RAREAGENTS: Advancing Rare Disease Care through LLM-Empowered Multi-disciplinary Team

Anonymous ACL submission

Abstract

Rare diseases, despite their low individual incidence, collectively impact around 300 million people worldwide due to the vast number of diseases. The involvement of multiple organs and systems, and the shortage of specialized doctors with relevant experience make diagnosing and treating rare diseases more challenging than common diseases. Recently, agents powered by large language models (LLMs) have demonstrated notable applications across various domains. In the medical field, some agent methods have outperformed direct prompts in question-answering tasks from medical examinations. However, current agent frameworks are not well-adapted to real-world clinical scenarios, especially those involving the complex demands of rare diseases. To bridge this gap, we introduce **RareAgents**, the first LLM-driven multi-disciplinary team framework designed specifically for the complex clinical context of rare diseases. *RareAgents* integrates advanced Multidisciplinary Team (MDT) coordination, memory mechanisms, and medical tools utilization, leveraging Llama-3.1-8B/70B as the base model. Experimental results show that *RareAgents* outperforms state-of-the-art domain-specific models, GPT-4o, and current agent frameworks in differential diagnosis and medication recommendation for rare diseases. Furthermore, we contribute a novel rare disease dataset, MIMIC-IV-EXT-RARE, to support further advancements in this field. Our code can be found at `https://anonymous.4open.science/r/AutoMDT-65EC`.

1 Introduction

Rare diseases are defined as disorders with low prevalence, typically affecting fewer than 1 in 2,000 individuals in Europe or fewer than 1 in 1,500 individuals in the United States (Valdez et al., 2016). Despite their rarity, more than 7,000 rare diseases have been identified, impacting approximately 300 million people worldwide (Nguengang Wakap et al., 2020). Rare diseases often present with complex and heterogeneous symptoms that overlap with common diseases. As a result, patients frequently experience several years of misdiagnosis, referred to as a "diagnostic odyssey" (Schieppati et al., 2008). Such delays not only limit access to timely and effective treatments but also cause the worsening of the disease. On the other hand, while deep learning models have shown promise in medication recommendation (Zhang et al., 2017; Shang et al., 2019b; Yang et al., 2021b), their performance for rare diseases remains suboptimal. Experimental studies on the MIMIC-III/IV datasets (Johnson et al., 2016, 2023) reveal that current state-of-the-art models for drug recommendation are substantially less effective for rare diseases than common ones (Zhao et al., 2024).

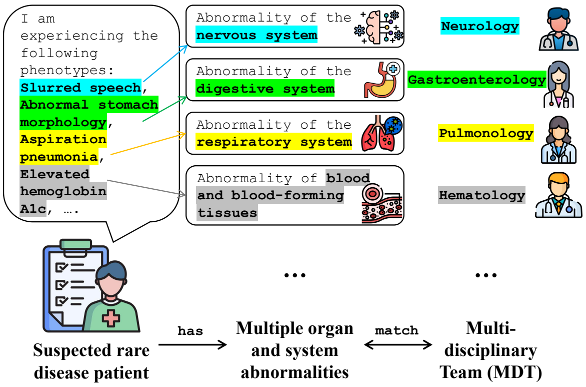

I am experiencing the following phenotypes: Slurred speech, Abnormal stomach morphology, Aspiration pneumonia, Elevated hemoglobin A1c, ….
Abnormality of the nervous system
Abnormality of the digestive system
Abnormality of the respiratory system
Abnormality of blood and blood-forming tissues
Neurology
Gastroenterology
Pulmonology
Hematology
…
…
Suspected rare disease patient
has
Multiple organ and system abnormalities
match
Multi-disciplinary Team (MDT)

Figure 1: A rare disease patient with multi-organ and system abnormalities necessitates multidisciplinary team for comprehensive diagnosis and treatment.

Large language models (LLMs), trained on the massive and diverse text corpora, have demonstrated remarkable potential across a wide range of natural language interaction tasks (Achiam et al., 2023; Dubey et al., 2024). In particular, LLM-based agents exhibit impressive capabilities in augmented reasoning and problem-solving within complex environments (Wang et al., 2024). In the domain of rare diseases, RareBench (Chen et al., 2024b) introduced the first benchmark to evaluate LLMs in phenotype extraction and differential diagnosis. Experimental results indicate that advanced LLMs, such as GPT-4 (Achiam et al., 2023), can achieve notable diagnostic accuracy under zero-shot settings, even outperforming human specialist physicians for certain rare diseases.

As shown in Figure 1, patients with rare diseases often experience the involvement of multiple organs and systems, necessitating multi-disciplinary team (MDT) care to integrate expertise from various specialties for more accurate diagnoses and personalized treatment plans (Xie et al., 2023). Although several multi-agent frameworks have been proposed for general medical applications (as summarized in Table 1), these methods primarily demonstrate improved performance in tasks like multiple-choice question answering (MCQA) (Tang et al., 2024; Jin et al., 2024) and basic question answering (QA) (Kim et al., 2024), where candidate options are provided or the decision-making is confined to limited and small scopes (Li et al., 2024b). These settings differ from the complexity and uncertainty of real-world clinical scenarios. Moreover, existing approaches tend to emphasize planning capabilities while placing less focus on the integration of memory and tool utilization. Additionally, the definition of different agent roles is frequently left to LLMs themselves, which can lead to potential hallucinations in medical contexts (Lee et al., 2023).

To address these challenges, we propose **RareAgents**, a patient-centered, personalized autonomous MDT framework tailored for real-world rare disease patients, fully leveraging the planning, memory, and tool-using capabilities of LLM agents. As illustrated in Figure 2, a patient first conveys his personal profile, including symptoms and diagnosis/treatment requests, to an **Attending Physician Agent**. Then, this agent assembles an MDT of specialists from a predefined pool of physician agents, designed with dynamic long-term memory and the ability to utilize specialized medical tools. This enhances the performance of LLMs in diagnosing and treating rare diseases, ultimately offering more accurate and personalized medical care for patients.

Overall, our contributions are three-fold: (1) We propose **RareAgents**, a novel patient-centered multi-disciplinary agent-based framework for enhanced diagnosis and treatment of rare diseases. Each physician agent within *RareAgents* is equipped with dynamic long-term memory, and can effectively utilize a wide range of medical tools, simulating the behavior of a human doctor. Additionally, *RareAgents* is a plug-and-play framework, easily extensible for various medical decision-making scenarios. (2) We evaluate *RareAgents* using Llama-3.1 models (8B and 70B), demonstrating superior diagnostic performance and improved accuracy in medication recommendations compared to state-of-the-art (SOTA) domain-specific models, GPT-4o, and existing medical agent frameworks. We also validate the effectiveness of each module within the *RareAgents* architecture. (3) To the best of our knowledge, this work first extends the medication recommendation task in MIMIC-IV to the LLM agent framework. Furthermore, we compile a rare disease medication recommendation dataset, MIMIC-IV-Ext-Rare, by mapping disease codes and applying rigorous filtering to MIMIC-IV data. This dataset contains 4,760 rare disease patients with 18,522 admission records, providing a valuable resource for the rare disease research community.

Method	Plan.	Mem.	Tools Using	Multi-Agent Roles	Scenario
MedAgents (Tang et al., 2024)	✓	✗	✗	LLM-generated	Medical MCQA
Agent Hospital (Li et al., 2024b)	✓	✓	✗	Pre-defined	Decision Making (MCQA)
MDAgents (Kim et al., 2024)	✓	✗	✗	LLM-generated	Decision Making (QA / VQA)
AgentMD (Jin et al., 2024)	✓	✗	✓	Single-Agent	Risk Prediction (MCQA)
RareAgents (ours)	✓	✓	✓	Pre-defined	Complex Clinical Tasks (Rare Disease Diagnosis & Treatment)

Table 1: Characteristics of different medical LLM agent methods: inclusion of planning, memory, and tool usage, along with role definition ways and target scenarios.

2 Related Work

2.1 LLM-based Agents

Large language models (LLMs) as agents have demonstrated remarkable capabilities in reasoning and decision-making within complex interactive environments (Liu et al., 2023). The concept of generative agents, which first simulated human behavior (Park et al., 2023), has evolved into sophisticated frameworks. LLM-based agents are typically composed of three key components: planning (Yao et al., 2023; Shinn et al., 2023), memory (Zhong et al., 2024), and tool-using (Nakano et al., 2021; Schick et al., 2023). Existing agent frameworks can be broadly categorized into two paradigms: single and multi-agent systems (Li et al., 2023). Among these, role-playing (Shanahan et al., 2023) is a widely adopted approach that assigns agents distinct personalities or roles, allowing them to adapt to specific task scenarios and adjust to diverse behaviors. LLM-based agents have shown significant potential in applications across domains such as education (Zhang et al., 2024b), finance (Yu et al., 2024), and healthcare (Mehandru et al., 2024).

2.2 Medical Agents

Med-Palm (Singhal et al., 2023) and Med-Gemini (Saab et al., 2024) have demonstrated promising single-agent capabilities as medical domain LLMs. Beyond this, MedAgents (Tang et al., 2024) introduces a multi-disciplinary collaboration framework for medical question-answering by leveraging the planning capabilities of multiple agents. MDAgents (Kim et al., 2024) adaptively adjusts to the difficulty of medical questions and extend to visual-question-answering tasks. AI Hospital (Fan et al., 2025) evaluates the performance of large language models (LLMs) as doctors in symptom collection, examination recommendation, and diagnostic decision-making. Agent Hospital (Li et al., 2024b) creates a virtual hospital environment that simulates task stratification within medical workflows. Furthermore, current applications of medical agents encompass a range of scenarios, including clinical triage (Lu et al., 2024), electronic health record reasoning (Shi et al., 2024), and medical imaging analysis (Li et al., 2024a).

2.3 AI Models for Rare Diseases

Most AI diagnostic models for rare diseases primarily rely on phenotypic and genotypic information (Javed et al., 2014; Robinson et al., 2020), utilizing statistical and machine learning approaches (Köhler et al., 2009, 2017; Yang et al., 2015; Peng et al., 2016; Jia et al., 2018; Zhai et al., 2023). RareBERT (Prakash et al., 2021) introduces a Transformer-based (Vaswani et al., 2017) model to identify rare disease patients. In the realm of LLMs, dynamic few-shot prompting methods (Chen et al., 2024b) have been explored to enhance diagnostic performance for rare diseases. RAREMed (Zhao et al., 2024) focuses on addressing fairness in drug recommendation systems and proposes a novel approach to improving therapeutic recommendations for rare disease patients. PhenoBrain (Mao et al., 2025) designs a workflow for phenotype extraction and differential diagnosis, enabling an end-to-end diagnostic process based on patients' electronic health records (EHRs).

3 Problem Formulation and Datasets

3.1 Definition of Rare Disease Tasks

Current medical agent frameworks typically formulate tasks as **multiple-choice questions or limited-answer problems**. However, real-world clinical scenarios are far more complex. To better simulate these conditions, we provide only the patient's profile records $\mathcal{R}$ and ask the agent to make decisions $\mathcal{A}$ based on the specific task demands ($query$). For rare disease diagnosis and treatment, we define the following task scenarios:

Differential Diagnosis The goal of differential diagnosis for rare diseases is to identify a specific rare disease by distinguishing it from other disorders with similar symptoms. This task focuses on **phenotype-based differential diagnosis**. Specifically, the patient's profile $\mathcal{R}$ is represented as a set of symptoms ($\{s_n\}$): $\mathcal{R} = \{s_1, s_2, \ldots, s_n \mid query = diagnosis\}$. **No candidate disease list is provided, nor is it explicitly stated that the patient has a rare disease.** The agent relies solely on the symptom information to reason and output the most likely diagnoses (e.g., the top 10 potential diseases): $\mathcal{A}_{diagnosis} = \{d_1, d_2, \ldots, d_{10}\}$.

Medication Recommendation This task involves patients who may have multiple admission visits for extended medical treatments. During each visit, the patient's profile $\mathcal{R}$ comprises a sequence of diagnosed diseases ($\{d_j\}$) and procedures ($\{p_k\}$), along with a full set of available medications $\mathcal{M}$: $\mathcal{R} = \{\{d_i\}_{i=1}^{j}; \{p_i\}_{i=1}^{k}; \mathcal{M} \mid query = treatment\}$, where $\mathcal{M}$ can include hundreds of drugs (e.g., $|\mathcal{M}| = 122$). The objective is to give the optimal combination of medications to match the patient's treatment needs (**exponential complexity**): $\mathcal{A}_{treatment} = \{m_1, m_2, \ldots, m_l\} \subset \mathcal{M}$.

3.2 Datasets

This research uses two publicly available datasets, RareBench (Chen et al., 2024b) and MIMIC-IV (Johnson et al., 2023), for distinct tasks. RareBench is primarily employed for rare disease differential diagnosis, whereas MIMIC-IV supports various medical tasks, including medication recommendation. From MIMIC-IV, we derive MIMIC-IV-Ext-Rare, a specialized dataset for medication recommendations tailored to rare disease patients. Detailed statistics for both datasets are presented in Table 2.

RareBench-Public (Chen et al., 2024b)		MIMIC-IV-Ext-Rare (Johnson et al., 2023)	
Type of Clinical Task	Differential Diagnosis	Type of Clinical Task	Medication Recommendation
Patient Data Source	Multi-center	Patient Data Source	BIDMC of Boston
# of Rare Disease Patients	1,197	# of Visits / # of Rare Disease Patients	18,522 / 4,760
# of Rare Diseases	498	Disease / Procedure / Medication Space Size	8,922 / 3,920 / 122
Symptom / Disease Space Size	17,232 / 9,260	Avg. / Max # of Visits	3.89 / 74
Avg. / Max # of Symptoms per Case	12.66 / 96	Avg. / Max # of Diseases per Visit	16.99 / 39
Avg. / Max # of Diseases per Case	1.42 / 26	Avg. / Max # of Procedures per Visit	2.82 / 32
Avg. / Max # of Cases per Disease	3.40 / 148	Avg. / Max # of Medications per Visit	11.27 / 65

Table 2: Statistics of RareBench-Public and MIMIC-IV-Ext-Rare datasets.

