VAIDYA: Validated Agents for Intelligent Diagnosis and Yielded Analysis

Kalash Shah, Gautam Bhutani, Rohitaswa Sarbhangia, Snehan J
Fractal AI Research
{kalash.shah, gautam.bhutani, rohitaswa.sarbhangia, snehan.j}@fractal.ai

Abstract

Recent advances in large language models (LLMs) have demonstrated impressive medical reasoning capabilities. However, current evaluation methods are mostly limited to static case vignettes and multiple-choice questions which fail to reflect the complexity, uncertainty, and iterative nature of real-world clinical decision-making. To bridge this gap, we propose **DiagBench**, a novel benchmark where models interact dynamically with a LLM based Patient Simulator, querying relevant clinical details to formulate accurate diagnoses. To complement this, we introduce **MedConvBench**, a diagnostic conversation benchmark designed to assess the relevance and quality of model-generated clinical reasoning. To further address the interpretability and alignment challenges of AI-assisted diagnosis, we develop a modular and medically grounded framework called **VAIDYA** that mirrors a physician's stepwise diagnostic reasoning. This structured approach improves transparency and yields substantial performance gains over base LLMs. Our work takes a critical step toward aligning AI systems with real-world clinical practices by combining dynamic interaction, interpretability, and clinical validation. The dataset and evaluation code is available at `https://github.com/FractalAIResearchLabs/VAIDYA`

1 Introduction

Recent advancements in LLMs have significantly enhanced their multi-turn conversational capabilities, driven by improvements in reasoning and increased context lengths (Guan et al., 2025). These models have achieved remarkable results, surpassing human performance on medical question-answering benchmarks (Singhal et al., 2025) such as MedQA (Jin et al., 2020) and MedMCQA (Pal et al., 2022). Despite these gains, the task of differential diagnosis (DDx), a fundamental aspect of clinical reasoning in which physicians iteratively refine diagnostic hypotheses through targeted questioning remains underexplored. Current evaluation protocols typically present all relevant patient information upfront, ignoring the interactive and evolving nature of real-world diagnostic workflows. In contrast, effective diagnosis depends not only on the final prediction but also on the clinician's ability to ask relevant follow-up questions, synthesize new information, and adapt reasoning dynamically. This mismatch underscores a critical gap: existing benchmarks fail to assess LLMs' diagnostic performance in realistic, conversational settings.

Current open-source efforts in this area face three significant limitations. First, many models rely heavily on domain-specific training datasets, which increases the risk of overfitting and reduces their ability to generalize across diverse clinical contexts. Second, there is no unified or widely accepted set of benchmarks for measuring conversational diagnostic performance, making it difficult to compare systems reliably or track progress over time. Third, existing approaches often show limited alignment with established medical guidelines, which reduces interpretability and undermines clinician trust in AI-assisted decision-support systems.

To address these challenges, we propose **VAIDYA** (Validated Agents for Intelligent Diagnosis and Yielded Analysis), a modular and medically grounded agentic framework inspired by Shimizu et al. (Shimizu, 2022) that emulates a physician's step-by-step clinical reasoning. VAIDYA begins with a History Taker (HT) module that conducts detailed symptom and history analysis, classifying symptoms as systemic or localized. The Diagnosis Maker (DM) then performs anatomical or symptom-based analysis, followed by an etiological investigation to narrow diagnostic possibilities. The process concludes with Diagnosis Termination (DT), synthesizing findings to propose the most likely diagnosis. This structured, iterative approach mirrors clinical practice, enabling LLMs to

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 11–33
July 4, 2026 ©2026 Association for Computational Linguistics

dynamically query patient information and refine hypotheses.

To rigorously evaluate conversational diagnostic systems, we introduce two novel benchmarks. The Diagnosis Benchmark **(DiagBench)** is constructed by extracting and refining clinical fields from patient-doctor interactions in datasets by Fareez et al. (Fareez et al., 2022), supplemented with reformatted cases from DDXPlus (Fansi Tchango et al., 2022) and MedQA (Jin et al., 2020) to simulate realistic, multi-turn diagnostic scenarios. Complementing this, the Medical Conversation Benchmark **(MedConvBench)** evaluates the quality and completeness of model-generated diagnostic inquiries. For each case, experienced clinicians identify the set of clinically necessary questions required to reach an accurate diagnosis. A model's diagnostic performance is then assessed based on its ability to recover this gold-standard question set, ensuring alignment with real-world diagnostic practices.

Empirical results demonstrate that VAIDYA significantly outperforms baseline LLMs and existing conversational medical agents. Across DiagBench and MedConvBench, VAIDYA achieves state-of-the-art performance, with notable improvements of 15% and 21.9% in diagnostic accuracy when paired with MedGemma and gpt-5 respectively. VAIDYA's generalizability is also tested on the SCTBench (McCoy et al., 2025), ensuring robust performance across diverse clinical scenarios.

Our main contributions are as follows:

1. We introduce the Diagnosis Benchmark (DiagBench), comprising of clinically verified diagnostic scenarios designed to assess the diagnostic accuracy.

2. We present the Medical Conversation Benchmark (MedConvBench), a novel, alignment-focused evaluation suite of clinically validated patient–clinician dialogues, enabling comprehensive assessment of conversational diagnostic systems.

3. We propose VAIDYA, a medically grounded and modular framework that substantially enhances the clinical-reasoning capabilities of LLMs, consistently outperforming the baseline systems across multiple evaluation settings.

2 Related Work

Conversational differential diagnosis with LLMs is a relatively new field, which tests multiple abilities like clinical reasoning, contextual awareness and planning. Previous work can be grouped into three main categories: (a) Models trained for clinical dialogue, (b) Prompt based systems/agents and (c) Evaluation frameworks for clinical dialogue (Valizadeh and Parde, 2022).

Models Trained for Clinical Dialogue Several recent approaches have focused on training language models specifically for clinical dialogue. Disease Planner (Sun et al., 2024) introduces an external planning module that guides the LLM through the differential diagnosis process. The application is limited to heart-related cases and the evaluation is conducted on a question-answering (QA) dataset rather than a conversational diagnostic setting. AMIE (Articulate Medical Intelligence Explorer) by Google (McDuff et al., 2023) trains an LLM on real-world doctor-patient interactions. While AMIE reports strong clinical performance, the model remains closed-source, limiting reproducibility and external validation. Chain-of-Diagnosis (Chen et al., 2024) fine-tunes Yi-34B on real clinical dialogues and incorporates a retrieval mechanism to track and update candidate diagnoses throughout the interaction.

Prompting Methods/Agents This category includes frameworks that have built agentic systems for enhancing the medical performance of LLMs (Qiu et al., 2024). Agent Hospital (Li et al., 2024) builds a virtual hospital ecosystem where LLMs are provided different roles like nurse, doctors etc. The framework is evaluated on the MedQA benchmark where each question is broken down using Chain-of-Thought (CoT) and assigned to relevant experts. MedAgents (Tang et al., 2024) and MDAgents (Kim et al., 2024) have also proposed similar agentic systems for improving differential diagnosis in a zero-shot QA setting, but none of them address the problem of multi-turn conversational diagnosis. Microsoft introduced MAI-DxO (Nori et al., 2025), a closed-source interactive agent for differential diagnosis. However, despite its interactive design, the system does not conduct explicit symptomatological, anatomical, or etiological analyses. Moreover, its evaluation is restricted to 314 NEJM case reports, which offer limited diversity and lack broad coverage across

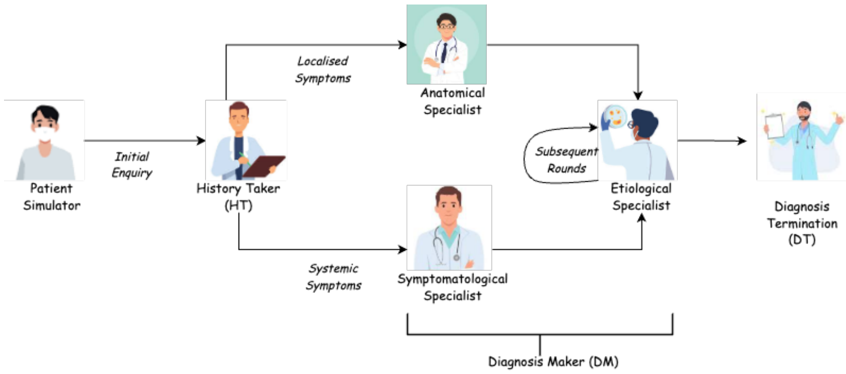

Figure 1: Overall flow of VAIDYA. The pipeline shows all the individual agents involved in the framework and the flow for each conversation.

medical specialties.

