
MedChain: Bridging the Gap Between LLM Agents and Clinical Practice with Interactive Sequence

Jie Liu¹, Wenxuan Wang², Zizhan Ma³, Guolin Huang⁴, Yihang Su³,
Kao-Jung Chang^{6,7}, Haoliang Li¹, Linlin Shen^{4,5}, Michael Lyu³, Wenting Chen^{1*}

¹The City University of Hong Kong ²Renmin University of China

³The Chinese University of Hong Kong ⁴Shenzhen University

⁵Guangdong Provincial Key Laboratory of Intelligent Information Processing

⁶National Yang Ming Chiao Tung University ⁷Taipei Veterans General Hospital

Project: <https://github.com/ljwztc/MedChain>

Abstract

Clinical decision making (CDM) is a complex, dynamic process crucial to health-care delivery, yet it remains a significant challenge for artificial intelligence systems. While Large Language Model (LLM)-based agents have been tested on general medical knowledge using licensing exams and knowledge question-answering tasks, their performance in the CDM in real-world scenarios is limited due to the lack of comprehensive benchmark that mirror actual medical practice. To address this gap, we present MedChain, a dataset of 12,163 clinical cases that covers five key stages of clinical workflow. MedChain distinguishes itself from existing benchmarks with three key features of real-world clinical practice: personalization, interactivity, and sequentiality. Further, to tackle real-world CDM challenges, we also propose MedChain-Agent, an AI system that integrates a feedback mechanism and a MedCase-RAG module to learn from previous cases and adapt its responses. MedChain-Agent demonstrates remarkable adaptability in gathering information dynamically and handling sequential clinical tasks, significantly outperforming existing approaches.

1 Introduction

At the intersection of artificial intelligence and healthcare lies one of medicine's most complex challenges: Clinical Decision Making (CDM). In healthcare delivery, CDM demands not only the integration of diverse data sources and continuous assessment of evolving clinical scenarios, but also evidence-based judgments for diagnosis and treatment [46]. While crucial for optimal patient care, this intricate process imposes significant cognitive demands on healthcare professionals, making it an ideal candidate for AI assistance [42].

Recent advances in Large Language Model (LLM)-based agents [34, 49, 15, 44, 16, 71] have emerged as an effective solution for complex decision-making tasks, from software development [38] to office automation [57]. In the medical domain, these LLMs have demonstrated impressive performance on medical licensing exams [45, 36] and knowledge-based assessments [13, 10, 30]. While LLMs have consistently scored well above passing thresholds in these benchmark [45], it is crucial to recognize that these assessments fall short of capturing the complexity of real-world CDM, where errors can cascade through multiple decision stages, as illustrated in Figure 1. Based on our analysis, CDM exhibits three key characteristics.

*Wenting Chen is the corresponding author.

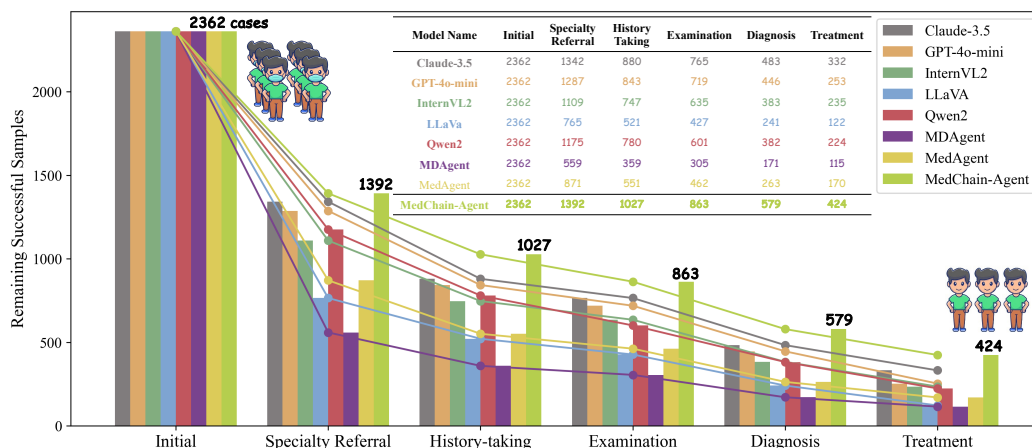


Figure 1: **Demonstration of error propagation of CDM in MedChain.** Starting with 2,362 initial cases, the diagram illustrates how diagnostic errors cascade through five clinical stages. Cases with incorrect diagnoses carry forward problematic information to subsequent stages, leading to a cumulative decrease in accuracy. After completing the treatment phase, we count cases that maintain correctness through each consecutive phase up to the each stage. Our MedChain-Agent achieves best performance in CDM comparing with other SOTA methods.

Firstly, these benchmarks rarely account for patient-specific information such as past medical history and present illness [36], which significantly influence clinical decisions in real clinical scenarios. This omission fails to capture the nuanced context that often shapes *personalized* diagnosis. **Secondly**, unlike real clinical scenarios where decisions build upon previous steps, existing benchmarks present clinical tasks as independent problems [40], missing the critical interdependencies in the diagnostic process. In reality, clinical decision-making is a *sequential* process where each step is contingent upon the preceding ones, and an error in one stage can profoundly impact subsequent decisions. **Thirdly**, most benchmarks present all relevant information upfront, providing a static, and comprehensive dataset [51]. However, real clinical workflow demand multiple rounds of dynamic information gathering through ongoing patient *interaction*.

MedChain: To address these critical gaps, we introduce MedChain, a novel benchmark designed to evaluate LLM-based agents in real-world clinical scenarios. Specifically, MedChain comprises 12,163 diverse cases spanning 19 medical specialties and 156 sub-categories, including 7,338 medical images with corresponding reports. Each case progresses through five crucial stages: specialty referral, history-taking, examination, diagnosis, and treatment. Unlike existing benchmarks, MedChain uniquely emphasizes three key features. 1) *Personalization*: Each case incorporates detailed patient-specific information. At first, agents are provided with only the patient's chief complaint and basic information. 2) *Interactivity*: Information must be actively gathered through dynamic consultation from patient. 3) *Sequentiality*: Decisions at each stage influence subsequent steps. Only after agent sequentially completes all five stages, the overall diagnostic process is evaluated.

MedChain-Agent: Given the novel features and challenges presented by this benchmark, existing agent frameworks struggle to address these issues adequately. To overcome these limitations, we propose MedChain-Agent, a multi-agent collaborative framework that enables LLMs with feedback mechanism and MedCase-RAG to dynamically gather information and handle sequential clinical tasks. Specifically, MedChain-Agent facilitates a synergistic interplay among three specialized agent types: General Agents for task-specific expertise, a Summarizing Agent for insight synthesis, and a Feedback Agent for iterative refinement. This multi-layered, iterative approach ensures decisions are products of thorough analysis and diverse perspectives. Additionally, to address the multifaceted nature of CDM, which demands the integration of evidence-based research, and patient-specific factors, we incorporate a novel MedCase-RAG module into our MedChain-Agent framework. Unlike conventional medical RAG methods, MedCase-RAG dynamically expand its database and employs a structured approach to data representation, mapping each medical case into a 12-dimensional feature vector. This system enables efficient retrieval of relevant cases and helps the model make informed decisions.

