RAREAGENTS: Autonomous Multi-disciplinary Team for Rare Disease Diagnosis and Treatment

Xuanzhong Chen[1*], Ye Jin[2*], Xiaohao Mao[1], Lun Wang[2], Shuyang Zhang[2†], Ting Chen[1†]

[1]Tsinghua University
[2]Peking Union Medical College Hospital
cxz23@mails.tsinghua.edu.cn, tingchen@tsinghua.edu.cn

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 decision-support tool 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 diagnosis and treatment for rare diseases. Furthermore, we contribute a novel rare disease dataset, MIMIC-IV-EXT-RARE, to facilitate further research in this field.

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, Ouyang, and Bolen 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 (Shang et al. 2019b; Yang et al. 2021b), their performance for rare diseases remains suboptimal. Experimental studies on the MIMIC-III and MIMIC-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).

Large language models, trained on massive and diverse text corpora, demonstrate 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. 2024) 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 symptoms affecting multiple organs and systems, indicating the critical need for expertise from multiple related specialties to achieve accurate diagnosis and personalized treatment plans (Xie et al. 2023). In clinical practice, this integrated approach is known as multi-disciplinary team (MDT) care, with the central objective of synthesizing their diagnostic insights and therapeutic proposals from diverse experts to formulate a comprehensive management strategy that prioritizes treatment steps and resolves potential conflicts among recommendations.

Although several multi-agent frameworks have been proposed for general medical applications (as summarized in Table 1), these methods primarily show 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). For these tasks, candidate options are provided or the decision-making is confined to limited and small scopes (Li et al. 2024b), but these settings differ significantly from the complex real-world clinical scenarios. Moreover, existing approaches tend to emphasize planning capabilities while focusing less on the integration, which can be achieved through memory usage and tool utilization. Additionally, the definition of dif-

*These authors contributed equally.
†Corresponding authors.

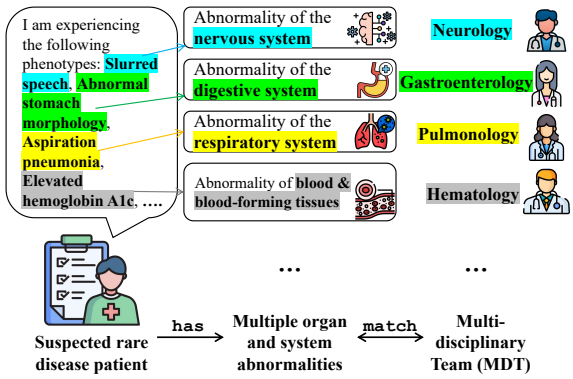

Figure 1: A rare disease patient with multi-organ / multi-system abnormalities necessitates a multi-disciplinary team for comprehensive diagnosis and treatment.

ferent agent roles is frequently left to LLMs themselves, leading to potential hallucinations in medical contexts (Lee, Bubeck, and Petro 2023).

To address these challenges, we propose **RareAgents**, a patient-centered autonomous MDT framework customized for real-world rare disease patients, taking advantage of 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. This dataset contains 4,760 rare disease patients with 18,522 admission records, providing a valuable resource for the rare disease research community.

Method	Planning	Memory	Tool Use	Multi-Agent Roles	Medical Scenario
MedAgents (Tang et al. 2024)	✓	✗	✗	LLM-generated	MCQA
Agent Hospital (Li et al. 2024b)	✓	✓	✗	Pre-defined	MCQA
MDAgents (Kim et al. 2024)	✓	✗	✗	LLM-generated	MCQA / VQA
AgentMD (Jin et al. 2024)	✓	✗	✓	Single-Agent	MCQA
RareAgents (ours)	✓	✓	✓	Pre-defined	Open-ended Complex QA

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

Related Work

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), memory (Zhong et al. 2024), and tool-using (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, McDonell, and Reynolds 2023) is a widely adopted approach that assigns agents distinct personalities or roles, allowing them to adapt to specific task scenarios. LLM-based agents have shown significant potential in applications across domains such as education (Zhang et al. 2024), finance, and healthcare (Mehandru et al. 2024).

