF, or the U.S. Government or any of its agencies.

## References

- [1] 2022. PaddleOCR. <https://github.com/PaddlePaddle/PaddleOCR>.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [3] Jong-Myon Bae. 2014. The clinical decision analysis using decision tree. *Epidemiology and health* 36 (2014).
- [4] Mousumi Banerjee, Evan Reynolds, Hedvig B Andersson, and Brahmajee K Nallamothu. 2019. Tree-based analysis: a practical approach to create clinical decision-making tools. *Circulation: Cardiovascular Quality and Outcomes* 12, 5 (2019), e004879.
- [5] Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346* (2023).

- [6] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2022. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280* (2022).
- [7] Martin Juan José Bucher and Marco Martini. 2024. Fine-Tuned Small LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. *arXiv preprint arXiv:2406.08660* (2024).
- [8] Kaixin Chen, Yongzhi Huang, Yicong Chen, Haobin Zhong, Lihua Lin, Lu Wang, and Kaishun Wu. 2022. Lisee: A headphone that provides all-day assistance for blind and low-vision users to reach surrounding objects. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–30.
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [10] Zachary Englhardt, Chengqian Ma, Margaret E Morris, Xuhai Xu, Chun-Cheng Chang, Lianhui Qin, Xin Liu, Shwetak Patel, Vikram Iyer, et al. 2023. From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. *arXiv preprint arXiv:2311.13063* (2023).
- [11] Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. AI Hospital: Interactive Evaluation and Collaboration of LLMs as Intern Doctors for Clinical Diagnosis. *arXiv preprint arXiv:2402.09742* (2024).
- [12] Thomas M File, DJ Sexton, MD Aronson, and L Park. 2013. Acute bronchitis in adults. *UpToDate* (2013).
- [13] Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. LLMs accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*. PMLR, 82–100.
- [14] Google Corporation. [n. d.]. FitBit for Activity Trackers & More. <https://www.fitbit.com/global/us/home>
- [15] Zhenfeng He, Yuqiang Han, Zhenqiu Ouyang, Wei Gao, Hongxu Chen, Guandong Xu, and Jian Wu. 2022. DialMed: A Dataset for Dialogue-based Medication Recommendation. In *Proceedings of the 29th International Conference on Computational Linguistics*. 721–733.
- [16] Aritra Hota, Soumyajit Chatterjee, and Sandip Chakraborty. 2024. Evaluating Large Language Models as Virtual Annotators for Time-series Physical Sensing Data. *arXiv preprint arXiv:2403.01133* (2024).
- [17] Yongzhi Huang, Kaixin Chen, Yandao Huang, Lu Wang, and Kaishun Wu. 2021. A portable and convenient system for unknown liquid identification with smartphone vibration. *IEEE Transactions on Mobile Computing* 22, 4 (2021), 1894–1911.
- [18] Yongzhi Huang, Kaixin Chen, Yandao Huang, Lu Wang, and Kaishun Wu. 2021. Vi-liquid: unknown liquid identification with your smartphone vibration. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 174–187.
- [19] Sijie Ji, Xinzhe Zheng, and Chenshu Wu. 2024. Hargpt: Are llms zero-shot human activity recognizers? *arXiv preprint arXiv:2403.02727* (2024).
- [20] 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 (2021), 6421.
- [21] Mingyu Jin, Qinkai Yu, Chong Zhang, Dong Shu, Suiyuan Zhu, Mengnan Du, Yongfeng Zhang, and Yanda Meng. 2024. Health-LLM: Personalized Retrieval-Augmented Disease Prediction Model. *arXiv preprint arXiv:2402.00746* (2024).
- [22] Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024. Position Paper: What Can Large Language Models Tell Us about Time Series Analysis. *arXiv preprint arXiv:2402.02713* (2024).
- [23] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2567–2577.
- [24] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866* (2024).
- [25] Daniel Kuemper, Thorben Iggena, Ralf Toenjes, and Elke Pulvermueller. 2018. Valid. IoT: A framework for sensor data quality analysis and interpolation. In *Proceedings of the 9th ACM Multimedia Systems Conference*. 294–303.
- [26] Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. 2023. MedDM: LLM-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441* (2023).
- [27] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* 36 (2024).
- [28] Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten de Rijke. 2021. Semi-supervised variational reasoning for medical dialogue generation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 544–554.
- [29] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhui Chen. 2024. Long-context LLMs Struggle with Long In-context Learning. *arXiv preprint arXiv:2404.02060* (2024).
- [30] Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. MedDG: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 447–459.