Our contributions are summarized as follows:

- We represent the first effort to propose a CDM benchmark, MedChain, providing a holistic assessment of diagnostic capabilities of LLM-based agents, closely reflecting real-world patient care.
- We propose a multi-agent framework based on the characteristics of CDM, called MedChain-Agent. This system enables efficient retrieval of relevant cases and helps the model make informed decisions.
- Through extensive experiments, we compare the performance of existing works on MedChain and the superiority of MedChain-Agent in CDM and reliability.

2 Related Works

2.1 Evaluation of LLM in Medicine

Benchmarking plays a vital role as a key performance indicator, directing model improvements, pinpointing weaknesses, and shaping the course of model evolution. The evaluation of LLMs in medicine has primarily focused on testing general medical knowledge through structured assessments [33]. Leading benchmarks such as MultiMedQA [45] integrate various medical QA datasets (e.g., MedQA [19], MedMCQA [36]), emphasizing performance on medical licensing examination materials. Other benchmarks like PubMedQA [20] focus on research-oriented queries, while several Chinese medical benchmarks [55, 2] evaluate models through multiple-choice questions from medical licensing exams. While these benchmarks effectively assess general medical knowledge, they fail to capture three critical aspects of real-world clinical decision-making (see Appendix Table 6), i.e., personalization in patient care, the interactive nature of clinical consultations, and the sequential dependency of medical decisions, where each step builds upon previous findings.

Recently, several benchmarks [59, 35, 28, 26, 24, 8] have been proposed to evaluate LLMs across diverse clinical scenarios and tasks, including information extraction, text summarization, and clinical outcome prediction. However, these benchmarks primarily consist of independent question-answer pairs, where each task is evaluated in isolation. Sequential decision-making is critical in medical practice, as each patient experiences a continuous journey from initial triage through treatment to recovery. Therefore, it is essential to evaluate how LLMs perform throughout this entire clinical pipeline. Our work distinguishes itself from existing benchmarks by focusing on sequential decision-making within interactive environments, specifically evaluating LLM performance in realistic clinical scenarios that require executing the complete patient care workflow.

2.2 LLM-based Agent in Medicine

LLM-based agents have demonstrated significant potential across various medical applications [53, 32], encompassing tasks such as medical examination questions, clinical diagnoses, and treatment plans. Recent research has explored different approaches: Agent Hospital [25] provides medical scenario simulation, while several frameworks [66, 47, 22] focus on specific medical stages with multi-agent architectures. Some works target specialized aspects, such as CoD [4] for interpretable diagnostics and EHRagent [43] for electronic health records (EHRs) analysis. Others, like Almanac Copilot [67], assist clinicians with EMR-specific tasks. AI Hospital [11] explores interactive clinical scenarios, but it falls short in handling multi-modal medical imaging and lacks a comprehensive benchmark for evaluating multi-agent performance. To enhance these agents' capabilities, researchers have integrated Retrieval-Augmented Generation (RAG), as demonstrated by MIRAGE [64]'s search-enhanced framework and Medical Graph RAG [61]'s knowledge-based approach.

However, current approaches face two major limitations. First, existing frameworks focus on isolated medical tasks rather than providing seamless integration across different clinical stages, making them insufficient for complex scenarios requiring effective inter-stage communication [14]. Second, current medical RAG systems' reliance on chunk-based indexing leads to context inconsistency and computational inefficiencies [18, 12], highlighting the need for more sophisticated approaches to medical knowledge integration.

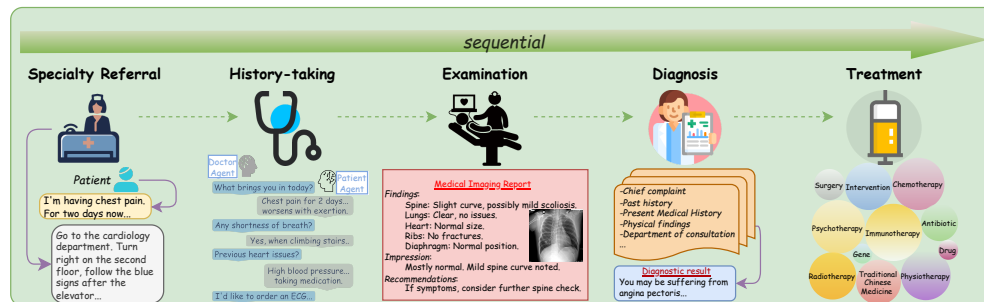


Figure 2: **MedChain Pipeline.** The MedChain is composed of a sequential medical process, including specialty referral, history-taking, examination, diagnosis, and treatment.

3 MedChain Benchmark

Overview. We introduce MedChain, a comprehensive clinical decision-making benchmark designed to simulate real-world scenarios. Built upon 12,163 Electronic Health Records (EHRs) spanning 19 medical specialties and 156 sub-categories, including 7,338 medical images with reports, MedChain uniquely emphasizes three key characteristics:

- **Personalization:** Each case includes detailed patient profiles that influence decision-making
- **Sequentiality:** Cases involve multiple interconnected decision-making stages
- **Interactivity:** Information must be actively gathered through dynamic consultation

3.1 Data Collection

Data Source and Processing. Our dataset is sourced from the Chinese medical website “iiYi”², which provides over 20,000 validated clinical cases spanning 28 disease categories. These cases are verified by professional doctors and have undergone de-identification to ensure patient privacy. We obtained formal permission from the website administrators to use the data for scientific research purposes. Each case typically contains the patient’s chief complaint, medical history, examination results, treatment process, and other relevant information, which insure the **personalization** of MedChain. Following the government standards³ and Medical Subject Headings⁴, we extracted and organized key information including patient basics, chief complaints, specialty referrals, examinations, imaging reports, diagnoses, and treatments. Cases with incomplete information were removed, resulting in 12,163 high-quality cases.

Quality Control. To ensure the highest standards of data integrity and clinical relevance in our benchmark, we implemented a rigorous quality control process involving a panel of five senior physicians, each with over 10 years of clinical experience. Our evaluation process examined a random sample of 6,000 cases (49.3% of the dataset). We developed a standardized scoring system that presents physicians with comprehensive case information alongside six binary quality dimensions: disease prevalence, clinical relevance, accuracy of patient history, appropriateness of diagnostic procedures, correctness of diagnosis, and suitability of treatment recommendations. Physicians evaluate each dimension through yes/no responses, with cases satisfying all dimensions considered valid.

The quality assessment yielded strong results, with 94.7% of evaluated cases meeting or exceeding our quality thresholds. Dimension-specific quality rates ranged from 92.9% to 97.2%, demonstrating consistently high standards across all evaluation criteria. Inter-rater reliability analysis produced a Cohen’s kappa coefficient of 0.82, indicating substantial agreement among our expert reviewers. Cases that failed to meet the thresholds (5.3%) were either revised or excluded from the final dataset to maintain benchmark integrity.