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 adjust 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).

AI Models for Rare Diseases Most AI diagnostic models for rare diseases primarily rely on phenotypic and geno-

RAREBENCH-PUBLIC (Chen et al. 2024)		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; processing details are in the supplement.

typic information (Javed, Agrawal, and Ng 2014; Robinson et al. 2020), utilizing statistical and machine learning approaches (Yang, Robinson, and Wang 2015; Köhler et al. 2017; Zhai et al. 2023). RareBERT (Prakash et al. 2021) introduces a Transformer-based model to identify rare disease patients. In the realm of LLMs, dynamic few-shot prompting methods (Chen et al. 2024) have been explored to enhance diagnostic performance. RAREMed (Zhao et al. 2024) focuses on 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).

Problem Formulation and Datasets

Definition of Rare Disease Tasks

As indicated in Table 1, 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 predict 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**): $\mathcal{A}_{treatment} = \{m_1, m_2, \ldots, m_l\} \subset \mathcal{M}$.

Datasets

This research uses two publicly available datasets, RareBench (Chen et al. 2024) 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 RareBench is a multi-center dataset comprising rare disease patient data from Europe, China, and Canada. It is specifically designed to evaluate LLMs' performance in the rare disease domain (Chen et al. 2024). We utilize 1,197 publicly available 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 curated for medication recommendation tasks in rare disease contexts.

Overview of RAREAGENTS

This section introduces the proposed **RareAgents** framework for rare disease diagnosis and treatment. Figure 2 provides the overview of the framework, respectively. Given space limitations, we provide detailed algorithm pseudocode, case studies and extensive comparisons in the supplementary material. The *RareAgents* framework is composed of three core modules: **(1) Multi-disciplinary Team (MDT) Collaboration:** The attending physician agent selects the most relevant specialists from a predefined special-

