MEDDxAgent: A Unified Modular Agent Framework for Explainable Automatic Differential Diagnosis

Daniel Rose[1], **Chia-Chien Hung**[2], **Marco Lepri**[2],
Israa Alqassem[2], **Kiril Gashteovski**[2,3] and **Carolin Lawrence**[2]
[1]University of California, Santa Barbara
[2]NEC Laboratories Europe, Heidelberg, Germany
[3]CAIR, Ss. Cyril and Methodius University of Skopje, North Macedonia
danielrose@ucsb.edu
{Chia-Chien.Hung, Marco.Lepri,
Israa.Alqassem, Kiril.Gashteovski, Carolin.Lawrence}@neclab.eu

Abstract

Differential Diagnosis (DDx) is a fundamental yet complex aspect of clinical decision-making, in which physicians iteratively refine a ranked list of possible diseases based on symptoms, antecedents, and medical knowledge. While recent advances in large language models (LLMs) have shown promise in supporting DDx, existing approaches face key limitations, including single-dataset evaluations, isolated optimization of components, unrealistic assumptions about complete patient profiles, and single-attempt diagnosis. We introduce a **M**odular **E**xplainable **DDx Agent** (**MEDDxAgent**) framework designed for interactive DDx, where diagnostic reasoning evolves through *iterative learning*, rather than assuming a complete patient profile is accessible. MEDDxAgent integrates three modular components: (1) an orchestrator (DDxDriver), (2) a history taking simulator, and (3) two specialized agents for knowledge retrieval and diagnosis strategy. To ensure robust evaluation, we introduce a comprehensive DDx benchmark covering respiratory, skin, and rare diseases. We analyze single-turn diagnostic approaches and demonstrate the importance of iterative refinement when patient profiles are not available at the outset. Our broad evaluation demonstrates that MEDDxAgent achieves over 10% accuracy improvements in interactive DDx across both large and small LLMs, while offering critical explainability into its diagnostic reasoning process.

1 Introduction

Differential Diagnosis (DDx) is a crucial step in medical decision-making, where doctors systematically narrow down the most likely diagnosis from a range of possible diseases (Rhoads et al., 2017). In real-world clinical practice, DDx is essential because it accounts for uncertainty in the diagnosis (Henderson et al., 2012). It's also incredibly challenging given the large number of potential diseases, rapidly evolving medical knowledge, and the fact that symptoms and antecedents can point to multiple diseases (Winter et al., 2024). Expert clinicians rely on pattern recognition and past experience to narrow down potential diseases. However, the complexity and variability of real-world clinical presentations have prompted recent research into computational frameworks that use large language models (LLMs) to improve the DDx process (Fansi Tchango et al., 2022a; Zhou et al., 2024).

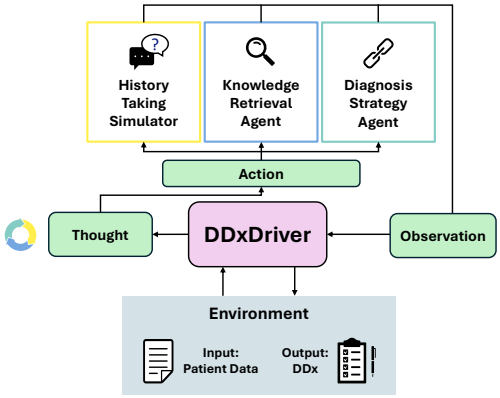


Figure 1: **MEDDxAgent** facilitates differential diagnosis by iteratively narrowing down a patient's possible disease. DDxDriver acts as the central orchestrator. It receives an interactive environment via a simulator (History Taking) and can access two agents (Knowledge Retrieval, Diagnosis Strategy).

Though LLM-based systems have shown promise in improving diagnostic assistance, existing methods face several limitations: (1) reliance on *single-dataset evaluations*, limiting the generalizability across diverse patient populations and disease categories (Alam et al., 2023); (2) focus on *optimizing a single diagnostic component* (e.g., diagnosis strategy *only*) (McDuff et al., 2023), without an integrated approach to enhance multiple

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13803–13826
July 27 - August 1, 2025

phases of the diagnostic process;[1] (3) assumption of *complete patient profiles* upfront (i.e., with all symptoms and antecedents) (Wu et al., 2024) and *single-turn* paradigm (Zelin et al., 2024), diverging from the reality that DDx is an investigative process, requiring follow-up actions to gather information (Li et al., 2024b); (4) lack of *iterative learning*, preventing diagnosis updates over successive interactions – an essential aspect of real-world diagnostic decision-making; (5) an *over-reliance on medical QA benchmarks* (Zhang et al., 2024) for medical applications, which do not accurately reflect the complexities of real-world DDx tasks.

We target these gaps and propose a **M**odular **E**xplainable **DDx Agent** (**MEDDxAgent**) framework (see Figure 1). It consists of (1) DDxDriver that acts as the central orchestrator; (2) a history taking simulator which enables an iterative environment; and (3) two individual agents – knowledge retrieval and diagnosis strategy – to support the diagnostic process. We advance the task of automatic DDx with the following contributions: (i) We propose a modular, multi-faceted DDx agent framework (MEDDxAgent), integrating a history taking simulator and two diagnostic agents (knowledge retrieval, diagnosis strategy), which enables extensible and explainable decision-making processes. (ii) We introduce an orchestrator (DDxDriver) as a unified interface, ensuring *iterative learning* and interactive optimizations between agents, as well as monitoring of the decision-making process. (iii) We build a new DDx benchmark incorporating three diagnostic sources with different disease categories: DDxPlus (Fansi Tchango et al., 2022b) (*respiratory*), iCRAFT-MD (Li et al., 2024b) (*skin*), and RareBench (Chen et al., 2024b) (*rare*). This allows for a more comprehensive diagnostic scope than in existing work. (iv) We evaluate MEDDxAgent in a more challenging but realistic scenario – *interactive differential diagnosis*, and demonstrate its effectiveness by achieving over 10% points improvements in accuracy (i.e., GTPA@1) for both large (70B) and small (8B) LLMs. (v) The code is publicly available.[2]

2 MEDDxAgent Overview

Our proposed MEDDxAgent framework (see Figure 1 and a detailed version in Figure 2) comprises a central orchestrator (DDxDriver), a history taking simulator, and two specialized diagnostic agents dedicated to knowledge retrieval and diagnosis strategy. Both the simulator and diagnostic agents communicate exclusively with the DDxDriver, which monitors, stores, maintains, and updates patient information and ranked differential diagnoses. This central role also positions the DDxDriver to coordinate an iterative feedback loop, wherein observations from each agent are leveraged to enhance and refine subsequent agent calls with agent instructions. In the following, we introduce the design of simulator (§ 2.1), agents (§ 2.2), orchestrator (DDxDriver) (§ 2.3), and iterative learning mechanism (§ 2.4).

2.1 Simulator

History taking is a critical first step in differential diagnosis, where clinicians gather essential information by asking patients questions about their symptoms, medical history, and lifestyle factors. In real-world clinical settings, a full patient profile is rarely available at the outset (Li et al., 2024b) – doctors typically start with only partial information (e.g., age, gender, chief complaint). The process of interactive DDx allows clinicians to gather more patient information and refine their diagnostic hypotheses before making a follow-up decision.

To simulate such an interactive environment, we introduce a history taking simulator. We initialize the simulator with two LLMs (Wu et al., 2023) in our experiments. The first LLM simulates the patient and receives access to the full patient profile. The second LLM simulates the doctor and receives an initial patient profile and optionally a set of conversational goals defined by DDxDriver (*action*). During the interactions, the doctor role asks questions relevant to the diagnosis process, and the patient role provides answers based on its patient profile. The interaction continues until either the conversational goals are achieved or a predefined stopping criterion (e.g., maximum number of questions) is reached. Once the conversation concludes, the dialogue history is forwarded to DDxDriver.

2.2 Agents

Knowledge Retrieval Agent. This agent aids the diagnostic process by retrieving relevant medical knowledge from external sources, such as scientific literature, medical databases, and clinical guidelines. This is particularly critical for diagnosing rare or complex conditions where external knowl-

[1]The DDx process typically involves three key components: history taking, knowledge retrieval, and diagnosis strategy (Cook and Décary, 2020; Kavanagh et al., 2024).

[2]https://github.com/nec-research/meddxagent

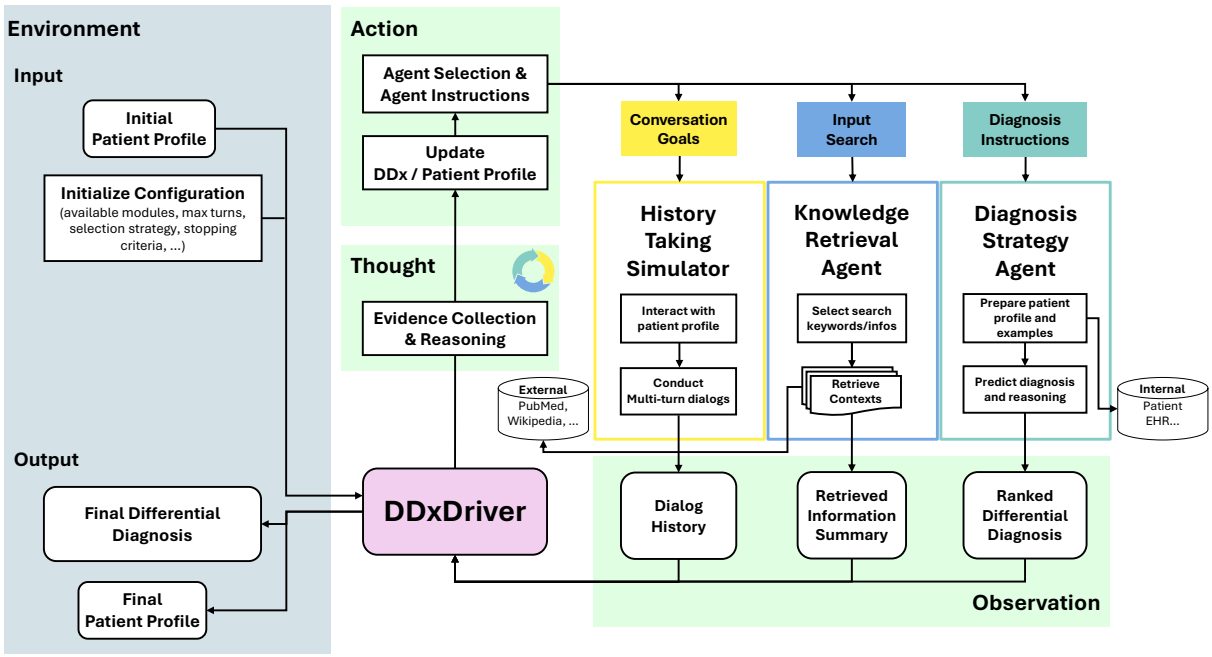


Figure 2: The architecture of the MEDDxAgent framework. MEDDxAgent unifies a central orchestrator (DDxDriver), a simulator (History Taking) and two agents (Knowledge Retrieval, Diagnosis Strategy). The framework follows the ReAct (Yao et al., 2023) paradigm (*thought*, *action*, *observation*), enabling sequential reasoning and action steps with transparent logging of all interactions through the iterative learning process.

edge (as compared to internal knowledge learned by LLMs' training data) is required to enhance clinical reasoning with validated information.