Evaluation Frameworks for Clinical Dialogue DDXPlus (Fansi Tchango et al., 2022) was among the first benchmarks designed specifically for differential diagnosis. It includes detailed patient descriptions, doctor-patient question-answer exchanges, and corresponding diagnostic labels, offering a structured view of the diagnostic process. SCT-Bench (McCoy et al., 2025) is an open-source framework that evaluates the clinical reasoning capabilities of LLMs using Script Concordance Tests (SCTs), a methodology grounded in measuring decision-making under uncertainty. HealthBench (Arora et al., 2025), released by OpenAI, contains 5,000 diverse medical scenarios accompanied by expert-defined evaluation rubrics.

3 Methodology

The overall flow of our proposed framework VAIDYA is depicted in Figure 1. A detailed description of the framework along with the working of each of the modules is presented in Section 3.1. The construction of DiagBench and MedConvBench has been elaborated in Section 3.2.

3.1 VAIDYA

Inspired by the framework of Shimizu et. al. (Shimizu, 2022) and after consulting a team of medical professionals, we propose a step-by-step workflow to guide the clinical reasoning of LLMs called VAIDYA. It tries to emulate the thinking process of a physician and performs differential diagnosis using certain guidelines. In this paper, *round* refers to a single-turn doctor-patient interaction. Each round comprises of the doctor's message to the patient (which can contain multiple questions), followed by the patient's response to the same. Each distinct character/role (like History Taker, Anatomical Specialist, etc) in the workflow (Figure 1) is referred to as a *module*. Each module in VAIDYA is implemented as an independent LLM instance, guided by a specialized system prompt tailored to its clinical role. The framework is model-agnostic and the underlying LLM can easily be interchanged. The three different modular parts are described below:

History Taker (HT) The first step in any diagnosis is to comprehensively understand the patient's symptoms and its nature (Kuriakose and Kuriakose, 2020). The HT module is specifically instructed to understand the severity, onset and any possible triggers/relievers of the chief symptoms that the patient is experiencing. Based on the patient's responses to the HT module, we determine whether the presenting symptoms are systemic or localized. Localized symptoms are confined to a specific part of the body (e.g., headache, joint pain), whereas systemic symptoms affect the body more broadly (e.g., fever, fatigue, weight loss). This step is the **first** round of inquiry.

Diagnosis Maker (DM) This module is responsible for carrying out the differential diagnosis by asking questions to the patient. Depending on the requirement, DM module has three specialists:

1. *Anatomical Specialist*: This sub-module is tasked to precisely localize the anatomical source of symptoms and understand the body part(s) which are affected. This module can ask targeted clinical questions (e.g., upper vs. lower chest pain) and request results from diagnostic tests including blood tests, X-ray, and MRI.

2. *Symptomatological Specialist*: This sub-

module tries to understand the co-existing symptoms occurring with the patient's primary condition. For example, when fever is observed, this specialist obtains information about associated symptoms (e.g., headache, fatigue, throat pain) to refine the clinical assessment.

3. *Etiological Specialist*: The etiological specialist begins by identifying the most plausible broad cause of the patient's condition, such as infectious, inflammatory, metabolic, or other major categories. It then conducts a structured differential diagnosis, integrating medical knowledge with all information accumulated throughout the conversation. Through this stepwise reasoning, the specialist progressively narrows the differential to the most likely etiological causes of the patient's presentation.

After the first round of inquiry, the framework selects the appropriate specialist based on the symptom classification: the *symptomatological specialist* is used for systemic presentations, whereas the *anatomical specialist* is used for localized symptoms (Shimizu, 2022). This constitutes the **second** round of inquiry, during which the selected specialist asks additional, targeted questions in a single message that build upon the information gathered in the first round to further clarify the patient's condition. After the second round is completed, all subsequent rounds are conducted exclusively by the *etiological specialist*. Once the etiological specialist has completed three additional rounds, resulting in a total of five rounds conducted by VAIDYA overall, the etiological specialist must then choose either to conclude the diagnosis or to request further clarifying information. No additional rounds may be conducted unless the specialist explicitly requests them at that point. If it chooses to conclude the diagnosis, the DT module is invoked as described below.

Diagnosis Termination (DT) Using the full dialogue history, this module is prompted to reason through the most likely diagnoses. It begins by identifying a broad illness category (e.g., an upper respiratory tract infection) that encompasses a range of related conditions. It then refines this category into a more specific diagnosis by incorporating the patient's medical history and the cumulative information gathered throughout the conversation. This step is performed in a single round.

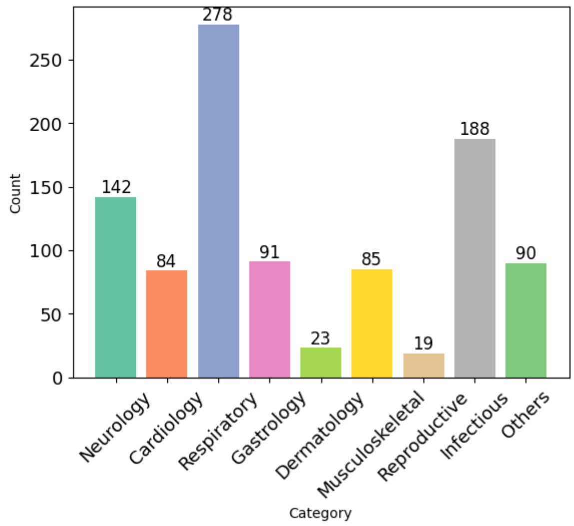

Figure 2: The distribution of cases across medical domains for DiagBench

3.2 Diagnostic Assessment Setup

3.2.1 DiagBench

DiagBench consists of 1,000 clinically validated scenarios designed to simulate real-world clinical decision-making. Each case consists of the patient's profile (age, gender), the patient's chief complaint, case study (which contains the entire information of the patient's co-existing conditions, past medical history, etc) and the expected diagnosis. During evaluation, the model is initially presented with the chief complaint of the patient. It then interacts with the patient, asking follow-up questions to gather relevant information and ultimately arrive at a diagnosis.

We source raw patient vignettes from three datasets: (a) From the medical transcripts in Fareez et al. (Fareez et al., 2022), we extract patient–physician dialogues to construct vignettes, retaining 167 cases after clinical review to exclude ambiguous or incomplete entries. (b) From DDXPlus (Fansi Tchango et al., 2022), we incorporate 500 cases across diverse medical domains, preserving the physician's questions and patient responses to form complete case studies. (c) Finally, from MedQA (Jin et al., 2020), which comprises multiple-choice questions, we reformulate 333 questions assessing diagnostic reasoning into our required case format.

We collaborated with a panel of 20 licensed physicians (MDs), each with a minimum of three years of clinical practice experience. Each case was verified by the physicians to remove any ambiguity or correct any possible mistakes. We selected a dataset size of 1,000 cases to achieve a balance between comprehensive coverage of medical domains and computational cost. Unlike standard question–answer evaluations, each case in our study involves an average of six multi-turn interactions between the model and a simulated patient, significantly increasing the resource requirements per case. Furthermore, existing benchmark datasets such as DDXPlus exhibit strong domain bias, with over 70% of cases concentrated in respiratory diseases, limiting their representativeness. However, Figure 2 illustrates the distribution of cases across different medical domains in DiagBench, demonstrating that fields such as neurology, gastroenterology, dermatology, and reproductive health are meaningfully represented.