- [31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [32] Sumit Majumder and M Jamal Deen. 2019. Smartphone sensors for health monitoring and diagnosis. *Sensors* 19, 9 (2019), 2164.
- [33] Izet Masic, Milan Miokovic, and Belma Muhamedagic. 2008. Evidence based medicine—new approaches and challenges. *Acta Informatica Medica* 16, 4 (2008), 219.
- [34] Ashley McKeane and David K Sherman. 2023. Diagnostic uncertainty in patients, parents, and physicians: a compensatory control theory perspective. *Health Psychology Review* 17, 3 (2023), 439–455.
- [35] Mike A Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, et al. 2024. Transforming wearable data into health insights using large language model agents. *arXiv preprint arXiv:2406.06464* (2024).
- [36] Ashley ND Meyer, Traber D Giardina, Lubna Khawaja, and Hardeep Singh. 2021. Patient and clinician experiences of uncertainty in the diagnostic process: current understanding and future directions. *Patient Education and Counseling* 104, 11 (2021), 2606–2615.
- [37] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005* (2022).
- [38] Jingping Nie, Hanya Shao, Yang Fan, Qijia Shao, Haoxuan You, Matthias Preindl, and Xiaofan Jiang. 2024. LLM-based Conversational AI Therapist for Daily Functioning Screening and Psychotherapeutic Intervention via Everyday Smart Devices. *arXiv preprint arXiv:2403.10779* (2024).
- [39] Jingping Nie, Hanya Shao, Minghui Zhao, Stephen Xia, Matthias Preindl, and Xiaofan Jiang. 2022. Conversational ai therapist for daily function screening in home environments. In *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*. 31–36.
- [40] PubMed. 2024. PubMed. <https://pubmed.ncbi.nlm.nih.gov/>
- [41] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617* (2023).
- [42] Vinicius G Sirtoli, Mokhtar Liamini, Lucas T Lins, Mathieu Lessard-Tremblay, Glenn ER Cowan, Ricardo J Zednik, and Ghyslain Gagnon. 2023. Removal of motion artifacts in capacitive electrocardiogram acquisition: A review. *IEEE Transactions on Biomedical Circuits and Systems* 17, 3 (2023), 394–412.
- [43] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- [44] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654* (2024).
- [45] Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings. 2022. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *arXiv preprint arXiv:2201.07284* (2022).
- [46] Kenneth J Turner. 2009. Abstraction and analysis of clinical guidance trees. *Journal of Biomedical Informatics* 42, 2 (2009), 237–250.
- [47] UpToDate. 2024. UpToDate: Trusted, evidence-based solutions for modern healthcare. <https://www.wolterskluwer.com/en/solutions/uptodate>
- [48] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024. Beyond Direct Diagnosis: LLM-based Multi-Specialist Agent Consultation for Automatic Diagnosis. *arXiv preprint arXiv:2401.16107* (2024).
- [49] Junda Wang, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. JMLR: Joint Medical LLM and Retrieval Training for Enhancing Reasoning and Professional Question Answering Capability. *arXiv preprint arXiv:2402.17887* (2024).
- [50] Xiaoyan Wang, Amy Chused, Noémie Elhadad, Carol Friedman, and Marianthi Markatou. 2008. Automated knowledge acquisition from clinical narrative reports. In *AMIA Annual Symposium Proceedings*, Vol. 2008. American Medical Informatics Association, 783.
- [51] Yubo Wang, Xueguang Ma, and Wenhui Chen. 2023. Augmenting black-box llms with medical textbooks for clinical question answering. *arXiv preprint arXiv:2309.02233* (2023).
- [52] Haiyang Wu, Kaiwei Liu, Siyang Jiang, Zhihe Zhao, Zhenyu Yan, and Guoliang Xing. 2024. Demo Abstract: CaringFM: An Interactive In-home Healthcare System Empowered by Large Foundation Models. In *2024 23rd ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 255–256.
- [53] Jing Xiao, Tat-Seng Chua, and Hang Cui. 2004. Cascading use of soft and hard matching pattern rules for weakly supervised information extraction. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. 542–548.
- [54] Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024. Penetrative ai: Making llms comprehend the physical world. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*. 1–7.
- [55] Lilin Xu, Chaojie Gu, Rui Tan, Shibo He, and Jiming Chen. 2023. MESEN: Exploit Multimodal Data to Design Unimodal Human Activity Recognition with Few Labels. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 1–14.
- [56] Ming Xu. 2023. MedicalGPT: Training Medical GPT Model. <https://github.com/shibing624/MedicalGPT>.
- [57] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305* (2023).