Upon activation, the agent receives a search query formulated by DDxDriver, based on the current patient profile and provisional DDx list. It extracts the key medical concepts from the query as structured keywords, then conducts a targeted search in external databases. We consider two primary sources: Wikipedia and PubMed[3], with the former providing concise summaries of top-ranked pages, while the latter retrieves abstracts of full-access articles. The retrieved knowledge is synthesized into an evidence-based summary, which ensures that the diagnostic reasoning process has access to up-to-date, relevant medical knowledge.

Diagnosis Strategy Agent. This agent is responsible for generating, refining, and ranking possible diagnoses based on the information prepared by DDxDriver. There are two distinct modes that can be chosen for the diagnosis strategy agent. First, in the zero-shot setting, the LLM predicts the most probable diagnoses solely based on the current patient. This approach is straightforward but may have limited accuracy for complex or rare conditions. Second, in the few-shot setting, the diagnosis strategy agent utilizes additional patient cases to guide its predictions, enabling more context-aware diagnostic reasoning. We explore two variations. First, in a standard few-shot approach, a fixed set of patient examples are selected as references and provided to the model alongside the current patient profile. Second, the dynamic few-shot approach improves upon this by selecting reference cases based on similarity metrics, ensuring that the most relevant patients are included. Patient similarity is determined using embedding-based retrieval, with two embeddings (i.e., BioClinicalBERT (Alsentzer et al., 2019), BGE (Xiao et al., 2024)) evaluated to match patients with similar profiles.

We also integrate Chain-of-Thought (CoT) reasoning (Wei et al., 2022), guiding the model to explicitly reason through intermediate clinical steps before predicting a diagnosis. CoT can be combined with both standard and dynamic few-shot approaches, with a stepwise rationale for each diagnosis. Inspired by MedPrompt (Nori et al., 2023b), we extend CoT by incorporating structured, example-driven reasoning, where each reference case includes both a diagnosis and an associated CoT explanation. The integration of CoT enables the system to better handle complex cases with diagnostic uncertainty, such as specific skin diseases in iCRAFT-MD which share common symptoms.

[3] https://pubmed.ncbi.nlm.nih.gov/

Once the model completes the diagnostic inference, the ranked list of differential diagnoses is returned to DDxDriver, which further refines or finalizes the diagnosis through iterative updates. By distinguishing between zero-shot and few-shot inference strategies, plus dynamic adaptation through embeddings and reasoning techniques, the diagnosis strategy agent aims to enhance both accuracy and generalizability.

2.3 Orchestrator

Inspired by the concept of a unified interface layer from previous work (Gioacchini et al., 2024), we introduce DDxDriver as the central coordination hub in the MEDDxAgent framework (Figure 1). DDxDriver enables modular compatibility between the diagnostic agents and benchmark datasets, with minimal adaptation efforts. DDxDriver uses the ReAct paradigm (Yao et al., 2023) – which combines step-by-step reasoning (*thought*) with decision-making (*action*) and feedback processing (*observation*). At each step, DDxDriver obtains the information from the *environment* (input/output) and the results from the previous state of the simulator and agents (*observation*, if it exists), then reasons about the current state of evidence (*thought*) and generates agent-specific instructions (*action*) based on the current state of the patient profile. It dispatches these instructions to the selected simulator/agent, executes, and then updates the patient profile with newly obtained information (*action*). Beyond execution management, DDxDriver serves four primary functions. First, it manages the patient profile, storing and maintaining all relevant clinical information, including demographics, medical history, symptoms, and evolving diagnostic rankings. Second, it schedules and dispatches diagnostic actions, dynamically determining which simulator/agent to invoke next based on the evolving diagnostic context. Third, it ensures traceability by logging all interactions, including inputs, outputs, and intermediate reasoning steps, thereby providing transparency in the decision-making process. Finally, it enforces stopping criteria by monitoring diagnostic convergence and applying configurable thresholds, such as the number of iterations or the stabilization of ranked diagnoses.

2.4 Iterative Learning Mechanism

Diagnoses in the real world are rarely made in a single step. They are refined through multiple interactions with patients, clinical data, and external knowledge. To mirror this process, the *iterative learning* mechanism is designed to avoid relying on any single diagnostic agent or static decision process. We implement two settings: (i) *fixed iteration*, and (ii) *dynamic iteration*. *Fixed iteration* cycles through the history taking simulator, knowledge retrieval agent, and diagnosis strategy agent in order until the predefined stopping criterion is met (e.g., n iterations). In contrast, the *dynamic iteration* process lifts constraints on the predetermined execution order, allowing the DDxDriver to adapt dynamically during the differential diagnosis process. After each observation, the DDxDriver reasons about which component – history taking simulator, knowledge retrieval agent, or diagnosis strategy agent – to call next based on up-to-date observations (i.e. updated patient profile, medical documents, predicted DDx). For instance, if the current diagnosis indicates a rare condition for which it needs clarifying details, the system may invoke the knowledge retrieval agent to search for specialized information. Similarly, if the retrieved knowledge introduces ambiguity, the system may loop back to the history taking simulator to clarify new symptoms or risk factors. This allows for flexible decision-making, opening up the opportunity for the diagnostic process to dynamically adjust as new information becomes available. The iterative learning mechanism allows MEDDxAgent to continuously refine diagnosis while offering transparent insights into its reasoning process. A detailed workflow example is illustrated in Figure 3.

3 Experimental Setup

3.1 DDx Benchmark

We introduce a comprehensive DDx benchmark integrating three datasets – DDxPlus, iCraft-MD, and RareBench, covering *respiratory*, *skin*, and *rare* diseases for a robust assessment of diagnostic performance. This addresses limitations of prior work, which often relies on a single dataset and single-turn evaluation for differential diagnosis. DDxPlus (Fansi Tchango et al., 2022b) provides a large-scale, structured dataset with 1.3 million synthetic respiratory patient cases across 49 respiratory-related pathologies. iCraft-MD (Li et al., 2024b) includes 394 skin diseases, adapting static dermatological clinical vignettes (from original Craft-MD dataset (Johri et al., 2024, 2025))

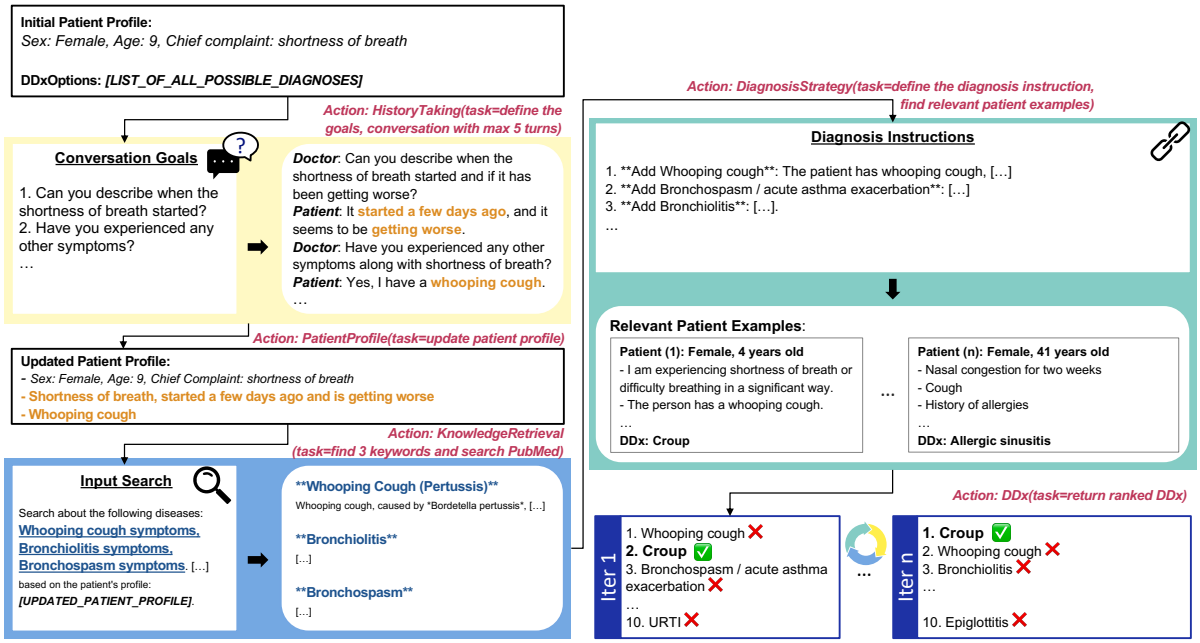


Figure 3: An illustrated DDxPlus (Fansi Tchango et al., 2022b) example with the MEDDxAgent framework. Given the initial patient profile and list of diagnosis options, DDxDriver determines the goals and actions for the simulator (History Taking) and agents (Knowledge Retrieval, Diagnosis Strategy), updating the patient profile, and returning the ranked DDx. Each step is logged for transparency, enabling iterative refinement and learning.

into an interactive setting[4] – the system is only provided with partial patient information and is expected to proactively ask questions and gather information. RareBench (Chen et al., 2024b) expands DDxPlus with 421 rare diseases. We select three subsets from RareBench – RAMEDIS (Europe), MME (Canada), and PUMCH (China) – to ensure diversity in regional representation.

