DrAgent: Empowering Large Language Models as Medical Agents for Multi-hop Medical Reasoning

Fenglin Liu[1], **Zheng Li**[2*], **Hongjian Zhou**[1], **Qingyu Yin**[2], **Jingfeng Yang**[2], **Xin Liu**[2], **Zhengyang Wang**[2], **Xianfeng Tang**[2], **Shiyang Li**[2], **Xiang He**[2], **Ruijie Wang**[2], **Bing Yin**[2], **Xiao Gu**[1], **Lei Clifton**[3], **David Clifton**[1,4*]

[1]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford [2]Amazon

[3]Applied Digital Health, Nuffield Department of Primary Care Health Sciences, University of Oxford

[4]Oxford-Suzhou Centre for Advanced Research, University of Oxford, Suzhou 215123, Jiangsu, China

Abstract

Although large language models (LLMs) have demonstrated outperforming human experts in medical examinations, it remains challenging to adopt LLMs in real-world clinical decision-making that typically involves multi-hop medical reasoning. Common practices include prompting commercial LLMs and fine-tuning LLMs on medical data. However, in the clinical domain, using commercial LLMs raises privacy concerns regarding sensitive patient data. Fine-tuning competitive medical LLMs for different tasks usually requires extensive data and computing resources, which are difficult to acquire, especially in medical institutions with limited infrastructure. We propose DrAgent, which can build LLMs as *agents* to deliver accurate medical *d*ecision-making and *r*easoning. In implementation, we take a lightweight LLM as the backbone to collaborate with diverse clinical tools. To make efficient use of data, DrAgent introduces recursive curriculum learning to optimize the LLM in an easy-to-hard progression. The results show that our approach achieves competitive performance on diverse datasets.

1 Introduction

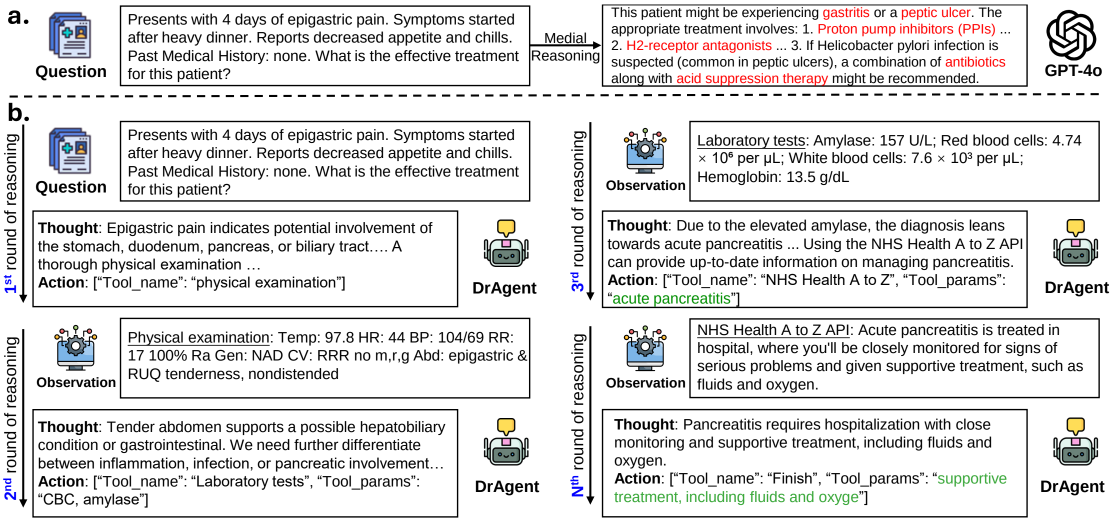
Recently, inspired by the impressive capabilities of Large Language Models (LLMs) (Zhao et al., 2023; Yang et al., 2023a) in understanding and generating human language, such as the GPT-series (OpenAI, 2023a) and the LLaMA-series (Touvron et al., 2023a,b), the application of LLMs in healthcare to assist clinicians has attracted extensive research interest. Existing efforts typically fine-tune publicly available LLMs, e.g., LLaMA-series, on massive medical datasets to develop medical LLMs (Liu et al., 2025b). Representative examples include MEDITRON (Chen et al., 2023b), Clinical Camel (Toma et al., 2023), and PMC-LLaMA (Wu et al., 2023). Notably, Med-Gemini (Yang et al., 2024a; Saab et al., 2024), developed by fine-tuning Gemini (Team et al., 2023) on more than 7 million data samples, and GPT-4-MedPrompt (Nori et al., 2023), which prompts GPT-4 (OpenAI, 2023a) with elaborately designed prompts, demonstrate performance surpassing that of experts on the United States Medical Licensing Examination (Jin et al., 2021).

However, these medical LLMs are typically evaluated on close-ended examination-style QA (Liu et al., 2025b; He et al., 2023) that does not reflect real-world clinical practice where multi-hop reasoning are required. For example, the process of treatment recommendations could involve the following: 1) Clinicians need to analyze the patient's symptoms and history of present illness. 2) When a patient's condition is complex, preliminary examinations (e.g., physical examination or blood tests) are often required for further assessment. 3) Sometimes, preliminary examinations do not provide enough information for an accurate diagnosis. In such cases, clinicians may order additional tests (e.g. CT scans or microbiology examinations) based on previous results, continuing until a reliable diagnosis can be made. 4) When clinicians recommend treatment plans, such as medications, they typically consider potential drug-drug interactions with the patient's current medications and account for the patient's specific circumstances (e.g. pregnancy) to evaluate possible side effects. Thus, real-world clinical practice diverges from the structured nature of exam-taking that existing medical LLMs are evaluated on. Existing research has shown that such LLMs perform poorly on complex clinical decision-making, e.g., medical code querying (Soroush et al., 2024), new drug understanding (Liu et al., 2024), and clinical diagnostics (Hager et al., 2024).

A potential solution for enabling LLMs to deliver accurate results across various medical tasks is fine-tuning LLMs on target medical data and

[*] Corresponding authors.
amzzhe@amazon.com; david.clifton@eng.ox.ac.uk

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15656–15668
November 4-9, 2025 ©2025 Association for Computational Linguistics



Presents with 4 days of epigastric pain. Symptoms started after heavy dinner. Reports decreased appetite and chills. Past Medical History: none. What is the effective treatment for this patient?

This patient might be experiencing gastritis or a peptic ulcer. The appropriate treatment involves: 1. Proton pump inhibitors (PPIs) ... 2. H2-receptor antagonists ... 3. If Helicobacter pylori infection is suspected (common in peptic ulcers), a combination of antibiotics along with acid suppression therapy might be recommended.