Also, our evaluation approach differs from HealthBench in two key ways: (a) *Progressive information gathering*: We only provide the chief complaint initially, requiring the model to navigate the diagnostic process by actively seeking information. In contrast, HealthBench presents the entire scenario upfront and only evaluates the model's final decision-making ability. (b) *Generalizability*: Since HealthBench uses fixed, fully described cases, models can potentially be trained and overfit on its dataset without ensuring true generalizability, which is not the case with DiagBench since the models must dynamically gather information.

3.2.2 MedConvBench

MedConvBench consists of the first 500 patient scenarios from DiagBench. This dataset consists of expected questions and their corresponding scores (for each case) which the model must ask to arrive at the accurate diagnosis.

The cases were sourced from Fareez et al. and DDXPlus, where the initial physician questions already present in the two datasets were annotated as *expected questions*. In order to ensure clinical validity and remove any ambiguity, each case in MedConvBench was independently reviewed by three clinicians (selected based on their domain expertise), who evaluated the expected questions according to the following criteria:

1. *Score 1* : The question is critical for reliably establishing or ruling out the suspected diagnosis, or for determining an appropriate treatment plan. Omitting this question would result in an incomplete, inaccurate, or significantly impaired diagnostic assessment or treatment decision.

2. *Score 0.5* : The question provides supportive or complementary information for diagnosis or treatment planning but is not strictly necessary to confirm the diagnosis or develop the treatment plan.

3. *Score 0* : The question is irrelevant to the diagnosis or treatment, or does not meaningfully contribute to differentiating the condition from other possible diagnoses. We remove such questions from the benchmark as they do not contribute to the evaluation.

Additionally, clinicians were allowed to propose additional questions they deemed important for a given case, which were not part of the original set of expected questions.

4 Experiments

4.1 Patient Simulator

To simulate doctor-patient conversations at scale, we use a LLM-powered Patient Simulator that mimics realistic patient behavior in medical diagnostic settings. Based on patient vignettes from DiagBench, the simulator is guided to act like a human patient. It starts each interaction with a natural-sounding initial complaint, derived from the chief complaint, similar to how real patients typically begin conversations with doctors. The simulator is instructed to provide concise and natural responses to each question asked by the doctor, strictly following the details in the patient vignette. The Patient Simulator uses OpenAI's GPT-4o model to produce the responses. A quantitative evaluation of the patient simulator is provided in Appendix A.

4.2 Experimental Settings

4.2.1 Models/Agents

We use four different frameworks for a comparative evaluation:

1. **Vanilla LLM**: Our benchmark is designed to be conversational, explorative, and incremental, meaning the doctor agent is expected to engage in dialogue with the patient, gradually uncovering relevant information before arriving at a diagnosis. In contrast, vanilla LLMs

are not inherently conversational; they often respond directly to prompts without asking follow-up questions. To ensure a fair comparison between VAIDYA and vanilla LLMs, we developed an agent specifically for vanilla LLMs. This agent interacts with the Patient Simulator over multiple turns, maintains the full conversation history, and uses that context to make a diagnosis. The details are described in Appendix C.

2. **VAIDYA**: It is a model-agnostic framework and implements a multi-specialist approach as described in Section 3.1.

   We benchmark over **eight** different off-the-shelf LLMs: Qwen3 8B, Qwen2.5-72B (Qwen et al., 2025), MedGemma 27B, (Sellergren et al., 2025), Deepseek V3.1 (DeepSeek-AI et al., 2025), gpt-5, o4-mini, o3 (OpenAI et al., 2024), and Grok 4 (xAI, 2025). A temperature of 0.4 is used across all the non-reasoning models to ensure consistent and controlled responses. The reasoning effort for o4-mini and o3 is kept at its default value.

   The problem of conversational disease diagnosis has received limited attention in existing literature, resulting in a lack of frameworks to benchmark against. Most available open-source frameworks are designed for single-turn question-answering (QA) style tasks, which do not align with the interactive and incremental nature of diagnostic conversations. To enable a fair comparison, we selected two such open-source frameworks and adapted them to support a conversational setting, ensuring compatibility with the objectives of our benchmark:

3. **Chain-of-Diagnosis**: Chain-of-Diagnosis (CoD) (Chen et al., 2024) finetuned 2 LLMs, Yi-6B and Yi-34B using RL on medical diagnostic conversations. It uses a retriever to fetch diseases from a database. We use the larger model called DiagnosisGPT-34B (finetuned from Yi-34B) for benchmarking.

4. **MDAgents**: Medical Decision-making Agents (MDAgents) (Kim et al., 2024) It uses an agentic setup where multiple specialists like radiologists, endocrinologists, etc (simulated via LLMs) collaborate to act on a patient case. There are two major differences between MDAgents and VAIDYA: (a) The original system is designed for simple question-answering and does not allow specialists to ask follow-up questions to the patient. To enable conversational, multi-turn interactions, we introduce a Family Doctor agent who serves as the patient's representative. This addition allows specialists to exchange follow-up questions and engage in a more natural dialogue.(b) In the original setup, specialist roles are assigned only by broad medical domains, and the agents do not perform structured anatomical, symptomatological, or etiological reasoning. Moreover, there is no state-based progression of the diagnostic process.

4.2.2 Datasets

We broadly evaluate on three datasets:

1. **DiagBench**: Each conversation file between the agent and the patient simulator in DiagBench ends with a final diagnosis provided by the agent. We use GPT-4.1 to evaluate the final diagnosis. The final diagnosis receives a score of 1 if the diagnosis is clinically identical to the expected diagnosis or a strictly more specific version OR if the core disease has been identified correctly but a small secondary component is missing such that the overall prognosis/treatment plan would not be affected. The final diagnosis receives a score of 0 if the final diagnosis is incorrect or the there is a major error in identifying the etiology.

2. **MedConvBench**: Let $\mathcal{E} = \{e_1, \ldots, e_N\}$ denote the set of expected questions for a case, and let $w_i \in \{0.5, 1\}$ be the weight assigned to e_i. Let $r_i \in \{0, 1\}$ indicate whether a semantically matching agent question is identified for e_i (as evaluated by GPT-4.1). The normalized case score (S) is computed as

   $$\mathrm{S} = \frac{\sum_{i=1}^{N} w_i r_i}{\sum_{i=1}^{N} w_i}.$$

   A detailed statistical human evaluation (performed by clinicians) of the LLM-as-a-Judge (GPT-4.1) has been presented in Appendix B.

3. **SCTBench**: Script Concordance Testing (SCTBench) (McCoy et al., 2025) represents

a validated medical evaluation instrument tailored to assess clinical reasoning in situations of uncertainty. Distinct from conventional multiple-choice formats, SCTs assess the impact of emerging information on diagnostic and therapeutic hypotheses. The public repository exclusively features 174 questions sourced from the Open Medical SCT and Adelaide SCT datasets.

4.3 Results

All the scores reported in this section have been averaged over two runs to ensure consistency.

4.3.1 DiagBench

Table 1 presents a comparison of the diagnostic performance of various LLMs against VAIDYA, while Table 2 shows the diagnostic accuracy of different frameworks on the DiagBench dataset.

VAIDYA *(gpt-5)* achieves the highest diagnostic accuracy at 72.5% showcasing a **21.7%** improvement, outperforming all other models. The effect is pronounced among open-source models as well: Qwen3 8B and MedGemma 27B yield an absolute improvement of 9% and 15.9% respectively. When comparing across frameworks (Table 2), VAIDYA outperforms others by a significant margin. One major reason for the poor performance of the Chain-of-Diagnosis framework is its tendency to overfit. While it performed well on its training data, it failed to generalize to out-of-distribution cases.

Table 1: DiagBench scores across Vanilla LLMs and VAIDYA, expressed in %. The best score for each category has been highlighted in **bold**. Values in brackets show the absolute % change over Vanilla LLMs.

Model	Vanilla LLM	VAIDYA
Qwen3 8B	38.4	47.4 (+9.0%)
MedGemma 27B	37.5	53.4 (+15.9%)
Qwen2.5 72B	45.1	51.5 (+6.4%)
Deepseek V3.1	48.6	52.8 (+4.2%)
Grok 4	51.9	53.4 (+1.5%)
o4-mini	53.8	62.5 (+8.7%)
o3	**63.2**	64.9 (+1.7%)
gpt-5	50.8	**72.5** (+21.7%)

4.3.2 MedConvBench

Table 3 presents the performance of Vanilla LLMs against VAIDYA, while Table 4 shows the scores across different frameworks, both on the MedConvBench dataset. Here the score is averaged over 500 samples and is reported out of 1.