To enable a consistent evaluation across datasets, we standardize each dataset into a structured format: (i) optional initial patient information (e.g., age, sex, chief complaint); (ii) full patient profile (complete list of symptoms and antecedents); and (iii) full set of possible diseases for differential diagnosis. This refinement enhances diagnostic consistency and supports the evaluation of interactive DDx. We sample 100 patients from each dataset at a fixed random seed, due to the cost of experiments and excessive time for reasoning steps. Detailed dataset statistics are in Appendix A.

3.2 Evaluation Metrics

To evaluate diagnostic performance, we employ three metrics. First, we compute the *average rank* of the correct disease, which represents the model's ability to position the correct diagnosis closer to the top. If the diagnosis does not appear in the top-10 position, we assign a rank of 11. Second, we use *GTPA@k* (Ground Truth Pathology Accuracy) (Fansi Tchango et al., 2022b), which measures whether the ground truth diagnosis appears within the top-k predicted diagnoses. Third, we introduce a new metric suitable for the iterative setting: *average progress rate* (Δ Progress). Inspired by AgentQuest (Gioacchini et al., 2024), it tracks changes in rank r of the ground truth pathology in the differential diagnosis. For each patient case i, we average the progress in rank $(r_{i,t} - r_{i,t+1})$ over N iterations of differential diagnosis, then aggregate over M patients. This metric quantifies how effectively the system refines and converges on the correct diagnosis over successive iterations:

$$\Delta\text{Progress} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N_i - 1}\sum_{t=1}^{N_i-1}\left(r_{i,t} - r_{i,t+1}\right)\right)$$

3.3 Models and Tasks

We evaluate on GPT-4o (version: 2024-11-20) (Hurst et al., 2024), Llama3.1-70B and Llama3.1-8B (Dubey et al., 2024) across all tasks, ensuring a comparison of LLMs at varying scales. Our experiments are conducted in two setups: (1) optimizing individual agents; and (2) interactive

[4]Interactive DDx is a more complex information-seeking setup, since in the real world the full patient profile might not be accessible initially.

differential diagnosis. In the first task, we evaluate the two agents (knowledge retrieval, diagnosis strategy) in a single-turn setting. This allows us to isolate the effectiveness of the reasoning mechanisms without the confounding factor of incomplete information. In the second task, we assess MEDDxAgent's performance at interactive DDx, comparing it against the single-turn diagnostic agents and history taking simulator. Interactive differential diagnosis, as suggested by Li et al. (2024b), is a challenging yet realistic scenario, where only initial patient information is available – without a complete list of symptoms and antecedents. This setup highlights how limited information constrains the single-turn setting (i.e., no iteration), compared to MEDDxAgent's iterative interactions, which refine and enhance the diagnostic process.

3.4 Hyperparameters and Optimization

For the knowledge retrieval agent, we limit searches to a maximum of three medical keywords per query. Wikipedia is used as an open-access resource, while PubMed retrieval is restricted to full-text articles from commercially licensed sources,[5] ensuring that retrieved information is clinically validated and relevant to the diagnostic task. For the diagnosis strategy agent, we take 5 examples for few-shot learning. For dynamic few-shot, we use BioClinicalBERT (BERT) (Alsentzer et al., 2019) and BGE-BASE-EN-V1.5 (BAII) (Xiao et al., 2024) embeddings, based on the structure proposed by Wu et al. (2024). Specifically, it uses L2 distance on normalized embeddings, a similar setting to cosine similarity. With the history taking simulator, we create an iterative environment, which we evaluate at 5, 10, and 15 maximum questions. This is based on prior clinical studies that indicate physicians typically ask fewer than 15 questions per consultation (Ely et al., 1999). This ensures that our model operates within a realistic range, capturing essential patient details without excessive interaction. To evaluate MEDDxAgent's iterative learning, we select the optimized history taking simulator and diagnostic agents and experiment on interactive DDx. Our setup is inspired by previous work (Johri et al., 2025), which demonstrates that updating the patient profile with new history-taking dialogue significantly enhances performance. We experiment with 1 to 3 iterations, with 5 questions per iteration. This aligns with the history-taking simulator setting (5 questions per iteration, max 15 for 3 iterations). Additionally, we set the DDxDriver's instruction for each agent and simulator to a list of length 10.

4 Evaluation Results

We experiment on two configurations: (1) optimizing individual agents (§ 4.1), by determining the best settings for knowledge retrieval and diagnosis strategy agents; and (2) interactive differential diagnosis (§ 2.4), where the optimized agents are used to assess MEDDxAgent's performance in the interactive DDx setup.

4.1 Optimizing Individual Agents

We first explore the optimal single-turn configuration for the knowledge retrieval and diagnosis strategy agents, before integrating them into the iterative setup. For this, we provide the full patient profile as in previous work (Wu et al., 2024; Chen et al., 2024b), and present the results in Table 1. For the knowledge retrieval agent, PubMed performs slightly better overall than Wikipedia, especially for Rarebench, which demands more complex disease information. For the diagnosis strategy agent, the best setting varies by dataset. Namely, dynamic few-shot with BAII embeddings performs the best on DDxPlus and RareBench, where relevant patient examples offer reliable contextual cues to likely diseases. In contrast, iCraft-MD benefits more from zero-shot CoT, which enables structured reasoning through complex clinical vignettes. Few-shot learning often decreases performance for iCraft-MD because each patient vignette is distinct, so additional examples can introduce noise. Based on the above findings, we select the following configurations for the iterative scenario:[6] PubMed for knowledge retrieval agent; few-shot (dynamic BAII) for DDxPlus and RareBench, and zero-shot (CoT) for iCraft-MD for diagnosis strategy agent.

4.2 Interactive Differential Diagnosis

We now evaluate the more challenging task of interactive DDx, where we begin with limited patient information and the history taking simulator enables the interactive environment (Table 2). The process is initialized with the number of turns n of the history taking simulator, coupled with either (a) knowledge retrieval agent only (KR), (b) diagnosis strategy agent only (DS), or (c) MEDDxAgent

[5] We use MediaWiki API: https://en.wikipedia.org/w/api.php and BIOPYTHON https://biopython.org/.

[6] We do not run all possible settings in the interactive environment due to cost reasons.

	DDxPlus			iCraft-MD			RareBench		
	GTPA@1 ↑	GTPA@5 ↑	Avg Rank ↓	GTPA@1 ↑	GTPA@5 ↑	Avg Rank ↓	GTPA@1 ↑	GTPA@5 ↑	Avg Rank ↓
GPT-4o									
Retrieval (PubMed)	0.69	0.90	2.27	0.68	**0.79**	3.23	0.45	0.72	3.92
Retrieval (Wiki)	0.69	0.90	2.24	**0.69**	**0.79**	**3.22**	0.45	0.74	4.00
Zero-shot (Standard)	0.69	0.90	2.21	0.68	0.77	3.37	0.46	0.72	3.99
Zero-shot (CoT)	0.71	0.92	2.10	0.68	0.77	3.35	0.47	0.69	4.02
Few-shot (Standard, Dyn_BAII)‡	0.96	**1.00**	1.06	0.62	0.72	3.85	0.79	**0.91**	**2.03**
Few-shot (CoT, Dyn_BERT)	0.96	**1.00**	1.05	0.64	0.73	3.68	0.81	**0.91**	2.04
Few-shot (CoT, Dyn_BAII)	**0.97**	**1.00**	**1.03**	0.60	0.70	4.00	**0.82**	0.88	2.11

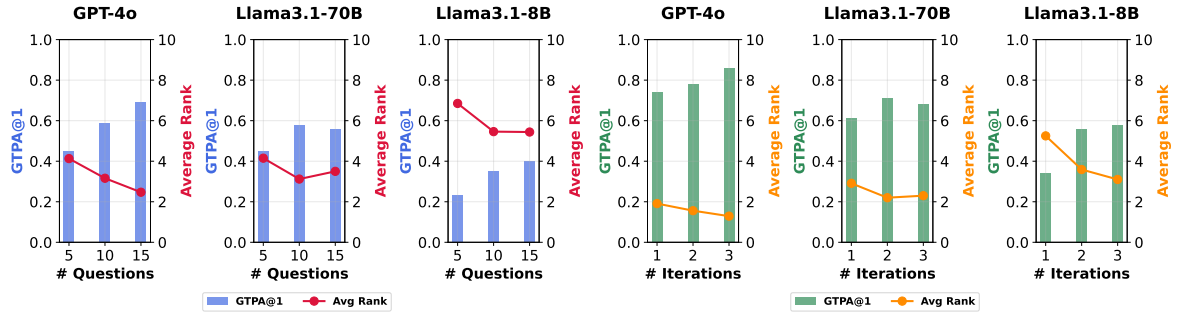
Table 1: Results in the non-interactive setting for the knowledge retrieval agent (*upper*) and the diagnosis strategy agent (*bottom*). ‡ Only Few-shot (Standard, Dyn_BAII) results are recorded, since the method is consistently better than Dyn_BERT. All models exhibit similar trends. To give a more concise overview, we only report GPT-4o here. The full set of results can be found in Table 7 in Appendix.