RareBench-Publich RareBench is a multi-center dataset comprising rare disease patient data from Europe, China, and Canada. It is specifically designed to evaluate the performance of LLMs in the rare disease domain (Chen et al., 2024b). We utilize 1,197 publicly available rare disease cases, each with at least three symptom codes and corresponding diagnostic information extracted from electronic health records (EHRs).

MIMIC-IV-Ext-Rare MIMIC-IV (version 3.0) contains EHR data from the Beth Israel Deaconess Medical Center (BIDMC) in the United States, spanning 2008 to 2022 (Johnson et al., 2023), with disease codes following ICD-9 and ICD-10 standards. We map these codes to rare disease identifiers from OMIM[1] and Orphanet[2], extracting patients with multiple hospital admissions while excluding cases with incomplete information. This yields MIMIC-IV-Ext-Rare, a dataset of 4,760 rare disease patients with 18,522 admission EHRs specifically curated for medication recommendation tasks in rare disease contexts.

4 Overview of RareAgents.

This section introduces the proposed **RareAgents** framework for rare disease diagnosis and treatment. Figures 2 and 5 provide an overview of the framework and a step-by-step example of its pipeline, respectively. The *RareAgents* framework is composed of three core modules: **(1) Multi-disciplinary Team Collaboration:** The attending physician agent selects the most relevant specialists from a predefined specialist pool based on the patient's clinical information to form an MDT. These special physician agents engage in multiple rounds of discussion to reach a consensus on the diagnosis and treatment plan. **(2) Dynamic Long-term Memory:** Each agent, whether the attending physician or a specialist, maintains a personalized long-term memory. These memories, built from past consultation processes, serve as dynamic experience bases that can be retrieved and updated continuously to assist decision-making. **(3) Medical Tool Utilization:** Throughout the reasoning process, all physician agents can access and utilize various diagnostic and treatment tools to support and enhance their decision-making capabilities. The complete algorithm is detailed in Appendix A.1.

4.1 Multi-disciplinary Team Collaboration

Previous implementations of LLM-based MDTs often have the LLMs autonomously define the roles and responsibilities of various specialists (Tang et al., 2024; Kim et al., 2024). In contrast, our approach mirrors real-world clinical practice by leveraging specialist departments commonly involved in rare disease cases (Xie et al., 2023). Under human specialist physicians' guidance, we constructed a Specialist Pool ($\mathcal{SP}$), which consists of 41 distinct clinical departments. Detailed definitions are provided in Appendix A.2. The entire MDT consultation process is divided into three stages: **(i) MDT Formation** (Line 3-6 of Algorithm 1): The attending physician agent assembles a patient-centric MDT. **(ii) Expert Consensus** (Line 15-22 of Algorithm 1): Specialist agents within the MDT engage in multi-turn discussions (up to a maximum of R rounds) to reach a consensus opinion $\mathcal{O}(\mathcal{R})$ based on patient's information $\mathcal{R}$. **(iii) Report Generation** (Line 25 of Algorithm 1): The attending physician agent synthesizes the opinions from all MDT members to generate a final discussion report $\mathcal{DR}$, where

$$\mathcal{DR} = \text{SUMMARY}\Big(\bigcup_{r=0}^{R} \bigcup_{s\in\text{MDT}} \mathcal{O}_s^{(r)}(\mathcal{R})\Big). \tag{1}$$

4.2 Dynamic Long-term Memory

In real-world clinical practice, physicians rely on both personal experience and historical patient

---

[1] `https://omim.org/`

[2] `https://www.orpha.net/`

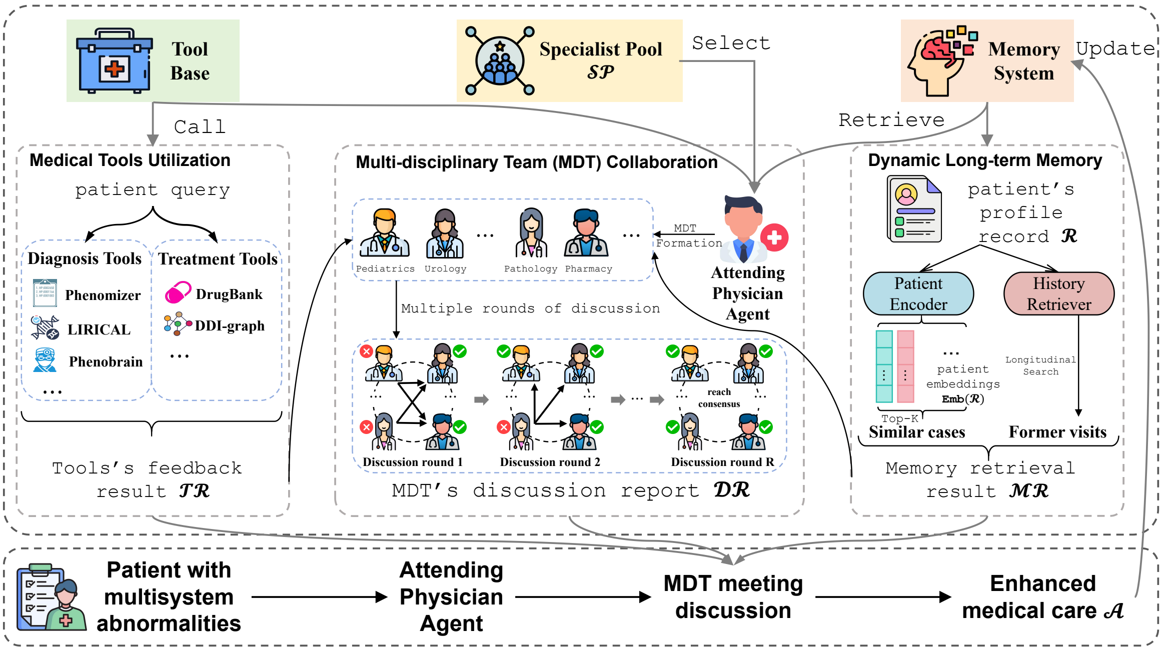

Tool Base
Specialist Pool $\mathcal{SP}$
Select
Memory System
Update
Call
Retrieve
Medical Tools Utilization
patient query
Diagnosis Tools
Treatment Tools
Phenomizer
LIRICAL
Phenobrain
...
DrugBank
DDI-graph
...
Tools's feedback result $\mathcal{TR}$
Multi-disciplinary Team (MDT) Collaboration
Pediatrics
Urology
Pathology
Pharmacy
MDT Formation
Attending Physician Agent
Multiple rounds of discussion
reach consensus
Discussion round 1
Discussion round 2
Discussion round R
MDT's discussion report $\mathcal{DR}$
Dynamic Long-term Memory
patient's profile record $\mathcal{R}$
Patient Encoder
History Retriever
patient embeddings Emb($\mathcal{R}$)
Longitudinal Search
Top-K
Similar cases
Former visits
Memory retrieval result $\mathcal{MR}$
Patient with multisystem abnormalities
Attending Physician Agent
MDT meeting discussion
Enhanced medical care $\mathcal{A}$

Figure 2: Overview of the **RareAgents** framework: For patients involving abnormalities in multiple organs and systems, the Attending Physician Agent selects specialists from a predefined pool to form an MDT, which reaches a consensus through iterative discussions. Each physician agent is equipped with: a) Dynamic long-term memory to retrieve and update similar cases or prior records; and b) Tools to assist in diagnosis and treatment decisions.

records within the healthcare system for decision-making (Trafton, 2018). Llama 3.1 (Dubey et al., 2024), with its expanded context window from 8K to 128K tokens, provides a significantly larger capacity for developing long-term memory. Inspired by these, we design a **dynamic long-term memory mechanism** for the physician agents in *RareAgents*, enabling them to store, retrieve, and update memories like human physicians (Line 12, 26, 31-33 of Algorithm 1). Agents can facilitate personalized diagnosis and treatment based on historical interactions. For diagnosis, we use the rare disease patient embeddings ($Emb(*)$) from RareBench (Chen et al., 2024b) to dynamically retrieve the top-k most similar cases from the patient database. In subsequent experiments, we select $k = 5$. For treatment, we leverage the longitudinal nature of patient records in the MIMIC-IV-Ext-Rare dataset, where each patient may have multiple admission records. During the n-th admission, the physician agent retrieves the patient's records from the previous $n - 1$ visits. Denote $\mathcal{MR}$ as the result of dynamic long-term memory retrieval, where

$$\mathcal{MR}_{diagnosis}\left(\mathcal{R}\right) = \arg\max_{\text{Top-K}}\left(Emb(\mathcal{R})\right),$$

$$\mathcal{MR}_{treatment}(\mathcal{R}^{(n)}) = \mathcal{R}^{(1:n-1)} \cup \mathcal{A}^{(1:n-1)}_{treatment}. \tag{2}$$

4.3 Medical Tools Utilization

Physicians frequently use various tools to assist decision-making in clinical practice (Kawamoto et al., 2005). Similarly, the physician agents in *RareAgents* have access to diagnostic and therapeutic tools to enhance their clinical reasoning capabilities (Line 14, 27 of Algorithm 1). Llama 3.1's built-in tool integration and function-calling capabilities enable the agents to interact with external environments dynamically (Dubey et al., 2024). In this research, diagnostic tools include Phenomizer (Köhler et al., 2009, 2017), LIRICAL (Robinson et al., 2020), and Phenobrain (Mao et al., 2025), all of which are accessible via APIs or web interfaces. Therapeutic tools are knowledge bases like DrugBank (for drug information) (Wishart et al., 2008) and DDI-graph (for drug-drug interaction relationships). Detailed tool functions are provided in Appendix A.4. Let $\mathcal{T} = \{T_1, T_2, ...\}$ denote the set of medical tool functions, and $\mathcal{TR}$ represents the aggregated output from the tools' feedback, where

$$\mathcal{TR} = \text{CONCAT}\Big(\bigcup_{T_i \in \mathcal{T}} T_i(\mathcal{R})\Big). \tag{3}$$

Finally, *RareAgents* synthesize the results from specialist consensus, dynamic long-term memory, and tools' feedback to generate the final decision $\mathcal{A}$:

$$\mathcal{A} = \text{LLM}(\mathcal{R}, \mathcal{DR}, \mathcal{MR}, \mathcal{TR}). \tag{4}$$

Model	Diagnosis on RAREBENCH-PUBLIC				Treatment on MIMIC-IV-EXT-RARE			
	Hit@1	Hit@3	Hit@10	MR(↓)	Jaccard	F1	DDI(↓)	#MED
General & Medical LLMs (zero-shot CoT)								
GPT-4o	0.4169	0.5815	0.7068	2.0	0.3282	0.4693	0.0907	12.10
GPT-3.5	0.3968	0.5079	0.6007	3.0	0.2277	0.3451	0.0856	8.72
UltraMedical-70B	0.4002	0.5639	0.6424	2.0	0.2606	0.3922	0.0739	13.08
OpenBioLLM-70B	0.3885	0.5388	0.6182	2.0	0.1504	0.2465	0.0615	14.73
UltraMedical-8B	0.3425	0.4294	0.4787	>10	0.1613	0.2549	0.0840	9.14
OpenBioLLM-8B	0.1495	0.1763	0.1997	>10	0.0997	0.1715	**0.0519**	20.96
o1-like LLMs (zero-shot)								
DeepSeek-R1-Distill-Llama-70B	0.3509	0.5221	0.6291	3.0	0.2924	0.4267	0.0901	11.87
DeepSeek-R1-Distill-Llama-8B	0.3158	0.4511	0.5171	6.0	0.2109	0.3251	0.0803	9.05
Baichuan-M1-14B	0.3175	0.5313	0.6241	3.0	0.2188	0.3381	0.0734	11.36
HuatuoGPT-o1-70B	0.3584	0.5305	0.6232	3.0	0.2536	0.3819	0.0837	10.57
Llama-3.1-8B-Instruct (Medical Agent framework)								
Single-Agent	0.3041	0.4578	0.5698	5.0	0.2104	0.3229	0.0951	9.68
MedAgents	0.3734	0.4879	0.5698	4.0	0.2285	0.3505	0.0997	9.60
MDAgents	0.3233	0.4453	0.5271	7.0	0.2311	0.3539	0.0715	10.92
RareAgents (MDT only)	0.3826	0.5013	0.6007	3.0	0.2376	0.3630	0.0957	11.74
RareAgents	0.4511	0.5647	0.7377	2.0	0.3052	0.4475	0.0820	12.98
Llama-3.1-70B-Instruct (Medical Agent framework)								
Single-Agent	0.3751	0.5397	0.6658	3.0	0.2543	0.3736	0.0907	10.97
MedAgents	0.4010	0.5163	0.6449	3.0	0.2607	0.3905	0.0974	11.20
MDAgents	0.4042	0.5640	0.6586	2.0	0.2961	0.4349	0.0813	12.41
RareAgents (MDT only)	0.4177	0.5455	0.6800	2.0	0.3089	0.4468	0.0950	13.40
RareAgents	**0.5589**	**0.6867**	**0.7811**	**1.0**	**0.4108**	**0.5563**	0.0796	13.17

Table 3: Results of general, medical and o1-like LLMs, as well as medical agent frameworks on the differential diagnosis task of RareBench-Public and the medication recommendation task of MIMIC-IV-Ext-Rare. **Bold** indicates the best performance, while underlined denotes the second-best performance.

Model	Hit@1	Hit@3	Hit@10	MR(↓)
Phenomizer	0.0844	0.2072	0.3835	>10
LIRICAL	0.1637	0.2840	0.4152	>10
BASE_IC	0.2047	0.3434	0.5322	8.0
Phen2Disease	0.2105	0.3266	0.5129	10.0
Phenobrain	0.2857	0.4670	0.6341	4.0
RareAgents (Llama-3.1-8B)	0.4511	0.5647	0.7377	2.0
RareAgents (Llama3.1-70B)	**0.5589**	**0.6867**	**0.7811**	**1.0**

Table 4: Performance of SOTA Models for Differential Diagnosis.