Presents with 4 days of epigastric pain. Symptoms started after heavy dinner. Reports decreased appetite and chills. Past Medical History: none. What is the effective treatment for this patient?

Thought: Epigastric pain indicates potential involvement of the stomach, duodenum, pancreas, or biliary tract…. A thorough physical examination …
Action: [“Tool_name”: “physical examination”]

Physical examination: Temp: 97.8 HR: 44 BP: 104/69 RR: 17 100% Ra Gen: NAD CV: RRR no m,r,g Abd: epigastric & RUQ tenderness, nondistended

Thought: Tender abdomen supports a possible hepatobiliary condition or gastrointestinal. We need further differentiate between inflammation, infection, or pancreatic involvement…
Action: [“Tool_name”: “Laboratory tests”, “Tool_params”: “CBC, amylase”]

Laboratory tests: Amylase: 157 U/L; Red blood cells: 4.74 × 10^6 per µL; White blood cells: 7.6 × 10^3 per µL; Hemoglobin: 13.5 g/dL

Thought: Due to the elevated amylase, the diagnosis leans towards acute pancreatitis ... Using the NHS Health A to Z API can provide up-to-date information on managing pancreatitis.
Action: [“Tool_name”: “NHS Health A to Z”, “Tool_params”: “acute pancreatitis”]

NHS Health A to Z API: Acute pancreatitis is treated in hospital, where you'll be closely monitored for signs of serious problems and given supportive treatment, such as fluids and oxygen.

Thought: Pancreatitis requires hospitalization with close monitoring and supportive treatment, including fluids and oxygen.
Action: [“Tool_name”: “Finish”, “Tool_params”: “supportive treatment, including fluids and oxyge”]

Figure 1: **a.** The answer generated by GPT-4o (OpenAI, 2023a); **b.** Our DrAgent is organized into three components (Thought, Action, and Observation) for collaborating with diverse medical tools to deliver effective multi-hop medical reasoning and decision-making. Red-colored and green-colored text indicate the wrong and correct diagnoses and treatments, respectively. GPT-4o provides an answer in just one round of reasoning, which is usually neither appropriate nor reliable in real-world clinical decision-making, resulting in wrong diagnoses and treatment recommendations. In contrast, our DrAgent is designed, as shown in the above example, to make decisions.

tasks. However, fine-tuning LLMs forces them to learn diverse and complex tasks, and therefore not only imposes huge learning pressure on them, but also requires vast amounts of data (ranging from 1 million (Ferber et al., 2024; Han et al., 2023; Singhal et al., 2023; Liu et al., 2023a) to 80 billion (Chen et al., 2023b)) and computational resources, creating barriers to adoption in medical institutions with limited technical infrastructure. Another solution is to design prompts for closed-source LLMs such as GPT-4 (Nori et al., 2023), but such prompt engineering and strict privacy regulations pose an adoption threshold in hospital settings (Soroush et al., 2024).

We propose an efficient solution, DrAgent, that can build lightweight LLMs as medical agents to collaborate with diverse medical tools, as shown in Figure 1. In implementation, we first build a tool library consisting of 16 diverse medical tools (as shown in Table 1), i.e., medical procedures, medical models, and medical APIs. DrAgent emulates real-world clinical practice to perform multi-hop medical reasoning in an iterative process. At each round, DrAgent relies on the historical reasoning process and previous results to select and execute a medical tool, ultimately arriving at accurate and confident answers. To efficiently train our lightweight DrAgent using limited data, we propose Recursive Curriculum Learning (RCL) to achieve competitive performance with larger LLMs. During training, RCL enables LLMs to progressively learn multi-hop medical reasoning in an easy-to-hard fashion and recursively learn experience from unseen cases, similar to the human learning curve: 1) It starts with simple medical reasoning (e.g., single-hop reasoning) data; 2) It then attempts to process more difficult medical reasoning data, where multi-hop reasoning is required; 3) It performs reasoning for unseen cases; and learns to incorporate the experience from unseen cases to further boost performance, i.e., the successfully reasoned cases are used for continued training, and the unsuccessfully reasoned cases can be used for reflection training. Such a practice has been shown to be a better solution than the common approach of uniformly sampling training data from limited medical data (Liu et al., 2021). Meanwhile, learning from unseen cases can reduce models' reliance on labeled data and enable the use of unlabeled data for training.

Overall, our method has the following advantages:

- DrAgent is designed to learn how to select the correct medical tools to collaborate with to perform multi-hop medical reasoning, without requiring fine-tuning on various downstream tasks with large amounts of medical data.

- DrAgent is capable of fully utilizing existing validated clinical tools and data, instead of performing extensive fine-tuning, thus saving costs and representing a critical step toward sustainable AI (Krishnan et al., 2023).

- We evaluate the performance of our DrAgent on a wide range of tasks, demonstrating superior performance compared to existing methods in handling diverse tasks.

2 Related Work

Medical Large Language Models Different medical LLMs (Liu et al., 2025b) that adapt general LLMs to the medical domain have been proposed to assist clinicians in decision-making (Thirunavukarasu et al., 2023; Patel and Lam, 2023; Zhou et al., 2025). However, existing works usually ignore not only the real-world complex decision-making that involves multi-hop medical reasoning (Hager et al., 2024; Liu et al., 2025b), but also the use of existing deployed medical tools to support reasoning. Meanwhile, it has shown that LLMs are ineffective at dealing with complex clinical tasks (Soroush et al., 2024; Liu et al., 2024). In this work, we propose the efficient DrAgent that collaborates with existing medical tools and enables LLMs to efficiently deal with a range of complex tasks.

AI Agents AI agents are designed to perceive their environment and make decisions to achieve specific goals (Xie et al., 2024). Recently, the superior performance of LLMs has improved agents' capabilities in interacting with the environment using language (Wang et al., 2024).

1) *Prompting* closed-source LLMs as agents: Most existing works (Shen et al., 2024; Yang et al., 2023b; Hu et al., 2024; Jin et al., 2024; Wang et al., 2025) focus on using prompting techniques to leverage closed-source LLMs (such as GPT (OpenAI, 2023a)) to make decisions. For example, MedAgents (Tang et al., 2024; Kim et al., 2024) prompts GPT-4 in a role-playing setting to take on multiple different roles, such as cardiologist and pulmonologist, and then involves them in discussions. However, the heavy reliance on prompts for customized roles makes it difficult and unstable to customize the agent's behavior (Liu et al., 2023c; Qiao et al., 2024). Additionally, deploying closed-source models in clinical settings raises privacy concerns over sensitive patient data.