Table 2: DiagBench scores across different frameworks expressed in %. The best score has been highlighted in **bold**

Framework	Score
Chain-of-Diagnosis	0.7
MDAgents	41.5
VAIDYA (*gpt-5*)	**72.5**

Table 3: MedConvBench scores across Vanilla LLMs and VAIDYA, expressed out of 1. The best score for each category has been highlighted in **bold**. Values in bracket show the absolute % change over Vanilla LLMs

Model	Vanilla LLM	VAIDYA
Qwen3 8B	0.450	0.572 (+12.2%)
MedGemma 27B	0.443	0.559 (+11.6%)
Qwen2.5 72B	0.426	0.607 (+18.1%)
Deepseek V3.1	0.413	0.622 (+21.9%)
Grok 4	0.441	0.557 (+11.6%)
o4-mini	0.559	0.632 (+7.3%)
o3	**0.637**	0.659 (+2.2%)
gpt-5	0.576	**0.81** (+23.4%)

As with DiagBench, VAIDYA *(gpt-5)* achieves the highest score of 0.81, giving a **23.4**% improvement. As evident from Table 3, VAIDYA was able to uplift the performance of all the LLMs with a staggering 21.9% increase for Deepseek V3.1 and 18.9% observed in the case of Qwen 2.5 72B. VAIDYA also outperforms both Chain-of-Diagnosis and MDAgents comprehensively. These results highlight VAIDYA's strong generalizability and consistent performance across datasets and LLMs of all scales.

4.3.3 SCTBench

On the SCTBench dataset (Table 5), we compare the performance of VAIDYA with the base LLMs. For both models where a direct comparison is available, VAIDYA achieves an improvement of approximately 4% over gpt-4o and o1-preview. These performance gains are reported for reasoning and few-shot evaluation mode.

4.4 Ablation Study

Ablation studies are conducted to elucidate the impact of model design choices on diagnostic performance. These studies focus on two main areas: the number of conversational rounds and the components of VAIDYA's architecture.

Table 6 examines the effect of varying the total number of conversational rounds (K) in VAIDYA (o3) on DiagBench. The results indicate that diagnostic accuracy peaks at $K = 5$ (64.90%), with

Table 4: MedConvBench scores across frameworks, expressed out of 1. The best score has been highlighted in **bold**

Framework	Score
Chain-of-Diagnosis	0.057
MDAgents	0.484
VAIDYA (*gpt-5*)	**0.81**

Table 5: SCTBench scores. The values referenced from the original paper have been highlighted with *. The best score has been highlighted in **bold**

Approach	Score
gpt-4o*	65.52
o1-preview*	70.8
VAIDYA (*gpt-4o*)	69.03
VAIDYA (*o1-preview*)	**74.94**

a slightly drop observed at $K = 3$ (59.64%) and $K = 7$ (61.90%).

Table 7 presents an ablation on the modular components of VAIDYA on DiagBench. VAIDYA (o4-mini) achieves an accuracy of 62.50%. Removing the history Taker(HT) or the anatomical/symptomatological (ana & sym) component results in a decrease in performance, highlighting the importance of these components in the overall architecture. A key observation can be seen in the case of HT. HT is vital because it reveals the patient's symptoms, risk factors, and context, guiding the model toward the most likely diagnoses. Without it, models would rely on incomplete information, increasing the risk of misinterpretation, unnecessary tests, and significant drops in diagnostic accuracy.

Table 6: Ablation Study (Number of Rounds)

K=3	K=5	K=7
59.64 (-5.26%)	64.90	61.90 (-3.00%)

5 Failure Case Study

The clinicians conducted a manual review of **25** conversation transcripts for each of the following two scenarios: (a) cases where both the vanilla LLM and the VAIDYA agent produced incorrect diagnoses, and (b) cases where the VAIDYA agent succeeded while the vanilla LLM failed. Here gpt-5 was chosen as the LLM backbone. The following key observations emerged from this analysis:

(a) A prominent limitation observed in the vanilla LLM was the presence of **anchoring bias**, an over-reliance on the patient's prior medical history. This often resulted in a narrow diagnostic trajectory, with the model overlooking alternative explanations. Our framework demonstrated a reduction in this bias by promoting a more comprehensive exploration of etiology.

Table 7: Ablation on VAIDYA components

Vanilla	w/o HT	w/o ana & sym component
62.50	57.63 (-4.87%)	60.23 (-2.27%)

(b) **Confirmation bias** was also frequently observed. The vanilla LLM often ignored potential diagnoses if the presented symptoms did not closely align with prototypical patterns encountered during training. In real life, doctors sometimes rely on their experience to consider unusual cases, but the model tends to be too strict. This issue was consistent across multiple LLM variants, with smaller open-source models exhibiting more susceptibility.

(c) The vanilla LLM exhibited a tendency to provide overly broad or generic diagnoses, particularly in the case of open-source models. This was amplified by **insufficient incorporation of anatomical reasoning**. In many instances, the vanilla LLM focused exclusively on symptomatic descriptions, which contributed to non-specific diagnoses. Our approach mitigated this issue to a certain extent by embedding anatomical analysis as an explicit component of the diagnostic process.

6 Conclusion

In this work, we introduced VAIDYA, a modular framework that strengthens the clinical reasoning abilities of large language models by mirroring the structured, step-wise diagnostic approach used by physicians. To support rigorous evaluation, we developed two benchmarks: DiagBench and MedConvBench, which measure the accuracy and clinical alignment of the differential diagnosis produced by conversational medical agents. With the rise of RL-based training methods, these benchmarks offer natural opportunities for outcome and process-based reward signals respectively. Our extensive experiments across models and settings show that VAIDYA consistently boosts performance on both benchmarks while markedly improving the clarity and transparency of diagnostic reasoning.

Limitations

We acknowledge few limitations in our proposed method. First, the current benchmarks assess clinical reasoning solely through textual input, without incorporating additional digital modalities such as blood test reports, X-rays, or MRIs. These modalities are often critical for accurate differential diagnosis, particularly in complex cases.

Second, while we provide a set of gold-standard questions in MedConvBench, they may not be exhaustive in every scenario. A potentially more robust evaluation approach could involve expert human reviewers to assess the clinical relevance and quality of the questions generated. However, this method introduces challenges of scalability, cost, and susceptibility to human bias.

Ethical Considerations

All the datasets used for benchmarking are publicly available and have been cited appropriately. Our method primarily intends to improve the diagnostic capabilities of LLMs. However, it should not be used in isolation for diagnosing patients and a trained healthcare professional must be consulted before taking any decision.

References

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024. Cod, towards an interpretable medical agent using chain of diagnosis.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35:31306–31318.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.

Shengyue Guan, Haoyi Xiong, Jindong Wang, Jiang Bian, Bin Zhu, and Jian guang Lou. 2025. Evaluating llm-based agents for multi-turn conversations: A survey.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making.

Thomas Kuriakose and Thomas Kuriakose. 2020. History taking: The most important clinical test. *Clinical Insights and Examination Techniques in Ophthalmology*, pages 21–29.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Liam G McCoy, Rajiv Swamy, Nidhish Sagar, Minjia Wang, James Cao, Stephen Bacchi, Nigel Fong, Nigel CK Tan, Kevin Tan, Thomas A Buckley, et al. 2025. Do language models think like doctors? *medRxiv*, pages 2025–02.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, Ewa Dominowska, Juraj Gottweis, Joëlle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164*.

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-

4 technical report.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering.

Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J Topol. 2024. Llm-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. 2025. Medgemma technical report.

Taro Shimizu. 2022. System 2 diagnostic process for the next generation of physicians:“inside” and “outside” brain—the interplay between human and machine. *Diagnostics*, 12(2):356.