Model	Jaccard	PRAUC	F1	DDI(↓)	#MED
LR	0.3564	0.6254	0.5020	0.0686	8.30
LEAP	0.2959	0.4484	0.4341	**0.0485**	5.92
RETAIN	0.3527	0.5814	0.5056	0.0626	13.08
G-Bert	0.4030	0.6481	0.5554	0.0751	14.61
GAMENet	0.3731	0.6250	0.5195	0.0650	10.00
SafeDrug	0.3903	0.6213	0.5426	0.0733	12.88
COGNet	0.3883	0.5839	0.5367	0.0751	14.28
MICRON	0.3887	0.5048	0.5417	0.0729	12.91
MoleRec	0.3975	0.6323	0.5498	0.0714	12.15
RAREMed	0.3800	**0.6506**	0.5268	0.0622	8.75
RareAgents (Llama3.1-70B)	**0.4108**	-	**0.5563**	0.0796	13.17

Table 5: Results of SOTA Models for Medication Recommendation.

5 Experimental Setup and Main Results

5.1 Evaluation Metrics

Differential Diagnosis The diagnostic task is evaluated using two primary metrics: top-k recall (Hit@k, where k=1, 3, 10) and median rank (MR). Hit@k measures diagnostic accuracy by checking if the actual disease is among the top-k predictions, while MR represents the median position of the correct diagnosis across all cases.

Medication Recommendation The therapeutic task is assessed with four metrics: Jaccard coefficient (Jaccard), F1-score (F1), Drug-Drug Interaction rate (DDI), and the average number of recommended medications (#MED). Jaccard measures the overlap between the recommended and ground truth medication sets, normalized by their union. F1 quantifies recommendation precision and recall, with higher values indicating better performance. DDI reflects the frequency of potential adverse interactions among recommended drugs, with lower values indicating safer prescriptions. #MED evaluates the consistency between the number of recommended medications and those prescribed by clinicians. Detailed formulas for all metrics are provided in Appendix A.5.

5.2 Baselines

Domain-specific SOTA models For the differential diagnosis task, the domain-specific SOTA models include Phenomizer (Köhler et al., 2009, 2017), LIRICAL (Robinson et al., 2020), BASE_IC, Phen2Disease (Zhai et al., 2023), and Phenobrain (Mao et al., 2025). For the medication recommendation task, we leverage ten models: Logistic Regression (LR), LEAP (Zhang et al., 2017), RETAIN (Choi et al., 2016), G-Bert (Shang et al., 2019a), GAMENet (Shang et al., 2019b), SafeDrug (Yang et al., 2021b), COGNet (Wu et al., 2022), MICRON (Yang et al., 2021a), MoleRec (Yang et al., 2023), and RAREMed (Zhao et al., 2024). Notably, these models for medication recommendation require training on the dataset. We conduct **5-fold cross-validation** based on the number of patients in MIMIC-IV-Ext-Rare and report the average results. In each fold, 20% of the data is used as the test set, while the remaining 80% is split into 80% training and 20% validation subsets. Appendix A.7.2 provides additional details on these baselines and their configurations.

General, Medical and o1-like LLMs General LLMs include the latest version of GPT-4o and GPT-3.5-turbo-0125 (Achiam et al., 2023). The medical LLMs include OpenBioLLM (Ankit Pal, 2024) and UltraMedical (Zhang et al., 2024a), both fine-tuned on medical datasets using Llama-3 (8B and 70B). O1-like LLMs include DeepSeek-R1-Distill-Llama (8B and 70B) (Guo et al., 2025), Baichuan-M1-14B, and HuatuoGPT-o1-70B (Chen et al., 2024a). All of these models are evaluated

Model	Diagnosis on RareBench-Public				Treatment on MIMIC-IV-Ext-Rare			
	Hit@1	Hit@3	Hit@10	MR(↓)	Jaccard	F1	DDI(↓)	#MED
Llama-3.1-8B-Instruct								
w/o MDT	0.4394 (↓ 2.6%)	**0.5973**	0.7343	**2.0**	0.2856	0.4244	0.0850 (↑ 3.7%)	12.91
w/o Memory	0.3952 (↓ 12.4%)	0.5581	0.6951	3.0	0.2422	0.3689	**0.0723** (↓ 11.8%)	12.94
w/o Tools	0.4361 (↓ 3.3%)	0.5113	0.7143	3.0	0.2644	0.3951	0.1012 (↑ 23.4%)	11.92
RareAgents	**0.4511**	0.5647	**0.7377**	**2.0**	**0.3052**	**0.4475**	0.0820	12.98
Llama-3.1-70B-Instruct								
w/o MDT	0.5171 (↓ 7.5%)	0.6416	0.7377	**1.0**	0.3828	0.5292	0.0859 (↑ 7.9%)	13.04
w/o Memory	0.4336 (↓ 22.4%)	0.5564	0.6976	2.0	0.3185	0.4584	0.0884 (↑ 11.1%)	13.20
w/o Tools	0.5221 (↓ 6.6%)	0.6558	0.7469	**1.0**	0.3662	0.5090	0.0961 (↑ 20.7%)	13.25
RareAgents	**0.5589**	**0.6867**	**0.7811**	**1.0**	**0.4108**	**0.5563**	**0.0796**	13.17

Table 6: Ablation study results for the impact of each module in *RareAgents* on diagnosis and treatment performance. Red percentages indicate the relative performance drop when the corresponding module is removed from *RareAgents*.

in a zero-shot setting with the temperature parameter set to 0. Non-o1-like LLMs utilize Chain-of-Thought (CoT) (Wei et al., 2022) to enhance reasoning.

Open-Source Medical Multi-Agents For open-source medical multi-agent frameworks, we select MedAgents (Tang et al., 2024) and MDAgents (Kim et al., 2024), both implemented initially using GPT-4 APIs. We have adapted them to operate on the local Llama-3.1 models.

5.3 Main Results

Table 3, 4, and 5 present the performance of all models on RareBench-Public for differential diagnosis and MIMIC-IV-Ext-Rare for medication recommendation. Detailed case studies are provided in Appendix A.9.

Differential Diagnosis RareAgents (Llama-3-70B) outperform all baselines across all evaluation metrics. Even though RareAgents (Llama-3-8B) ranks second in some metrics such as Top-1 Recall (Hit@1), demonstrating significant improvements over other medical agent frameworks. Interestingly, LLMs' performance already surpasses that of domain-specific SOTA models. Among the fine-tuned models, UltraMedical performs better than the base Llama-3.1, while OpenBioLLM shows a decline in performance. This suggests that fine-tuned models may not generalize well to all medical tasks, because their effectiveness is highly dependent on the fine-tuning data and methods.

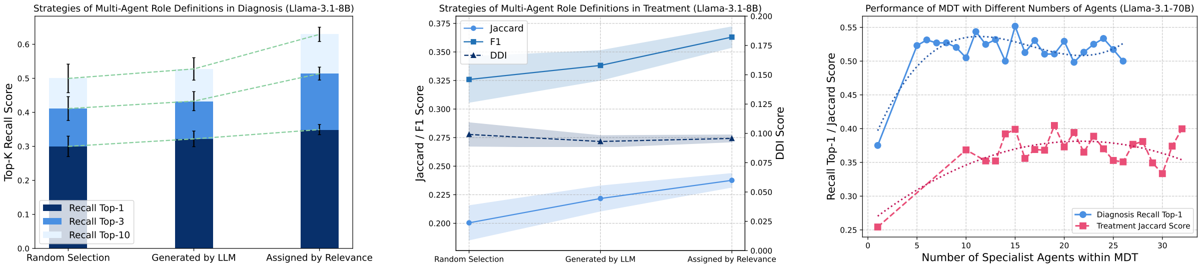
Medication Recommendation Table 3 summarizes the results of the top five domain-specific SOTA models, with full results for all ten models presented in Appendix A.8. *RareAgents* (Llama-3-70B) achieves the best performance across all metrics except for Drug-Drug Interaction (DDI). The dataset's inherent DDI and average number of medications recommended per case (#MED) are 0.0755 and 11.27, respectively. While OpenBioLLM achieves the lowest DDI rate, it performs poorly in Jaccard and F1. Its higher #MED indicates a tendency to recommend more irrelevant medications. For other metrics, existing LLMs and multi-agent frameworks remain inferior to the performance of domain-specific SOTA models trained on the dataset. Notably, *RareAgents* demonstrates competitive performance through a plug-and-play framework without additional training.

6 Analysis and Discussion

Given the space constraints in the main text, this section primarily conducts comprehensive experiments to assess the advancement of multidisciplinary team collaboration within *RareAgents*. For more detailed analyses regarding the optimal number of cases in memory and the individual contributions of each tool, please refer to Appendix A.3.

6.1 Ablation Study

RareAgents consists of three key components: Multi-disciplinary Team (MDT) collaboration, dynamic long-term memory, and medical tools utilization. To quantify the contribution of each module, we conduct ablation experiments by removing one component at a time. The results are presented in Table 6. The findings reveal that removing any single component leads to a performance decline to varying degrees. Among them, when the memory module is removed, the performance drop is most significant. This is attributed to the complexity of rare diseases. The memory module provides the necessary context, helping the model distinguish rare conditions from more common ones, thus avoiding the pitfalls of a cold start in reasoning. In the medication recommendation task, the



(a) Diagnosis MDT settings (b) Treatment MDT settings (c) Agents Number in MDT

Figure 3: A detailed exploration of the advanced nature of multidisciplinary team collaboration within *RareAgents*.

removal of the tools module results in a significant increase in DDI rate. This is because drug knowledge bases deliver detailed professional guidance, which effectively reduces drug-drug interactions and enhances medication safety.

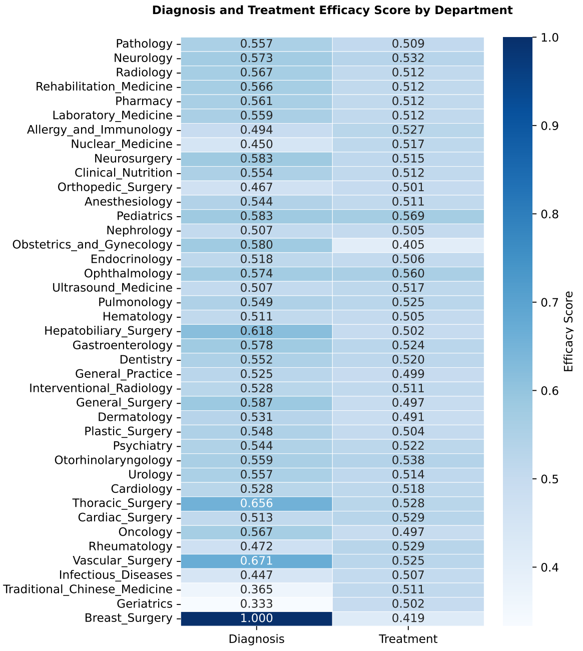
6.2 Advanced Nature of MDT in RareAgents

To further evaluate the efficacy of MDT within *RareAgents*, we conduct experiments solating the MDT component from the memory mechanisms and external tools. As reported in Table 3, the configuration **RareAgents(MDT only)** consistently outperforms other medical agents. This result highlights not only the robustness of the MDT within *RareAgents* but also its pivotal role in navigating the intricate challenges posed by rare diseases.

Moreover, we explore three strategies for assigning specialist roles: (1) autonomously generated by the LLM (Tang et al., 2024; Kim et al., 2024), (2) randomly selected from a predefined specialist pool, and (3) assigned based on the most relevant departmental expertise. All strategies employ the same number of specialist agents. Figures 3a and 3b demonstrate that assigning specialists based on departmental relevance consistently outperforms the other two strategies. This advantage arises from the expert-curated role definitions, which are grounded in domain-specific knowledge and enable deeper contextual understanding.

6.3 Specialists Collaboration Patterns

To gain deeper insights into the collaboration and contribution distribution of specialist agents within *RareAgents*, we conduct a detailed analysis of their interaction patterns. Figure 7 shows that, on average, *RareAgents* (Llama-3.1-70B) engage 12.53 specialists for differential diagnosis and 22.22 for medication recommendation. As shown in Figure 3c, MDT performance peaks around these agent numbers. Furthermore, we analyzed the efficacy scores of each specialist agent, measured by the average diagnostic Recall@1 and medication recommendation F1 across assigned cases. As shown in Figure 4, over 90% of specialists achieved efficacy scores above 0.5 in both tasks, confirming their necessity and effectiveness.


Diagnosis and Treatment Efficacy Score by Department

	Diagnosis	Treatment
Pathology	0.557	0.509
Neurology	0.573	0.532
Radiology	0.567	0.512
Rehabilitation_Medicine	0.566	0.512
Pharmacy	0.561	0.512
Laboratory_Medicine	0.559	0.512
Allergy_and_Immunology	0.494	0.527
Nuclear_Medicine	0.450	0.517
Neurosurgery	0.583	0.515
Clinical_Nutrition	0.554	0.512
Orthopedic_Surgery	0.467	0.501
Anesthesiology	0.544	0.511
Pediatrics	0.583	0.569
Nephrology	0.507	0.505
Obstetrics_and_Gynecology	0.580	0.405
Endocrinology	0.518	0.506
Ophthalmology	0.574	0.560
Ultrasound_Medicine	0.507	0.517
Pulmonology	0.549	0.525
Hematology	0.511	0.505
Hepatobiliary_Surgery	0.618	0.502
Gastroenterology	0.578	0.524
Dentistry	0.552	0.520
General_Practice	0.525	0.499
Interventional_Radiology	0.528	0.511
General_Surgery	0.587	0.497
Dermatology	0.531	0.491
Plastic_Surgery	0.548	0.504
Psychiatry	0.544	0.522
Otorhinolaryngology	0.559	0.538
Urology	0.557	0.514
Cardiology	0.528	0.518
Thoracic_Surgery	0.656	0.528
Cardiac_Surgery	0.513	0.529
Oncology	0.567	0.497
Rheumatology	0.472	0.529
Vascular_Surgery	0.671	0.525
Infectious_Diseases	0.447	0.507
Traditional_Chinese_Medicine	0.365	0.511
Geriatrics	0.333	0.502
Breast_Surgery	1.000	0.419

Figure 4: Heatmap of Efficacy Scores in Diagnostic and Treatment Tasks Across Specialist Departments.