2) *Fine-tuning* open-source LLMs as agents: This type of work (Liu et al., 2025a, 2023b; Yang et al., 2024b; Chen et al., 2023a; Li et al., 2024) collects the instruction-following data to fine-tune open-source LLMs (e.g. LLaMA) (Touvron et al., 2023a). Fine-tuning the customized behavior of LLMs can enhance their ability to understand instructions and make decisions, achieving results comparable to those of closed-source LLMs (Liu et al., 2025a; Qiao et al., 2024). However, current efforts are limited to the general (non-medical) domains (Liu et al., 2025a; Li et al., 2024). To this end, we propose to fine-tune LLMs as medical agents to improve their ability to handle multi-hop medical reasoning. Furthermore, we propose Recursive Curriculum Learning (RCL) to alleviate the reliance of fine-tuning LLMs on a large amount of data and model parameters.

3 Approach

3.1 Tool Library

Our DrAgent uses a broad range of medical procedures, medical APIs, and medical models as medical tools in our tool library, as shown in Table 1.

• Physical Examinations: Assessments are conducted to evaluate a patient's physical health through observation, palpation, auscultation, and other diagnostic techniques.

• Laboratory Tests: Analytical procedures conducted on blood, urine, or tissue samples to identify disease, measure organ function, or monitor treatment progress.

• Microbiology Tests: Identifying infectious agents, such as bacteria, viruses, fungi, and parasites, through culture, microscopy, or molecular techniques.

• Radiology Tests: Imaging techniques, such as X-rays, CT scans, MRCP scans, and ultrasounds, used to visualize internal body structures and diagnose medical conditions.

• UK NHS Health API: Providing clinical-standard condition information sourced from the UK National Health Service[1] to support healthcare decision-making.

• UK NHS Medicine API: Providing clinical-standard medication data, including usage, dosage, and side effects, from the UK National Health Service[2].

We detail the medical models in Appendix A. The above comprehensive tool library allows DrAgent to address a wide range of medical reasoning tasks. Note that our tool library is agnostic to the

[1] https://www.nhs.uk/conditions/

[2] https://www.nhs.uk/medicines/

Types	Tools	Sources
Procedures	Physical Examinations	MIMIC-IV (Johnson et al., 2023)
	Laboratory Tests	MIMIC-IV (Johnson et al., 2023)
	Microbiology Tests	MIMIC-IV (Johnson et al., 2023)
	Radiology Tests	MIMIC-IV (Johnson et al., 2023)
APIs	UK NHS Health	UK National Health Service Health[2]
	UK NHS Medicine	UK National Health Service Medicines[3]
Task-specific Models	Question Answering	BioLinkBERT (Yasunaga et al., 2022)
	Drug Recommendation	BioLinkBERT (Yasunaga et al., 2022)
	Drug Adverse Reaction	BioLinkBERT (Yasunaga et al., 2022)
	Drug-drug Interaction	BioLinkBERT (Yasunaga et al., 2022)
	Named Entity Recognition	BioLinkBERT (Yasunaga et al., 2022)
	Relation Extraction	BioLinkBERT (Yasunaga et al., 2022)
	Document classification	BioLinkBERT (Yasunaga et al., 2022)
	Radiology Report Generation	Transformer (Vaswani et al., 2017)
	Clinical Note Summarization	Transformer (Vaswani et al., 2017)

Table 1: Medical tools included in the tool library of our DrAgent.

type of medical tools, allowing users to include additional tools, such as clinical calculators, medical equipment, and medical (wearable) devices.

3.2 DrAgent

As shown in Figure 1, given the input question X, the goal of DrAgent is to provide a correct answer Y based on the multi-hop reasoning $R_N = \{r_1, r_2, ..., r_N\}$, where N denotes the number of reasoning hops. Our DrAgent can be defined as

$$Y = \text{DrAgent}(X, R_N) \tag{1}$$

For multi-hop medical reasoning, to help our DrAgent better organize the reasoning process, we follow previous work (Yao et al., 2023) in generating the reasoning in the format of Thought-Action-Observation, i.e., $r_i = (t_i, a_i, o_i)$.

Thought DrAgent analyzes the historical reasoning process and selects the tool from the tool library required for the current reasoning. Thought is the core decision-making component of DrAgent. At the i_{th} round of reasoning, it analyzes the current problem X and historical reasoning $R_{i-1} = \{r_1, r_2, ..., r_{i-1}\}$ to decide whether a specific tool is needed to proceed with the reasoning or if a final answer can be generated directly. Thought at the i_{th} round of reasoning t_i can be defined as:

$$t_i = \text{DrAgent}(X, R_{i-1}) \tag{2}$$

Thought plans the reasoning process, enabling DrAgent to efficiently solve medical multi-hop reasoning.

Action After Thought makes a decision, Action is responsible for executing the selected tool and parsing the required parameters (if needed) from the problem context or prior reasoning steps. Thus, Action at the i_{th} round of reasoning a_i is:

$$a_i = \text{DrAgent}(X, R_{i-1}, t_i) \tag{3}$$

Action serves as DrAgent's execution engine, ensuring precise tool usage and accurate parameter handling to support the reasoning process.

Observation Observation is defined as the external information (e.g., tool outputs). Observation stores the output results of invoked tools (e.g. drug recommendations, diagnostic information, or laboratory test results) and provides them to Thought for reasoning updates. We define Observation as:

$$o_i = a_i[\text{“Tool\_name”}](\text{“Tool\_params”}) \tag{4}$$

We can see that Observation acts as a bridge between DrAgent and the external environment, ensuring smooth and efficient information flow during multi-hop reasoning.

Iteratively performing the above steps enables DrAgent to perform accurate and efficient multi-hop reasoning in complex medical scenarios.

3.3 Recursive Curriculum Learning

Existing work (Chen et al., 2023a; Zeng et al., 2023; Li et al., 2024) usually requires massive labeled training data[3] (i.e., instruction fine-tuning (IFT)), which, however, is not easy to obtain for multi-hop reasoning. Although some researchers (Chen et al., 2023a; Zeng et al., 2023) attempt to use GPT-4 (OpenAI, 2023a) to generate large amounts of training (reasoning) data, it is hard to control the quality of the generated data (Qiao et al., 2024). We propose Recursive Curriculum Learning (RCL), which, during training, enables LLMs to gradually progress from easy samples to more complex ones and recursively obtain more (pseudo-labeled) training data using limited labeled data: (1) first starting with simple and easily reasoned samples; (2) then attempting to handle harder samples that involve multi-hop reasoning; (3) learning to perform reasoning for unseen cases; (4) finally incorporating the experience from unseen cases to further boost performance.