²<https://www.iiyi.com>

³National Standards for Medical Items

⁴<https://www.ncbi.nlm.nih.gov/mesh/1000048>

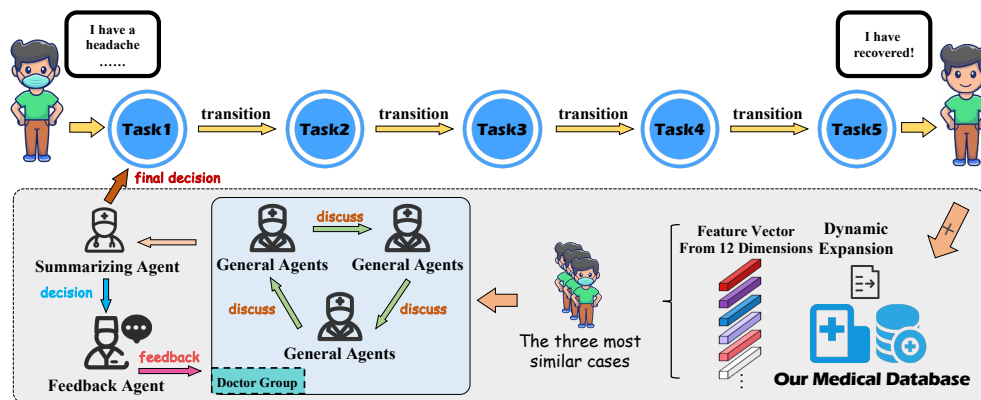


Figure 3: **MedChain-Agent framework.** Depicts a cyclical feedback medical multi-task system, where decisions are supported by retrieving similar past cases from a medical database.

3.2 Clinical Workflow Simulation

Sequential Stages. MedChain simulates the complete clinical workflow, comprising five sequential tasks, each representing a different stage of the clinical decision-making process, as shown in Figure 2. The results from each stage serve as inputs for the subsequent stage, creating a dependency where later decisions are influenced by the quality of earlier ones. This design guarantees the **sequentiality** of MedChain, mimicking the interconnected nature of real-world clinical decision-making processes. The pipeline consists of: 1) *Specialty Referral*: Assessment of case urgency and appropriate department selection; 2) *History-taking*: Dynamic information gathering through doctor-patient dialogue; 3) *Examination*: Medical image analysis and report generation; 4) *Diagnosis*: Comprehensive diagnosis based on accumulated information; 5) *Treatment*: Treatment plan formulation considering patient-specific factors.

The construction pipeline of MedChain is sketched out here. The standardization process and the format of each pipeline can be refer to Appendix A. We demonstrate an example in Appendix Figure 4 and Figure 5.

3.3 Interaction Environment

To simulate authentic doctor-patient consultations, we developed an **interactive** environment where LLM-based agents must actively gather information through dynamic interactions. We employ Gemma2 (9b) [50] as the patient agent, initializing it with pre-defined case information while withholding the actual diagnosis. Each tested LLM serves as a doctor simulator. This setup enables the agent to provide symptom information and respond to inquiries in a manner that mirrors real patient experiences, drawing inspiration from standardized patients in medical education [1]. We conduct systematic evaluation with senior physician to validate the effectiveness of patient agent. The details can be found in Appendix A.2.2.

3.4 Benchmark Evaluation

Given the complexity of medical decision-making, we developed a comprehensive evaluation framework that goes beyond simple binary assessments. The standard answer for each stage comprises multiple key points. The evaluated agent's score increases with the number of standard points addressed, while irrelevant content leads to score reduction. For each task, we employ specific evaluation metrics: Task 1 (Specialty Referral) uses accuracy and Intersection over Union (IoU) to compare predictions with ground truth. Task 2 (History-taking) calculates IoU between predicted and ground truth examination items. Task 3 leverages DocLens [63] for assessing image interpretation quality. Task 4 employs a carefully designed prompt based on National Health Commission of China guidelines to evaluate diagnostic accuracy. Task 5 measures IoU between predicted and ground truth examination items. We elaborate the evaluation metrics in Appendix A.

4 MedChain-Agent Framework

4.1 Multi-agent System with Feedback

The MedChain-Agent framework introduces a multi-agent system that simulates the complex, interconnected nature of medical decision-making. This system integrates specialized agents that contribute distinct expertise to the diagnostic and treatment process. We sketch the main content here and more details please refer to Appendix B.

4.1.1 Agent Roles

Our framework comprises three agent types:

General Agents: Task-specific agents recruited based on domain requirements. For specialty referral tasks, these agents possess expertise in medical specialties and triage protocols, enabling effective patient routing. They engage in collaborative discussions that mirror real-world medical consultations.

Summarizing Agent: Acts as a central coordinator, consolidating insights from general agents into coherent decisions. This agent synthesizes collective expertise and delivers final recommendations, similar to a senior physician's role in clinical settings.

Feedback Agent: Maintains decision quality through continuous evaluation and refinement. This agent assesses outputs, provides targeted feedback, and initiates improvements when necessary, creating a self-correcting decision process.

4.1.2 Decision Making with Feedback

The decision-making process begins with general agents analyzing patient data and engaging in structured discussions. The summarizing agent then consolidates these assessments into preliminary decisions, which undergo review by the feedback agent. When issues are identified, the feedback agent initiates an iterative refinement cycle, with general agents reconsidering their assessments and the summarizing agent adjusting decisions accordingly. This process continues until reaching consensus or completing a preset number of iterations.

4.2 MedCase-RAG

To enhance decision-making capabilities, we developed MedCase-RAG, a specialized Retrieval-Augmented Generation system for medical applications. Unlike traditional medical RAG systems, our approach employs structured case representation using feature vectors that encode 12 key clinical dimensions: Age, Sex, Chief Complaint, Symptoms, Medical History, Physical Examination, Laboratory Tests, Imaging Reports, Diagnosis, Treatment Plan, Primary Department, and Secondary Department. This feature vectors is extracted by text-embedding-ada-002 from openai. When processing new cases, MedCase-RAG performs similarity searches, identifying the three most similar cases using cosine similarity metrics. This approach provides agents with comprehensive reference points for decision-making. The MedCase-RAG also features dynamic knowledge base expansion, incorporating resolved cases as pseudo-data. This mechanism enables continuous learning from new clinical experiences, adapting to evolving medical knowledge and practices.

5 Experiments

5.1 Experimental Setup

We split the dataset into training, validation, and testing sets with a ratio of 7:1:2. Our study evaluates both single-agent and multi-agent systems. For base LLM, we test two closed-source models (gpt-4o-mini [34], and claude-3.5-sonnet [49]) and four open-source models (InternVL2-8b [6], llava-llama-3-8b-v1_1 [27], HuaTuoGPT [68], and Qwen2-7B-Instruct [52]), with model weights obtained from official Hugging Face repositories. In the single-agent evaluation, we compare with zero-shot manner, few-shot manner, CoT [58], Self-consistency [56] and RAG [23]. In the multi-agent evaluation, we compare MedChain-Agent against DyLAN [31], AutoGen [62], MedAgent [48] and MDAgent [21]. All agent framework are based on InternVL2-8b [6]. The deployment was conducted using the LMDeploy framework [7]. All tests executed on NVIDIA A100 GPUs featuring

Table 1: **Evaluation of various LLM-based agent in MedChain.** The best performance for each task is highlighted in **bold**.