7 Conclusion

This paper presents **RareAgents**, a patient-centered framework designed to facilitate personalized diagnosis and treatment for rare diseases through the integration of multidisciplinary team collaboration, dynamic long-term memory, and medical tools. As a plug-and-play framework, *RareAgents* demonstrates superior performance on Llama-3.1 (8B and 70B), surpassing domain-specific state-of-the-art models, general, medical and o1-like LLMs, as well as medical multi-agent frameworks. Furthermore, we contribute MIMIC-IV-Ext-Rare, a curated dataset specifically for rare disease patients, providing a valuable resource for future research.

Limitations

In this study, we utilize pre-extracted symptoms and clinical codes from patients' original electronic health records (EHRs), focusing solely on textual data. While this approach provides valuable insights, it is important to acknowledge that the diagnosis and treatment of rare diseases often benefit from integrating multimodal data, such as medical imaging and genotypic information. Although *RareAgents* is designed as a flexible, plug-and-play framework that supports easy integration and adaptation, the current implementation does not include domain-specific fine-tuning of the underlying large language models (LLMs) for medical or rare disease contexts. In future work, we plan to address these limitations by incorporating more comprehensive patient data and performing targeted fine-tuning on state-of-the-art open-source LLMs, to achieve more accurate diagnostic and therapeutic outcomes.

Ethical Considerations

Licenses In our research, we utilize two public datasets, adhering to the highest ethical standards. Both datasets are free from content that could compromise patient privacy or disclose personally identifiable information. The RareBench dataset (Chen et al., 2024b) is released under the Creative Commons Attribution 4.0 International License (CC BY). MIMIC-IV (Johnson et al., 2023) is governed by the PhysioNet Credentialed Health Data License 1.5.0. The MIMIC-IV-Ext-Rare dataset, derived from MIMIC-IV, is strictly intended for research purposes and will be shared on PhysioNet under the same licensing terms as the original dataset.

Potential Risks Although *RareAgents* has demonstrated promising performance in tasks related to differential diagnosis and medication recommendations for rare diseases on real-world datasets such as RareBench and MIMIC-IV, further validation on external datasets and feedback from medical professionals are essential. Additionally, it is important to acknowledge the inherent limitations of LLMs, including potential biases and hallucinations in their outputs. Therefore, it is critical to emphasize that LLMs should currently be viewed only as supplementary tools. For clinical decision-making, guidance from qualified medical professionals is indispensable, especially for specific diagnostic or treatment decisions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024a. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

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*, pages 4850–4861.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29.

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*.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Asif Javed, Saloni Agrawal, and Pauline C Ng. 2014. Phen-gen: combining phenotype and genotype to analyze rare disorders. *Nature methods*, 11(9):935–937.

Jinmeng Jia, Ruiyuan Wang, Zhongxin An, Yongli Guo, Xi Ni, and Tieliu Shi. 2018. Rdad: a machine learning system to support phenotype-based rare disease diagnosis. *Frontiers in genetics*, 9:587.

Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W John Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, et al. 2024.

Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *arXiv preprint arXiv:2402.13225*.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. 2005. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj*, 330(7494):765.

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. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sebastian Köhler, Marcel H Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N Robinson. 2009. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464.

Sebastian Köhler, Nicole A Vasilevsky, Mark Engelstad, Erin Foster, Julie McMurry, Ségolène Aymé, Gareth Baynam, Susan M Bello, Cornelius F Boerkoel, Kym M Boycott, et al. 2017. The human phenotype ontology in 2017. *Nucleic acids research*, 45(D1):D865–D876.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. 2024a. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: communicative agents for" mind" exploration of large language model society. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 51991–52008.

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

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations*.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024. Triageagent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–5764.

Xiaohao Mao, Yu Huang, Ye Jin, Lun Wang, Xuanzhong Chen, Honghong Liu, Xinglin Yang, Haopeng Xu, Xiaodong Luan, Ying Xiao, et al. 2025. A phenotype-based ai pipeline outperforms human experts in differentially diagnosing rare diseases using ehrs. *npj Digital Medicine*, 8(1):68.

Nikita Mehandru, Brenda Y Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J Butte, and Ahmed Alaa. 2024. Evaluating large language models as agents in the clinic. *NPJ digital medicine*, 7(1):84.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Stéphanie Nguengang Wakap, Deborah M Lambert, Annie Olry, Charlotte Rodwell, Charlotte Gueydan, Valérie Lanneau, Daniel Murphy, Yann Le Cam, and Ana Rath. 2020. Estimating cumulative point prevalence of rare diseases: analysis of the orphanet database. *European journal of human genetics*, 28(2):165–173.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Jiajie Peng, Hansheng Xue, Yukai Shao, Xuequn Shang, Yadong Wang, and Jin Chen. 2016. Measuring phenotype semantic similarity using human phenotype ontology. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 763–766. IEEE.

PKS Prakash, Srinivas Chilukuri, Nikhil Ranade, and Shankar Viswanathan. 2021. Rarebert: transformer architecture for rare disease patient identification using administrative claims. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 453–460.

Peter N Robinson, Vida Ravanmehr, Julius OB Jacobsen, Daniel Danis, Xingmin Aaron Zhang, Leigh C Carmody, Michael A Gargano, Courtney L Thaxton, Guy Karlebach, Justin Reese, et al. 2020. Interpretable clinical genomics with a likelihood ratio paradigm. *The American Journal of Human Genetics*, 107(3):403–417.

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*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessí, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: language models can teach themselves to use tools. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 68539–68551.

Arrigo Schieppati, Jan-Inge Henter, Erica Daina, and Anita Aperia. 2008. Why rare diseases are an important medical and social issue. *The Lancet*, 371(9629):2039–2041.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019a. Pre-training of graph augmented transformers for medication recommendation. In *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019b. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1126–1133.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May Dongmei Wang. 2024. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22315–22339.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 8634–8652.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

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.

Anne Trafton. 2018. Doctors rely on more than just data for medical decision making. *Science Daily*, 20.

Rodolfo Valdez, Lijing Ouyang, and Julie Bolen. 2016. Public health and rare diseases: oxymoron no more. *Preventing chronic disease*, 13.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

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.

David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl_1):D901–D906.

Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. 2022. Conditional generation net for medication recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 935–945.

Jing Xie, Ye Jin, Min Shen, Limeng Chen, and Shuyang Zhang. 2023. A patient-centric, coordinated care model for rare diseases: The multidisciplinary consultation program at peking union medical college hospital. *NEJM Catalyst Innovations in Care Delivery*, 4(s1).

Chaoqi Yang, Cao Xiao, Lucas Glass, and Jimeng Sun. 2021a. Change matters: Medication change prediction with recurrent residual networks. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*, pages 3728–3734. International Joint Conferences on Artificial Intelligence.

Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021b. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*, pages 3735–3741. International Joint Conferences on Artificial Intelligence.

Hui Yang, Peter N Robinson, and Kai Wang. 2015. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nature methods*, 12(9):841–843.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. 2023. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *Proceedings of the ACM Web Conference 2023*, pages 4075–4085.

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

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W Suchow, Zhenyu Cui, Rong Liu, et al. 2024. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Weiqi Zhai, Xiaodi Huang, Nan Shen, and Shanfeng Zhu. 2023. Phen2disease: a phenotype-driven model for disease and gene prioritization by bidirectional maximum matching semantic similarities. *Briefings in Bioinformatics*, 24(4):bbad172.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. 2024a. Ultramedical: Building specialized generalists in biomedicine. *arXiv preprint arXiv:2406.03949*.

Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. 2017. Leap: learning to prescribe effective and safe treatment combinations for multimorbidity. In *proceedings of the 23rd ACM SIGKDD international conference on knowledge Discovery and data Mining*, pages 1315–1324.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024b. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*.

Zihao Zhao, Yi Jing, Fuli Feng, Jiancan Wu, Chongming Gao, and Xiangnan He. 2024. Leave no patient behind: Enhancing medication recommendation for rare disease patients. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 533–542.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

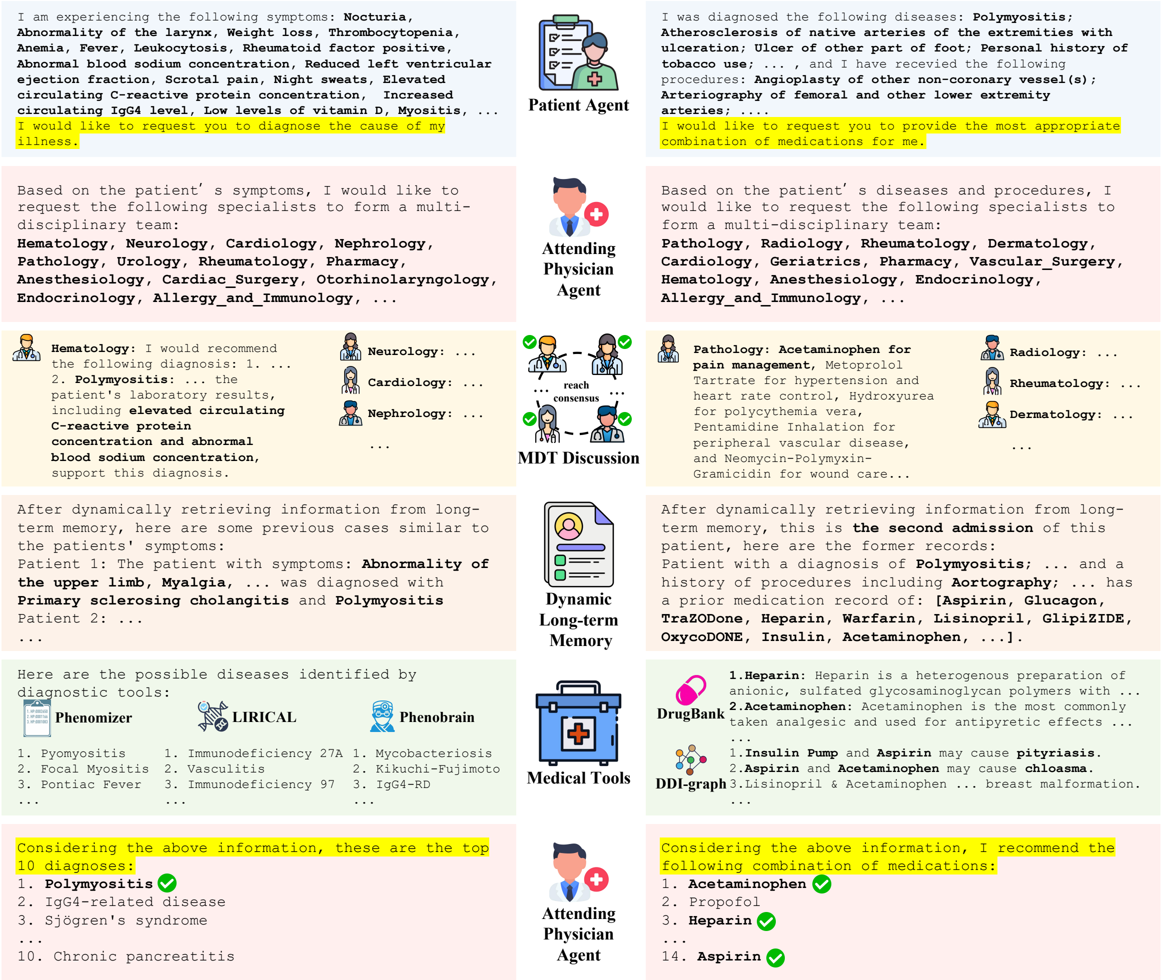
A Appendix

A.1 RAREAGENTS Algorithm

Algorithm 1: Overview of RAREAGENTS

```
Input: $\mathcal{R}$: patient profile, $APA$: attending physician
       agent, $\mathcal{SP}$: specialist pool, $\mathcal{T}$: medical tool
       functions, $R$: the maximum number of rounds
Output: Final decision $\mathcal{A}$ for rare disease patient
Initialize $f^c \leftarrow False$, $r \leftarrow 0$, $\mathcal{O}^{(0)}(\mathcal{R}) \leftarrow \varnothing$
// Multi-Disciplinary Team (MDT) formation
MDT $\leftarrow$ LLM$_{APA}(\mathcal{R}, \mathcal{SP})$
// Specialist physician agents in MDT initialization
foreach $spa \in MDT$ do
    $f^c_{spa} \leftarrow$ False, $\Delta\mathcal{O}^{(0)}_{spa} \leftarrow \mathcal{R}$
while not $f^c$ and $r < R$ do
    $f^c \leftarrow True$
    foreach $s$ in MDT do
        if not $f^c_s$ then
            // Dynamic self-memory retrieval
            $\mathcal{MR}_s \leftarrow$ LLM$_s.retrieve(\mathcal{R})$
            // Medical tools utilization
            $\mathcal{TR}_s \leftarrow$ LLM$_s.useTools(\mathcal{T}(\mathcal{R}))$
            $\mathcal{O}^{(r)}_s \leftarrow$ LLM$_s(\Delta\mathcal{O}^{(r)}_s, \mathcal{MR}_s, \mathcal{TR}_s)$
            // Opinion synchronization
            $\Delta\mathcal{O}^{(r+1)}_s \leftarrow$ MEETING$(\mathcal{O}^{(r)}_s$, MDT$)$
            if $\Delta\mathcal{O}^{(r+1)}_s = \varnothing$ then
                $f^c_s \leftarrow True$
                $\mathcal{O}^{(r)}(\mathcal{R}) \leftarrow \mathcal{O}^{(r)}(\mathcal{R}) \cup \mathcal{O}^{(r)}_s$
            else
                $f^c \leftarrow False$
    $\mathcal{O}^{(r+1)}(\mathcal{R}) \leftarrow \mathcal{O}^{(r)}(\mathcal{R})$, $r \leftarrow r + 1$
// Summary of MDT's final discussion result
$\mathcal{DR} \leftarrow$ LLM$_{APA}.summary(\mathcal{O}^{(r)}(\mathcal{R}))$
$\mathcal{MR} \leftarrow$ LLM$_{APA}.retrieve(\mathcal{R})$
$\mathcal{TR} \leftarrow$ LLM$_{APA}.useTools(\mathcal{T}(\mathcal{R}))$
// Final Decision
$\mathcal{A} \leftarrow$ LLM$_{APA}(\mathcal{R}, \mathcal{DR}, \mathcal{MR}, \mathcal{TR})$
// Agents memory update
LLM$_{APA}.update(\mathcal{R}, \mathcal{A})$
foreach $spa \in MDT$ do
    LLM$_{spa}.update(\mathcal{R}, \mathcal{A})$
return $\mathcal{A}$
```

Algorithm 1 provides a detailed overview of the *RareAgents* framework, in which the attending physician agent coordinates the entire patient-centered medical decision-making process by integrating three essential components: **multidisciplinary team collaboration, dynamic long-term memory, and medical tools**, ensuring the delivery of personalized diagnostic and treatment plans.