[1]https://omim.org/

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

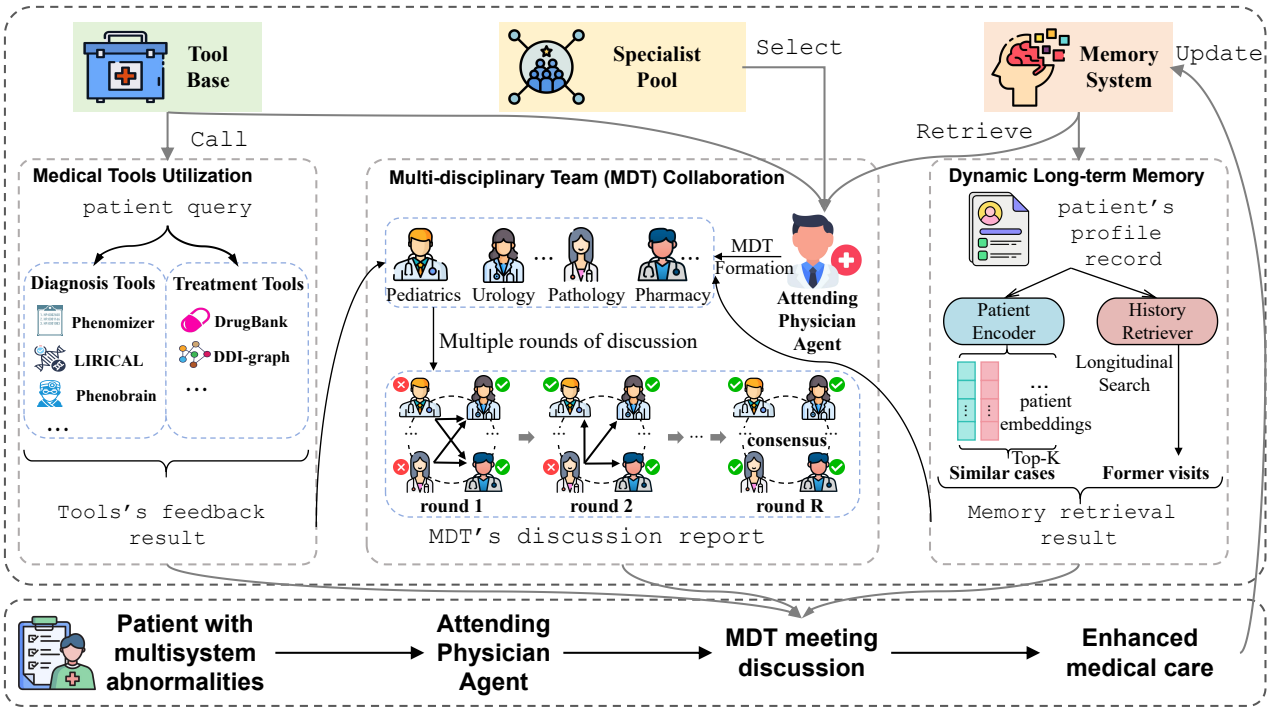

Figure 2: Overview of the **RareAgents**: For patients with multiple organ / multiple system abnormalities, the Attending Physician Agent selects specialists from a predefined pool to form an MDT, which reaches consensus through iterative discussions. Each physician agent is equipped with: a) Dynamic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

ist 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 online rare disease diagnostic programs and treatment tools to enhance their decision-making capabilities.

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 the supplementary material. The entire MDT consultation process is divided into three stages: **(i) MDT Formation** : The attending physician agent assembles a patient-centric MDT. **(ii) Expert Consensus**: 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**: The attending physician agent synthesizes the opinions from all MDT members to generate a final discussion report $\mathcal{DR}$, where

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

Dynamic Long-term Memory

In real-world clinical practice, physicians rely on both personal experience and historical patient 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. 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. 2024) 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

Model	Diagnosis on RAREBENCH				Treatment on MIMIC-IV-RARE			
	Hit@1	Hit@3	Hit@10	MR↓	Jaccard	F1	DDI↓	#MED
Domain-specific SOTA (Diagnosis / Treatment)								
Phenomizer / LEAP	0.0844	0.2072	0.3835	>10	0.2959	0.4341	**0.0485**	5.92
LIRICAL / G-Bert	0.1637	0.2840	0.4152	>10	0.4030	0.5554	0.0751	14.61
BASE_IC / SafeDrug	0.2047	0.3434	0.5322	8.0	0.3903	0.5426	0.0733	12.88
Phen2Disease / MoleRec	0.2105	0.3266	0.5129	10.0	0.3975	0.5498	0.0714	12.15
Phenobrain / RAREMed	0.2857	0.4670	0.6341	4.0	0.3800	0.5268	0.0622	8.75
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)								
DS-R1-Distill-Llama-70B	0.3509	0.5221	0.6291	3.0	0.2924	0.4267	0.0901	11.87
DS-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: Integrated benchmarking results. The first block compares domain-specific baselines, where models are listed as "*Diagnosis Model / Treatment Model*" corresponding to the left and right metrics respectively. **Bold** indicates the best performance, and underlined indicates the second best across the entire table.

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

$$\begin{aligned} \mathcal{MR}_{diagnosis}\left(\mathcal{R}\right) &= \arg\max_{\text{Top-K}}\left(Emb(\mathcal{R})\right), \\ \mathcal{MR}_{treatment}(\mathcal{R}^{(n)}) &= \mathcal{R}^{(1:n-1)} \cup \mathcal{A}^{(1:n-1)}_{treatment}. \end{aligned} \tag{2}$$

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. 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. 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) and DDI-graph (for drug-drug interaction relationships). Detailed tool functions are provided in the supplementary material. 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

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

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

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

Experimental Setup and Main Results

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

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 4: Ablation study results for the impact of each module in *RareAgents*.

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 of metrics are provided in the supplementary material.

Baselines

Domain-specific SOTA models For the differential diagnosis task, the domain-specific SOTA models include Phenomizer (Köhler et al. 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, 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. Supplementary material provides additional details on these baselines and their configurations.

General and Medical LLMs General LLMs include GPT-4o and GPT-3.5-turbo-0125. The medical LLMs include OpenBioLLM and UltraMedical, both fine-tuned on medical datasets using Llama-3 (8B and 70B). O1-like LLMs include DeepSeek-R1-Distill-Llama (8B and 70B), Baichuan-M1-14B, and HuatuoGPT-o1-70B. All of these models are evaluated 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.