Framework	Methods	Specialty Referral		History-taking	Examination	Diagnosis	Treatment	Average
		Level 1	Level 2					
Base LLM	GPT-4o-mini [34]	0.5449	0.2871	0.3399	0.5112	0.4177	0.3930	0.4156
	GPT-o3	0.5495	0.2866	0.3493	-	0.4891	0.3343	0.4017
	Claude-3.5-sonnet [49]	0.5681	0.3050	0.3562	0.5018	0.4207	0.4053	0.4262
	MedGemma [41]	0.5063	0.1975	0.3315	0.6324	0.4668	0.3558	0.4105
	LLaVA [27]	0.3240	0.0730	0.3182	0.5165	0.3667	0.1060	0.2841
	Qwen2 [52]	0.4975	0.2215	0.4226	0.4829	0.4530	0.2193	0.3828
	InternVL2 [52]	0.4811	0.1935	0.4645	0.4490	0.4367	0.2903	0.3859
	Baichuan [9]	0.2959	0.0532	0.3264	-	0.4297	0.2591	0.2728
	HuaTuoGPT [68]	0.0707	0.0207	0.3465	-	0.4163	0.1780	0.2064
	MedReason [60]	0.4602	0.1258	0.3315	-	0.4741	0.2993	0.3381
Single-agent	FineMedLM-o1 [65]	0.1502	0.0124	0.3969	-	0.2772	0.1345	0.1942
	Zero-shot	0.4811	0.1935	0.3450	0.4994	0.4572	0.2817	0.3763
	Few-shot	0.5584	0.2481	0.4870	0.3758	0.4382	0.3553	0.4105
	CoT [58]	0.5698	0.1484	0.3750	0.6396	0.4382	0.3421	0.4189
	RAG [23]	0.5927	0.2467	0.4732	0.6408	0.4167	0.4524	0.4704
	Self-consistency [56]	0.5143	0.2319	0.4213	0.4144	0.4198	0.3441	0.3910
Multi-agent	DyLAN [31]	0.4415	0.1731	0.4434	0.4980	0.3863	0.2983	0.3734
	AutoGen [62]	0.5228	0.2347	0.4528	0.4559	0.4250	0.3590	0.4084
	MedAgent [48]	0.3830	0.2039	0.4454	0.4456	0.4102	0.3673	0.3759
	MDAgent [21]	0.2398	0.1343	0.4240	0.4983	0.3973	0.3620	0.3426
	MDAgent + RAG [21]	0.4754	0.2063	0.4412	0.5346	0.4198	0.4371	0.4190
	MedChain-Agent	0.5873	0.3505	0.5836	0.6566	0.4807	0.4613	0.5200

Table 2: **Performance Comparison of LLM-based Agents across Other Diagnosis Benchmarks.** Implementation of agent frameworks based on InternVL2, evaluated on multiple medical benchmarks including MedQA [19], PubMedQA [20], PathVQA [17], and MedBullets [3].

Framework	Method	MedQA	PubMedQA	PathVQA	MedBullets	Average
Single-agent	Zero-shot	0.426	0.668	0.449	0.490	0.508
	Few-shot	0.477	0.648	0.448	0.503	0.519
	CoT	0.470	0.714	0.465	0.500	0.537
	Self-consistency	0.460	0.688	0.482	0.500	0.533
Multi-agent	MedAgents	0.501	0.622	0.569	0.435	0.532
	MDAgents	0.435	0.744	0.582	0.422	0.546
	AutoGen	0.395	0.656	0.568	0.448	0.517
	DyLAN	0.414	0.610	0.540	0.448	0.503
	MedChain-Agent	0.462	0.746	0.621	0.474	0.576

80GB of memory. To enhance output stability and reliability across all experiments, we consistently set the temperature parameter to 0. The experiments were conducted in Chinese.

5.2 Benchmark Performance Results

The results of our evaluation in the MedChain are presented in Table 1. Our analysis yields two significant insights:

(1) *Sequential decision-making tasks continue to pose significant challenges, even for advanced models.* For instance, within the single-agent frameworks, GPT-4o-mini and InternVL2 achieve average scores of 0.4156 and 0.3859, respectively. These results indicate that despite their sophistication, these models struggle to maintain consistent performance across the sequential stages of clinical decision-making, highlighting the inherent difficulty of these tasks.

(2) *The integration of the MedChain-Agent framework with open-source LLMs demonstrates significant superiority over proprietary models like GPT-4o-mini.* The substantial performance gain observed with MedChain-Agent (average score of 0.5200) implies that our framework can leverage the strengths of open-source LLMs to achieve superior outcomes. This suggests that open-source

Table 3: **Ablation Study for Key Components for MedChain-Agent.** This table presents the performance impact of sequentially removing the Feedback mechanism and MedCase-RAG from the full MedChain-Agent framework.

Feedback	MedCase-RAG	Specialty referral		History-taking	Examination	Diagnosis	Treatment	Average
		Level 1	Level 2					
		0.5523	0.2228	0.3285	0.6369	0.4724	0.3915	0.4341
✓		0.5739	0.2906	0.4222	0.6377	0.4299	0.4209	0.4692
	✓	0.5928	0.3353	0.5801	0.6488	0.4699	0.4568	0.5140
✓	✓	0.5873	0.3505	0.5836	0.6566	0.4804	0.4613	0.5200

Table 4: **Generalizability of MedChain-Agent with Various Base LLM.** We apply the MedChain-Agent framework to different base LLMs to validate its generalizability. Performance changes are highlighted with light green for improvements and light red for decreases. Results demonstrate that the MedChain-Agent framework consistently brings performance gains across different base LLMs.

Base LLM	Specialty Referral		History-taking	Examination	Diagnosis	Treatment	Average
	Level 1	Level 2					
GPT-4o-mini	0.6065 (+0.0616)	0.2302 (-0.0569)	0.6205 (+0.2806)	0.6279 (+0.1167)	0.4975 (+0.0798)	0.4329 (+0.0399)	0.5026 (+0.0870)
HuaTuoGPT	0.2554 (+0.1847)	0.0097 (-0.0110)	0.3722 (+0.0257)	-	0.4159 (-0.0004)	0.1450 (-0.0330)	0.2396 (+0.033)
Qwen2	0.5818 (+0.0843)	0.2781 (+0.0566)	0.5962 (+0.1736)	0.6534 (+0.1705)	0.4628 (+0.0098)	0.4315 (+0.2122)	0.5006 (+0.1178)
InternVL2	0.5873 (+0.1062)	0.3505 (+0.1570)	0.5836 (+0.1191)	0.6566 (+0.2076)	0.4806 (+0.0439)	0.4613 (+0.1710)	0.5200 (+0.1341)

models, when enhanced with our framework, are not only competitive but can also excel in handling intricate medical decision-making tasks.

5.3 Evaluation in Existing Diagnosis Datasets

Moreover, we evaluate MedChain-Agent on several well-established medical QA datasets, including MedQA [19], PubMedQA [20], PathVQA [17], and MedBullets [3]. As shown in Table 2, our framework demonstrates strong performance compared to both single-agent baselines and multi-agent alternatives. MedChain-Agent achieves the highest average score (0.576). This consistent superiority across various medical QA benchmarks further validates the effectiveness of our framework, even on simpler, more structured tasks that differ from our real-world clinical scenarios.