I am experiencing the following symptoms: **Nocturia, Abnormality of the larynx, Weight loss, Thrombocytopenia, Anemia, Fever, Leukocytosis, Rheumatoid factor positive, Abnormal blood sodium concentration, Reduced left ventricular ejection fraction, Scrotal pain, Night sweats, Elevated circulating C-reactive protein concentration, Increased circulating IgG4 level, Low levels of vitamin D, Myositis,** ... I would like to request you to diagnose the cause of my illness.

I was diagnosed the following diseases: **Polymyositis; Atherosclerosis of native arteries of the extremities with ulceration; Ulcer of other part of foot; Personal history of tobacco use;** ... , and I have recevied the following procedures: **Angioplasty of other non-coronary vessel(s); Arteriography of femoral and other lower extremity arteries;** ....
I would like to request you to provide the most appropriate combination of medications for me.

Based on the patient' s symptoms, I would like to request the following specialists to form a multi-disciplinary team:
Hematology, Neurology, Cardiology, Nephrology, Pathology, Urology, Rheumatology, Pharmacy, Anesthesiology, Cardiac_Surgery, Otorhinolaryngology, Endocrinology, Allergy_and_Immunology, ...

Based on the patient' s diseases and procedures, I would like to request the following specialists to form a multi-disciplinary team:
Pathology, Radiology, Rheumatology, Dermatology, Cardiology, Geriatrics, Pharmacy, Vascular_Surgery, Hematology, Anesthesiology, Endocrinology, Allergy_and_Immunology, ...

Hematology: I would recommend the following diagnosis: 1. ... 2. **Polymyositis:** ... the patient's laboratory results, including **elevated circulating C-reactive protein concentration and abnormal blood sodium concentration,** support this diagnosis.

Neurology: ...

Cardiology: ...

Nephrology: ...

...

Pathology: Acetaminophen for pain management, Metoprolol Tartrate for hypertension and heart rate control, Hydroxyurea for polycythemia vera, Pentamidine Inhalation for peripheral vascular disease, and Neomycin-Polymyxin-Gramicidin for wound care...

Radiology: ...

Rheumatology: ...

Dermatology: ...

...

After dynamically retrieving information from long-term memory, here are some previous cases similar to the patients' symptoms:
Patient 1: The patient with symptoms: **Abnormality of the upper limb, Myalgia,** ... was diagnosed with **Primary sclerosing cholangitis** and **Polymyositis**
Patient 2: ...
...

After dynamically retrieving information from long-term memory, this is **the second admission** of this patient, here are the former records:
Patient with a diagnosis of **Polymyositis;** ... and a history of procedures including **Aortography;** ... has a prior medication record of: **[Aspirin, Glucagon, TraZODone, Heparin, Warfarin, Lisinopril, GlipiZIDE, OxycoDONE, Insulin, Acetaminophen,** ...**].**

Here are the possible diseases identified by diagnostic tools:

Phenomizer
1. Pyomyositis
2. Focal Myositis
3. Pontiac Fever
...

LIRICAL
1. Immunodeficiency 27A
2. Vasculitis
3. Immunodeficiency 97
...

Phenobrain
1. Mycobacteriosis
2. Kikuchi-Fujimoto
3. IgG4-RD
...

1.**Heparin:** Heparin is a heterogenous preparation of anionic, sulfated glycosaminoglycan polymers with ...
2.**Acetaminophen:** Acetaminophen is the most commonly taken analgesic and used for antipyretic effects ...
...

1.**Insulin Pump** and **Aspirin** may cause **pityriasis.**
2.**Aspirin** and **Acetaminophen** may cause **chloasma.**
3.Lisinopril & Acetaminophen ... breast malformation.
...

Considering the above information, these are the top 10 diagnoses:
1. **Polymyositis** ✓
2. IgG4-related disease
3. Sjögren's syndrome
...
10. Chronic pancreatitis

Considering the above information, I recommend the following combination of medications:
1. **Acetaminophen** ✓
2. Propofol
3. **Heparin** ✓
...
14. **Aspirin** ✓

Figure 5: Illustrative examples of our proposed **RareAgents** framework for diagnosing and treating rare diseases are presented. The left panel features a patient from RareBench-Public requiring differential diagnosis based on symptoms, while the right panel shows a case from MIMIC-IV-Ext-Rare involving medication recommendations based on diseases and procedures. *RareAgents* integrates multidisciplinary team (MDT) collaboration, dynamic long-term memory, and the utilization of medical tools to provide patient-centered, personalized care.

Pediatrics
Urology
Hematology
Radiology
Neurosurgery
Rheumatology
Psychiatry
Pulmonology
Dentistry
Endocrinology
Allergy and Immunology
Nuclear Medicine
Hepatobiliary Surgery
Plastic Surgery
Interventional Radiology
Cardiology
Thoracic Surgery
Clinical Nutrition
Vascular Surgery
Anesthesiology
Laboratory Medicine
Pathology
Neurology
Obstetrics and Gynecology
Ophthalmology
General Surgery
Dermatology
Geriatrics
Orthopedic Surgery
Cardiac Surgery
Traditional Chinese Medicine
Nephrology
Oncology
General Practice
Gastroenterology
Infectious Diseases
Rehabilitation Medicine
Pharmacy
Ultrasound Medicine
Otorhinolaryngology
Breast Surgery

Table 7: List of 41 Departments from Specialist Pool $\mathcal{SP}$.

A.2 Predefined Specialist Pool

Under the guidance of human specialist physicians, we established a **Predefined Specialist Pool** ($\mathcal{SP}$) consisting of 41 different medical specialties, as listed in Table 7. Specifically, each specialist agent is characterized by two primary attributes: **description** and **system_message**. When forming a multidisciplinary team (MDT), the attending physician agent selects specialists based on the content of their descriptions, while the system_message serves as the system prompt for the chosen agent.

Pediatrics

<description>

A Pediatrician specializes in the medical care of infants, children, and adolescents, focusing on their healthy growth, development, and managing childhood illnesses.

< system_message>

As a Pediatrician, you are dedicated to the medical care of infants, children, and adolescents. Your expertise involves diagnosing and treating a wide range of illnesses and conditions specific to younger patients. Your role is crucial in promoting healthy growth and development, providing preventative care, and offering guidance to parents on the well-being of their children. Demonstrate proficiency in pediatric assessments, immunizations, and managing both acute and chronic conditions. Communicate effectively with children and their families to ensure a comprehensive understanding of treatment plans and health maintenance strategies.

Figure 6: Example of Agent Role for Pediatrics in the Specialist Pool.

Figure 6 illustrates an example of a pediatric specialist agent, while detailed definitions for the remaining 40 specialties can be found in the code repository.

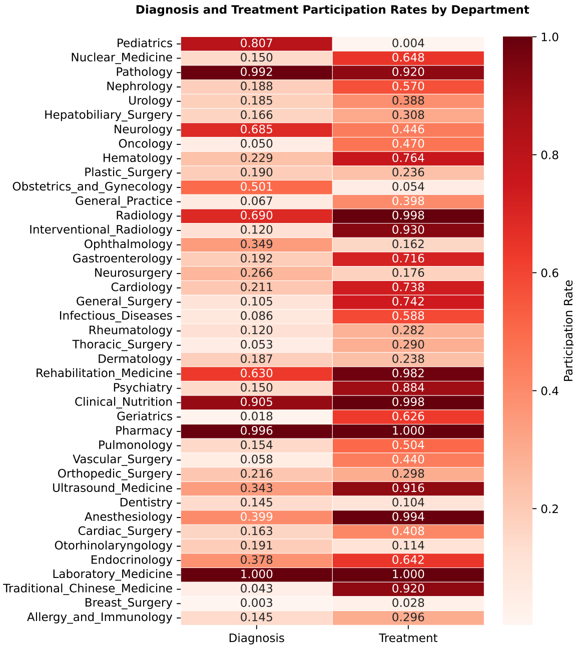
A.3 More Detailed Analysis

A.3.1 Specialists Collaboration Patterns

Figure 7 displays the participation patterns of specialist agents within *RareAgents* (Llama-3.1-70B). In the diagnostic task, the relatively even distribution of contributions across specialties underscores the necessity of diverse clinical insights for accurately identifying rare diseases. In contrast, the treatment task sees more concentrated input from select experts, reflecting the need for deep, specialized knowledge to navigate complex therapeutic protocols and ensure optimal patient outcomes. Furthermore, the structured role assignments facilitate better collaboration among agents, allowing them to leverage complementary expertise and provide a more holistic clinical assessment.

A.3.2 More Ablation Study about Memory Mechanism

As shown in Table 1, existing medical multi-agent frameworks, such as MedAgents (Tang et al., 2024) and MDAgents (Kim et al., 2024), do not incorporate a memory mechanism. To assess the impact of this feature, we performed an ablation study comparing conditions with and without the memory module under the Single-Agent baseline. The results, presented in the Table 8, indicate that the inclusion of a memory mechanism yields a noticeable performance improvement. Specifically, while the memory-enhanced agent outperforms the baseline, RareAgents achieves the highest performance, underscoring the combined benefits of memory and multi-agent collaboration. This suggests that the memory mechanism not only aids in retaining contextual information but also complements the synergy among agents, which is crucial in rare disease diagnosis and treatment.



	Diagnosis	Treatment
Pediatrics	0.807	0.004
Nuclear_Medicine	0.150	0.648
Pathology	0.992	0.920
Nephrology	0.188	0.570
Urology	0.185	0.388
Hepatobiliary_Surgery	0.166	0.308
Neurology	0.685	0.446
Oncology	0.050	0.470
Hematology	0.229	0.764
Plastic_Surgery	0.190	0.236
Obstetrics_and_Gynecology	0.501	0.054
General_Practice	0.067	0.398
Radiology	0.690	0.998
Interventional_Radiology	0.120	0.930
Ophthalmology	0.349	0.162
Gastroenterology	0.192	0.716
Neurosurgery	0.266	0.176
Cardiology	0.211	0.738
General_Surgery	0.105	0.742
Infectious_Diseases	0.086	0.588
Rheumatology	0.120	0.282
Thoracic_Surgery	0.053	0.290
Dermatology	0.187	0.238
Rehabilitation_Medicine	0.630	0.982
Psychiatry	0.150	0.884
Clinical_Nutrition	0.905	0.998
Geriatrics	0.018	0.626
Pharmacy	0.996	1.000
Pulmonology	0.154	0.504
Vascular_Surgery	0.058	0.440
Orthopedic_Surgery	0.216	0.298
Ultrasound_Medicine	0.343	0.916
Dentistry	0.145	0.104
Anesthesiology	0.399	0.994
Cardiac_Surgery	0.163	0.408
Otorhinolaryngology	0.191	0.114
Endocrinology	0.378	0.642
Laboratory_Medicine	1.000	1.000
Traditional_Chinese_Medicine	0.043	0.920
Breast_Surgery	0.003	0.028
Allergy_and_Immunology	0.145	0.296

Figure 7: Heatmap of Participation Rates in Diagnostic and Treatment Tasks Across Specialist Departments.

A.3.3 Random vs. Dynamic of Memory Setting

To evaluate the efficacy of the dynamic retrieval mechanism in long-term memory, we compare it with a baseline approach that randomly selects an equivalent number of cases. The results are shown in Figure 8, the dynamic memory mechanism achieves significant performance improvements by retrieving only a small number of highly relevant records (e.g., just 5 cases), emphasizing that precise context outweighs sheer volume. In contrast, random retrieval offers limited utility even when retrieving a vast number of cases.

Further analysis of the dynamic long-term memory for diagnosis reveals that the optimal diagnostic performance occurs at $k = 5$. As the number of cases increases, the diagnostic performance may decline. This is because cases with similar symptoms are often retrieved within the first five instances, and incorporating more cases may introduce additional hallucinations in the LLM, reducing diagnostic reliability.

A.3.4 Effectiveness of Different Medical Tools

We further assess the individual contributions of tools in a single-agent setup, where the agent is restricted to using one tool at a time. The results, shown in Figures 9, indicate that each tool independently enhances the agent's performance, with more effective tools contributing greater enhancements. Their combined use further reinforces overall results by addressing different aspects of the task, highlighting the complementary nature of these resources. Combining all tools delivers the best overall results. Moreover, incorporating the DDI-graph significantly reduces the DDI rate in treatment, promoting medication safety and efficacy.