Main Results

Table 3 presents the performance of all models on RareBench-Public for differential diagnosis and MIMIC-IV-Ext-Rare for medication recommendation.

Differential Diagnosis RareAgents (Llama-3.1-70B) outperform all baselines across all evaluation metrics. Even though RareAgents (Llama-3.1-8B) ranks second in some metrics such as Top-1 Recall (Hit@1), it demonstrates significant improvements over other medical agent frameworks. Interestingly, LLMs' performance already surpasses that of domain-specific SOTA models. Among the fine-tuned LLMs, 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 *RareAgents* (Llama-3-70B) achieves the best performance across all metrics except for DDI. The dataset's inherent DDI and average number of drugs 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.

Analysis and Discussion

Ablation Study

RareAgents consists of three key components: Multi-disciplinary Team (MDT) collaboration, dynamic long-term

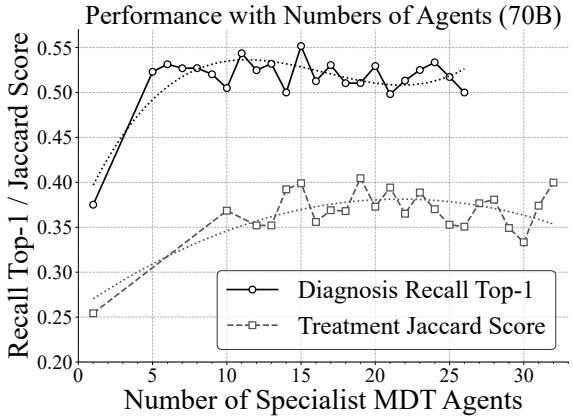

Figure 3: Agents Number in MDT.

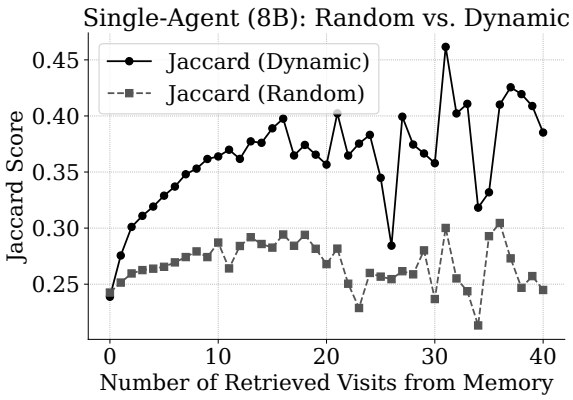

Figure 5: Jaccard by memory settings.

memory, and medical tools utilization. To quantify the contribution of each module, we conduct ablation experiments by removing one component at a time, with results shown in Table 4. 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 LLM distinguish rare conditions from more common ones, thus avoiding the pitfalls of a cold start in reasoning. In the medication recommendation task, the removal of the tools module results in a significant increase in DDI rate. This is because drug knowledge bases deliver specialized pharmacological insights, which effectively reduce DDI and enhance prescription safety.

Advanced Nature of MDT in RareAgents

To further evaluate the efficacy of MDT within *RareAgents*, we conduct experiments isolating 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. On average, *RareAgents* (Llama-3.1-70B) engage 12.53 specialists for differential diagnosis and 22.22 for medication recommendation. As shown in Figure 3, MDT performance peaks around these agent numbers.

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. Figure 4 demonstrates 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.

To evaluate the effectiveness of the dynamic retrieval mechanism in long-term memory, we compare it with a baseline that randomly selects the same number of cases. Using the medication recommendation task as an example (Figure 5), dynamic memory mechanism significantly outperforms the baseline, showing that precise contextual relevance matters more than sheer retrieval volume. In contrast, random retrieval yields limited gains, even with more cases.