5.4 Ablation Studies and Discussion

(1) *Ablation Study for Key components in MedChain-Agent:* To assess our framework’s components, we conduct ablation studies as shown in Table 3. Both the Feedback mechanism and MedCase-RAG module demonstrate significant individual contributions to overall performance. The Feedback mechanism alone improves the average score from 0.4341 to 0.4692, with notable gains in History-taking and Level 2 Specialty referral. MedCase-RAG shows stronger individual impact, boosting the average score to 0.5140, with substantial improvements in History-taking and Specialty referral tasks. While MedCase-RAG excels in diagnostic phases, the Feedback mechanism appears more beneficial for Treatment tasks. When combined, these components show synergistic effects, achieving the highest average performance (0.5200) and optimal scores across five of six evaluated tasks, validating their complementary roles in enhancing clinical reasoning capabilities.

(2) *Ablation Study for Three Key Characteristics in MedChain:* To validate the effectiveness of personalization, interactivity, and sequentiality within our benchmark, we conduct an ablation study as shown in Table 5. We systematically remove each characteristic and observe its impact on model performance across Diagnosis and Treatment tasks. ‘w/o Person’ means all detailed patient profiles are omitted from the input, resulting in a lack of personalized information that makes correct diagnosis more challenging for the model. ‘w/o Seq’ means we use the ground truth from the previous stage as input to the next stage, rather than using the model’s previous output, which simplifies the benchmark. ‘w/o Inter’ means we directly provide all patient examination results as input without requiring the agent to autonomously inquire about the patient’s condition, also simplifying the benchmark. Removing patient-specific information (w/o Person.) consistently degrades performance across all models in diagnosis tasks (with drops ranging from 2.51% to 10.31%), demonstrating that personalized information is crucial for accurate clinical decision-making.

Table 5: **The ablation study for three key characteristics in MedChain.** This table presents the impact of personalization, interactivity, and sequentiality on Diagnosis and Treatment tasks. The arrows $\uparrow\downarrow$ next to settings indicate expected performance change direction. The arrows next to results show actual changes (highlighted in gray when matching expectations).

Setting	Model	Diagnosis	Treatment
Full	MedAgent	0.4106	0.3673
	MDAgent	0.3959	0.3620
	gpt-4o-mini	0.4157	0.3930
	InternVL2	0.4378	0.4472
	MedChain-Agent	0.4802	0.4613
a) w/o Person. $\downarrow$	MedAgent	0.3075 $\downarrow$	0.3754 $\uparrow$
	MDAgent	0.3283 $\downarrow$	0.3109 $\downarrow$
	gpt-4o-mini	0.3906 $\downarrow$	0.3406 $\downarrow$
	InternVL2	0.3702 $\downarrow$	0.2527 $\downarrow$
	MedChain-Agent	0.4159 $\downarrow$	0.4310 $\downarrow$

Setting	Model	Diagnosis	Treatment
b) w/o Seq. $\uparrow$	MedAgent	0.4030 $\downarrow$	0.4456 $\uparrow$
	MDAgent	0.4497 $\uparrow$	0.4418 $\uparrow$
	gpt-4o-mini	0.4522 $\uparrow$	0.4423 $\uparrow$
	InternVL2	0.4481 $\uparrow$	0.2903 $\downarrow$
	MedChain-Agent	0.4807 $\uparrow$	0.4743 $\uparrow$
c) w/o Inter. $\uparrow$	MedAgent	0.3129 $\downarrow$	0.4109 $\uparrow$
	MDAgent	0.3998 $\uparrow$	0.3627 $\uparrow$
	gpt-4o-mini	0.4663 $\uparrow$	0.4003 $\uparrow$
	InternVL2	0.4550 $\uparrow$	0.3173 $\downarrow$
	MedChain-Agent	0.4634 $\downarrow$	0.5207 $\uparrow$

Interestingly, when removing the sequential dependency between stages (*w/o Seq.*), most models show improved performance, indicating that sequential decision-making poses greater challenges that better reflect real-world clinical scenarios. Similarly, the improved performance observed after removing interactive information gathering (*w/o Inter.*) confirms the effectiveness of interactivity in our benchmark design. These results collectively suggest that while both sequentiality and interactivity make the benchmark more challenging, they are essential components that better simulate the complexity of real-world clinical decision-making processes.

(3) *Generalizability for MedChain-Agent*: To evaluate the generalizability of our framework, we apply MedChain-Agent to various base LLMs as shown in Table 4. The results demonstrate substantial performance improvements across most models and tasks. InternVL2 shows the most significant enhancement with a 13.41% average improvement, followed by Qwen2 (11.78%) and GPT-4o-mini (8.70%). Even with HuaTuoGPT, which has relatively lower baseline performance, our framework still achieves a 3.33% improvement. Notably, the improvements are particularly pronounced in complex tasks such as History-taking and Examination, where gains of up to 28.06% are observed. These consistent enhancements across diverse models and tasks strongly validate the framework’s robust generalizability and effectiveness in medical decision-making scenarios.

6 Conclusion

In this paper, we introduced MedChain, a novel benchmark for evaluating LLM-based agents in clinical decision-making that authentically reflects real-world medical practice through three essential characteristics: personalization, interactivity, and sequentiality. Our comprehensive dataset encompasses 12,163 diverse clinical cases across 19 medical specialties, including 7,338 medical images with corresponding reports, providing a robust foundation for evaluating AI systems in complex healthcare scenarios. To address the challenges presented by this benchmark, we also introduced MedChain-Agent, a multi-agent framework enhanced by feedback mechanisms and MedCase-RAG, which demonstrates superior performance across various clinical tasks. This work establishes new benchmarks for evaluating medical AI systems and provides practical solutions for enhancing their clinical decision-making capabilities. As AI continues to evolve in healthcare applications, frameworks that can navigate the complexities of real-world clinical scenarios will be increasingly valuable for improving patient care and supporting healthcare professionals.

7 Limitations

This paper has two primary limitations that offer avenues for future research:

1) *Data Source Diversity*: The MedChain benchmark is constructed from 12,163 electronic health records obtained from the Chinese medical website “iiYi.” Although this dataset is extensive and covers 19 medical specialties and 156 sub-categories, it is derived from a single source. Additionally,

there exists a notable imbalance in data distribution across different medical specialties, with some departments having significantly more cases than others. This imbalance may introduce bias in model evaluation and limit performance in underrepresented specialties. In our future work, we will incorporate additional data sources from different regions or healthcare systems to further enrich the dataset and address the specialty imbalance.