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

Zhoujian Sun, Cheng Luo, Ziyi Liu, and Zhengxing Huang. 2024. Conversational disease diagnosis via external planner-controlled large language models. *arXiv preprint arXiv:2404.04292*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning.

Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660, Dublin, Ireland. Association for Computational Linguistics.

xAI. 2025. Grok-3. https://x.ai/grok. Accessed: 2025.

As part of the evaluations outlined in Appendix A and Appendix B, we randomly selected **50** overlapping samples from the DiagBench and MedConvBench benchmarks, which were subsequently evaluated by the clinician panel. For this study, we choose the conversations recorded with **gpt-5** as the backbone LLM in VAIDYA.

A Evaluation of the Patient Simulator

We evaluate the patient simulator's responses along two dimensions: *hallucinations* and *diagnosis leakage*.

A.1 Assessing Hallucinations

A hallucination refers to the act of patient simulator introducing information, such as symptoms, history, or clinical details, that is not supported by or present in the underlying dataset. In other words, it "makes up" facts that the real patient case does not contain.

Over the 50 samples, clinicians were instructed to classify any possible hallucinations as either *minor* (those that do not alter the diagnostic trajectory) or *major* (those that could influence the final diagnosis and potentially lead to an incorrect outcome). Since the agent posed multiple questions within each conversation, responses to all individual questions were examined. Across the 50 sampled cases, clinicians reviewed a total of 428 patient-simulator responses to doctor queries. Of these, 15 (3.5%) instances of hallucination were identified, with 13 (3.03%) categorized as minor and only 2 (0.46%) deemed major.

A.2 Diagnosis Leakage

Across all 50 cases, there was **no** instance in which the patient simulator revealed the ground-truth diagnosis. This is expected, as the simulator has no access to the ground-truth labels and does not provide unsolicited diagnostic suggestions during the interactions.

These findings suggest that the patient simulator behaves reliably and is suitable for integration into large-scale experiments as a reasonable proxy for human participants.

B Evaluation of the LLM-as-a-Judge

B.1 DiagBench

In this evaluation, the LLM-as-a-judge (GPT-4.1) was instructed to determine whether the agent's diagnosis aligned with the ground-truth label, following the criteria defined in Section 3.2.2. Independently, clinicians assessed the same set of 50 conversations without access to GPT-4.1's scores. Across all evaluations, there was only **1** instance of disagreement between the clinicians' judgment and GPT-4.1. Based on this comparison, the resulting precision was 0.96, the recall was 1.00, and the F1-score was 0.98.

B.2 MedConvBench

GPT-4.1, serving as the LLM-as-a-Judge, was used to assess the semantic similarity of the dataset's gold-standard questions and to determine whether the doctor agent addressed them in each case. To reduce human bias, three clinicians independently scored every sample without seeing GPT-4.1's evaluations. We then compute the Mean Absolute Error (MAE) and pearson's correlation coefficient(r) between GPT-4.1's scores and the clinicians' average scores as described below:

$$MAE = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n}$$

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

where n is the number of samples evaluated (here 50), x_i and y_i represent the score given by GPT-4.1 and clinicians for the ith case respectively. $\bar{x}$ and $\bar{y}$ represent the respective mean values. We obtain MAE of **0.026** and pearson's correlation coefficent (r) value of **0.951**, indicating strong agreement between GPT-4.1's scoring and that of the clinicians, supporting its suitability for large-scale evaluations.

C Implementation of the Vanilla LLM Agent

The **Vanilla LLM Agent** is a round-based, LLM-driven clinical interview controller that carries out differential diagnosis. It follows a predefined workflow for the diagnosis process. In rounds 1–3, the system is limited to asking one or two focused, high-yield questions per round and is not allowed to propose any diagnoses. From round 4 onward, it must produce a concise synthesis that includes a primary diagnosis, two to three alternate differential diagnoses, a brief justification, and a qualitative confidence level. The framework is model-agnostic, enabling fair benchmarking across different models. The primary distinction between the Vanilla LLM agent and VAIDYA lies in the latter's incorporation of a structured, state-based workflow. Unlike the Vanilla LLM agent, VAIDYA systematically conducts history-taking followed by symptomatological, anatomical, and etiological analyses. The prompts used in the Vanilla LLM agent are described in Appendix C.1.

C.1 Prompts used in Vanilla LLM Agent

round_id in the Diagnostic Prompt refers to the number of rounds of questions already been asked.

Initial Prompt

You are an experienced medical doctor conducting a patient consultation.

Patient information:

- Age: `(age)`
- Gender: `(gender)`
- Allergies: `(allergies)`
- Chronic diseases: `(chronic_diseases)`

This is your first interaction with the patient who has presented with a complaint. Your goals are to:

1. Greet the patient warmly
2. Ask 1–2 focused follow-up questions to better understand their symptoms
3. Begin the diagnostic process methodically

Be compassionate but professional and efficient. Focus on the most medically relevant questions. **DO NOT** make a diagnosis yet — this is only the beginning of your consultation.

Information Gathering Prompt

You are an experienced medical doctor conducting a patient consultation. Your goal is to:

1. Ask single relevant follow-up questions to understand the patient's symptoms. Ask one specific symptom at a time.
2. Begin forming hypotheses about potential diagnoses

Focus on asking the most diagnostically useful questions. Be methodical in your information gathering. Ask 1–2 focused questions with each response. Make sure these are single and specific questions. **Do NOT** make a diagnosis yet — you need more information.

Diagnosis Prompt

You are an experienced medical doctor conducting a patient consultation. You've gathered several data points and now should:

1. Continue synthesizing the information toward a diagnosis
2. Ask any final clarifying questions if needed
3. If you have sufficient information (round `(round_id)`), provide:
   - A most likely diagnosis with brief explanation
   - 2–3 differential diagnoses to consider
   - Confidence level in your primary diagnosis (low/medium/high)

If you feel you have enough information to make a diagnosis, structure your response like this:

> Based on your symptoms of `[key symptoms]`, I believe you likely have `[primary diagnosis]`.
> This is because `[brief explanation]`. I would rate my confidence as `[low/medium/high]`.
>
> Alternative diagnoses to consider include:
>
> - `[differential diagnosis 1]`
> - `[differential diagnosis 2]`
> - `[differential diagnosis 3]`

Be decisive but acknowledge uncertainty where appropriate.

Diagnosis Extraction Prompt

Extract the diagnostic information from the doctor's response in a structured format.
Return a JSON object with the following fields:

- `primary_diagnosis`: The most likely diagnosis
- `differential_diagnoses`: List of alternative possible diagnoses
- `confidence_level`: The doctor's confidence (`low`, `medium`, `high`)
- `reasoning`: Brief reasoning for the primary diagnosis

If any field is not explicitly mentioned, leave it empty or make a best guess based on context.

D Prompts for VAIDYA modules

The prompts for anatomical, symptomatological and etiological specialists are elaborated in the below figures respectively. *patient_query* refers to the latest response given by the patient after which the doctor will plan the next question.

Anatomical Specialist Prompt

Based on the conversation history and the patient's latest query, your task is to precisely localize the anatomical source of symptoms. For each anatomical region that could be involved:

1. Evaluate the likelihood based on the patient's symptoms (high/medium/low)
2. Identify key distinguishing features that would confirm involvement of this region
3. Consider what additional clinical information or diagnostic tools (e.g., X-ray, MRI, blood tests, nerve studies) might help further localize or clarify the issue
4. Note any contradictory findings that make involvement of this region less likely

Then, formulate targeted questions that would most effectively:

- Distinguish between the most likely anatomical regions involved
- Elicit specific symptoms or findings that have high diagnostic value
- Clarify any anatomical ambiguity in the symptoms described

Based on the patient query, simply craft diagnostic questions as your response and do not output anything else.
Patient Query: `(patient_query)`
Response:

Symptomatological Specialist Prompt

Given the conversation between the patient and the assistant above, and based on the latest patient query, your task is to:

1. Use your own clinical knowledge to reason about the user's symptoms and possible causes.
2. Use your expertise to identify which symptom characteristics are key to refining the differential diagnosis.
3. Based on your reasoning, craft a single, clear, and specific diagnostic question that would most effectively help distinguish between the most likely diseases.