A.4 Medical tools

Phenomizer Phenomizer is a sophisticated web-based application designed to assist clinicians in the differential diagnosis of rare genetic diseases. We simulated interactions with the Phenomizer platform using a Python script. Specifically, the script selects "symmetric" as the similarity measure from the menu, sequentially inputs the symptom codes (The Human Phenotype Ontology codes), and retrieves the diagnostic results. For each predicted disease, the output includes an associated p-value. In our framework, *RareAgents* utilizes a tool function to obtain the top 10 diseases predicted by Phenomizer.

LIRICAL LIRICAL (LIkelihood Ratio Interpretation of Clinical AbnormaLities) is a locally deployable tool that we executed using Java 17. It provides posterior probabilities for predicted diseases. Similarly, *RareAgents* retrieves the top 10 diagnostic results generated by LIRICAL through a function call.

Phenobrain Phenobrain is a rare disease diagnostic tool available both as a web application and through an API. For each predicted disease, it provides a corresponding prediction score. *RareAgents* interacts with Phenobrain via a dedicated API-access function to retrieve the top 10 predicted diseases.

DrugBank DrugBank is a comprehensive, freely accessible online database that provides detailed information on drugs and their targets, including chemical properties, pharmacology, mechanisms of action, and molecular biology data. In the medication recommendation task, for drugs appearing in the candidate list, *RareAgents* retrieves more detailed information about each drug from DrugBank via function calls.

DDI-Graph DDI-Graph (Drug-Drug Interaction Graph) is a graph-based model representing drug-drug interactions (DDIs), where drugs are represented as nodes and interactions as edges. *RareAgents* can query this graph data via function calls to retrieve pairwise interaction relationships between drugs. We utilize the same version of DrugBank and DDI-Graph as employed by RAREMed (Zhao et al., 2024).

A.5 Metric Formulation

Top-k Recall In the differential diagnosis task, let D_{ground_truth} and $D_{predicted}$ represent the actual diagnoses provided by physicians and the list of diseases predicted by the model, respectively. The Top-k Recall is defined as 1 if and only if $D_{ground\_truth} \in D_{predicted}^{[:Top-K]}$. When there are multiple ground truth diagnoses, only the highest-ranked one in the predicted list is considered.

Jaccard coefficient In the medication recommendation task, let M_{ground_truth} and $M_{predicted}$ represent the actual recommendations provided by physicians and the predictions made by the model, respectively. The Jaccard score is computed as follows:

$$Jaccard = \frac{|M_{ground\_truth} \cap M_{predicted}|}{|M_{ground\_truth} \cup M_{predicted}|}.$$

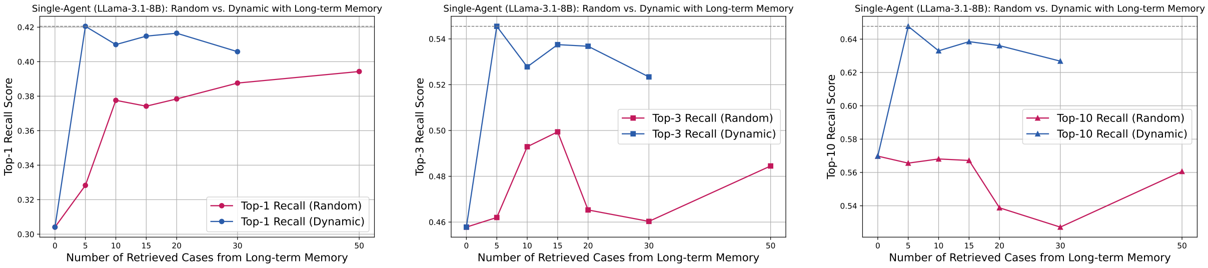
F1-score The recall and precision are formulated as:

$$Recall = \frac{|M_{ground\_truth} \cap M_{predicted}|}{|M_{ground\_truth}|},$$

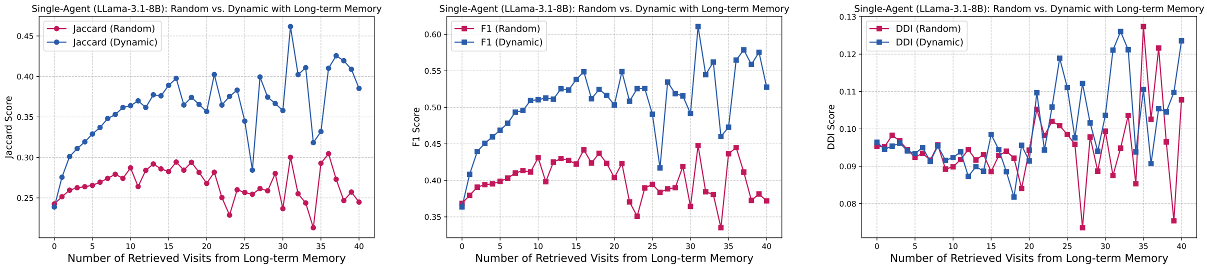
$$Precision = \frac{|M_{ground\_truth} \cap M_{predicted}|}{|M_{predicted}|}.$$

Model	Diagnosis on RAREBENCH-PUBLIC				Treatment on MIMIC-IV-EXT-RARE			
	Hit@1	Hit@3	Hit@10	MR(↓)	Jaccard	F1	DDI(↓)	#MED
Llama-3.1-8B-Instruct								
Single Agent (w/o Memory)	0.3041	0.4578	0.5698	5.0	0.2104	0.3229	0.0951	9.68
Single Agent (w/ Memory)	0.4002	0.5063	0.6558	3.0	0.2458	0.3813	0.0928	11.73
RareAgents	**0.4511**	**0.5647**	**0.7377**	**2.0**	**0.3052**	**0.4475**	**0.0820**	12.98
Llama-3.1-70B-Instruct								
Single Agent (w/o Memory)	0.3751	0.5397	0.6658	3.0	0.2543	0.3736	0.0907	10.97
Single Agent (w/ Memory)	0.4887	0.6057	0.6934	2.0	0.3345	0.4731	0.0891	13.27
RareAgents	**0.5589**	**0.6867**	**0.7811**	**1.0**	**0.4108**	**0.5563**	**0.0796**	13.17

Table 8



(a) Top-1 Recall by memory settings. (b) Top-3 Recall by memory settings. (c) Top-10 Recall by memory settings.



(d) Jaccard Score by memory settings. (e) F1 Score by memory settings. (f) DDI Score by memory settings.

Figure 8: A detailed exploration of each module in *RareAgents*: (i) MDT roles, (ii) Memory, (iii) Tools.

The F1-score is then calculated as:

$$F1 = \frac{2 \cdot Recall \times Precision}{Recall + Precision}.$$

DDI rate Let A_{ddi} denote the adjacency matrix of the DDI-Graph, where $A_{ddi}^{(jk)} = 1$ indicates the presence of an interaction between drugs j and k, and $A_{ddi}^{(jk)} = 0$ indicates no interaction. The DDI rate score is calculated as follows:

$$DDI = \frac{\sum_{j,k \in M_{predicted}} A_{ddi}^{(jk)}}{\sum_{j,k \in M_{predicted}} 1}.$$

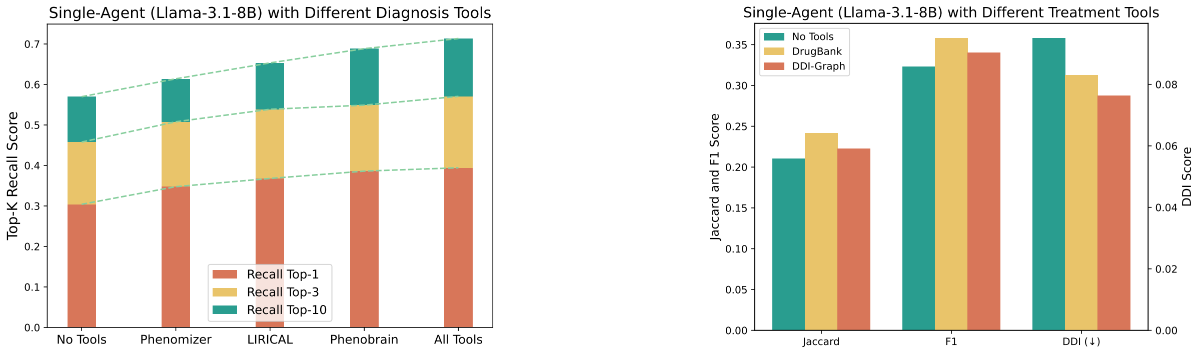
#MED Let N_p denote the total number of patients, $N_v^{(i)}$ represent the number of visits for the i-th patient, and $M_{predicted}^{(i,j)}$ indicate the set of medications recommended by the model for the j-th visit of the i-th patient. Then,

$$\#MED = \frac{\sum_{i=1}^{N_p} \sum_{j=1}^{N_v^{(i)}} |M_{predicted}^{(i,j)}|}{\sum_{i=1}^{N_p} \sum_{j=1}^{N_v^{(i)}} 1}.$$

A.6 Evaluation Details

For the medication recommendation task, since a predefined list of 122 candidate drugs is provided, the large language model's outputs are matched against drug names using regular expressions for precise identification.

For the differential diagnosis task, since no information about candidate diseases is provided, the large language model's outputs may differ in expression (e.g., abbreviations or aliases) from the standard diagnoses provided by physicians. To address this, we use GPT-4o as an external evaluation model to assess whether the standard diagnosis appears within the top 10 predicted diseases. If the standard diagnosis is found, GPT-4o also identifies the corresponding rank. The evaluation prompt is as follows:



(a) Treatment Performance under five Tools settings.

(b) Treatment Performance under three Tools settings.

Figure 9: Detailed Analysis of the Tool Module in *RareAgents*: Utilization of Single Tool Enhances Performance.

Diagnosis Evaluation Prompt

You are a specialist in the field of rare diseases.
I will now give you ten predicted diseases if the predicted diagnosis is in the standard diagnosis. Please output the predicted rank, otherwise output "No", only output "No" or "1-10" numbers, if the predicted disease has multiple conditions, only output the top rank.
Output only "No" or one number, no additional output.
Predicted diseases: `{{predict_diagnosis_list}}`
Standard diagnosis:`{{golden_diagnosis}}`

A.7 Details of Experiment Settings

RareAgents is built on the Llama-3.1 (8B and 70B) (Dubey et al., 2024) and runs via vLLM (with parameters temperature=0 and seed=42) (Kwon et al., 2023) on a system with 4 NVIDIA A100 GPUs, each with 80 GB memory. In the differential diagnosis task, *RareAgents* (Llama-3.1-70B) requires an average time of 36.31 seconds for multidisciplinary consultation per patient. In the medication recommendation task, the average time for multidisciplinary consultation per patient is 72.26 seconds.

A.7.1 Task Prompt

Below are the attending physician agent's system prompt and task-specific prompts. Full prompt configurations can be found in the code repository.

Attending Physician Agent

You are a highly experienced physician. You will be provided with a complex clinical case that may involve atypical presentations or rare conditions. Carefully review the patient's symptoms, history, and any other relevant information.

Differential Diagnosis Prompt

Based on the symptoms of the patient, list the diagnosis separately at the end in the following format:
DIAGNOSIS:
1. <Diagnosis 1>
2. <Diagnosis 2>
3. <Diagnosis 3>
...
10. <Diagnosis 10>

Medication Recommendation Prompt

Based on the diagnosis and procedures provided, please give the most appropriate combination of medications. Select medications only from the given list. List each medication on a separate line using the following format:
TREATMENT:
1. <Medication 1>
2. <Medication 2>
3. <Medication 3>
...
Only include medications from the provided list.

A.7.2 Domain-Specific Baselines

Phen2Disease (Zhai et al., 2023) is a phenotype-driven model that ranks diseases and genes based on bidirectional maximum-matching semantic similarity calculations between patient and disease phenotype sets. In this study, we used the default settings of Phen2Disease, focusing exclusively on the disease ranking results. BASE_IC is a semantic similarity computation method based on Information Content (IC). It evaluates the informativeness of phenotypic terms by assessing their specificity and frequency, aiming to improve phenotype-disease matching and ranking. Descriptions of Phenomizer (Köhler et al., 2009, 2017), LIRICAL(Robinson et al., 2020), and Phenobrain (Mao et al., 2025) are provided in Appendix A.4.

Logistic Regression (LR) is an instance-based classifier with L_2 regularization, where inputs are represented as a multi-hot vector of length $|Diseases| + |Procedures|$.

LEAP (LEArn to Prescribe) (Zhang et al., 2017) models medication recommendation as a sequential decision-making problem, using a recurrent decoder to capture drug dependencies. The hyperparameters in this study are set as follows: learning rate = 2e-4, number of epochs = 30.

RETAIN (Choi et al., 2016) employs a two-level neural attention mechanism to analyze electronic health record (EHR) data, identifying key historical visits and clinical variables that influence predictions. The hyperparameters are set as follows: learning rate = 2e-4, number of training epochs = 40.

G-Bert (Shang et al., 2019a) integrates hierarchical information from medical ontologies with EHR data for pretraining on medical code representation and drug recommendation tasks. The hyperparameters are set as follows: learning rate = 1e-4, dimension of node embeddings = 64, number of pretraining epochs = 100.

GAMENet (Shang et al., 2019b) combines a Graph Convolutional Network (GCN) and a Graph Augmented Memory Module to integrate a Drug-Drug Interaction (DDI) knowledge graph with longitudinal patient EHR data. The hyperparameters are set as follows: learning rate = 2e-4, dimension of node embeddings = 64, number of training epochs = 200.

SafeDrug (Yang et al., 2021b) encodes the molecular structure of drugs using a global Message Passing Neural Network (MPNN) and a local bipartite learning module. The hyperparameters are set as follows: learning rate = 5e-4, dimension of node embeddings = 64, number of training epochs = 200.