2) Patient Interaction Simulation: In our interactive environment, the patient responses during the history-taking stage are generated by the Gemma 2 language model. While this approach ensures consistency and control in evaluating the LLM-based agent, the real patient interactions can be more varied and complex. Future work could explore more advanced patient simulators or incorporate real dialogue data to capture a wider range of communication styles and behaviors.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China under Grant 82261138629 and 12326610, and Guangdong Provincial Key Laboratory under Grant 2023B1212060076

References

- [1] Howard S. Barrows. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *aamc. Academic Medicine*, 68:443–51, 1993.
- [2] Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. Medbench: A large-scale chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17709–17717, 2024.
- [3] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*, 2024.
- [4] Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*, 2024.
- [5] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyang Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *arXiv preprint arXiv:2408.03361*, 2024.
- [6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.
- [7] LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. <https://github.com/InternLM/lmdeploy>, 2023.
- [8] Meidan Ding, Jipeng Zhang, Wenxuan Wang, Haiqin Zhong, Xiaoqin Wang, Xinheng Lyu, Wenting Chen, and Linlin Shen. Eagle: Expert-guided self-enhancement for preference alignment in pathology large vision-language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14603–14619, 2025.
- [9] Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiayuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, et al. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*, 2025.
- [10] Alexander V Eriksen, Sören Möller, and Jesper Ryg. Use of gpt-4 to diagnose complex clinical cases, 2023.

- [11] Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv preprint arXiv:2402.09742*, 2024.
- [12] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.
- [13] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 9(1):e45312, 2023.
- [14] Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, page 895–943, Feb 2022.
- [15] Yu Gu, Xiang Deng, and Yu Su. Don’t generate, discriminate: A proposal for grounding language models to real-world environments. In *ACL*, 2023.
- [16] Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems*, 36:79081–79094, 2023.
- [17] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.
- [18] Yizheng Huang and Jimmy Huang. A survey on retrieval-augmented text generation for large language models, 2024.
- [19] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- [20] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.
- [21] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Adaptive collaboration strategy for llms in medical decision making. *arXiv preprint arXiv:2404.15155*, 2024.
- [22] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeon-hoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. Mdagents: An adaptive collaboration of llms for medical decision-making, 2024.
- [23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [24] Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, et al. Towards a holistic framework for multimodal llm in 3d brain ct radiology report generation. *Nature Communications*, 16(1):2258, 2025.
- [25] Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents, 2024.
- [26] Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, et al. Wsi-llava: A multimodal large language model for whole slide image. *arXiv preprint arXiv:2412.02141*, 2024.
- [27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

- [28] Jie Liu, Wenxuan Wang, Su Yihang, Jingyuan Huang, Yudi Zhang, Cheng-Yi Li, Wenting Chen, Xiaohan Xing, Kao-Jung Chang, Linlin Shen, et al. Asclepius: A spectrum evaluation benchmark for medical multi-modal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24181–24201, 2025.
- [29] Mianxin Liu, Jinru Ding, Jie Xu, Weiguo Hu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, et al. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *arXiv preprint arXiv:2407.10990*, 2024.
- [30] Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. A comprehensive evaluation of multi-modal large language models for endoscopy analysis. *arXiv preprint arXiv:2505.23601*, 2025.
- [31] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*, 2023.
- [32] Xinheng Lyu, Yuci Liang, Wenting Chen, Meidan Ding, Jiaqi Yang, Guolin Huang, Daokun Zhang, Xiangjian He, and Linlin Shen. WSI-Agents: A Collaborative Multi-Agent System for Multi-Modal Whole Slide Image Analysis . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15964. Springer Nature Switzerland, September 2025.
- [33] Zizhan Ma, Wenxuan Wang, Guo Yu, Yiu-Fai Cheung, Meidan Ding, Jie Liu, Wenting Chen, and Linlin Shen. Beyond the leaderboard: Rethinking medical benchmarks for large language models. *arXiv preprint arXiv:2508.04325*, 2025.
- [34] OpenAI. Gpt-4v(ision) system card. Technical report, OpenAI, 2023.
- [35] Zetian Ouyang, Yishuai Qiu, Linlin Wang, Gerard De Melo, Ya Zhang, Yanfeng Wang, and Liang He. Climedbench: A large-scale chinese benchmark for evaluating medical large language models in clinical scenarios. *arXiv preprint arXiv:2410.03502*, 2024.
- [36] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.
- [37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [38] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, 2024.
- [39] Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384, 2024.
- [40] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.
- [41] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.
- [42] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. " the human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 99–109, 2020.

- [43] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May D Wang. Ehragent: Code empowers large language models for complex tabular reasoning on electronic health records. *arXiv preprint arXiv:2401.07128*, 2024.
- [44] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [45] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [46] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17, 2020.
- [47] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning, 2024.
- [48] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023.
- [49] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [50] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- [51] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):A10a2300138, 2024.
- [52] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [53] Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. A survey of LLM-based agents in medicine: How far are we from baymax? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10345–10359, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [54] Wenxuan Wang, Yihang Su, Jingyuan Huan, Jie Liu, Wenting Chen, Yudi Zhang, Cheng-Yi Li, Kao-Jung Chang, Xiaohan Xin, Linlin Shen, et al. Asclepius: A spectrum evaluation benchmark for medical multi-modal large language models. *arXiv preprint arXiv:2402.11217*, 2024.
- [55] Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*, 2023.
- [56] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- [57] Zilong Wang, Yuedong Cui, Li Zhong, Zimin Zhang, Da Yin, Bill Yuchen Lin, and Jingbo Shang. Officebench: Benchmarking language agents across multiple applications for office automation. *arXiv preprint arXiv:2407.19056*, 2024.

- [58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [59] Chaoyi Wu, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards evaluating and building versatile large language models for medicine. *npj Digital Medicine*, 8(1):58, 2025.
- [60] Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.
- [61] Junde Wu, Jiayuan Zhu, and Yunli Qi. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation, 2024.
- [62] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.
- [63] Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. Doclens: Multi-aspect fine-grained medical text evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–679, 2024.
- [64] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine, 2024.
- [65] Hongzhou Yu, Tianhao Cheng, Ying Cheng, and Rui Feng. Finemedlm-o1: Enhancing the medical reasoning ability of llm from supervised fine-tuning to test-time training. *arXiv preprint arXiv:2501.09213*, 2025.
- [66] Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning, 2024.
- [67] Cyril Zalka, Joseph Cho, Gracia Fahed, Rohan Shad, Michael Moor, Robyn Fong, Dhamanpreet Kaur, Vishnu Ravi, Oliver Aalami, Roxana Daneshjou, et al. Almanac copilot: Towards autonomous electronic health record navigation. *arXiv preprint arXiv:2405.07896*, 2024.
- [68] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023.
- [69] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [70] Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [71] Yuchen Zhuang, Xiang Chen, Tong Yu, Saayan Mitra, Victor Bursztn, Ryan A Rossi, Somdeb Sarkhel, and Chao Zhang. Toolchain*: Efficient action space navigation in large language models with a* search. In *The Twelfth International Conference on Learning Representations*.

Appendix for MedChain

Abstract.

Appendix A describes the process of standardizing and organizing the dataset for the MedChainbenchmark.

Appendix B provides a detailed explanation of the MedChain-Agent framework, including its implementation, feedback mechanism, and the novel Retrieval-Augmented Generation (RAG) approach used to enhance decision-making.

Appendix C lists the additional details for experiment.

Appendix D discuss the difference between MedChain and several similar related works.