	DDxPlus			iCraft-MD			RareBench		
	GTPA@1 ↑	Avg Rank ↓	Δ Progress	GTPA@1 ↑	Avg Rank ↓	Δ Progress	GTPA@1 ↑	Avg Rank ↓	Δ Progress
GPT-4o									
KR (*n*=0)	0.18	7.33	-	0.15	8.27	-	0.07	9.07	-
DS (*n*=0)	0.27	6.01	-	0.18	7.87	-	0.11	8.38	-
KR (*n*=5)	0.52	3.32	-	0.49	5.36	-	0.40	5.27	-
DS (*n*=5)	0.72	2.14	-	0.40	5.55	-	0.50	4.94	-
MEDDx (*iter*=1, *n*=5)	0.74	1.91	0.00	0.52	4.93	0.00	0.51	4.37	0.00
MEDDx (*iter*=2, *n*=10)	0.78	1.56	+0.32	**0.54**	**4.71**	+0.26	**0.56**	4.10	+0.13
MEDDx (*iter*=3, *n*=15)	**0.86**	**1.29**	+0.32	**0.54**	4.80	+0.17	0.50	**4.09**	+0.16
Llama3.1-70B									
KR (*n*=0)	0.19	7.58	-	0.13	8.19	-	0.09	9.13	-
DS (*n*=0)	0.17	7.28	-	0.11	8.74	-	0.20	6.81	-
KR (*n*=5)	0.39	5.03	-	0.34	6.86	-	0.29	5.86	-
DS (*n*=5)	0.50	2.89	-	0.24	7.33	-	0.23	5.77	-
MEDDx (*iter*=1, *n*=5)	0.61	2.91	0.00	0.29	7.05	0.00	0.39	5.05	0.00
MEDDx (*iter*=2, *n*=10)	**0.71**	**2.20**	+0.41	0.37	**6.26**	+0.07	**0.48**	4.48	+0.75
MEDDx (*iter*=3, *n*=15)	0.68	2.30	+0.17	**0.42**	6.31	+0.26	**0.48**	**4.30**	+0.44
Llama3.1-8B									
KR (*n*=0)	0.20	7.49	-	0.11	**8.86**	-	**0.11**	8.58	-
DS (*n*=0)	0.16	8.45	-	0.03	10.37	-	0.04	8.52	-
KR (*n*=5)	0.21	7.42	-	0.09	9.48	-	0.04	9.69	-
DS (*n*=5)	0.23	5.77	-	0.03	10.08	-	0.06	8.64	-
MEDDx (*iter*=1, *n*=5)	0.34	5.25	0.00	0.11	9.38	0.00	0.08	8.47	0.00
MEDDx (*iter*=2, *n*=10)	0.56	3.59	+1.73	**0.14**	9.22	+0.22	0.09	**8.11**	+0.44
MEDDx (*iter*=3, *n*=15)	**0.58**	**3.10**	+1.23	0.12	9.07	+0.17	0.07	8.56	+0.38

Table 2: Interactive experiment performance across 3 datasets without *full* patient profile, with KR: knowledge retrieval agent; DS: diagnosis strategy agent; n is the number of turns of the simulator; MEDDx uses KR+DS.

with both knowledge retrieval and diagnosis agents (MEDDx). At $n = 0$, the simulator has not yet learned any patient information, and performance drops significantly from observing the full patient profile (Table 1). For GPT-4o in RareBench, the knowledge retrieval agent (KR)'s GTPA@1 drops from 0.45 to 0.07. Similarly, the diagnosis strategy agent (DS) drops from 0.46 (zero-shot) to 0.11. This simple baseline showcases that previous evaluations do not hold well in the interactive setup with initially limited patient information. Already for $n = 5$, we find a large boost in performance for both KR and DS. These findings emphasize the importance of history taking for diagnostic precision. We illustrate the trend for changing n in Figure 4 and find that gains also plateau around n=10-15 questions, reinforcing the optimal balance between information gathering and diagnostic efficiency (Ely et al., 1999).

Finally, we run MEDDxAgent, which calls KR+DS in the *fixed iteration* pipeline (§ 2.4). MEDDxAgent exhibits clear improvements over the KR and DS baselines for $n = 5$, supporting our hypothesis that all three modules are important for interactive DDx. It also improves significantly over the history taking baselines, as we illustrate in Figure 4. MEDDxAgent is also capable of improving upon the zero-shot setting with the full patient profile (Table 1). For DDxPlus, GTPA@1 for GPT-4o and Llama3.1-70B rise from 0.69 to 0.86 and from 0.54 to 0.71, respectively. For Llama3.1-8B, the trend continues for DDxPlus but inconsistently for iCraft-MD and RareBench, highlighting the importance of model scale. Notably, MEDDxAgent



(a) (b)

Figure 4: Results of DDxPlus compared between (a) history taking simulator, and (b) MEDDxAgent, over the number of questions and iterations. For brevity, the results of iCraft-MD and RareBench are in Appendix C.

improves over successive iterations, though the optimal number of iterations (2, 3) depends on the dataset and LLM. The values of Δ are consistently positive, indicating that MEDDxAgent iteratively increases the rank of the ground-truth diagnosis over time. Δ Progress also varies by dataset and model, offering explainable insight to the diagnosistic improvement of MEDDxAgent. The overall results show that MEDDxAgent can operate well in the challenging, realistic setup of interactive DDx. Additionally, MEDDxAgent logs all intermediate reasoning, action, and observations, providing critical insight into its DDx process (Figure 3).

5 Analysis

Fixed vs. Dynamic Iterations. A key feature of MEDDxAgent is its iterative DDx process, which operates in *fixed* or *dynamic* iteration. Our experiments (see Table 8 in Appendix) show that fixed iteration consistently outperforms dynamic iteration in both accuracy and system efficiency.[7] Fixed iteration ensures a structured sequence where all modules – history taking simulator, knowledge retrieval agent, and diagnosis strategy agent, are utilized in each cycle, preventing over-reliance on a single component. In contrast, dynamic iteration, which allows DDxDriver to choose the component at each step, introduces some suboptimal decision-making. We observe that Llama3.1-* models, for instance, frequently favor the history taking simulator rather than leveraging the knowledge retrieval or diagnosis strategy agents, leading to redundant questioning rather than efficient diagnostic reasoning. Despite this, our findings demonstrate the general applicability of MEDDxAgent for dynamic iteration and highlight future work toward optimizing dynamic iteration for interactive DDx.

Error Analysis. To better understand the struggles of MEDDxAgent, we conduct error analysis on cases where it failed to reach the correct diagnosis efficiently. We emphasize that our MEDDxAgent's logging of intermediate logic greatly enhances our understanding and explanations of failure cases. First, in RareBench, over-reliance on few-shot examples often misprioritizes frequent conditions over rarer diseases, as some rare conditions are underrepresented in knowledge retrieval databases. Second, in the first iteration, MEDDxAgent tends to prioritize the knowledge retrieval while overlooking few-shot patient examples with similar profiles. This is then mitigated when further iterations help to refine and enhance the DDx process. Third, larger models (GPT-4o, Llama3.1-70B) benefit more from iterative refinement, while smaller models (Llama3.1-8B) plateau after the second iteration, especially for iCraft-MD and RareBench. Fourth, recent studies have shown that domain-specific medical LLMs often underperform general-purpose models in diagnostic tasks (Nori et al., 2023b,a; Maharjan et al., 2024). We hypothesize this occurs because such models are typically fine-tuned on specific tasks or disease specialities, limiting their general instruction-following capability and diagnostic accuracy. Our experiments confirm this hypothesis, demonstrating that medical LLMs fine-tuned on tasks such as QA and open-ended medical chat (Zhang et al., 2024) perform worse than general LLMs of similar sizes (see Appendix D). Addressing these challenges can further enhance MEDDxAgent's diagnostic accuracy.

[7] On average, fixed iteration is 1.2x-1.7x faster than dynamic iteration.

6 Related Work

6.1 LLM-based Methods

Researchers have studied the capabilities of LLMs for automatic diagnosis (Mizuta et al., 2024). One line of work has found that the performance of LLMs is comparable to the performance of physicians (Hirosawa et al., 2023; Rutledge, 2024), or that the performance of the physicians themselves is improved when they use LLMs (Ten Berg et al., 2024). However, researchers have also observed that LLMs struggle to perform this task when applied on rarer diseases or on more unusual cases (Fabre et al., 2024; Shikino et al., 2024). For this reason we assemble a benchmark that measures performance for different rarity levels.

Many of the current methods are typically targeting either automatic standard diagnosis or automatic differential diagnosis in an end-to-end manner. For example, there are methods that use Chain-of-Thought strategies (Wu et al., 2023; Savage et al., 2024; Nachane et al., 2024), reinformcement learning (Fansi Tchango et al., 2022a), fine-tuning LLMs (Alam et al., 2023; Saab et al., 2024; Reese et al., 2024), preranking-reranking methods (Sun et al., 2024) or specifically trained neural networks (Liu et al., 2020; Hwang et al., 2022). Such LLM-based methods do not allow for modularity, posing difficulties for integrating specific modules that solve sub-problems within the diagnosis process.

6.2 Agent-based Methods

Recent work has shifted from standalone LLMs to multi-agent frameworks, enhancing efficiency by enabling external tools, assigning specialized roles to each agent to accomplish complex tasks more efficiently. In medical applications, agent-based methods streamline clinical workflows to improve diagnostic accuracy. KG4Diagnosis (Zuo et al., 2024) integrated LLMs with knowledge graphs (KGs) for medical diagnosis. However, its static KG dependence makes expanding diagnoses for rare diseases difficult, and the lack of iterative refinement limits adaptability to evolving clinical cases. Wu et al. (2024) presented StreamBench to evaluate the continuous improvement of LLM agents in streaming environments via simulated feedback. However, it evaluates on full patient profiles and a single DDx dataset, limiting its generalizability. Among recent advances, multi-turn diagnostic frameworks such as CoD (Chen et al., 2024a) and MediQ (Li et al., 2024b) have placed particular emphasis on modeling the history-taking process through iterative dialogue. In these systems, agents engage in multi-turn interactions to simulate how clinicians gather patient history: asking follow-up questions and incrementally collecting information needed for differential diagnosis. The focus on iterative, dialogue-based history taking brings agent-based systems closer to real-world clinical reasoning, where uncertainty is managed through successive inquiry and response. Other frameworks, such as AMIE (Tu et al., 2024), AMSC (Wang et al., 2024), AgentHospital (Li et al., 2024a), MedAgents (Tang et al., 2024) and MedAgentBench (Jiang et al., 2025), address different aspects of clinical interactions. However, these approaches suffer from (i) lack of iterative refinement, relying on single-turn reasoning; (ii) limited evaluation, focusing on a single dataset or a single diagnostic component, limiting diagnostic generalization; (iii) assuming full patient profiles as the input, which does not reflect real-world interactive differential diagnosis in which clinicians collect information iteratively. We address these aspects in our work, proposing a more challenging interactive setting and a modular, iterative agent framework.