COGNet (Wu et al., 2022) is a conditional generative network for drug recommendation that dynamically determines whether to copy drugs from prior recommendations or predict new drugs based on the patient's current diagnosis and historical records. The hyperparameters are set as follows: learning rate = 1e-3, max number of sentences in beam search = 4, dimension of node embeddings = 64, number of training epochs = 100.

MICRON (Yang et al., 2021a) is a recurrent residual learning model for drug change prediction, capturing patient state transitions through efficient residual health representations. The hyperparameters are set as follows: learning rate = 2e-4, dimension of node embeddings = 64, number of training epochs = 40.

MoleRec (Yang et al., 2023) is a molecule substructure-aware model for drug combination recommendation. The hyperparameters are set as follows: learning rate = 5e-4, dropout ratio = 0.7, dimension of node embeddings = 64, number of training epochs = 50.

RAREMed (Zhao et al., 2024) is a drug recommendation model designed for rare disease patients, employing a pretraining-finetuning paradigm. It incorporates self-supervised tasks (sequence alignment prediction and self-reconstruction) and a unified Transformer encoder for input sequences to capture complex relationships between diseases and medication codes. The hyperparameters are set as follows: learning rate = 1e-5, dropout ratio = 0.3, dimension of node embeddings = 512, number of pretraining epochs = 20.

A.8 Detailed Baseline Results of Medication Recommendation

The results of ten domain-specific state-of-the-art models for medication recommendation are presented in Table 5. **RareAgents (Llama-3.1-70B)** achieves superior performance in terms of Jaccard and F1 scores. Since the large language model generates recommended drug names directly, without assigning a candidate probability to each prediction, the PRAUC metric cannot be computed.

A.9 Case Studies

This section provides two concrete examples demonstrating the differences in performance between Single-Agent and *RareAgents*, both based on the Llama-3.1-70B model.

A.9.1 Case 1 (Differential Diagnosis)

The first example involves a patient diagnosed with **Brugada Syndrome**. The Single-Agent (Llama-3.1-70B) performs a step-by-step analysis of the patient's symptoms, systematically categorizing them. However, despite its detailed analysis, it fails to include the correct diagnosis within its top 10 predicted diseases. In contrast, *RareAgents* (Llama-3.1-70B) leverages a multidisciplinary team discussion approach, dynamically retrieving similar patient cases from long-term memory and diagnostic tools' feedback. As a result, it successfully identifies the correct diagnosis as the top prediction.

A.9.2 Case 2 (Medication Recommendation)

The second example focuses on a patient requiring medication recommendations from the MIMIC-IV-Ext-Rare dataset. The Single-Agent (Llama-3.1-70B) produces a list of 21 recommended drugs, of which only 8 align with the actual prescriptions made by physicians, missing 13 critical medications. In contrast, *RareAgents* (Llama-3.1-70B), by incorporating multidisciplinary team discussions and dynamically retrieving prior patient visit records and treatment tools' feedback, achieves a significantly better result. It recommends 17 correct medications, with only 4 missing from the final recommendations.

Case 1

Patient Agent: I am experiencing the following symptoms: Urinary hesitancy, Urinary incontinence, Prostatitis, Syncope, Muscle weakness, Elevated circulating creatine kinase concentration, Exercise intolerance, Loss of consciousness, Pedal edema, ST segment elevation, Increased circulating creatine kinase MB isoform, Dysuria, Amaurosis fugax. I would like to request you to diagnose the cause of my illness.

Single-Agent (Llama-3.1-70B): Based on the symptoms provided, here's a detailed analysis and the top 10 most likely diagnoses:

The patient's symptoms can be broadly categorized into several groups: urinary symptoms (urinary hesitancy, urinary incontinence, prostatitis, dysuria), cardiovascular symptoms (syncope, loss of consciousness, ST segment elevation, increased circulating creatine kinase MB isoform), muscular symptoms (muscle weakness, elevated circulating creatine kinase concentration, exercise intolerance), and neurological symptoms (amaurosis fugax). Pedal edema could be related to either cardiovascular or renal issues.

Given the complexity and the combination of these symptoms, it's essential to consider conditions that could potentially explain multiple symptom groups simultaneously. Here's a step-by-step breakdown:

1. **Cardiovascular Symptoms**: The presence of ST segment elevation and increased circulating creatine kinase MB isoform strongly suggests myocardial infarction (heart attack) or acute coronary syndrome. However, the combination with other symptoms like muscle weakness and elevated creatine kinase concentration points towards conditions affecting both the heart and skeletal muscles.

2. **Muscular Symptoms**: Elevated creatine kinase levels and muscle weakness could indicate a myopathy. When combined with cardiovascular symptoms, conditions like myocarditis or cardiomyopathy with skeletal muscle involvement become plausible.

3. **Urinary Symptoms**: These could be indicative of a urinary tract infection, benign prostatic hyperplasia (BPH), or other conditions affecting the prostate or urinary system. However, in the context of the other symptoms, a systemic condition affecting multiple body systems is more likely.

4. **Neurological Symptoms**: Amaurosis fugax (temporary loss of vision in one eye) could be indicative of a transient ischemic attack (TIA) or a vascular issue affecting the eye.

Considering the combination of these symptoms, here are the top 10 most likely diagnoses:

DIAGNOSIS:

☒ 1. **Myocardial Infarction with Cardiac Sarcoidosis**: A condition where granulomatous inflammation affects the heart, potentially leading to heart failure, arrhythmias, and symptoms overlapping with those described.

☒ 2. **Duchenne Muscular Dystrophy with Cardiac Involvement**: Although typically diagnosed in childhood, some cases may present later, and cardiac involvement is common, which could explain the combination of muscle weakness, elevated CK, and cardiac symptoms.

☒ 3. **Myocarditis with Skeletal Muscle Involvement**: Inflammation of the heart muscle that could also affect skeletal muscles, explaining the elevated CK levels and muscle weakness.

☒ 4. **Cardiac Amyloidosis**: A condition where abnormal proteins deposit in the heart, potentially leading to heart failure and symptoms that could overlap with those described, including neuropathy and muscle weakness.

☒ 5. **Ehlers-Danlos Syndrome with Cardiac Involvement**: A genetic disorder affecting connective tissue, which could explain the diverse range of symptoms including cardiac issues, muscle weakness, and potentially the urinary symptoms.

☒ 6. **Mitochondrial Myopathy, Encephalopathy, Lactic Acidosis, and Stroke-like Episodes (MELAS) Syndrome**: A mitochondrial disorder that affects many areas of the body, including the heart, muscles, and nervous system.

☒ 7. **Fabry Disease**: A genetic disorder that results from the buildup of a particular type of fat in the body's cells, affecting many areas including the heart, kidneys, and nervous system.

☒ 8. **Acute Coronary Syndrome with Rhabdomyolysis**: A condition where a heart attack is complicated by muscle breakdown, leading to elevated CK levels and muscle weakness.

☒ 9. **Cardiac Sarcoidosis with Neurosarcoidosis**: Sarcoidosis affecting both the heart and nervous system, potentially explaining the amaurosis fugax and other neurological symptoms.

☒ 10. **Danon Disease**: A rare genetic disorder characterized by cardiomyopathy, skeletal myopathy, and other systemic symptoms.

Each of these conditions is rare and complex, and a definitive diagnosis would require extensive diagnostic testing, including genetic analysis, cardiac evaluation, muscle biopsy, and potentially other specialized tests.

Case 1

RareAgents (Llama-3.1-70B):

Attending Physician Agent: Based on the patient's symptoms, I would like to request the following specialists to form a multi-disciplinary team: Urology, Cardiology, Neurology, Pathology, Nuclear Medicine, Radiology, Interventional Radiology, Ophthalmology, Rehabilitation Medicine, Pharmacy, Vascular Surgery, Ultrasound Medicine, Anesthesiology, Cardiac Surgery, Laboratory Medicine.

`{{Multiple rounds of discussions by multidisciplinary team (MDT)}}`

...

Cardiology Physician Agent: Based on the patient's symptoms, medical history, and test results, I would recommend the following diagnosis and treatment plan: 1. **Acute Myocardial Infarction (AMI) with Rhabdomyolysis**: The patient's symptoms of chest pain, shortness of breath, and elevated troponin levels are consistent with an acute myocardial infarction. The presence of rhabdomyolysis, as indicated by elevated creatine kinase levels, suggests muscle damage. 2. **Myasthenia Gravis with Cardiac Involvement**: The patient's symptoms of muscle weakness, fatigue, and difficulty swallowing, along with the presence of anti-acetylcholine receptor antibodies, suggest myasthenia gravis. The cardiac involvement is indicated by the patient's symptoms of palpitations and shortness of breath....

Neurology Physician Agent: Based on the diagnostic tools' results, the most likely diagnosis is Brugada Syndrome 4, which is a genetic disorder that affects the heart's electrical system and can cause abnormal heart rhythms, including ventricular fibrillation. The patient's symptoms, such as syncope, muscle weakness, and elevated circulating creatine kinase concentration, are consistent with this diagnosis. However, it's important to note that the patient's symptoms also suggest other possible diagnoses, such as multiple sclerosis, muscular dystrophy, and myasthenia gravis, which should be considered in the differential diagnosis. Therefore, a comprehensive diagnostic workup, including genetic testing, electromyography, and imaging studies, is necessary to confirm the diagnosis and rule out other possible conditions.

Pathology Physician Agent: Based on the patient's symptoms, I would recommend the following diagnosis: 1. **Acute Myocardial Infarction (AMI) with Rhabdomyolysis**: The patient's symptoms of ST segment elevation, elevated circulating creatine kinase concentration, and increased circulating creatine kinase MB isoform suggest an acute myocardial infarction. The presence of muscle weakness, exercise intolerance, and elevated creatine kinase levels also suggest rhabdomyolysis. 2. **Myasthenia Gravis with Cardiac Involvement**: The patient's symptoms of muscle weakness, exercise intolerance, and syncope could be indicative of myasthenia gravis, an autoimmune disorder that affects the neuromuscular junction. The presence of cardiac involvement is suggested by the ST segment elevation and elevated creatine kinase levels....

Urology Physician Agent: Based on the diagnostic tools' results, the most likely diagnosis is Brugada Syndrome 4, which is a genetic disorder that affects the heart's electrical system and can cause abnormal heart rhythms, including ventricular fibrillation. The patient's symptoms, such as syncope, muscle weakness, and ST segment elevation, are consistent with this diagnosis. However, it is essential to note that the diagnostic tools' results are not definitive, and a comprehensive evaluation by a healthcare professional is necessary to confirm the diagnosis. Additional testing, such as genetic testing, imaging studies, and electrophysiological studies, may be necessary to determine the underlying cause of the patient's symptoms. In conclusion, based on the diagnostic tools' results, Brugada Syndrome 4 is the most likely diagnosis, but further evaluation and testing are necessary to confirm the diagnosis and rule out other potential causes of the patient's symptoms.

...

Case 1

RareAgents (Llama-3.1-70B):

`{{Dynamic retrieval of patients with similar symptoms from long-term memory}}`

Memory Retrieve Results: Consider these previous cases for reference:

1. Patient with symptoms: Trismus, Hyperhidrosis, Syncope, Vomiting, Loss of consciousness, Right bundle branch block, ST segment elevation was diagnosed with Brugada syndrome.
2. Patient with symptoms: Urinary urgency, Urinary hesitancy, Urinary incontinence, Syncope, Stroke, Slurred speech, Constipation, Cerebral atrophy, Dysdiadochokinesis, Abnormal spinal cord morphology, Positive Romberg sign, Abnormal cerebral white matter morphology, Increased intracranial pressure, Cerebral ischemia, Increased CSF protein concentration, EMG: neuropathic changes, Babinski sign, Difficulty climbing stairs, Loss of consciousness, Lower limb muscle weakness, Abnormal prostate morphology, Sleep apnea, CSF pleocytosis, Dysesthesia, Abasia, Abnormal male reproductive system physiology, Schwannoma, Pollakisuria, Dysuria, Neoplasm of the lung, Erectile dysfunction, Abnormal lumbar spine morphology was diagnosed with Multiple system atrophy.
3. Patient with symptoms: Syncope, Palpitations, ST segment elevation was diagnosed with Brugada Syndrome.
4. Patient with symptoms: Syncope, Cardiomyopathy, Ventricular fibrillation, Cardiac arrest, Right bundle branch block, Myocarditis was diagnosed with Cateeholaminergic polymorphic ventricular tachycardia.
5. Patient with symptoms: Ventricular fibrillation, Pneumonia, Dyspnea, Exercise intolerance, Viral hepatitis, ST segment elevation, Cough was diagnosed with Brugada syndrome.