As shown in Algorithm 1, we first sort the labeled datasets according to the number of reasoning rounds (i.e., reasoning hops), enabling the model

[3] Here, the labeled data represent the reasoning process data R from the input X to the output Y.

Algorithm 1 Recursive Curriculum Learning.

```
Input: The labeled reasoning data $D^l_j = \{X^l, Y^l, R^l_j\}$, consisting of $j$-hop reasoning ($j \in [1; N]$). The unlabeled data without reasoning $D^{\text{u}} = \{X^u, Y^u\}$
Output: A medical agent for multi-hop medical reasoning with recursive curriculum learning.
 1: Sort $D^l_j$ based on the number of reasoning rounds $j$;
 2: for $k = 1$ to $N$ do
 3:     Train the model with the training data $D^l_{j \leq k}$;
 4:     The model performs inference on the unlabeled training data $D^{\text{u}}$, obtaining the answers $Y^*$ and reasoning $R^*_j$ given the input questions $X^u$.
 5:     if $Y^* = Y^u$ and $j \leq k$ then
 6:         Reasoning data: $D^l_{j \leq k} \leftarrow \text{Add}(\{X^u, Y^u, R^*_j\})$
 7:     else if $Y^* \neq Y^u$ then
 8:         Reflection data: $F \leftarrow \text{Add}(\{X^u, Y^u, R^*_j\})$
 9:     end if
10: end for
11: repeat
12:     Train the model with the updated reasoning data $D^l_j$ to perform reasoning;
13:     Train the model with the generated reflection data $F$ to perform reflection;
14: until Model converge.
```

to be trained in an easy-to-hard fashion (Line 1). Then, at the k_{th} round of training, DrAgent is first trained on a small set of labeled reasoning data $\{X^l, Y^l, R^l_{j\leq k}\}$, i.e., medical questions with correct reasoning processes and answers, where the reasoning hops j are no greater than k. We train the model by using the widely-used instruction fine-tuning (Line 3), defined as:

$$Y^l = \text{DrAgent}(X^l, R^l_{j\leq k}) \quad (5)$$

Through limited high-quality labeled data for training, our DrAgent obtains the initial ability to perform medical reasoning with $j \leq k$ hops. The trained model then performs inference (Line 4) on unlabeled unseen data (without reasoning processes) $\{X^u, Y^u\}$ to generate answers Y^* and reasoning processes R^*_j, where j can vary with any number, generating varying reasoning rounds. Next, we take those with correct answers as the pseudo-labeled reasoning data, which are added to the original labeled dataset (Lines 5-6). Meanwhile, as shown in Lines 7-8, we collect the data with incorrect answers to form a reflection dataset. Finally, the resulting larger training dataset is used to continue training DrAgent (Line 12); The reflection dataset is used to train the model to learn reflection (Xie et al., 2024) (Line 13). By repeating these steps we can iteratively generate new reasoning and reflection data and gradually train a better model to achieve desirable multi-hop medical reasoning results, using limited labeled data.

Downstream Tasks	Datasets	Metric
Single-hop Medical Reasoning		
Question Answering (QA)	PubMedQA	Accuracy
Drug Recommendation (DR)	HealthCareMagic	F1
Drug Adverse Reaction (DAR)	ADE-Corpus-v2	Accuracy
Drug-drug Interaction (DDI)	DDI-Corpus	Accuracy
Named Entity Recognition (NER)	BC5-Disease	F1 entity-leve
Relation Extraction (RE)	GAD	Micro F1
Document Classification (DC)	HoC	Micro F1
Radiology Report Generation (RRG)	MIMIC-CXR	ROUGE-L
Discharge Summarization (DS)	MIMIC-III	ROUGE-L
Multi-hop Medical Reasoning		
Treatment Recommendation	MIMIC-IV	Accuracy

Table 2: Overview of tasks and datasets used in our experiments.

4 Experiments

In this section, we provide the main results and analyses to show the effectiveness of our approach.

4.1 Datasets, Metrics, and Settings

Table 2 shows an overview of the datasets and metrics we used for evaluation.

4.1.1 Single-hop Medical Reasoning Data

To evaluate the performances of the DrAgent, we first adopt existing datasets, i.e., PubMedQA (Jin et al., 2019), HealthCareMagic (Yunxiang et al., 2023), ADE-Corpus-v2 (Gurulingappa et al., 2012), DDI-Corpus (Herrero-Zazo et al., 2013), BC5-Disease (Li et al., 2016), GAD (Becker et al., 2004), HoC (Baker et al., 2016), MIMIC-CXR (Johnson et al., 2019), and MIMIC-III (Johnson et al., 2016).

4.1.2 Multi-hop Medical Reasoning Data

We then apply our DrAgent in real-world clinical data from the MIMIC-IV database (Johnson et al., 2023) that contains real patient cases from 300,000 patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA, between 2008 and 2019. It includes all recorded measurement data, such as physical examinations, laboratory tests, microbiological tests, radiology tests, diagnoses, surgeries, and treatment information. In our evaluation, the model is required to provide treatment recommendations based on the patient's condition. The model may need to rely on one or more of the following procedures: physical examinations, laboratory tests, microbiological tests, radiology tests, or others, to give accurate answers (Hager et al., 2024).

4.1.3 Settings

DrAgent adopts the Phi-3.5-mini-instruct (Abdin et al., 2024) model with 3.8 billion parameters as