7 Conclusions

Existing approaches to automatic differential diagnosis rely on single-dataset evaluations, assume fully observed patient profiles, focus on optimizing isolated diagnostic components, or diagnose in a single attempt. We introduce MEDDxAgent, a modular and explainable framework that enhances automatic differential diagnosis through iterative learning. MEDDxAgent integrates a history taking simulator, two agents (knowledge retrieval, diagnosis strategy), and an orchestrator (DDxDriver) to tackle the more challenging and realistic scenario of interactive differential diagnosis, where patient profiles are initially incomplete. Its modular design enables systematic evaluation of optimal agent configurations, while intermediate logging and a novel average progress rate metric provide critical transparency into its reasoning process. Experimental results demonstrate that interactive differential diagnosis is significantly more challenging, allowing MEDDxAgent to iteratively refine predictions and outperform simpler, single-turn approaches. We hope this framework fosters continued progress in developing more adaptive and effective models for automatic differential diagnosis.

Acknowledgments

We would like to thank Andreas Ripke for the support of the infrastructure and the anonymous reviewers for their valuable feedback.

Limitations

While our proposed MEDDxAgent framework advances the task of automatic differential diagnosis through a modular, explainable, and interactive approach, we acknowledge certain limitations. (1) Model Selection: Our evaluation focuses on Llama models (8B and 70B) and GPT-4o. These models demonstrate strong instruction-following capabilities, so findings may not generalize to all LLM architectures. We also evaluate state-of-the-art medical-domain LLMs (Zhang et al., 2024) but find that these instruction-tuned models underperform the general-purpose LLMs (e.g., Llama3.1 (Dubey et al., 2024)), likely due to diverse instruction-following behavior. Further exploration is needed for models with different architectures, training paradigms, or domain adaptations. (2) Language Coverage: our framework is evaluated primarily on an English DDx benchmark, limiting its applicability to non-English-speaking regions, where medical terminology, case presentations, and healthcare practices may differ significantly. Extending the framework for multilingual and cross-lingual diagnostic tasks remains an important direction (Qiu et al., 2024; García-Ferrero et al., 2024). (3) Multimodality: Medical information often relies on multimodal data, such as medical imaging and videos, laboratory tests, electronic health records, and genomic/pathology data. Our DDx Benchmark is soley text-based, which limits the scope of its applicability to clinical decision making. Future work should explore multimodal agentic frameworks (Schmidgall et al., 2024; Kim et al., 2024) and reasoning over multimodal sequential data (Rose et al., 2023; Himakunthala et al., 2023; Zhu et al., 2024). (4) Benchmark Dataset Selection: We introduce a DDx benchmark encompassing respiratory, skin, and rare diseases. Though our benchmark offers greater diversity than prior work, it does not cover all medical specialties or real-world patient distributions. Expanding datasets to reflect a wider range of diseases, demographics, and clinical settings would improve generalizability. (5) MEDDxAgent requires significant communication among components and tools (i.e., DDxDriver, agents, patient profiles, web search, ...), allowing higher latency and computational cost for more thorough reasoning and improved diagnostic accuracy. Future deployments could reduce latency through strategies like parallelism, caching, memory optimization, model selection, and prompt engineering. (6) Training and Deployment: Although MEDDxAgent provides all intermediate reasoning logs and citations, proper physician training is essential to avoid over-reliance on AI-generated diagnoses, which may contain inaccuracies or hallucinations.

Despite these limitations, we hope this work contributes to advancing automatic DDx with a challenging yet realistic setup – interactive differential diagnosis. We hope that future research builds on top of our findings to include more languages, modalities, datasets, and a deeper exploration of diagnostic processes to enhance the applicability and effectiveness of interactive diagnostic models.

References

Mohammad Mahmudul Alam, Edward Raff, Tim Oates, and Cynthia Matuszek. 2023. DDxt: Deep generative transformer models for differential diagnosis. In *Deep Generative Models for Health Workshop NeurIPS 2023*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024a. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. 2024b. Rarebench: Can llms serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 4850–4861, New York, NY, USA. Association for Computing Machinery.

Chad E. Cook and Simon Décary. 2020. Higher order thinking about differential diagnosis. *Brazilian Journal of Physical Therapy*, 24(1):1–7.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

John W Ely, Jerome A Osheroff, Mark H Ebell, George R Bergus, Barcey T Levy, M Lee Chambliss, and Eric R Evans. 1999. Analysis of questions

asked by family doctors regarding patient care. *Bmj*, 319(7206):358–361.

B.L. Fabre, M.A.F. Magalhaes Filho, P.N. Aguiar, F.M. da Costa, B. Gutierres, W.N. William, and A. Del Giglio. 2024. Evaluating gpt-4 as an academic support tool for clinicians: a comparative analysis of case records from the literature. *ESMO Real World Data and Digital Oncology*, 4:100042.

Arsene Fansi Tchango, Rishab Goel, Julien Martel, Zhi Wen, Gaetan Marceau Caron, and Joumana Ghosn. 2022a. Towards trustworthy automatic diagnosis systems by emulating doctors' reasoning with deep reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24502–24515.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022b. Ddxplus: A new dataset for automatic medical diagnosis. In *Advances in Neural Information Processing Systems*, volume 35, pages 31306–31318.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177, Torino, Italia. ELRA and ICCL.

Luca Gioacchini, Giuseppe Siracusano, Davide Sanvito, Kiril Gashteovski, David Friede, Roberto Bifulco, and Carolin Lawrence. 2024. AgentQuest: A modular benchmark framework to measure progress and improve LLM agents. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 185–193, Mexico City, Mexico. Association for Computational Linguistics.

Mark C Henderson, Lawrence M Tierney Jr, and Gerald W Smetana. 2012. The patient history: An evidence-based approach to differential diagnosis. *McGraw Hill Professional*.

Vaishnavi Himakunthala, Andy Ouyang, Daniel Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, and William Wang. 2023. Let's think frame by frame with VIP: A video infilling and prediction dataset for evaluating video chain-of-thought. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 204–219, Singapore. Association for Computational Linguistics.

Takanobu Hirosawa, Ren Kawamura, Yukinori Harada, Kazuya Mizuta, Kazuki Tokumasu, Yuki Kaji, Tomoharu Suzuki, and Taro Shimizu. 2023. ChatGPT-Generated Differential Diagnosis Lists for Complex Case–Derived Clinical Vignettes: Diagnostic Accuracy Evaluation. *JMIR Medical Informatics*, 11.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

In-Chang Hwang, Dongjun Choi, You-Jung Choi, Lia Ju, Myeongju Kim, Ji-Eun Hong, Hyun-Jung Lee, Yeonyee E Yoon, Jun-Bean Park, Seung-Pyo Lee, et al. 2022. Differential Diagnosis of Common Etiologies of Left Ventricular Hypertrophy Using a Hybrid CNN-LSTM Model. *Scientific Reports*, 12(1):20998.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, Andrew Y Ng, and Jonathan H Chen. 2025. Medagentbench: Dataset for benchmarking llms as agents in medical applications. *arXiv preprint arXiv:2501.14654*.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, et al. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, pages 1–10.

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. CRAFT-MD: A conversational evaluation framework for comprehensive assessment of clinical LLMs. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Sallianne Kavanagh, Katie Van Der Tuijn, and Amie Bain. 2024. Principles of diagnostic reasoning. *Pharmaceutical Journal*, 312(7982).

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai "Orson" Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452. Curran Associates, Inc.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024a. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024b. Mediq: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. 2020. A Deep Learning System for Differential Diagnosis of Skin Diseases. *Nature medicine*, 26(6):900–908.

Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. 2024. Openmedlm: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14(1):14156.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. 2023. Towards Accurate Differential Diagnosis with Large Language Models. *arXiv preprint arXiv:2312.00164*.

Kazuya Mizuta, Takanobu Hirosawa, Yukinori Harada, and Taro Shimizu. 2024. Can chatgpt-4 evaluate whether a differential diagnosis list contains the correct diagnosis as accurately as a physician? *Diagnosis*, (0).

Saeel Sandeep Nachane, Ojas Gramopadhye, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. 2024. Few shot chain-of-thought driven reasoning to prompt LLMs for open-ended medical question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 542–573, Miami, Florida, USA. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

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

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Justin T Reese, Daniel Danis, J Harry Caufield, Tudor Groza, Elena Casiraghi, Giorgio Valentini, Christopher J Mungall, and Peter N Robinson. 2024. On the limitations of large language models in clinical diagnosis. *medRxiv*, pages 2023–07.

Jacqueline Rhoads, Julie C Penick, et al. 2017. *Formulating a Differential Diagnosis for the Advanced Practice Provider*. Springer Publishing Company.

Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. 2023. Visual chain of thought: bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*.

Geoffrey W Rutledge. 2024. Diagnostic Accuracy of GPT-4 on Common Clinical Scenarios and Challenging Cases. *Learning Health Systems*, 8(3):e10438.

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

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

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Kiyoshi Shikino, Taro Shimizu, Yuki Otsuka, Masaki Tago, Hiromizu Takahashi, Takashi Watari, Yosuke Sasaki, Gemmei Iizuka, Hiroki Tamura, Koichi Nakashima, et al. 2024. Evaluation of ChatGPT-Generated Differential Diagnosis for Common Diseases With Atypical Presentation: Descriptive Research. *JMIR Medical Education*, 10:e58758.

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. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Hidde Ten Berg, Bram van Bakel, Lieke van de Wouw, Kim E Jie, Anoeska Schipper, Henry Jansen, Rory D O'Connor, Bram van Ginneken, and Steef Kurstjens. 2024. ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation. *Annals of Emergency Medicine*, 83(1):83–86.

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*.

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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Benjamin Winter, Alexei Gustavo Figueroa Rosero, Alexander Loeser, Felix Alexander Gers, Nancy Katerina Figueroa Rosero, and Ralf Krestel. 2024. DDx-Gym: Online transformer policies in a knowledge graph based natural language environment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4438–4448, Torino, Italia. ELRA and ICCL.

Cheng-Kuang Wu, Wei-Lin Chen, and Hsin-Hsi Chen. 2023. Large language models perform diagnostic reasoning.

Cheng-Kuang Wu, Zhi Rui Tam, Chieh-Yen Lin, Yun-Nung Chen, and Hung yi Lee. 2024. Streambench: Towards benchmarking continuous improvement of language agents. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Charlotte Zelin, Wendy K Chung, Mederic Jeanne, Gongbo Zhang, and Chunhua Weng. 2024. Rare disease diagnosis using knowledge guided retrieval augmentation for chatgpt. *Journal of Biomedical Informatics*, 157:104702.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Hu Jinfang, Zhiyuan Liu, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shuang Zhou, Mingquan Lin, Sirui Ding, Jiashuo Wang, Genevieve B Melton, James Zou, and Rui Zhang. 2024. Interpretable differential diagnosis with dual-inference large language models. *arXiv preprint arXiv:2407.07330*.

Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. 2024. Emerge: Integrating rag for improved multimodal ehr predictive modeling. *arXiv preprint arXiv:2406.00036*.

Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. 2024. Kg4diagnosis: A hierarchical multi-agent llm framework with knowledge graph enhancement for medical diagnosis. *arXiv preprint arXiv:2412.16833*.

A Details of DDx Benchmark

Dataset	Domain	# Cases	# Diseases	Synthetic	License†
DDxPlus (Fansi Tchango et al., 2022b)	*respiratory*	1.3M	49	✓	CC-BY
iCraft-MD (Li et al., 2024b)	*skin*	140	394	✓	MIT
RareBench (Chen et al., 2024b)	*rare*	2,185	421	×	Apache-2.0

Table 3: Overview of the selected sources for constructing DDx benchmark. We consider three domains (i.e., disease categories) (*respiratory*, *skin*, *rare*) with different sizes of diagnosis options. All selected sources are applicable for *commercial* usage. †License: Creative Commons Attribution International License (CC-BY).

Datasets. To address the limitation of existing work, which often evaluates on a *single* dataset and diagnoses in a *single* turn, we construct a comprehensive DDx benchmark sourced from three datasets: DDxPlus, iCraft-MD, and RareBench, covering *respiratory*, *skin*, and *rare* diseases, respectively. The statistics of each dataset are presented in Table 3. DDxPlus (Fansi Tchango et al., 2022b) is a large-scale synthetic dataset, spanning 1.3 million patient cases across 49 respiratory-related pathologies, focusing on conditions where the chief complaint is related to cough, sore through, or breathing issues. As one of the largest structured DDx datasets, it provides both ground-truth diagnoses *and* ground truth ranked differential diagnosis lists, enabling effective few shot examples as well as a direct evaluation of predicting and refining DDx rankings. iCraft-MD (or interactive Craft-MD) (Li et al., 2024b) adapts static dermatological clinical vignettes from the original Craft-MD dataset (Johri et al., 2024, 2025) into an interactive setting. It consists of 140 dermatology cases, with 100 sourced from an online medical question bank and 40 designed by expert clinicians. RareBench (Chen et al., 2024b) further expands the diagnostic landscape by extending DDxPlus to include 421 rare diseases. We specifically select three regional subsets from Rarebench – RAMEDIS (Europe), MME (Canada), and PUMCH (China) – to ensure diversity in rare disease regional representation. Each of these datasets includes patient profiles with two core components: (1) symptom/antecedent data and (2) ground-truth pathology/disease.

Benchmark Compilation. To enable a consistent evaluation across datasets, we normalize each dataset into a structured format, each dataset is converted to include: (i) Optional initial information of the patient (e.g., age, sex, chief complaint); (ii) full patient profile (complete list of symptoms and medical history); (iii) full set of possible diseases for differential diagnosis. For DDxPlus, we inherit the format from StreamBench (Wu et al., 2024). In iCraft-MD, the initial information of the patient is provided as initial case details, whereas in Rarebench no initial patient information is available. We process iCraft-MD and RareBench to extract the full patient profile. One of the major challenges in iCraft-MD and Rarebench is the lack of predefined differential diagnosis options and the presence of redundant disease names. To address this, we iterate through all patient records and employ GPT-4o to generate a unique, non-redundant disease set for each patient case. As a result, we obtain 394 unique dermatological conditions for iCraft-MD and 102 rare diseases for the RareBench subset. This refinement step ensures that each patient's diagnostic process operates within a well-defined differential diagnosis structure, reducing ambiguity and improving evaluation reliability.

DDxPlus

Initial Patient Profile:
Age: 39
Sex: M
Chief Complaint: nasal congestion
Complete Patient Profile:
Sex: Male, Age: 39
- I am currently being treated or have recently been treated with an oral antibiotic for an ear infection.
- I have pain somewhere related to my reason for consulting.
- I have a fever (either felt or measured with a thermometer).
- I have nasal congestion or a clear runny nose.
- My vaccinations are up to date.
- On a scale of 0-10, the pain intensity is 6
- On a scale of 0-10, the pain's location precision is 8
- On a scale of 0-10, the pace at which the pain appear is 2
- The pain is:
* sensitive
* sharp
- The pain locations are:
* ear(R)
- The pain radiates to these locations:
* nowhere
Ground Truth Pathology: Acute otitis media
Ground Truth DDx:
1. Acute otitis media
2. URTI
3. Chagas

iCraft-MD

Initial Patient Profile:
Age: 61 years
Sex: male
Chief Complaint: A 61-year-old man presents with a 7-month history of lesions on his hands and arms
Complete Patient Profile:
- A 61-year-old man presents with a 7-month history of lesions on his hands and arms
- His medical history includes depression, hypertension, and hyperlipidemia
- He has no personal or family history of skin problems
- His skin lesions are not painful or itchy, and he is not bothered by their appearance
- He has not tried any treatments for the lesions
- Physical examination reveals a number of pink, annular plaques with smooth raised borders on the patient's dorsal forearms and hands
- On close inspection, small discrete papules are seen within the plaques.
Ground Truth Pathology: Localized granuloma annulare

RareBench

Initial Patient Profile:
N/A
Complete Patient Profile: - Hematuria
- Slurred speech
- Abnormality of the liver
- Dysphagia
- Drooling
- Abnormal caudate nucleus morphology
- Hand tremor
- Poor appetite
- Decreased circulating ceruloplasmin concentration
- Increased urinary copper concentration
- Kayser-Fleischer ring
Ground Truth Pathology: Wilson disease

B Prompt Design

We present the prompt design for history taking simulator, knowledge retrieval agent, diagnosis strategy agent, and DDxDriver in this section.

History Taking Simulator: Doctor

System Prompt:
`<SPECIALIST_PREFACE>`
Your job is to take medical history from a patient by asking them specific questions to determine their antecedents and symptoms, as well as narrow down the possible diseases they may be suffering from. `[...]`

You may receive this additional information to guide your dialogue:
- Initial Patient Information: Information the patient has already self-reported, such as chief complaint, age, sex, etc.
- Dialogue History: The conversation you and the patient have had so far, formatted as `'Doctor'` / `'Patient'` turns.
- Suggested Conversation Goals: Specific topics or questions to try to cover in the dialogue. You may also ask questions outside of these conversation goals; do not limit yourself to these.

You may either start, end, or continue the conversation, as explained below:
`[...]`

Response Instructions:
`[...]`

History Taking Simulator: Patient

System Prompt Act as a patient with the patient profile below engaging in a medical history taking with a doctor. `[...]`

You may receive this additional information to guide your dialogue:
- Initial Patient Information: Information you as the patient have already self-reported to the doctor, such as chief complaint, age, sex, etc.
- Dialogue History: The conversation you and the patient have had so far, formatted as `'Doctor'` / `'Patient'` turns.

When asked for information which is explicitly present in your patient profile (including as synonyms), either respond:
a. Positively ("Yes"...) if your patient profile explicitly indicates you have this antecedent/symptom
b. Negatively ("No"...) if your patient profile explicitly indicates that you do not have this antecedent/symptom

When asked for information which is not explicitly mentioned in your patient profile (including as synonyms), respond "I don't know."
`[...]`

Response instructions:
`[...]`

Knowledge Retrieval Agent

Keywords Prompt:
Your job is to assist in the creation of a differential diagnosis for a patient by searching for relevant information online. Given an input search from a user, break it up into a list of simplified keyword searches to find relevant medical information online.
Follow these steps:
[...]
Format example:
Input search:
<INPUT_SEARCH>
Keyword searches list:
[<KEYWORD_SEARCH_1>, <KEYWORD_SEARCH_2>]

Synthesis Prompt
You are a helpful research assistant to a doctor creating a differential diagnosis of a patient. Concisely answer the doctor's input search by analyzing and summarizing the relevant medical content in the search results. [...]

Inputs:
1. Doctor's Input Search: the search the doctor requested
- This search may contain multiple topics
2. Search results: the search results fetched
- You may only answer based on topics present in these search results
3. Diagnosis Options (optional): the possible diseases the patient may be suffering from
- If provided, use this exact terminology to refer to the diseases

Response Instructions:
[...]

Diagnosis Strategy Agent

System Prompt:
<SPECIALIST_PREFACE>
Given a patient's profile (a list of antecedents and symptoms), provide a ranked differential diagnosis of the <DDX_LENGTH> most likely diseases. You may be provided a list of diagnosis options you can choose from. You must use this exact disease terminology when referring to the diseases. If you aren't provided the diagnosis options, consider all possible diseases.

Your ranked differential diagnosis should have the possible diseases ranked from most likely to least likely.

You will also be provided with:
1. Previous Ranked Differential Diagnoses: [...]
2. Suggested Diagnosis Instructions (optional): [...]
3. Previous Search Content (optional): [...]
4. Patient profile: the known symptoms/antecedents of the patient
5: Patient examples (optional): [...]

[...]

Directly provide the ranked differential diagnosis of the <DDX_LENGTH> most likely diseases for the patient in the following format (without additional text before or after), with one diagnosis per line (replace [DIAGNOSIS_X] with the actual diagnosis name, and do not include the brackets themselves): [RANK_NUMBER]. [DIAGNOSIS]. I.e.:
1. [DIAGNOSIS_1]
2. [DIAGNOSIS_2]
...
Directly provide your response in the format specified, without additional text.

DDxDriver

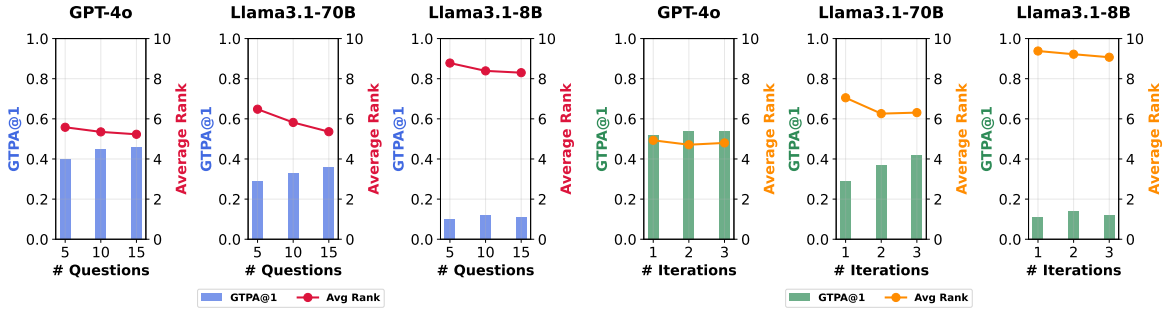
Fixed Iteration System Prompt:
Your job is to facilitate the process of differential diagnosis of a patient by concisely prompting medical agents.