`{{Feedback after the function calls the diagnostic tool}}`

Phenomizer: 1. 109820 BLADDER DIVERTICULUM, p-value: 0.0160. 2. MULTIPLE SCLEROSIS, SUSCEPTIBILITY TO, p-value: 0.0374. 3. MUSCLE GLYCOGENOSIS, X-LINKED, p-value: 0.0374. 4. #125310 CEREBRAL ARTERIOPATHY, AUTOSOMAL DOMINANT, WITH SUBCORTICAL INFARCTSAND LEUKOENCEPHALOPATHY; CADASIL;DEMENTIA, HEREDITARY MULTI-INFARCT TYPE;CASIL, p-value: 0.0881. 5. #616231 MYOPATHY, VACUOLAR, WITH CASQ1 AGGREGATES; VMCQA, p-value: 0.0881. 6. #616094 MUSCULAR DYSTROPHY-DYSTROGLYCANOPATHY (LIMB-GIRDLE), TYPE C, 12; MDDGC12;MUSCULAR DYSTROPHY-DYSTROGLYCANOPATHY, LIMB-GIRDLE, POMK-RELATED, p-value: 0.0881. 7. #611876 BRUGADA SYNDROME 4; BRGDA4, p-value: 0.1207. 8. #609620 SHORT QT SYNDROME 1; SQT1, p-value: 0.1207. 9. #612347 JERVELL AND LANGE-NIELSEN SYNDROME 2; JLNS2, p-value: 0.1207. 10. MUSCULAR DYSTROPHY, CARDIAC TYPE, p-value: 0.1207

LIRCAL: 1. Glycogen storage disease II, posterior probability: 72.30 %. 2. Danon disease, posterior probability: 0.00 %. 3. Polyglucosan body myopathy 1 with or without immunodeficiency, posterior probability: 0.00 %. 4. Myopathy, myofibrillar, 7, posterior probability: 0.00 %. 5. Muscular dystrophy, limb-girdle, autosomal recessive 25, posterior probability: 0.00 %. 6. Neuronal intranuclear inclusion disease, posterior probability: 0.00 %. 7. Progressive external ophthalmoplegia with mitochondrial DNA deletions, autosomal recessive 5, posterior probability: 0.00 %. 8. Progressive external ophthalmoplegia with mitochondrial DNA deletions, autosomal dominant 4, posterior probability: 0.00 %. 9. Lipodystrophy, congenital generalized, type 4, posterior probability: 0.00 %. 10. Mitochondrial DNA depletion syndrome 11, posterior probability: 0.00 %.

Phenobrain: 1. Cerebralautosomal dominant angiopathy with subcortical infarcts and leukoencephalopathy, score: 0.9998560115190785. 2. Malakoplakia, score: 0.9992800575953923. 3. Medium chain acyl-CoA dehydrogenase deficiency, score: 0.9992080633549316. 4. Eisenmenger syndrome, score: 0.9991360691144708. 5. Pure autonomic failure, score: 0.99906407487401. 6. Porphyria/Porphyria, score: 0.9987041036717063. 7. Adrenoleukodystrophy, score: 0.9985601151907847. 8. Multiple sclerosis, susceptibility to, score: 0.9983441324694025. 9. Acute intermittent porphyria/Porphyria, acute intermittent, score: 0.9983081353491721.

Case 1

RareAgents (Llama-3.1-70B):

`{{Comprehensive decision-making based on MDT discussions, memory retrieval results, and tool feedback}}`

Attending Physician Agent: Considering the above information, these are the top 10 diagnoses:

☑ 1. Brugada Syndrome.
2. Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 3
3. Multiple system atrophy; MSA/Multiple system atrophy/Multiple system atrophy 1, susceptibility to
4. Catecholaminergic polymorphic ventricular tachycardia, CPVT/Catecholaminergic polymorphic ventricular tachycardia/Ventricular tachycardia, catecholaminergic polymorphic, 1, with or without atrial dysfunction and/or dilated cardiomyopathy
5. NKX6-2-related autosomal recessive hypomyelinating leukodystrophy/Spastic ataxia 8, autosomal recessive, with hypomyelinating leukodystrophy
6. Muscular dystrophy, limb-girdle, autosomal recessive 25
7. Glycogen storage disease II
8. Danon disease
9. Polyglucosan body myopathy 1 with or without immunodeficiency
10. Myopathy, myofibrillar, 7

Case 2

Patient Agent: I was diagnosed the following diseases: Encounter for antineoplastic chemotherapy; Primary central nervous system lymphoma, unspecified site, extranodal and solid organ sites; Other specified disorders of metabolism; Diabetes insipidus; Nonspecific reaction to tuberculin skin test without active tuberculosis; Epilepsy, unspecified, without mention of intractable epilepsy; Dysthymic disorder; Arthropathy, unspecified, site unspecified; Sleep disturbance, unspecified; Personal history of other infectious and parasitic diseases; Long-term (current) use of steroids, and I have received the following procedures: Injection or infusion of cancer chemotherapeutic substance. I would like to request you to provide the most appropriate combination of medications for me.

`{{122 medication candidates}}`

Single-Agent (Llama-3.1-70B): TREATMENT:

☒ 1. *NF* Epirubicin
☒ 2. Mesna
☒ 3. Levothyroxine Sodium
☑ 4. Desmopressin Nasal
☒ 5. Phenoxybenzamine HCl
☒ 6. Metoprolol Tartrate
☑ 7. Lorazepam
☑ 8. LeVETiracetam
☒ 9. phenobarbital
☑ 10. Sertraline
☒ 11. prednisolone
☒ 12. fludrocortisone
☑ 13. Potassium Chloride
☒ 14. Potassium Acetate
☑ 15. Sodium Bicarbonate
☑ 16. Ondansetron
☑ 17. Prochlorperazine
☒ 18. Dantrolene Sodium
☒ 19. Melphalan
☒ 20. interferon alfa-2b 6 million unit/mL
☒ 21. Aldesleukin

`{{Absence of ground truth medication: Ranitidine, 0.9% Sodium Chloride, 5% Dextrosecandidates, Propranolol, Heparin, Lisinopril, DiphenhydrAMINE, Calcium Carbonate, traZODONE, OxycoDONE (Immediate Release), PredniSONE, Methotrexate, Isoniazid}}`

Case 2

RareAgents (Llama-3.1-70B):

Attending Physician Agent: Based on the patient's symptoms, I would like to request the following specialists to form a multi-disciplinary team: Nuclear Medicine, Pathology, Nephrology, Urology, Neurology, Oncology, Hematology, Radiology, Interventional Radiology, Neurosurgery, Infectious Diseases, Rheumatology, Rehabilitation Medicine, Psychiatry, Clinical Nutrition, Pharmacy, Orthopedic Surgery, Hematology, Dentistry, Anesthesiology, Endocrinology, Laboratory Medicine, Traditional Chinese Medicine, Allergy and Immunology.

`{{Multiple rounds of discussions by multidisciplinary team (MDT)}}`

...

Nuclear Medicine Physician Agent: Based on the patient's diagnosis and medical history, I will provide my expert opinion on the treatment. As the patient has been diagnosed with primary central nervous system lymphoma, methotrexate is a common chemotherapeutic agent used to treat this condition. As the patient has epilepsy, LeVETiracetam is an antiepileptic medication that can help control seizures. As the patient has diabetes insipidus, desmopressin nasal spray can help regulate fluid balance and reduce urine production. As the patient has dysthymic disorder, sertraline is an antidepressant medication that can help manage symptoms of depression....

Pathology Physician Agent: Based on the patient's diagnosis and medical history, the following medications are recommended: 1. **Methotrexate**: As the patient is undergoing antineoplastic chemotherapy, methotrexate is a common chemotherapeutic agent used to treat various types of cancer, including lymphoma. 2. **LeVETiracetam**: As the patient has epilepsy, LeVETiracetam is an antiepileptic medication that can help control seizures. 3. **Desmopressin Nasal**: As the patient has diabetes insipidus, desmopressin nasal spray can help regulate water balance in the body. 4. **Sertraline**: As the patient has dysthymic disorder, sertraline is an antidepressant medication that can help manage symptoms of depression....

Nephrology Physician Agent: **Treatment Plan:** **Chemotherapy:** Continue with the current antineoplastic chemotherapy regimen as prescribed by the oncologist. **Diabetes Insipidus:** Desmopressin Nasal Spray: 1-2 sprays (10-20 mcg) intranasally at bedtime, as needed, to control nocturia and polyuria. **Epilepsy:** LeVETiracetam (Keppra): 500-1000 mg orally twice daily, as needed, to control seizures. **Dysthymic Disorder:** Sertraline (Zoloft): 50-100 mg orally once daily, as needed, to manage depressive symptoms. **Pain Management:** Acetaminophen (Tylenol): 650-1000 mg orally every 4-6 hours, as needed, for pain relief. **Steroid-Induced Side Effects:** Monitor for signs of steroid-induced side effects, such as hyperglycemia, hypertension, and osteoporosis....

Urology Physician Agent: Based on the patient's diagnosis and medical history, I will provide a treatment plan that addresses the various conditions and medications. **Primary central nervous system lymphoma, unspecified site, extranodal and solid organ sites:** The patient is already receiving antineoplastic chemotherapy, which is the standard treatment for primary central nervous system lymphoma. I recommend continuing the chemotherapy regimen as prescribed. **Diabetes insipidus:** The patient is likely experiencing symptoms of diabetes insipidus due to the chemotherapy treatment. I recommend Desmopressin Nasal to manage the symptoms of diabetes insipidus. **Other specified disorders of metabolism:** The patient's metabolic disorders may be related to the chemotherapy treatment or underlying conditions. I recommend monitoring the patient's metabolic parameters and adjusting the treatment plan as needed. **Nonspecific reaction to tuberculin skin test without active tuberculosis:** The patient's reaction to the tuberculin skin test indicates a possible exposure to tuberculosis. I recommend monitoring the patient for signs of tuberculosis and considering prophylactic treatment if necessary....

...

Case 2

RareAgents (Llama-3.1-70B):

`{{Dynamic retrieval of patient's former visits from long-term memory}}`

Memory Retrieve Results: Consider these previous visits for reference:

First admission: Patient with a diagnosis of Encounter for antineoplastic chemotherapy; Diabetes insipidus; Diffuse large B-cell lymphoma, extranodal and solid organ sites; Calculus of kidney; Personal history of nicotine dependence; Anxiety disorder, unspecified; Insomnia, unspecified; Essential (primary) hypertension; Presence of artificial hip joint, bilateral and a history of procedures including Introduction of Other Antineoplastic into Central Vein, Percutaneous Approach has a prior medication record of: ['Magnesium Sulfate', '5% Dextrose', 'Sodium Bicarbonate', 'LamoTRIgine', 'Methotrexate', 'Heparin'].

Second admission: Patient with a diagnosis of Primary central nervous system lymphoma, unspecified site, extranodal and solid organ sites; Cerebral edema; Diabetes insipidus; Other convulsions; Unspecified essential hypertension; Hip joint replacement; Personal history of tobacco use; Personal history of tuberculosis and a history of procedures including Injection or infusion of cancer chemotherapeutic substance has a prior medication record of: ['Sodium Bicarbonate', '0.9% Sodium Chloride', 'Diazepam', 'LeVETiracetam', 'Heparin', '5% Dextrose', 'Dexamethasone', 'TraZODone', 'Acetaminophen', 'Methotrexate', 'Propranolol', 'Calcium Carbonate', 'Senna', 'Sertraline'].

Third admission: Patient with a diagnosis of Encounter for antineoplastic chemotherapy; Other malignant lymphomas, unspecified site, extranodal and solid organ sites; Diabetes insipidus; Unspecified essential hypertension; Hip joint replacement; Epilepsy, unspecified, without mention of intractable epilepsy; Anxiety state, unspecified; Sleep disturbance, unspecified and a history of procedures including Injection or infusion of cancer chemotherapeutic substance has a prior medication record of: ['traZODONE', 'Famotidine', 'Prochlorperazine', 'Dexamethasone', '0.9% Sodium Chloride', 'Sodium Bicarbonate', 'Clonazepam', 'Senna', 'Ondansetron', '5% Dextrose', 'Docusate Sodium', 'LeVETiracetam', 'Acetaminophen', 'Desmopressin Nasal', 'Lisinopril', 'Methotrexate'].

...

`{{Feedback after the function calls the therapeutic tool}}`

DrugBank: More detailed drug information is provided below:

Ondansetron: A competitive serotonin type 3 receptor antagonist. It is effective in the treatment of nausea and vomiting caused by cytotoxic chemotherapy drugs, including cisplatin, and has reported anxiolytic and neuroleptic properties.

Tolvaptan: Tolvaptan is used to treat low blood sodium levels (hyponatremia) associated with various conditions like congestive heart failure, cirrhosis, and syndrome of inappropriate antidiuretic hormones.

Ranitidine: A non-imidazole blocker of those histamine receptors that mediate gastric secretion (H2 receptors). It is used to treat gastrointestinal ulcers.

Penicillamine: Penicillamine is a pharmaceutical of the chelator class. The pharmaceutical form is D-penicillamine, as L-penicillamine is toxic (it inhibits the action of pyridoxine). It is an α-amino acid metabolite of penicillin, although it has no antibiotic properties.

Pyridoxine: Pyridoxine is the 4-methanol form of vitamin B6, an important water-soluble vitamin that is naturally present in many foods. As its classification as a vitamin implies, Vitamin B6 (and pyridoxine) are essential nutrients required for normal functioning of many biological systems within the body. While many plants and microorganisms are able to synthesize pyridoxine through endogenous biological processes, animals must obtain it through their diet.

...

DDI-graph: The following are potential drug interactions:

Prochlorperazine and Potassium Chloride may cause Myringitis.

Sertraline and Prochlorperazine may cause meibomianitis.

Mesna and Metoprolol Tartrate may cause corticosteroid therapy.

Metoprolol Tartrate and Potassium Chloride may cause corticosteroid therapy.

Sertraline and fludrocortisone may cause meibomianitis.

...

Case 2

RareAgents (Llama-3.1-70B):

`{{Comprehensive decision-making based on MDT discussions, memory retrieval results, and tool feedback}}`

Attending Physician Agent: Considering the above information, I recommend the following combination of medications:

☑ 1. Ondansetron
☑ 2. Methotrexate
☑ 3. LeVETiracetam
☑ 4. Sertraline
☑ 5. Propranolol
☑ 6. Sodium Bicarbonate
☑ 7. 5% Dextrose
☑ 8. 0.9% Sodium Chloride
☑ 9. Heparin
☒ 10. Acetaminophen
☑ 11. OxycoDONE (Immediate Release)
☑ 12. traZODONE
☑ 13. Desmopressin Nasal
☑ 14. Lisinopril
☒ 15. Dexamethasone
☒ 16. Clonazepam
☒ 17. Furosemide
☑ 18. Potassium Chloride Replacement (Oncology)
☑ 19. Calcium Carbonate
☑ 20. Lorazepam
☑ 21. PredniSONE

`{{Absence of ground truth medication: Ranitidine, Prochlorperazine, DiphenhydrAMINE, Isoniazid}}`