Types	Methods	# Params	Single-hop Medical Reasoning									Multi-hop Medical Reasoning				
			QA	DR	DAR	DDI	NER	RE	DC	RRG	DS	1-hop	2-hop	3-hop	4-hop	5-hop
LLMs	Phi-3.5-mini (Abdin et al., 2024)	3.8B	64.8	7.1	44.6	40.3	43.7	41.2	55.8	10.9	4.4	40.3	31.0	29.4	27.3	20.0
	Mistral (Jiang et al., 2023)	7B	69.4	5.8	27.1	30.4	46.8	44.3	59.6	13.2	5.3	42.7	32.1	29.8	27.2	25.0
	Meditron-7B (Chen et al., 2023b)	7B	61.6	8.7	30.6	34.5	46.5	43.3	57.9	12.5	5.9	42.9	33.9	30.7	26.4	28.7
	BioMistral (Labrak et al., 2024)	7B	66.4	10.2	52.6	43.9	48.8	48.5	64.3	14.2	6.6	43.9	34.5	32.4	27.3	33.3
	MedAlpaca (Han et al., 2023)	13B	65.6	11.0	55.6	45.8	49.2	44.5	59.4	11.7	3.5	45.8	41.4	38.2	36.4	33.3
	Meditron-70B (Chen et al., 2023b)	70B	70.6	12.8	58.4	47.0	54.3	49.6	69.6	13.3	7.7	47.0	51.7	47.1	45.5	40.0
	GPT-3.5-turbo (OpenAI, 2023c)	-	71.2	17.1	66.2	61.2	54.6	56.0	57.9	14.1	9.2	76.5	75.9	70.6	72.7	66.7
	GPT-4o (OpenAI, 2023b)	-	**82.6**	**23.7**	**74.7**	**70.4**	**71.7**	**67.9**	**74.8**	**25.1**	**18.8**	**84.6**	**81.3**	**78.4**	**75.9**	**70.1**
AI Agents	ReAct (Yao et al., 2023)	7B	71.2	14.6	60.6	62.5	63.1	60.8	67.9	24.8	21.0	72.3	43.1	42.0	40.5	39.6
	BOLAA (Liu et al., 2023c)	7B	74.8	16.8	64.2	67.3	65.7	61.2	72.4	29.3	24.1	73.8	45.1	44.0	43.6	43.5
	Chameleon (Lu et al., 2024)	7B	73.3	18.4	65.1	64.4	69.6	64.4	69.0	27.5	25.9	73.5	44.0	42.8	41.7	41.3
	MedAgents (Tang et al., 2024)	GPT-3.5	75.6	20.6	68.8	69.1	73.5	71.7	75.2	30.6	29.8	76.3	71.2	68.3	64.2	60.7
	DrAgent (Ours)	3.8B	**81.5**	**29.4**	**77.8**	**79.4**	**87.5**	**84.1**	**85.9**	**41.2**	**37.2**	**79.1**	**75.4**	**72.3**	**71.2**	**62.8**

Table 3: Comparison of our DrAgent with existing methods on diverse single-hop and one multi-hop medical reasoning datasets.

the backbone. We fine-tune all our models with LoRA (Hu et al., 2021) for 5 epochs. The rank of LoRA is set to 128, and the training batch size is set to 64. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer and adopt a cosine learning rate scheduler with a peak learning rate of 2e-4. All the training and inference experiments are conducted on four A100 GPUs. We use GPT-4o (OpenAI, 2023a) to generate 3,000 training data for each single-hop medical reasoning task, 10,000 training data for the multi-hop medical reasoning task. We regard the remaining data in the training set as unlabeled training datasets. To improve accuracy and reliability, we further introduce a reflection step (Xie et al., 2024) by prompting the method using “Please reflect your answer based on the history". The reflection dataset generated during training will instruct the model to consider all the historical reasoning and provide a reflection result. We run the experiments using random seeds and report the average results.

4.2 Single-hop Reasoning Results

For the baselines of performance comparison, we select existing representative LLMs including Phi-3.5-mini (Abdin et al., 2024), Mistral (Jiang et al., 2023), GPT-3.5-turbo (OpenAI, 2023c), GPT-4 (OpenAI, 2023b); state-of-the-art medical LLMs including Meditron (Chen et al., 2023b), BioMistral (Labrak et al., 2024), and MedAlpaca (Han et al., 2023); and LLM agents including ReAct (Yao et al., 2023), BOLAA (Liu et al., 2023c), Chameleon (Lu et al., 2024), and MedAgents (Tang et al., 2024). To ensure a fair comparison, we use the same settings to re-implement their performance on the downstream tasks. For previous agents, we adopt the LLaMA-2-7B (Touvron et al., 2023b) as their backbone.

Table 3 shows that our DrAgent achieves the desirable performance across all datasets and metrics, with the fewest model parameters. As expected, all existing LLMs exhibit poor performance on complex medical reasoning, supporting the motivation of our proposed DrAgent. By collaborating with medical tools, our DrAgent surpasses the currently popular LLMs (including the commercial LLM, GPT-3.5-turbo and GPT-4o) by comfortable margins in performance, with fewer parameters. Notably, on the drug recommendation task, the performance of DrAgent almost doubles that of state-of-the-art larger medical LLM, Meditron-70B. These promising results indicate the effectiveness of DrAgent at dealing with complex medical tasks by collaborating with validated medical tools.

4.3 Multi-hop Reasoning Results

Here we further evaluate the performance of DrAgent across multi-hop medical reasoning ranging from 1-hop to 5-hop reasoning. Table 3 summarizes the results of our model compared to both LLMs and LLM Agents. It shows that all methods' performance degrades as the reasoning depth increases, highlighting the challenges of multi-hop medical reasoning that are common in real-world clinical settings. It is, therefore, necessary to investigate multi-hop medical reasoning beyond close-ended examination-style reasoning done in existing work. GPT-4o achieves the highest performance in all cases, with a peak performance of 84.6% in 1-hop reasoning and maintaining superior accuracy across all hops. Our DrAgent, with only 3.8B parameters, consistently outperforms existing LLM agents, including ReAct, BOLAA, Chameleon, and MedAgents, across all hops. DrAgent achieves

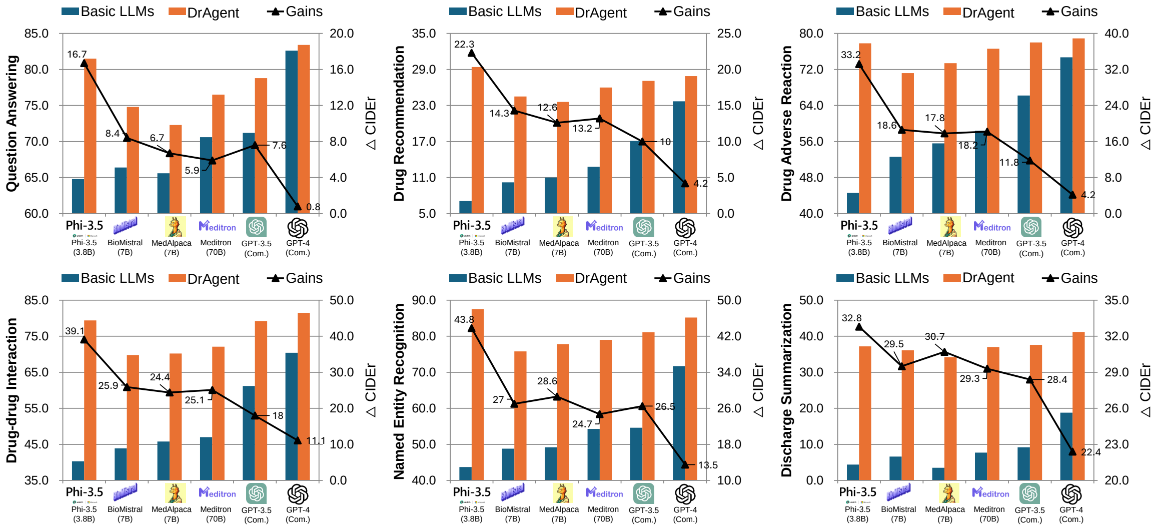


Figure 2: The generalization ability of our method. We report the performance of basic LLMs and our DrAgent, with varying numbers of model parameters. The performance gains in different LLMs are shown as the polyline on the right y-axis. The results show that our DrAgent consistently improves the performance of basic LLMs. The smaller the LLMs, the larger the gains.