You will be provided with:
1) Agent Descriptions. This includes:
a) Agent Function: A description of the function of medical agent.
b) Agent Prompt: A description of how to prompt the agent
2) Available Information: The available information you can extract from to prompt the agent. Do not invent new information. This may include:
a) Patient Initial Information
b) Patient Profile
c) Dialogue History
d) Previous RAG content
- External information found about diseases the patient may be suffering from
e) Previous Ranked Differential Diagnoses
f) Diagnosis Options
- These are the only diseases the patient may be suffering from.
- You must use the exact terminology in this list when referring to the diseases

Follow these steps to create a prompt for the medical agent:
1. Analyze the description of the medical agent and its input prompt. Note whether its input prompt is optional or mandatory.
2. Review the current information you were provided. Determine how this information can help the agent.
- You should only prompt based on this current information.
3. Follow the agent's input prompt description and design a prompt for this agent.
4. Respond with your agent prompt, nothing else.

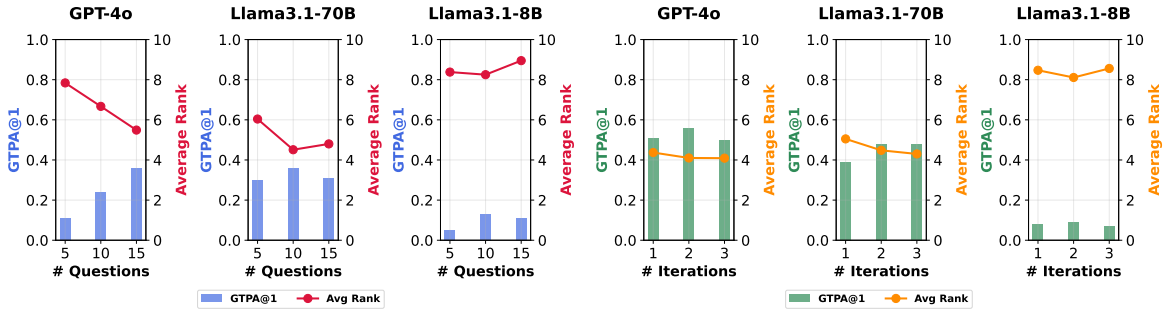
C Additional Analysis

Results of iCraft-MD (Figure 5) and RareBench (Figure 6) compared between (a) History Taking Simulator, and (b) MEDDxAgent over the selection of max questions (5, 10, 15), and number of iterations (1, 2, 3).



(a) (b)

Figure 5: Results of iCraft-MD (Li et al., 2024b) compared between (a) History Taking Simulator, and (b) MEDDxAgent.



(a) (b)

Figure 6: Results of RareBench (Chen et al., 2024b) compared between (a) History Taking Simulator, and (b) MEDDxAgent.

D Additional Experiments

D.1 History Taking Simulator

Model	Metric	DDxPlus			iCraft-MD			RareBench		
		5	10	15	5	10	15	5	10	15
GPT-4o	**GTPA@1**	0.45	0.59	0.69	0.40	0.45	0.46	0.11	0.24	0.36
	GTPA@3	0.60	0.73	0.82	0.51	0.53	0.53	0.22	0.36	0.48
	GTPA@5	0.72	0.83	0.88	0.57	0.57	0.60	0.35	0.47	0.59
	Avg Rank	4.13	3.16	2.47	5.58	5.35	5.23	7.84	6.67	5.49
Llama3.1-70B	**GTPA@1**	0.45	0.58	0.56	0.29	0.33	0.36	0.30	0.36	0.31
	GTPA@3	0.65	0.77	0.76	0.39	0.49	0.53	0.42	0.57	0.59
	GTPA@5	0.71	0.83	0.79	0.47	0.55	0.60	0.53	0.65	0.67
	Avg Rank	4.15	3.12	3.50	6.48	5.82	5.36	6.04	4.51	4.80
UltraMedical-70B	**GTPA@1**	0.45	0.52	0.50	0.23	0.27	0.22	0.40	0.43	0.44
	GTPA@3	0.62	0.68	0.69	0.27	0.31	0.29	0.58	0.61	0.59
	GTPA@5	0.70	0.73	0.73	0.29	0.34	0.31	0.61	0.67	0.67
	Avg Rank	4.70	4.30	4.68	8.12	7.61	7.99	5.01	4.59	4.54
Llama3.1-8B	**GTPA@1**	0.23	0.35	0.40	0.10	0.12	0.11	0.05	0.13	0.11
	GTPA@3	0.37	0.49	0.53	0.20	0.23	0.23	0.21	0.21	0.18
	GTPA@5	0.43	0.60	0.60	0.25	0.27	0.29	0.27	0.25	0.20
	Avg Rank	6.85	5.46	5.44	8.78	8.39	8.3	8.38	8.25	8.95
UltraMedical3.1-8B	**GTPA@1**	0.26	0.24	0.20	0.16	0.14	0.14	0.15	0.14	0.12
	GTPA@3	0.36	0.35	0.32	0.22	0.20	0.21	0.23	0.25	0.21
	GTPA@5	0.42	0.39	0.38	0.23	0.23	0.22	0.25	0.28	0.23
	Avg Rank	7.00	7.86	7.73	8.60	8.78	8.72	8.44	8.29	8.46

Table 4: History taking simulator performance across different datasets (DDxPlus, iCraftMD, Rarebench) with max questions (5, 10, 15), aggregated over 100 patients.

D.2 Knowledge Retrieval Agent

			DDxPlus		iCraft-MD		RareBench	
Model	**Source**	**Metric**	**RAG**	**Base**	**RAG**	**Base**	**RAG**	**Base**
GPT-4o	**PubMed**	**GTPA@1**	0.69	0.69	0.68	0.68	0.45	0.39
		GTPA@3	0.88	0.88	0.77	0.76	0.60	0.58
		GTPA@5	0.90	0.90	0.79	0.77	0.72	0.72
		Avg Rank	2.27	2.21	3.23	3.37	3.92	3.99
	Wiki	**GTPA@1**	0.69	0.69	0.69	0.68	0.45	0.39
		GTPA@3	0.88	0.88	0.77	0.76	0.58	0.58
		GTPA@5	0.90	0.90	0.79	0.77	0.74	0.72
		Avg Rank	2.24	2.21	3.22	3.37	4.00	3.99
Llama3.1-70B	**PubMed**	**GTPA@1**	0.56	0.54	0.44	0.40	0.38	0.39
		GTPA@3	0.77	0.74	0.56	0.56	0.62	0.59
		GTPA@5	0.79	0.78	0.63	0.64	0.75	0.77
		Avg Rank	3.42	3.53	4.72	4.87	3.96	4.05
	Wiki	**GTPA@1**	0.49	0.54	0.44	0.40	0.39	0.39
		GTPA@3	0.74	0.74	0.59	0.56	0.59	0.59
		GTPA@5	0.77	0.78	0.66	0.64	0.75	0.77
		Avg Rank	3.60	3.53	4.71	4.87	4.09	4.05
UltraMedical-70B	**PubMed**	**GTPA@1**	0.58	0.60	0.31	0.31	0.45	0.44
		GTPA@3	0.68	0.73	0.37	0.38	0.65	0.63
		GTPA@5	0.70	0.76	0.38	0.39	0.71	0.70
		Avg Rank	4.63	6.55	7.08	7.01	4.20	4.47
	Wiki	**GTPA@1**	0.58	0.6	0.31	0.31	0.44	0.44
		GTPA@3	0.68	0.73	0.38	0.38	0.64	0.63
		GTPA@5	0.70	0.76	0.39	0.39	0.70	0.70
		Avg Rank	4.38	6.55	7.01	7.01	4.25	4.47
Llama3.1-8B	**PubMed**	**GTPA@1**	0.42	0.48	0.29	0.27	0.35	0.33
		GTPA@3	0.58	0.63	0.38	0.37	0.55	0.54
		GTPA@5	0.67	0.69	0.44	0.43	0.59	0.57
		Avg Rank	4.50	5.25	6.93	7.02	5.33	5.45
	Wiki	**GTPA@1**	0.43	0.48	0.29	0.27	0.36	0.33
		GTPA@3	0.58	0.63	0.38	0.37	0.52	0.54
		GTPA@5	0.67	0.69	0.44	0.43	0.67	0.57
		Avg Rank	4.56	5.25	6.93	7.02	4.80	5.45
UltraMedical3.1-8B	**PubMed**	**GTPA@1**	0.27	0.33	0.19	0.27	0.21	0.22
		GTPA@3	0.39	0.48	0.24	0.37	0.46	0.35
		GTPA@5	0.46	0.51	0.26	0.43	0.48	0.42
		Avg Rank	6.81	6.88	8.38	7.02	6.23	7.02
	Wiki	**GTPA@1**	0.25	0.33	0.18	0.18	0.27	0.22
		GTPA@3	0.38	0.48	0.23	0.23	0.44	0.35
		GTPA@5	0.45	0.51	0.25	0.25	0.51	0.42
		Avg Rank	6.90	6.88	8.42	8.45	6.37	7.02

Table 5: Knowledge retrieval agent performance with different datasets with varying sources (PubMed, Wikipedia) and methods, aggregated over 100 patients.