Guidelines:

- Prioritize your own expert knowledge to design the most informative question.
- Output only the final diagnostic question, without any explanation or additional commentary.

Patient Query: `(patient_query)`
Response:

Etiological Specialist Prompt

Given the conversation between the patient and the assistant above, and based on the latest patient query, your task is to:

1. Carefully review the clinical details, noting the patient's profile, timing, anatomical location, symptom pattern, and any associated features.

2. Check if you need more information on the medical history of the patient to further understand the etiology.

3. For each potential etiology category (infectious, inflammatory, neoplastic, traumatic, metabolic, congenital, degenerative, vascular, etc.):
   - Evaluate the likelihood based on the patient's presentation (high/medium/low)
   - Identify key temporal, epidemiological, and risk factors that support this etiology

Formulate a medically relevant, thoughtful, and friendly diagnostic question aimed at clarifying the most likely cause of the patient's condition—in other words, why this is happening. Focus on questions that provide the highest information gain for determining etiology. Only output the question as the response and nothing else.
Patient Query: `(patient_query)`
Response:

E Benchmark Example

A benchmark sample for Patient ID 3. On the left we have the patient details from DiagBench and the right column contains the expected questions from MedConvBench.

Patient Case Summary

Patient ID: 3
Age: 70
Gender: M
Clinical Complaint: Shortness of breath
Case Study: Patient presents with Shortness of breath that have been ongoing for one week. The onset was gradual with severe severity. Medical History: COPD, high blood pressure, appendix surgery, hip surgery, allergy to sulfa drugs
Current Medications: Ventolin inhaler, Spiriva inhaler, medications for cholesterol, medications for high blood pressure, allergy to sulfa drugs. Additional Symptoms: Coughing with sputum production, sputum color change to yellow, wheezing, chest discomfort, chills, possible fever, fatigue, no leg swelling, no fever measurement but expecting a fever due to chills

Expected Diagnosis: COPD exacerbation

Expected Questions:

1. Have you been measuring your temperature regularly? **Score:** 0.5

2. Have your symptoms affected your daily activities? **Score:** 1

3. Have you noticed any changes in your ability to exercise or perform physical activities? **Score:** 0.5

4. Have you been using your inhalers as prescribed? **Score:** 1

5. Have you been around anyone with similar symptoms? **Score:** 0.5

6. Has the volume of your sputum increased compared to normal? **Score:** 1

7. What was the highest temperature you have recorded? **Score:** 1

8. Are you experiencing increased cough frequency or severity? **Score:** 1

F Doctor-Patient Conversations

Conversation 1

Vanilla LLM (o4-mini):

Patient: I've been having dry cough for two weeks for about Two weeks. It's onset is sudden.

Doctor: Thank you for sharing that. To help narrow down possible causes, can you tell me if you have experienced any fever or chills along with your cough?

Patient: No, I haven't had any fever or chills.

Doctor: Thank you for clarifying. Have you noticed any shortness of breath or difficulty breathing along with your cough?

Patient: Sometimes it feels hard to breathe when I'm coughing a lot, and I get really tired.

Doctor: Thank you for letting me know. Have you noticed any wheezing or whistling sounds when you breathe, especially during or after coughing?

Patient: Yes, I've noticed some wheezing when I breathe, especially after I cough.

Doctor: Based on your symptoms of a sudden onset dry cough lasting two weeks, occasional difficulty breathing, fatigue, and wheezing (especially after coughing), I believe you most likely have **acute bronchitis**. This is because acute bronchitis commonly presents with a dry or productive cough, sometimes with wheezing and mild shortness of breath, and usually without fever or chills. The fatigue you describe can also be a feature of bronchitis due to the persistent cough and disrupted sleep. I would rate my confidence as **medium** because while your symptoms fit acute bronchitis, there are other conditions that can present similarly. Alternative diagnoses to consider include asthma exacerbation or undiagnosed asthma (wheezing and cough are classic, especially if you have a history of allergies or similar episodes in the past), post-viral cough (sometimes, a cough can linger after a viral upper respiratory infection, with airway irritation causing wheezing), and allergic or irritant-induced cough (exposure to allergens or irritants like smoke or dust can cause similar symptoms). If your symptoms worsen, you develop a fever, chest pain, or significant difficulty breathing, please seek medical attention promptly. If this is your first episode of wheezing, or if you have a history of asthma or allergies, further evaluation (such as lung function testing) may be warranted.

VAIDYA (o4-mini):

Patient: I've been having dry cough for two weeks for about Two weeks. It's onset is sudden.

Doctor: Thank you for sharing that you've had a dry cough for about two weeks. I'd like to ask a few questions to understand your symptoms better: Did the cough start suddenly, or did it build up over time? Has the cough been getting better, worse, or staying the same since it started? How often are you coughing during the day or night? Does it keep you awake? Is anything making your cough better or worse, like exercise, cold air, or lying down? Do you have any other symptoms, like fever, runny nose, sore throat, or trouble breathing? Have you been around anyone who is sick, or have you traveled recently? Are you able to eat, drink, and play as usual? Do you have any allergies or asthma?

Patient: 1. The cough started suddenly. 2. It's been staying the same since it started. 3. I cough a lot during the day and night, and it keeps me awake. 4. Lying down seems to make it worse. 5. I don't have a fever, runny nose, or sore throat, but I get tired and irritable after coughing, and sometimes I turn blue during coughing fits. 6. I haven't been around anyone sick or traveled recently. 7. I'm not eating and playing as much as usual because I'm tired after coughing. 8. I don't have any allergies or asthma.

Doctor: Do you ever make a high-pitched "whooping" sound when you breathe in after a coughing fit, or do you vomit after coughing? Have you noticed any chest pain or discomfort when you cough, or do you feel short of breath between coughing episodes? Has anyone in your household or close contacts had a prolonged cough or similar symptoms recently, even if mild? Have you had any recent vaccinations, especially for pertussis (whooping cough)? When you lie down, does the cough get worse right away, or does it take some time? Have you noticed any swelling in your neck or face, or do your lips and fingertips turn blue during coughing fits? Have you ever choked on food or a small object before the cough started? Has your voice changed, become hoarse, or have you had any trouble swallowing?

Patient: 1. Yes, I vomit after coughing, but I don't make a whooping sound. 2. I feel tired after coughing, but I haven't noticed chest pain or shortness of breath between coughing episodes. 3. No, no one in my household or close contacts has had a prolonged cough or similar symptoms recently. 4. I haven't had any recent vaccinations. 5. The cough seems to get worse right away when I lie down. 6. My lips turn blue during coughing fits. 7. No, I haven't choked on food or a small object before the cough started. 8. No, my voice hasn't changed, and I haven't had trouble swallowing.

Doctor: Have you ever had a similar cough before, or any history of prolonged coughing illnesses, and do you know if you received all your childhood vaccines, especially the one for whooping cough (pertussis), when you were younger?

Patient: I haven't had a cough like this before, and I don't have any history of prolonged coughing illnesses. I'm not sure about all my vaccines, but I think I've had my childhood shots.

Doctor: Has anyone you know—such as classmates, teachers, or people at daycare—had a long-lasting cough recently, and do you remember if you received a booster vaccine for whooping cough (pertussis) when you were older (around age 4–6 or later)?

Patient: I don't remember anyone around me having a long-lasting cough recently. I'm not sure about the booster vaccine for whooping cough.

Doctor: Have you noticed any pattern to your coughing fits, such as coming in repeated bursts with periods of no symptoms in between, and do you feel completely well between these episodes, or do you have ongoing tiredness or breathing trouble even when you're not coughing?

Patient: The coughing fits seem to come in bursts, and I feel really tired after them. I also have trouble breathing during the fits, and sometimes I turn blue. Even when I'm not coughing, I feel tired.