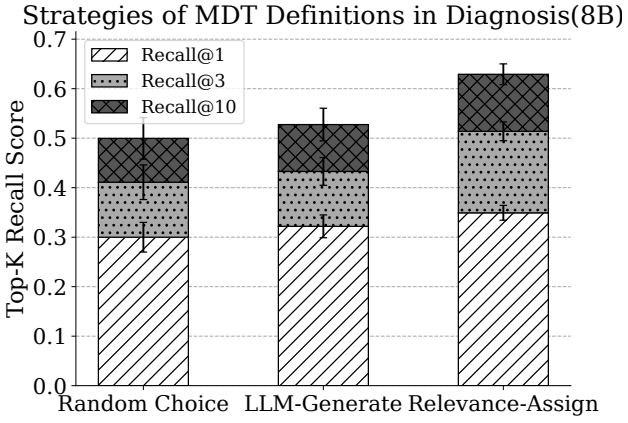

Figure 4: Diagnosis MDT settings

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 rare disease patients dataset, providing a valuable resource for future research.

Acknowledgments

This study was supported by grants from the National Science Foundation of China (T2541010), the National Key R&D Program of China (2024YFF1207100, 2024YFF1207103), and Beijing National Research Center for Information Science and Technology (BNRist). The funders had no roles in study design, data collection and analysis, publication decisions, or manuscript preparation.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chen, X.; Mao, X.; Guo, Q.; Wang, L.; Zhang, S.; and Chen, T. 2024. RareBench: Can LLMs Serve as Rare Diseases Specialists? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4850–4861.

Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; and Stewart, W. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fan, Z.; Wei, L.; Tang, J.; Chen, W.; Siyuan, W.; Wei, Z.; and Huang, F. 2025. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, 10183–10213.

Javed, A.; Agrawal, S.; and Ng, P. C. 2014. Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nature methods*, 11(9): 935–937.

Jin, Q.; Wang, Z.; Yang, Y.; Zhu, Q.; Wright, D.; Huang, T.; Wilbur, W. J.; He, Z.; Taylor, A.; Chen, Q.; et al. 2024. AgentMD: Empowering Language Agents for Risk Prediction with Large-Scale Clinical Tool Learning. *arXiv preprint arXiv:2402.13225*.

Johnson, A. E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

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

Kim, Y.; Park, C.; Jeong, H.; Chan, Y. S.; Xu, X.; McDuff, D.; Lee, H.; Ghassemi, M.; Breazeal, C.; and Park, H. W. 2024. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Köhler, S.; Vasilevsky, N. A.; Engelstad, M.; Foster, E.; McMurry, J.; Aymé, S.; Baynam, G.; Bello, S. M.; Boerkoel, C. F.; Boycott, K. M.; et al. 2017. The human phenotype ontology in 2017. *Nucleic acids research*, 45(D1): D865–D876.

Lee, P.; Bubeck, S.; and Petro, J. 2023. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13): 1233–1239.

Li, B.; Yan, T.; Pan, Y.; Luo, J.; Ji, R.; Ding, J.; Xu, Z.; Liu, S.; Dong, H.; Lin, Z.; et al. 2024a. MMedAgent: Learning to Use Medical Tools with Multi-modal Agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 8745–8760.

Li, G.; Al Kader Hammoud, H. A.; Itani, H.; Khizbullin, D.; and Ghanem, B. 2023. CAMEL: communicative agents for" mind" exploration of large language model society. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 51991–52008.

Li, J.; Wang, S.; Zhang, M.; Li, W.; Lai, Y.; Kang, X.; Ma, W.; and Liu, Y. 2024b. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. 2023. AgentBench: Evaluating LLMs as Agents. In *The Twelfth International Conference on Learning Representations*.

Lu, M.; Ho, B.; Ren, D.; and Wang, X. 2024. TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 5747–5764.

Mao, X.; Huang, Y.; Jin, Y.; Wang, L.; Chen, X.; Liu, H.; Yang, X.; Xu, H.; Luan, X.; Xiao, Y.; et al. 2025. A phenotype-based AI pipeline outperforms human experts in differentially diagnosing rare diseases using EHRs. *npj Digital Medicine*, 8(1): 68.