Methods	# Params	# Data	1-hop	2-hop	3-hop	4-hop	5-hop
Meditron-70B	70B	-	47.0	51.7	47.1	45.5	40.0
GPT-3.5-turbo	-	-	76.5	75.9	70.6	72.7	66.7
Phi-3.5-mini		Baseline	40.3	31.0	29.4	27.3	20.0
DrAgent	3.8B	1,000	65.7(↑25.4)	59.6(↑28.6)	61.7(↑32.3)	60.5(↑33.2)	54.1(↑34.1)
		5,000	73.6(↑33.3)	66.2(↑35.2)	67.3(↑37.9)	64.4(↑37.1)	59.7(↑39.7)
		10,000	**79.1(↑38.8)**	**75.4 (↑44.4)**	**72.3(↑42.9)**	**71.2(↑43.9)**	**62.8(↑42.8)**

Table 4: Effect of different number of labeled training data for multi-hop medical reasoning. Our method consistently outperforms existing baseline methods, while maintaining satisfactory performance across all reasoning levels.

79.1% accuracy on 1-hop reasoning and 75.4% on 2-hop reasoning, surpassing baseline agents by a significant margin. Even at deeper reasoning levels (3-hop, 4-hop, and 5-hop), DrAgent demonstrates strong performance, achieving 72.3%, 71.2%, and 62.8% accuracy, respectively, indicating its robustness for complex multi-hop reasoning. One notable strength of our DrAgent is its light requirement on computational resources (including training data and model parameters), which is desirable for real-world applications.

4.4 Generalization analysis

In this study, we introduce a method that allows LLMs to collaborate with existing medical tools to produce trustworthy, evidence-based medical outputs. Our framework is independent of specific LLM architectures / backbones, which means it can be applied across different models to enhance their performance on medical tasks.

We further explore how well our system generalizes and adapts to LLMs with different parameter scales. For evaluation, we include Phi-3.5-mini (Abdin et al., 2024), several leading medical LLMs such as BioMistral (Labrak et al., 2024), MedAlpaca (Han et al., 2023), and Meditron (Chen et al., 2023b), as well as advanced commercial models like GPT-3.5-turbo and GPT-4o.

To compare performance, we report the overall accuracy of (i) the original LLMs and (ii) the same LLMs enhanced with our DrAgent. As shown in Figure 2, our approach consistently achieves higher accuracy compared to the baseline models. Notably, models with fewer parameters gain larger improvements, indicating that our method is particularly beneficial for healthcare settings with limited computational resources.

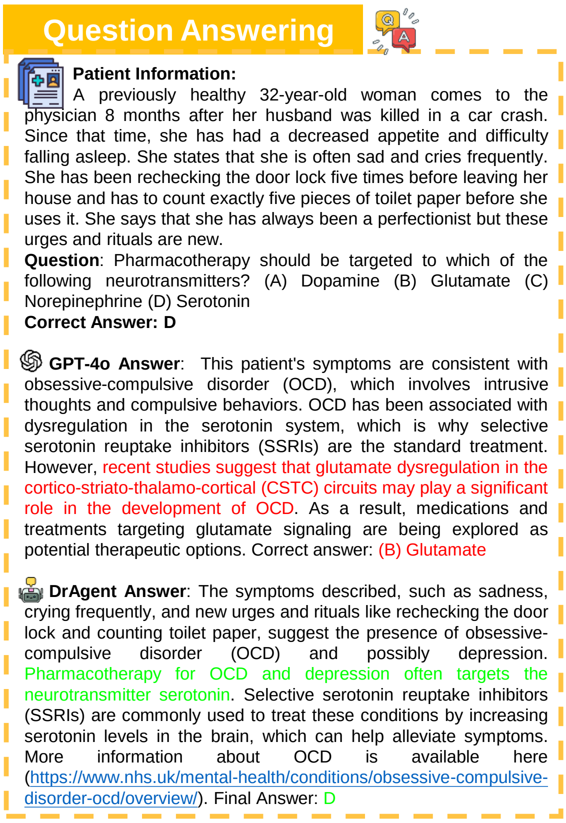
4.5 Robustness Analysis

To assess the robustness of our method to the number of training data, we evaluate the performances of DrAgent with respect to the increasing

Methods	Task Performance					Tool Selection Accuracy				
	1-hop	2-hop	3-hop	4-hop	5-hop	1-hop	2-hop	3-hop	4-hop	5-hop
Base (Phi-3.5-mini)	40.3	31.0	29.4	27.3	20.0	73.6	68.4	66.5	62.0	58.8
w/ Tools	68.6	60.5	53.7	47.4	41.2	84.5	80.2	75.3	71.6	74.0
w/ Thought	70.4	63.3	56.1	49.8	44.9	86.7	81.9	77.4	74.8	73.9
w/ Action	73.9	66.4	60.2	53.5	51.6	88.1	82.5	78.2	76.0	75.3
w/ Curriculum Learning	75.7	69.8	63.0	58.3	55.9	89.2	83.7	80.1	78.4	78.0
w/ RCL (DrAgent)	**79.1**	**75.4**	**72.3**	**71.2**	**62.8**	**91.7**	**87.4**	**86.6**	**85.3**	**80.4**

Table 5: Ablation study of DrAgent on multi-hop medical reasoning, showing both the task performance and the tool selection accuracy.

sizes of training data for multi-hop medical reasoning in Table 4. For comparison, we also show the performances of state-of-the-art medical LLM Meditron-70B (Chen et al., 2023b) and representative commercial LLM GPT-3.5-turbo (OpenAI, 2023c). DrAgent of different number of training data consistently outperforms basic LLM, indicating its robustness. Using limited (i.e. 1,000) number of training data, DrAgent helps a lightweight LLM (3.8B) to outperform larger medical LLM Meditron (70B), indicating not only its robustness but also its potential for application to new domains or tasks using a small amount of data.