A Benchmark Construction and Evaluation

A.1 Dataset Standardization

We employed a combination of large language models and human verification to label data across different tasks. Our methodology involves task-specific prompt construction and output matching to ensure data quality and diversity while maintaining alignment with real clinical case scenarios.

To ensure consistency and comparability across the benchmark, we standardized the classification of examination items into two main categories: Physical Examinations and Auxiliary Examinations. Physical Examinations include evaluations of various body systems and general health indicators, while Auxiliary Examinations encompass different imaging techniques and laboratory tests. We utilized GPT-4o to extract and classify examination items from each case, followed by manual verification to ensure accuracy. For medical imaging, we classified images into seven types, and manual review ensured the correctness of the classifications. Additionally, treatment items were extracted and categorized from each case. This standardization process ensures that the dataset is consistent, facilitating accurate and comparable evaluations of LLM performance. Figure 4 and Figure 5 demonstrate a case after standardization in English and Chinese. Figure 6 shows the statistics of different departments.

A.2 Tasks Details

This section provides a comprehensive description of the five specific tasks that make up the MedChainbenchmark. Each subsection elaborates on a particular task, detailing its input, output, and evaluation methods. Figure 9 visualizes the differences between MedChainand other methods.

A.2.1 Task 1: Specialty referral

The specialty referral task evaluates the LLM's ability to assess the urgency of a patient's condition and determine the appropriate department based on the patient's chief complaint. The input consists of the patient's chief complaint, and the output space includes 19 first-level departments and 156 second-level departments. The LLM must first assign the patient to one of the 19 first-level departments, then to one or more of the 156 second-level departments based on the primary symptoms.

Evaluation metrics include accuracy for first-level department assignment and Intersection over Union (IoU) for second-level department assignment. IoU is used for second-level departments to account for the possibility of multiple correct assignments and to reflect partial correctness, which can occur in complex cases. This metric better captures the nuanced nature of departmental referrals in clinical practice.

A.2.2 Task 2: History-taking

The history-taking task is designed to simulate doctor-patient communication, where the goal is to obtain relevant information and infer necessary examination items. We employ a multi-agent system to evaluate this process:

```

{
  "tags": {
    "department": ["Pediatrics", "Pediatric Immunology"],
    "disease": ["Uterine fibroids", "Constipation"]
  },
  "[Case Introduction]": {
    "Physical Examination": {
      "General Examination": "Heart and lungs normal",
      "Urogenital System Examination": "Cervical erosion grade II, uterus enlarged to the size of a duck egg, left adnexa palpable mass size of a goose egg"
    },
    "Auxiliary Examinations": {
      "CT": "CT plain scan: A mixed-density lesion in the middle of the pelvis, about 10.0cm×7.0cm×6.5cm, with uneven internal density, CT value -146~45HU, and a round high-density focus, CT value 623HU (Fig 1, 2), with smooth edges. Uterus enlarged, rounded, protruding slightly anteriorly, with uneven internal density, CT value 32~48HU. Intrauterine device visible in the uterine cavity, bladder compressed and displaced. Imaging diagnosis: Uterine fibroids, ovarian teratoma.",
      "Ultrasound": "Color Doppler: A 10.0cm×6.0cm cystic-solid mass visible anterior to the uterus, with heterogeneous internal echo. A medium-echo mass visible on the right side, with a strong echo focus inside, rich color blood flow can be drawn out inside and around. The left side is cystic, with chaotic internal echo, cord-like strong echo visible. Right ovary normal",
      "Blood Test": "Alkaline phosphatase: 89u/L (normal range 100~290u/L)",
      "Pathological Examination": "Tumor diameter 8.0cm, intact capsule, smooth surface; partly cystic diameter about 3.0cm, containing fat and hair, solid area gray-yellow on section, soft and brittle. Another uterine upper segment 5.0cm×5.0cm×4.0cm, intramural tumor tissue about 2.0cm in diameter, yellowish-white on section, hard and tough. Microscopy: Rich fibrous stroma separating long, wavy, parallel cell bands composed of columnar cells; elongated nuclei perpendicular to the long axis of the cell bands (Fig 4). Intracystic lipid seen, cyst wall with squamous epithelium and sebaceous glands (Fig 3). Another tumor cells spindle-shaped, rod-shaped nuclei, interlaced arrangement, cells without atypia (Fig 5). Immunohistochemistry: inhibin (-), ck (+), cy (+). Pathological diagnosis: Ovarian carcinoid (trabecular type) with mature cystic teratoma, uterine leiomyoma."
    },
    "Chief Complaint": ["Patient is 8 years old", "Constipation for nearly 1 year, weight loss of 20 pounds in the past half year. Regular menstruation, no frequent urination"],
    "Images": [
      {"filename": "133.png", "category": ["Pathological Examination"]},
      {"filename": "131.png", "category": ["CT"]},
      {"filename": "132.png", "category": ["Pathological Examination"]}
    ],
    "Imaging Reports": {
      "CT": "CT plain scan: A mixed-density lesion in the middle of the pelvis, about 10.0cm×7.0cm×6.5cm, with uneven internal density, CT value -146~45HU, and a round high-density focus, CT value 623HU (Fig 1, 2), with smooth edges. Uterus enlarged, rounded, protruding slightly anteriorly, with uneven internal density, CT value 32~48HU. Intrauterine device visible in the uterine cavity, bladder compressed and displaced. Imaging diagnosis: Uterine fibroids, ovarian teratoma.",
      "Pathological Examination": "Tumor diameter 8.0cm, intact capsule, smooth surface; partly cystic diameter about 3.0cm, containing fat and hair, solid area gray-yellow on section, soft and brittle. Another uterine upper segment 5.0cm×5.0cm×4.0cm, intramural tumor tissue about 2.0cm in diameter, yellowish-white on section, hard and tough. Microscopy: Rich fibrous stroma separating long, wavy, parallel cell bands composed of columnar cells; elongated nuclei perpendicular to the long axis of the cell bands (Fig 4). Intracystic lipid seen, cyst wall with squamous epithelium and sebaceous glands (Fig 3). Another tumor cells spindle-shaped, rod-shaped nuclei, interlaced arrangement, cells without atypia (Fig 5). Immunohistochemistry: inhibin (-), ck (+), cy (+). Pathological diagnosis: Ovarian carcinoid (trabecular type) with mature cystic teratoma, uterine leiomyoma."
    },
    "[Diagnosis and Treatment Process]": {
      "Initial Diagnosis": ["Uterine fibroids, ovarian teratoma."],
      "Treatment Course": ["Intramural fibroid nodules visible in the uterus; left ovary enlarged, multilocular, with intact and smooth capsule. Left fallopian tube normal, right adnexa normal. Subtotal hysterectomy and left adnexectomy performed. Pathological examination: Tumor diameter 8.0cm, intact capsule, smooth surface; partly cystic diameter about 3.0cm, containing fat and hair, solid area gray-yellow on section, soft and brittle. Another uterine upper segment 5.0cm×5.0cm×4.0cm, intramural tumor tissue about 2.0cm in diameter, yellowish-white on section, hard and tough. Microscopy: Rich fibrous stroma separating long, wavy, parallel cell bands composed of columnar cells; elongated nuclei perpendicular to the long axis of the cell bands (Fig 4). Intracystic lipid seen, cyst wall with squamous epithelium and sebaceous glands (Fig 3). Another tumor cells spindle-shaped, rod-shaped nuclei, interlaced arrangement, cells without atypia (Fig 5). Immunohistochemistry: inhibin (-), ck (+), cy (+). Pathological diagnosis: Ovarian carcinoid (trabecular type) with mature cystic teratoma, uterine leiomyoma."
    },
    "[Treatment Items]": ["Surgery"]
  }
}