D.3 Diagnosis Strategy Agent

		DDxPlus				iCraft-MD				RareBench			
Model	**Metric**	**None**	**Static**	**Dyn_BAII**	**Dyn_BERT**	**None**	**Static**	**Dyn_BAII**	**Dyn_BERT**	**None**	**Static**	**Dyn_BAII**	**Dyn_BERT**
Standard													
GPT-4o	**GTPA@1**	0.69	0.74	0.96	0.96	0.68	0.64	0.62	0.67	0.46	0.52	0.79	0.78
	GTPA@5	0.90	0.90	1.00	1.00	0.77	0.74	0.72	0.77	0.72	0.80	0.91	0.90
	Avg Rank	2.21	2.20	1.06	1.06	3.37	3.64	3.85	3.31	3.99	3.58	2.03	2.19
Llama3.1-70B	**GTPA@1**	0.54	0.57	0.86	0.84	0.40	0.38	0.40	0.40	0.39	0.42	0.73	0.72
	GTPA@5	0.78	0.80	0.95	0.94	0.64	0.63	0.63	0.62	0.77	0.71	0.87	0.87
	Avg Rank	3.53	3.41	1.59	1.68	4.87	5.15	5.02	4.96	4.05	4.29	2.44	2.44
UltraMedical-70B	**GTPA@1**	0.58	0.60	0.97	0.96	0.31	0.37	0.42	0.40	0.44	0.47	0.74	0.71
	GTPA@5	0.70	0.76	1.00	1.00	0.39	0.45	0.47	0.48	0.70	0.62	0.83	0.80
	Avg Rank	4.18	6.55	1.03	1.04	7.01	6.29	6.14	6.15	4.47	4.92	2.74	2.96
Llama3.1-8B	**GTPA@1**	0.45	0.48	0.97	0.97	0.27	0.25	0.21	0.22	0.33	0.39	0.71	0.70
	GTPA@5	0.68	0.69	1.00	1.00	0.43	0.44	0.42	0.40	0.57	0.63	0.83	0.81
	Avg Rank	9.00	5.25	1.03	1.04	7.02	6.78	6.93	7.32	5.45	4.76	2.80	2.94
UltraMedical3.1-8B	**GTPA@1**	0.26	0.33	0.85	0.81	0.18	0.16	0.18	0.15	0.22	0.24	0.60	0.57
	GTPA@5	0.45	0.51	0.89	0.93	0.25	0.26	0.26	0.24	0.42	0.36	0.73	0.63
	Avg Rank	6.86	6.88	3.04	2.09	8.45	8.52	8.24	8.70	7.02	7.21	3.66	4.66
Chain-of-Thought (CoT)													
GPT-4o	**GTPA@1**	0.71	0.72	0.97	0.96	0.68	0.64	0.60	0.64	0.47	0.57	0.82	0.81
	GTPA@5	0.92	0.92	1.00	1.00	0.77	0.72	0.70	0.73	0.69	0.77	0.88	0.91
	Avg Rank	2.10	1.98	1.03	1.05	3.35	3.79	4.00	3.68	4.02	3.48	2.11	2.04
Llama3.1-70B	**GTPA@1**	0.45	0.58	0.89	0.91	0.48	0.44	0.45	0.45	0.49	0.50	0.71	0.75
	GTPA@5	0.78	0.82	0.93	0.95	0.66	0.62	0.63	0.61	0.75	0.72	0.87	0.88
	Avg Rank	3.69	3.08	1.71	1.55	4.50	4.88	4.90	4.93	3.91	4.04	2.62	2.35
UltraMedical-70B	**GTPA@1**	0.47	0.47	0.96	0.93	0.26	0.33	0.34	0.34	0.39	0.35	0.69	0.32
	GTPA@5	0.57	0.63	1.00	0.99	0.26	0.42	0.38	0.41	0.62	0.43	0.78	0.47
	Avg Rank	5.46	6.70	1.04	1.17	8.35	6.78	7.11	6.80	5.05	6.75	3.36	6.53
Llama3.1-8B	**GTPA@1**	0.45	0.51	0.97	0.95	0.27	0.34	0.3	0.29	0.24	0.36	0.65	0.64
	GTPA@5	0.70	0.71	1.00	0.99	0.40	0.44	0.44	0.36	0.55	0.61	0.82	0.84
	Avg Rank	4.51	5.08	1.03	1.19	7.25	6.45	6.66	7.28	5.65	4.98	2.95	2.96
UltraMedical3.1-8B	**GTPA@1**	0.22	0.22	0.74	0.81	0.13	0.13	0.14	0.18	0.09	0.19	0.39	0.50
	GTPA@5	0.32	0.28	0.79	0.88	0.17	0.16	0.18	0.20	0.17	0.31	0.49	0.58
	Avg Rank	15.30	16.25	7.24	3.97	9.33	9.41	9.26	9.02	9.36	7.99	6.23	5.26

Table 6: Diagnosis strategy module performance across 3 datasets with different methods (Standard vs. CoT), aggregated over 100 patients.

D.4 Optimizing Knowledge Retrieval Agent vs. Diagnosis Strategy Agent

	DDxPlus			iCraft-MD			RareBench		
	GTPA@1 ↑	GTPA@5 ↑	Avg Rank ↓	GTPA@1 ↑	GTPA@5 ↑	Avg Rank ↓	GTPA@1 ↑	GTPA@5 ↑	Avg Rank ↓
GPT-4o									
Retrieval (PubMed)	0.69	0.90	2.27	0.68	**0.79**	3.23	0.45	0.72	3.92
Retrieval (Wiki)	0.69	0.90	2.24	**0.69**	**0.79**	**3.22**	0.45	0.74	4.00
Zero-shot (Standard)	0.69	0.90	2.21	0.68	0.77	3.37	0.46	0.72	3.99
Zero-shot (CoT)	0.71	0.92	2.10	0.68	0.77	3.35	0.47	0.69	4.02
Few-shot (Standard, Dyn_BAII)‡	0.96	**1.00**	1.06	0.62	0.72	3.85	0.79	**0.91**	**2.03**
Few-shot (CoT, Dyn_BERT)	0.96	**1.00**	1.05	0.64	0.73	3.68	0.81	**0.91**	2.04
Few-shot (CoT, Dyn_BAII)	**0.97**	**1.00**	**1.03**	0.60	0.70	4.00	**0.82**	0.88	2.11
Llama3.1-70B									
Retrieval (PubMed)	0.56	0.79	3.42	0.44	0.63	4.72	0.38	0.75	3.96
Retrieval (Wiki)	0.49	0.77	3.60	0.44	**0.66**	4.71	0.39	0.75	4.09
Zero-shot (Standard)	0.54	0.78	3.53	0.40	0.64	4.87	0.39	0.77	4.05
Zero-shot (CoT)	0.45	0.78	3.69	**0.48**	**0.66**	**4.50**	0.49	0.75	3.91
Few-shot (Standard, Dyn_BAII)‡	0.86	**0.95**	1.59	0.40	0.63	5.02	0.73	0.87	2.44
Few-shot (CoT, Dyn_BERT)	**0.91**	**0.95**	**1.55**	0.45	0.61	4.93	**0.75**	**0.88**	**2.35**
Few-shot (CoT, Dyn_BAII)	0.89	0.93	1.71	0.45	0.63	4.90	0.71	0.87	2.62
Llama3.1-8B									
Retrieval (PubMed)	0.42	0.67	4.50	0.29	**0.44**	6.93	0.35	0.59	5.33
Retrieval (Wiki)	0.43	0.67	4.56	0.29	**0.44**	6.93	0.36	0.67	4.80
Zero-shot (Standard)	0.45	0.68	9.00	0.27	0.43	7.02	0.33	0.57	5.45
Zero-shot (CoT)	0.45	0.70	4.51	0.27	0.40	7.25	0.24	0.55	5.65
Few-shot (Standard, Dyn_BAII)‡	**0.97**	**1.00**	**1.03**	0.21	0.42	6.93	**0.71**	0.83	**2.80**
Few-shot (CoT, Dyn_BERT)	0.95	0.99	1.19	0.29	0.36	7.28	0.64	**0.84**	2.96
Few-shot (CoT, Dyn_BAII)	**0.97**	**1.00**	**1.03**	**0.30**	**0.44**	**6.66**	0.65	0.82	2.95

Table 7: Full comparison of knowledge retrieval agent with diagnosis strategy agent, assuming that there are existing *full* patient profiles. ‡ Only Few-shot (Standard, Dyn_BAII) results are recorded, since the method is consistently better than Dyn_BERT.

D.5 MEDDxAgent

		DDxPlus						iCraft-MD						RareBench					
		Fixed			Dynamic			Fixed			Dynamic			Fixed			Dynamic		
Model	**Metric**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
GPT-4o	**GTPA@1**	0.74	0.78	0.86	0.74	0.76	0.81	0.52	0.54	0.54	0.44	0.52	0.52	0.51	0.56	0.50	0.35	0.43	0.46
	Avg Rank	1.91	1.56	1.29	2.00	1.62	1.48	4.93	4.71	4.80	4.85	4.57	4.56	4.37	4.10	4.09	6.14	4.62	4.24
	Avg Progress	0.00	0.32	0.32	-0.13	0.04	0.01	0.00	0.26	0.17	0.14	-0.06	0.01	0.00	0.13	0.16	-0.23	-0.15	-0.39
Llama3.1-70B	**GTPA@1**	0.61	0.71	0.68	0.53	0.61	0.60	0.29	0.37	0.42	0.24	0.30	0.31	0.39	0.48	0.48	0.32	0.37	0.46
	Avg Rank	2.91	2.20	2.30	2.96	2.89	2.73	7.05	6.26	6.31	7.08	6.77	6.82	5.05	4.48	4.30	5.44	4.66	4.19
	Avg Progress	0.00	0.41	0.17	0.00	0.04	0.02	0.00	0.07	0.26	0.10	0.02	0.00	0.00	0.75	0.44	-0.06	0.00	0.07
Llama3.1-8B	**GTPA@1**	0.34	0.56	0.58	0.47	0.58	0.54	0.11	0.14	0.12	0.03	0.04	0.04	0.08	0.09	0.07	0.06	0.10	0.18
	Avg Rank	5.25	3.59	3.10	5.00	3.82	3.92	9.38	9.22	9.07	10.11	9.95	9.91	8.47	8.11	8.56	8.01	7.54	7.21
	Avg Progress	0.00	1.73	1.23	0.00	0.00	0.00	0.00	0.22	0.17	0.00	0.00	0.00	0.00	0.44	0.38	0.00	0.00	0.00

Table 8: Iterative experiment performance compared between fixed iteration and dynamic iteration with 3 datasets.