- [58] Bufang Yang, Lixing He, Neiwen Ling, Zhenyu Yan, Guoliang Xing, Xian Shuai, Xiaozhe Ren, and Xin Jiang. 2024. EdgeFM: Leveraging Foundation Model for Open-set Learning on the Edge. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems (SenSys '23)*. Association for Computing Machinery, New York, NY, USA, 111–124.
- [59] Bufang Yang, Lixing He, Kaiwei Liu, and Zhenyu Yan. 2024. VIAssist: Adapting Multi-Modal Large Language Models for Users with Visual Impairments. In *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*. 32–37.
- [60] Bufang Yang, Le Liu, Wenxuan Wu, Mengliang Zhou, Hongxing Liu, and Xinbao Ning. 2023. BrainZ-BP: A Non-invasive Cuff-less Blood Pressure Estimation Approach Leveraging Brain Bio-impedance and Electrocardiogram. *IEEE Transactions on Instrumentation and Measurement* (2023).
- [61] Bufang Yang, Wenxuan Wu, Yitian Liu, and Hongxing Liu. 2022. A novel sleep stage contextual refinement algorithm leveraging conditional random fields. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–13.
- [62] Bufang Yang, Xilin Zhu, Yitian Liu, and Hongxing Liu. 2021. A single-channel EEG based automatic sleep stage classification method leveraging deep one-dimensional convolutional neural network and hidden Markov model. *Biomedical Signal Processing and Control* 68 (2021), 102581.
- [63] Huanqi Yang, Sijie Ji, Rucheng Wu, and Weitao Xu. 2024. Are You Being Tracked? Discover the Power of Zero-Shot Trajectory Tracing with LLMs! *arXiv preprint arXiv:2403.06201* (2024).
- [64] Junjie Ye, Xuanning Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420* (2023).
- [65] Sofia Yfantidou, Christina Karagianni, Stefanos Efstathiou, Athena Vakali, Joao Palotti, Dimitrios Panteleimon Giakatos, Thomas Marchioro, Andrei Kazlouski, Elena Ferrari, and Šarūnas Girdzijauskas. 2022. LifeSnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild. *Scientific Data* 9, 1 (2022), 663.
- [66] Yufeng Yu, Yuelong Zhu, Shijin Li, and Dingsheng Wan. 2014. Time series outlier detection based on sliding window prediction. *Mathematical problems in Engineering* 2014, 1 (2014), 879736.
- [67] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 9241–9250.
- [68] Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641* (2023).
- [69] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. 2023. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 10859–10885.
- [70] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. 2014. Human symptoms–disease network. *Nature communications* 5, 1 (2014), 4212.