Mehandru, N.; Miao, B. Y.; Almaraz, E. R.; Sushil, M.; Butte, A. J.; and Alaa, A. 2024. Evaluating large language models as agents in the clinic. *NPJ digital medicine*, 7(1): 84.

Nguengang Wakap, S.; Lambert, D. M.; Olry, A.; Rodwell, C.; Gueydan, C.; Lanneau, V.; Murphy, D.; Le Cam, Y.; and Rath, A. 2020. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European journal of human genetics*, 28(2): 165–173.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Prakash, P.; Chilukuri, S.; Ranade, N.; and Viswanathan, S. 2021. RareBERT: transformer architecture for rare disease patient identification using administrative claims. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 453–460.

Robinson, P. N.; Ravanmehr, V.; Jacobsen, J. O.; Danis, D.; Zhang, X. A.; Carmody, L. C.; Gargano, M. A.; Thaxton,

C. L.; Karlebach, G.; Reese, J.; et al. 2020. Interpretable clinical genomics with a likelihood ratio paradigm. *The American Journal of Human Genetics*, 107(3): 403–417.

Saab, K.; Tu, T.; Weng, W.-H.; Tanno, R.; Stutz, D.; Wulczyn, E.; Zhang, F.; Strother, T.; Park, C.; Vedadi, E.; et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Schick, T.; Dwivedi-Yu, J.; Dessí, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: language models can teach themselves to use tools. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 68539–68551.

Schieppati, A.; Henter, J.-I.; Daina, E.; and Aperia, A. 2008. Why rare diseases are an important medical and social issue. *The Lancet*, 371(9629): 2039–2041.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 623(7987): 493–498.

Shang, J.; Ma, T.; Xiao, C.; and Sun, J. 2019a. Pre-training of graph augmented transformers for medication recommendation. In *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence.

Shang, J.; Xiao, C.; Ma, T.; Li, H.; and Sun, J. 2019b. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1126–1133.

Shi, W.; Xu, R.; Zhuang, Y.; Yu, Y.; Zhang, J.; Wu, H.; Zhu, Y.; Ho, J.; Yang, C.; and Wang, M. D. 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*, 22315–22339.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Tang, X.; Zou, A.; Zhang, Z.; Li, Z.; Zhao, Y.; Zhang, X.; Cohan, A.; and Gerstein, M. 2024. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 599–621. Bangkok, Thailand: Association for Computational Linguistics.

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

Valdez, R.; Ouyang, L.; and Bolen, J. 2016. Public health and rare diseases: oxymoron no more. *Preventing chronic disease*, 13.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, R.; Qiu, Z.; Jiang, J.; Qi, G.; and Wu, X. 2022. Conditional generation net for medication recommendation. In *Proceedings of the ACM Web Conference 2022*, 935–945.

Xie, J.; Jin, Y.; Shen, M.; Chen, L.; and Zhang, S. 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).

Yang, C.; Xiao, C.; Glass, L.; and Sun, J. 2021a. Change Matters: Medication Change Prediction with Recurrent Residual Networks. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*, 3728–3734. International Joint Conferences on Artificial Intelligence.

Yang, C.; Xiao, C.; Ma, F.; Glass, L.; and Sun, J. 2021b. SafeDrug: Dual Molecular Graph Encoders for Recommending Effective and Safe Drug Combinations. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*, 3735–3741. International Joint Conferences on Artificial Intelligence.

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

Yang, N.; Zeng, K.; Wu, Q.; and Yan, J. 2023. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *Proceedings of the ACM Web Conference 2023*, 4075–4085.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*.

Zhai, W.; Huang, X.; Shen, N.; and Zhu, S. 2023. Phen2Disease: a phenotype-driven model for disease and gene prioritization by bidirectional maximum matching semantic similarities. *Briefings in Bioinformatics*, 24(4): bbad172.

Zhang, Z.; Zhang-Li, D.; Yu, J.; Gong, L.; Zhou, J.; Liu, Z.; Hou, L.; and Li, J. 2024. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*.

Zhao, Z.; Jing, Y.; Feng, F.; Wu, J.; Gao, C.; and He, X. 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*, 533–542.

Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; and Wang, Y. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19724–19731.