Question Answering

Patient Information: A previously healthy 32-year-old woman comes to the physician 8 months after her husband was killed in a car crash. Since that time, she has had a decreased appetite and difficulty falling asleep. She states that she is often sad and cries frequently. She has been rechecking the door lock five times before leaving her house and has to count exactly five pieces of toilet paper before she uses it. She says that she has always been a perfectionist but these urges and rituals are new.
Question: Pharmacotherapy should be targeted to which of the following neurotransmitters? (A) Dopamine (B) Glutamate (C) Norepinephrine (D) Serotonin
Correct Answer: D

GPT-4o Answer: This patient's symptoms are consistent with obsessive-compulsive disorder (OCD), which involves intrusive thoughts and compulsive behaviors. OCD has been associated with dysregulation in the serotonin system, which is why selective serotonin reuptake inhibitors (SSRIs) are the standard treatment. However, recent studies suggest that glutamate dysregulation in the cortico-striato-thalamo-cortical (CSTC) circuits may play a significant role in the development of OCD. As a result, medications and treatments targeting glutamate signaling are being explored as potential therapeutic options. Correct answer: (B) Glutamate

DrAgent Answer: The symptoms described, such as sadness, crying frequently, and new urges and rituals like rechecking the door lock and counting toilet paper, suggest the presence of obsessive-compulsive disorder (OCD) and possibly depression. Pharmacotherapy for OCD and depression often targets the neurotransmitter serotonin. Selective serotonin reuptake inhibitors (SSRIs) are commonly used to treat these conditions by increasing serotonin levels in the brain, which can help alleviate symptoms. More information about OCD is available here (https://www.nhs.uk/mental-health/conditions/obsessive-compulsive-disorder-ocd/overview/). Final Answer: D

Figure 3: Qualitative analysis of DrAgent, showing example of GPT-4o and our DrAgent. The red-colored text indicates errors or ‘hallucinations’, while the green-colored text indicates the correct analysis and answers. It shows that DrAgent not only generates correct answer but also provides evidence from the tools.

4.6 Quantitative Analysis

We conduct an ablation study of our proposed DrAgent, shown in Table 5, where we simultaneously evaluate task performance and tool selection accuracy. (i) All the proposed components, i.e. Tools, Thought, Action, Curriculum Learning, and Recursive Curriculum Learning (RCL), contribute to improvements in both task performance and tool selection accuracy over the base model (Phi-3.5-mini). For example, incorporating Tools improves the 1-hop task performance from 40.3 to 68.6 and the tool selection accuracy from 73.6 to 84.5, demonstrating the benefit of collaborating with medical tools. (ii) We observe that Action provides notable gains across multi-hop reasoning tasks, achieving 73.9 for 1-hop task performance and 66.4 for 2-hop task performance, while also boosting tool selection accuracy to 88.1 for 1-hop and 82.5 for 2-hop. These results suggest that the ability to execute actions effectively is critical for performance enhancement. (iii) Curriculum learning further enhances multi-hop reasoning, particularly for deeper reasoning chains. It increases 5-hop task performance from 41.2 to 55.9 and tool selection accuracy from 74.0 to 78.0, demonstrating that the easy-to-hard learning approach helps the model generalize better across complex reasoning tasks. (iv) The DrAgent incorporated with Recursive Curriculum Learning achieves the best performance across all settings, including 79.1 for 1-hop task performance and 62.8 for 5-hop task performance, with corresponding tool selection accuracies of 91.7 and 80.4, respectively. These results suggest that our DrAgent improves both task accuracy and tool usage efficiency across multi-hop reasoning scenarios. Overall, we observe the improved performance by each component in the ablation study, highlighting their individual importance.

4.7 Qualitative Analysis

Here we provide a qualitative analysis that examines how our DrAgent improves medical reasoning performance. Figure 3 shows that DrAgent not only provides accurate answers but also generates correct evidence that offers more trustworthy decision support than the current state-of-the-art LLMs. In contrast, GPT-4o gives incorrect answers, and worse, GPT-4o generates some factually incorrect content. Speficially, GPT-4o generates the statement "recent studies suggest that glutamate dysregulation in the cortico-striato-thalamo-cortical (CSTC) circuits may play a significant role in the development of OCD", where the "recent studies" mentioned do not exist, leading to an incorrect answer. Such fabricated content can mislead users, especially inexperienced clinicians. This is addressed by our DrAgent, which, by invoking tools such as the NHS Health API, provides a correct analysis and answer. As a result, based on real-world clinical-standard knowledge, DrAgent correctly answers the question and provides relevant evidence.

5 Conclusions

We have introduced DrAgent, a lightweight large language model designed to address critical challenges in adopting LLMs for real-world medical decision-making and reasoning. Our approach emphasizes data efficiency, parameter efficiency, and compute efficiency, enabling effective collaboration with diverse clinical tools while overcoming privacy concerns and resource limitations commonly faced in the clinical domain. To alleviate the reliance on labeled training data, we propose recursive curriculum learning, which allows DrAgent to achieve desirable performance on complex medical reasoning tasks, even with limited data. Experiments across different datasets, including multi-hop medical reasoning tasks, highlight the effectiveness of our method.

Overall, our work highlights the potential of hybrid systems that combine LLMs with validated medical tools, circumventing the computational and logistical barriers to training standalone AI models for every clinical scenario. By leveraging tools, we enhance the precision and reliability of AI-assisted decision-making while aligning with the practical needs of healthcare systems. Our findings advocate for a collaborative model of AI deployment in medicine, emphasizing scalability, resource efficiency, and clinical accuracy.

Limitations

Our work highlights the potential of DrAgent to improve the LLM capabilities in the medical domain, making high-performing, privacy-compliant AI tools feasible for resource-constrained medical institutions. Future directions include exploring more robust integrations with clinical tools, scaling to additional medical domains, and further evaluating the effectiveness of the system in real-world scenarios. In terms of technological contribution, we borrow the strengths of existing LLMs, which have been widely adopted in recent medical LLMs and VLMs (Nath et al., 2024; Xia et al., 2024a,b; Chen et al., 2024; Singhal et al., 2025; Labrak et al., 2024; Schmidgall et al., 2024). However, most existing works focus on improving their performance by performing extensive fine-tuning using large-scale datasets and computing resources, creating barriers to adoption in medical institutions with limited technical infrastructure. In this work, we propose enabling a lightweight LLM to deal with complex tasks by learning to collaborate with existing clinical tools.