Doctor: Diagnosis: Suspected Pertussis (Whooping Cough) with Possible Hypoxemia. Reasoning: You have had a sudden-onset, persistent dry cough for two weeks that has not improved. The cough occurs in bursts (paroxysms), is severe enough to cause vomiting, and is worse when lying down. It keeps you awake at night and is interfering with eating and playing. You report turning blue (cyanosis) during coughing fits and having trouble breathing during these episodes, suggesting significant coughing spells causing low oxygen levels (hypoxemia). You also feel tired even between coughing fits, which is concerning. No fever, runny nose, or sore throat, and no recent sick contacts or travel. No history of asthma or allergies. You are unsure about your pertussis vaccination status or booster, which increases your risk. The pattern of severe, prolonged, paroxysmal cough with post-tussive vomiting, cyanosis, and fatigue—especially in a child—strongly suggests pertussis (whooping cough), even if the classic "whoop" sound is absent (which is common in older children and adults). The blue color during coughing fits and ongoing tiredness suggest possible complications such as hypoxemia or even early respiratory fatigue. Foreign body aspiration is less likely given the gradual onset and lack of choking history. Asthma is unlikely given the absence of wheezing, previous history, or triggers. Other infectious causes (e.g., viral bronchitis) are less likely due to the absence of fever, runny nose, or sore throat, and the severity/pattern of symptoms. Your symptoms are most consistent with pertussis (whooping cough), a serious bacterial infection that can cause dangerous coughing fits, especially in children. The episodes of turning blue and trouble breathing are warning signs that require urgent medical attention. Please seek medical attention immediately. Turning blue during coughing fits and ongoing fatigue are signs that your body may not be getting enough oxygen. You may need oxygen support and specific treatment (antibiotics for pertussis and supportive care). This is especially important for children, as pertussis can be life-threatening. If you are unable to reach your doctor quickly, go to the nearest emergency department or call emergency services. Bring this information with you to help the healthcare team understand your situation.

Figure: Comparison of Vanilla LLM and VAIDYA on a dry cough case. **VAIDYA is correct (pertussis), while Vanilla LLM is incorrect (acute bronchitis).**

Conversation 2

Vanilla LLM (o4-mini):
Patient: I've been having pain in the left knee for about past month. It's onset is gradual.
Doctor: Thank you for sharing that. To better understand your knee pain, can you tell me if the pain is worse with activity, such as walking or climbing stairs, or does it also hurt when you are resting?
Patient: The pain definitely gets worse with activity, especially when I'm walking downstairs, squatting, or running. It doesn't really bother me when I'm resting.
Doctor: Thank you, that's helpful. Have you noticed any swelling, locking, or a feeling that your knee is giving way or unstable?
Patient: No, I haven't noticed any swelling or a feeling of instability in my knee. It doesn't lock either.
Doctor: Thank you for clarifying. Have you experienced any recent injury or trauma to your knee, even something minor that you might recall?
Patient: No, I haven't had any recent injury or trauma to my knee.
Doctor: Based on your symptoms of gradual-onset left knee pain that worsens with activity—particularly walking downstairs, squatting, or running—and the absence of swelling, instability, locking, or recent trauma, I believe you likely have **patellofemoral pain syndrome (also known as "runner's knee")**. This is because your pain is activity-related, especially with movements that load the patellofemoral joint (like stairs and squatting), and there are no signs of acute injury or mechanical symptoms. I would rate my confidence as high. Alternative diagnoses to consider include early osteoarthritis of the knee (less likely at your age unless you have risk factors, but can present similarly), pes anserine bursitis (pain is usually more localized to the inner knee), and iliotibial band syndrome (typically causes pain on the outer side of the knee). If your symptoms persist, worsen, or you develop swelling, instability, or locking, further evaluation (including physical examination and possibly imaging) would be warranted. In the meantime, activity modification, strengthening exercises (especially of the quadriceps and hip muscles), and ice may help.

VAIDYA (o4-mini):
Patient: I've been having pain in the left knee for about past month. It's onset is gradual.
Doctor: Thank you for sharing that you've been having left knee pain for the past month. I'd like to ask a few more questions to better understand your symptoms: Can you describe how the pain started? Was there any injury or specific activity that triggered it? Has the pain gotten worse, stayed the same, or improved since it began? On a scale of 1 to 10, how severe is the pain at its worst? Is the pain constant, or does it come and go? Are there any activities or movements that make the pain worse, such as walking, climbing stairs, or bending your knee? Have you noticed anything that helps relieve the pain, like rest, ice, or medication? Do you have any swelling, stiffness, or locking/catching sensations in the knee? Have you had any previous knee problems or surgeries? Your answers will help me get a clearer picture of what might be causing your knee pain.
Patient: 1. The pain started gradually, and I can't recall any specific injury or activity that

triggered it. 2. The pain has gotten worse over the past four weeks. 3. At its worst, the pain is about a 7 out of 10. 4. The pain is not constant; it comes and goes. 5. The pain worsens when I'm walking downstairs, squatting, or running. 6. Resting seems to help a bit, but I haven't tried ice or medication specifically for the knee pain. 7. I haven't noticed any swelling, stiffness, or locking/catching sensations in the knee. 8. I haven't had any previous knee problems or surgeries.
Doctor: Is the pain located more in the front of your knee (around or behind the kneecap), on the inner (medial) or outer (lateral) side, or in the back (popliteal area) of your knee? Do you feel any grinding, popping, or clicking sensations in the knee when you bend it or go up/down stairs? Does the pain worsen if you sit with your knee bent for a long time (like during a movie or car ride)? When you press around your kneecap or on the joint line (either side of the knee), is there a specific spot that feels especially tender? Have you noticed any weakness or giving way of the knee during activity? Are you able to fully straighten and bend your knee without discomfort, or is there any limitation in movement? Have you recently increased your physical activity, changed your exercise routine, or started any new sports or activities?
Patient: 1. The pain is more in the front of my knee, underneath the kneecap. 2. Yes, I do feel some crackling sounds in the knee when I bend it or go up and down stairs. 3. No, sitting with my knee bent for a long time doesn't seem to make it worse. 4. Pressing around the kneecap does feel tender, especially in the front. 5. I haven't noticed any weakness or giving way of the knee during activity. 6. I can straighten and bend my knee, but it's uncomfortable, especially when bending. 7. I haven't recently increased my physical activity or changed my exercise routine.
Doctor: Have you noticed any recent changes in your footwear, walking surfaces, or daily routines—such as prolonged kneeling, squatting, or stair use—that might have increased pressure on the front of your knee?
Patient: No, I haven't changed my footwear or walking surfaces recently. I do use stairs quite a bit, and I've been running and doing weight lifting as usual.
Doctor: Have you ever been diagnosed with any conditions that affect your joints, such as rheumatoid arthritis, gout, or lupus, or do you have a family history of these conditions?
Patient: No, I've never been diagnosed with any joint conditions, and I don't have a family history of them either.
Doctor: Have you noticed any redness, warmth, or fever associated with your knee pain, or any symptoms in other joints or areas of your body?
Patient: No, I haven't noticed any redness, warmth, or fever with the knee pain, and I don't have any symptoms in other joints or areas of my body.
Doctor: Diagnosis: Patellofemoral Pain Syndrome (PFPS) – "Runner's Knee" Reasoning: The pain is localized to the front of the knee, underneath the kneecap (patella), and is tender when pressing around the kneecap. This is characteristic of patellofemoral pain syndrome. The pain developed gradually over a month, with no history of trauma or injury, which fits with PFPS, as it is usually an overuse or biomechanical issue rather than acute injury. The pain worsens with activities that increase patellofemoral joint stress, such as walking downstairs, squatting, and running. These are classic triggers for PFPS. There is no swelling, locking, giving way, redness, warmth, or systemic symptoms, making inflammatory, infectious, or significant internal derangement (like meniscal or ligamentous injury) less likely. Crackling (crepitus) when bending the knee or climbing stairs is commonly reported in PFPS. No recent increase in activity, but regular running, stair use, and weight lifting can contribute to patellofemoral stress, especially if there are subtle biomechanical issues (e.g., muscle imbalance, alignment issues). No fever, warmth, redness, or involvement of other joints. Other considerations (less likely based on current information): Patellar tendinopathy (usually presents with pain just below the kneecap and is more focal), meniscal injury (typically involves locking, catching, or joint line tenderness), osteoarthritis (unlikely at a younger age without risk factors and usually involves stiffness and swelling), inflammatory or infectious arthritis (no systemic symptoms or warmth/redness). The

most likely diagnosis is patellofemoral pain syndrome (PFPS), a common cause of anterior knee pain associated with activities that load the patellofemoral joint. It is often seen in physically active individuals and is aggravated by running, squatting, and stair use. Next steps: Activity modification (reduce aggravating activities temporarily), ice and NSAIDs as needed for pain, physical therapy focused on strengthening quadriceps, hip abductors, and stretching, consider evaluation by a healthcare provider for confirmation and tailored management. If symptoms worsen, persist beyond several weeks, or if new symptoms such as swelling, locking, instability, or systemic symptoms develop, further evaluation (including imaging) may be warranted.

Figure: Comparison of Vanilla LLM and VAIDYA on a left knee pain case. **VAIDYA is correct (patellofemoral pain syndrome), while Vanilla LLM is also correct but provides less detailed reasoning.**

Conversation 3

Vanilla LLM (o4-mini):
Patient: I've been having pain in the medial aspect of the right knee following a soccer injury for about Two days. It's onset is sudden.
Doctor: Thank you for sharing that information. To better understand your symptoms: Can you describe the nature of the pain—does it feel sharp, dull, aching, or throbbing? Did you hear or feel a pop at the time of the injury?
Patient: 1. The pain feels more like an aching sensation. 2. I didn't hear or feel a pop when it happened.
Doctor: Thank you for clarifying. Is the pain worse with certain movements or activities, such as walking, bending the knee, or twisting? Have you noticed any swelling, bruising, or instability (feeling like the knee might give way) since the injury?
Patient: 1. Yes, the pain gets worse when I try to bend my knee or twist it. Walking is also a bit uncomfortable. 2. Yes, there's swelling on the medial aspect of my right knee, and it's also red. I haven't noticed any bruising or instability, though.
Doctor: Thank you for the details. Are you able to fully straighten and bend your knee, or is your range of motion limited by pain or swelling? Have you had any locking (where the knee gets stuck and won't move) or catching sensations in the knee since the injury?
Patient: 1. It's a bit difficult to fully straighten and bend my knee because of the pain and swelling. 2. No, I haven't experienced any locking or catching sensations in my knee.
Doctor: Based on your symptoms of sudden onset medial knee pain, swelling, redness, and difficulty with full range of motion following a soccer injury—without a popping sensation, instability, or locking—I believe you most likely have a **medial collateral ligament (MCL) sprain**. This is because the MCL is commonly injured in sports involving twisting or impact to the knee, especially on the medial (inner) side, and typically presents with pain, swelling, and difficulty moving the knee, but usually without locking or a popping sound. My confidence in this diagnosis is high. Alternative diagnoses to consider include: Medial meniscus injury (less likely in absence of locking/catching, but possible), Pes anserine bursitis (can cause medial knee pain and swelling, but less commonly acute after trauma), Medial tibial plateau bone bruise or contusion. If symptoms worsen, you develop instability, or you are unable to bear weight, further evaluation (including imaging) may be necessary. Rest, ice, compression, and elevation (RICE) are helpful, but I recommend seeing a healthcare provider for a physical examination and to discuss possible imaging or physical therapy.

VAIDYA (o4-mini):

Patient: I've been having pain in the medial aspect of the right knee following a soccer injury for about Two days. It's onset is sudden.
Doctor: Thank you for sharing that. I'd like to ask a few more questions to better understand your knee pain: Can you describe exactly how the injury happened during soccer? (e.g., twisting, direct hit, sudden stop) Has the pain gotten better, worse, or stayed the same since it started two days ago? On a scale of 1 to 10, how severe is the pain right now? Does anything make the pain worse, like walking, bending, or twisting your knee? Is there anything that helps relieve the pain, such as rest, ice, or medication? Have you noticed any swelling, bruising, or instability (like your knee giving way)? Are you able to put weight on your right leg and walk normally? Have you had any previous injuries to this knee? Your answers will help me guide you further.
Patient: 1. I was running and suddenly twisted my knee while trying to change direction quickly. 2. The pain has stayed about the same since it started. 3. Right now, I'd say the pain is around a 5 out of 10. 4. Walking and bending my knee definitely make the pain worse. 5. Resting and applying ice seem to help a bit. 6. Yes, there's some redness and swelling on the medial aspect of my knee. 7. I can put some weight on it, but walking normally is difficult. 8. No, I haven't had any previous injuries to this knee.
Doctor: When you twist your knee or try to walk, do you feel any sensation of the knee "giving way" or instability, or does it just hurt without feeling unstable? Is the pain very localized to one spot on the inner (medial) side of your knee, or does it spread out over a larger area? Can you fully straighten and bend your knee, or is your range of motion limited? If limited, is it due to pain, swelling, or a feeling of mechanical blockage (like something is catching inside the knee)? If you press along the joint line on the inner side of your knee, is there a specific spot that is especially tender? Have you noticed any clicking, popping, or locking sensations in your knee since the injury? Is the swelling mostly over the bony part just below the knee joint (tibia), directly at the joint line, or higher up toward the thigh? Are you able to bear weight better when your knee is straight versus when it is bent? Do you have any numbness, tingling, or weakness in your lower leg or foot?
Patient: 1. It just hurts without feeling unstable. 2. The pain is very localized to one spot on the inner side of my knee. 3. My range of motion is limited due to pain and swelling. 4. Yes, there's a specific spot along the joint line that's especially tender. 5. I haven't noticed any clicking, popping, or locking sensations. 6. The swelling is mostly directly at the joint line. 7. I can bear weight better when my knee is straight. 8. I don't have any numbness, tingling, or weakness in my lower leg or foot.
Doctor: Thank you for those details. Since your pain is very localized to the inner joint line of your knee, started suddenly after a twisting injury, and is associated with swelling and tenderness right at the joint line (without instability or locking), could you tell me if you notice increased pain when you try to rotate your lower leg outward (external rotation) while your knee is slightly bent, or when someone presses on the inner joint line while your leg is in this position? This can help clarify if the medial meniscus or the medial collateral ligament is most likely involved.
Patient: Yes, I do feel increased pain when I try to rotate my lower leg outward while my knee is slightly bent. It also hurts when someone presses on the inner joint line in that position.
Doctor: Have you noticed any episodes where your knee suddenly locks up and you can't fully straighten or bend it, even temporarily, since the injury?
Patient: No, I haven't noticed any locking in my knee since the injury.
Doctor: Have you had any fever, chills, or feeling generally unwell since your injury, or is the pain strictly related to movement and pressure on your knee?
Patient: No, I haven't had any fever or chills. The pain seems to be related to movement and pressure on my knee.
Doctor: Diagnosis: Suspected Medial Meniscus Tear (Right Knee) Reasoning: The patient experienced a sudden twisting injury while changing direction during soccer, a classic mechanism

for meniscal injury. Pain is very localized to the medial (inner) joint line of the right knee, which is typical for a medial meniscus injury. Pain began suddenly after the injury and has persisted for two days. There is swelling and redness over the medial joint line, and the area is especially tender to touch. Pain worsens with walking, bending, and especially with external rotation of the lower leg while the knee is flexed (a provocative test for meniscus injury). No locking, catching, or instability reported, which suggests the tear may not be large or causing a mechanical block. No fever, chills, or signs of systemic illness, making infection unlikely. No previous knee injuries. Less likely diagnoses: Medial Collateral Ligament (MCL) Sprain (possible but less likely), Pes Anserine Bursitis, Fracture or Osteochondral Injury. Recommendation: The patient should avoid activities that worsen pain, continue rest and ice, and seek evaluation by a healthcare provider for a physical examination and possibly imaging (MRI) to confirm the diagnosis and guide management. If severe swelling, inability to bear weight, or mechanical symptoms (locking, catching) develop, urgent evaluation is warranted.

Figure: Comparison of Vanilla LLM and VAIDYA on a right knee injury case. **VAIDYA is incorrect (medial meniscus tear), while Vanilla LLM is correct (MCL sprain).**