Ethic Statements

We only use public data secondary and do not recruit any human research participants for this study. Our study was conducted on public datasets, in which all protected health information (e.g., patient name, sex, gender, and date of birth) is officially de-identified for all datasets used in our experiments. It means that the deletion of Protected Health Information (PHI) from structured data sources (e.g., database fields that provide age, genotypic information, past and current diagnosis and treatment categories) is performed in compliance with the Health Insurance Portability and Accountability Act (HIPAA) standards in order to facilitate public access to the datasets.

Acknowledgments

DAC was funded by an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; and the InnoHK Hong Kong Centre for Cerebro-cardiovascular Engineering (COCHE); and was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and the Pandemic Sciences Institute at the University of Oxford. The Applied Digital Health (ADH) group at the Nuffield Department of Primary Care Health Sciences is supported by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration Oxford and Thames Valley at Oxford Health NHS Foundation Trust. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. FL was funded by the Clarendon Fund and the Magdalen Graduate Scholarship. We sincerely thank all the editors and reviewers for their constructive comments and suggestions that substantially improved this paper.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440.

Kevin G Becker, Kathleen C Barnes, Tiffani J Bright, and S Alex Wang. 2004. The genetic association database. *Nature genetics*, 36(5):431–432.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. 2023a. Fireact: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and 1 others. 2024. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023b. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Dyke Ferber, Isabella C Wiest, Georg Wölflein, Matthias P Ebert, Gernot Beutel, Jan-Niklas Eckardt, Daniel Truhn, Christoph Springfeld, Dirk Jäger, and Jakob Nikolas Kather. 2024. Gpt-4 for information retrieval and comparison of medical oncology guidelines. *NEJM AI*, 1(6):AIcs2300235.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and 1 others. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ziniu Hu, Ahmet Iscen, Chen Sun, Kai-Wei Chang, Yizhou Sun, David Ross, Cordelia Schmid, and Alireza Fathi. 2024. Avis: Autonomous visual information seeking with large language model agent. *Advances in Neural Information Processing Systems*, 36.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W John Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, and 1 others. 2024. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *arXiv preprint arXiv:2402.13225*.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

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

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

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

Gokul Krishnan, Shiana Singh, Monika Pathania, Siddharth Gosavi, Shuchi Abhishek, Ashwin Parchani, and Minakshi Dhar. 2023. Artificial intelligence in clinical medicine: catalyzing a sustainable global healthcare paradigm. *Frontiers in Artificial Intelligence*, 6:1227091.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and 1 others. 2024. Mmedagent: Learning to use medical tools with multi-modal agent. *arXiv preprint arXiv:2407.02483*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Fenglin Liu, Shen Ge, and Xian Wu. 2021. Competence-based multimodal curriculum learning for medical report generation. In *Annual Meeting of the Association for Computational Linguistics*.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024. Large language models are poor clinical decision-makers: A comprehensive benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13696–13710, Miami, Florida, USA. Association for Computational Linguistics.

Fenglin Liu, Jinge Wu, Hongjian Zhou, Xiao Gu, Soheila Molaei, Anshul Thakur, Lei Clifton, Honghan Wu, and David A Clifton. 2025a. Riskagent: Autonomous medical ai copilot for generalist risk prediction. *arXiv preprint arXiv:2503.03802*.

Fenglin Liu, Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Yining Hua, Peilin Zhou, and 1 others. 2025b. Application of large language models in medicine. *Nature Reviews Bioengineering*, pages 1–20.

Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, and 1 others. 2023a. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, and 1 others. 2023b. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*.

Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, and 1 others. 2023c. Bolaa: Benchmarking and orchestrating llm-augmented autonomous agents. *arXiv preprint arXiv:2308.05960*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2024. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yucheng Tang, Pengfei Guo, and 1 others. 2024. Vila-m3: Enhancing vision-language models with medical expert knowledge. *arXiv preprint arXiv:2411.12915*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI. 2023a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2023b. Gpt-4 technical report. *Preprint at https://arxiv.org/abs/2303.08774*.

OpenAI. 2023c. https://platform.openai.com/docs/models/gpt-3-5-turbo.

Sajan B Patel and Kyle Lam. 2023. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108.

Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Jiang, Chengfei Lv, and Huajun Chen. 2024. AutoAct: Automatic agent learning from scratch for QA via self-planning. In *ACL*, pages 3003–3021.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

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

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

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

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. 2024. Large language models are poor medical coders—benchmarking of medical code querying. *NEJM AI*, 1(5):AIdbp2300040.

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 *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *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, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024a. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024b. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023a. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, and 1 others. 2024a. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2024b. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

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)*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. 2023. Agenttuning: Enabling generalized agent abilities for llms. *arXiv preprint arXiv:2310.12823*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Hongjian Zhou, Fenglin Liu, Jinge Wu, Wenjun Zhang, Guowei Huang, Lei Clifton, David Eyre, Haochen Luo, Fengyuan Liu, Kim Branson, and 1 others. 2025. A collaborative large language model for drug analysis. *Nature Biomedical Engineering*, pages 1–12.

A Medical Models

- Drug-drug Interaction: Analyzing potential interactions between medications using BioLinkBERT-Large (Yasunaga et al., 2022) to ensure safety.
- Drug Adverse Reaction: Identifying and predicting potential negative reactions caused by medications using BioLinkBERT-Large (Yasunaga et al., 2022) for improved patient care.
- Drug Recommendation: Providing tailored medication suggestions based on patient-specific conditions using BioLinkBERT-Large (Yasunaga et al., 2022).
- Named Entity Recognition: Extracting relevant medical entities, such as diseases, drugs, and procedures, from clinical texts using BioLinkBERT-Large (Yasunaga et al., 2022).
- Relation Extraction: Identifying relationships between medical entities (e.g., symptoms and diseases) in texts using BioLinkBERT-Large.
- Document Classification: Categorizing clinical documents or reports based on content using BioLinkBERT-Large (Yasunaga et al., 2022).
- Question Answering: Providing answers to medical queries using BioLinkBERT-Large (Yasunaga et al., 2022) for enhanced decision support.
- Radiology Report Generation: Automatically generating detailed radiology reports using Transformer-BASE (Yasunaga et al., 2022).
- Clinical Note Summarization: Summarizing lengthy clinical notes into concise, informative summary using Transformer-BASE (Yasunaga et al., 2022).