```

Figure 4: **Case Report "77_Ovarian Carcinoid with Mature Cystic Teratoma: A Case Report."**

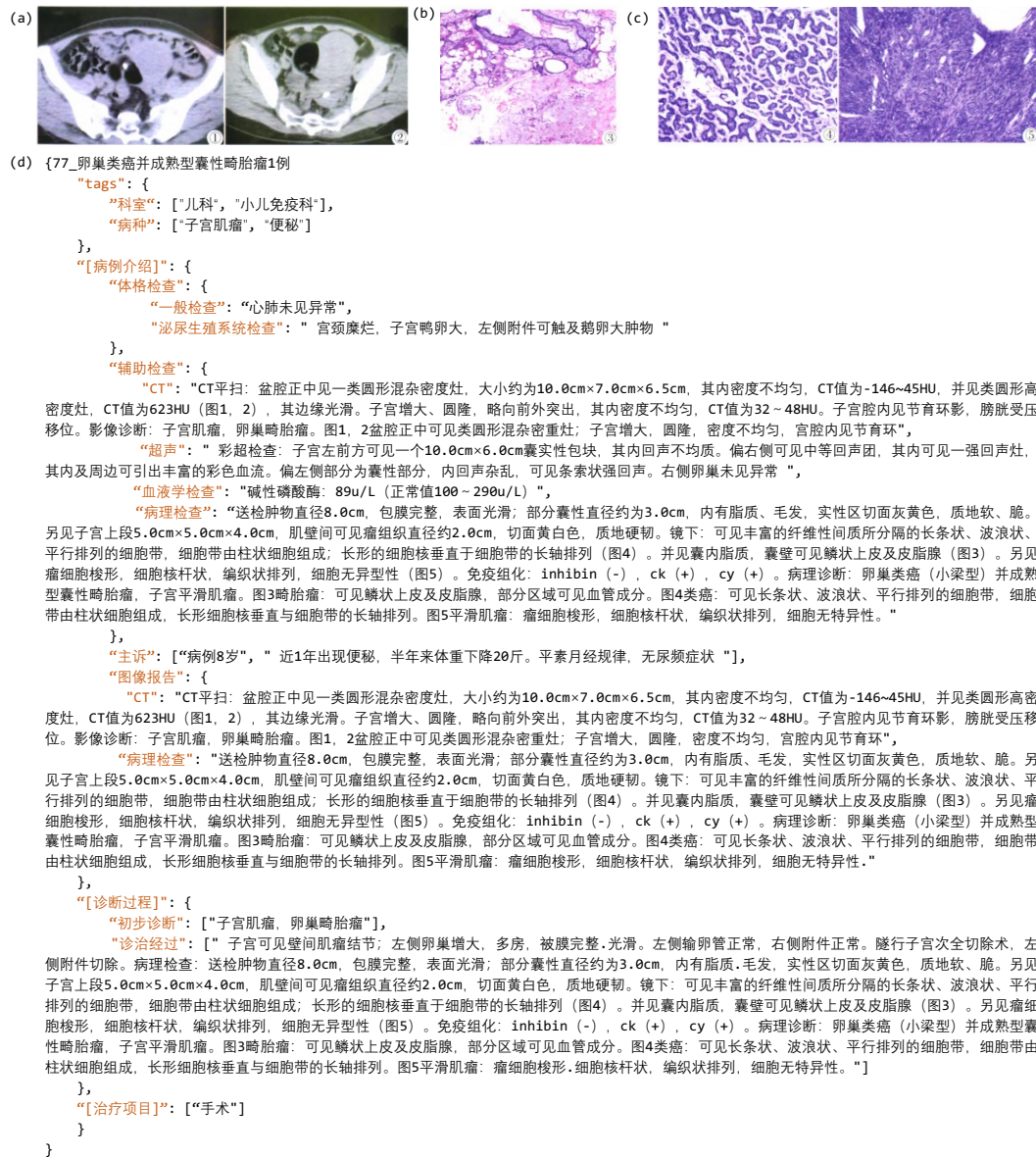


Figure 5: Case Report Chinese Version with Corresponding Medical Imaging. "77_Ovarian Carcinoid with Mature Cystic Teratoma: A Case Report." (a), (b) and (c) Medical Imaging. (d) Chinese version.

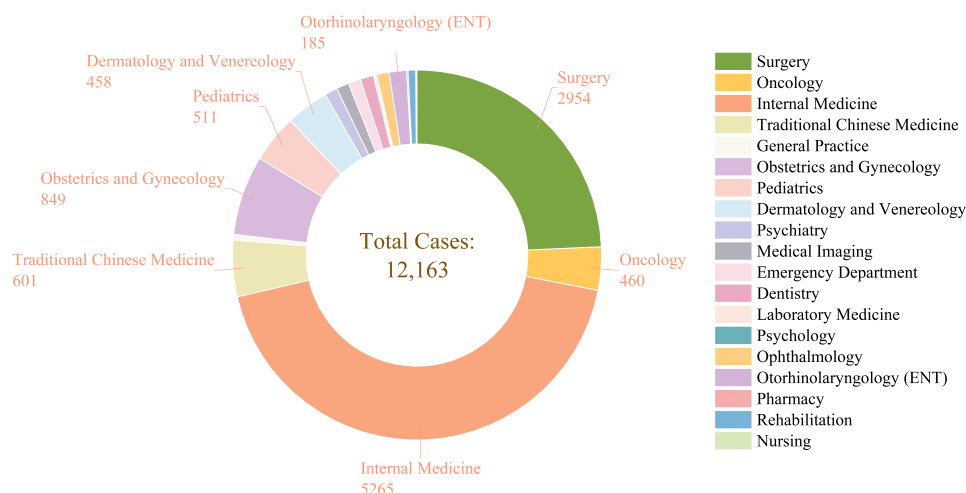


Figure 6: Multi-departmental Distribution

- **Doctor Agent:** The LLM being evaluated plays this role, asking questions and suggesting examinations based on the patient's responses.
- **Patient Agent:** A local large model (such as Google's gemma2/9b) simulates the patient, responding based on pre-defined case information. This includes the patient's chief complaint, medical history, and examination results.

Prompt for Patient: *You are to role-play as a Standardized Patient. Here is your case information: {patient information}. You need to answer the doctor's questions directly based on the case information (do not fabricate doctor-patient dialogues). Note that unless the doctor explicitly asks about physical examination and auxiliary examination findings, please do not proactively mention or inquire about physical examination and auxiliary examination related content. If the doctor asks about content that does not exist in the case, please indicate that you don't know and avoid fabricating information. At all times, remember that you are only playing the role of a standardized patient.*

Prompt for Doctor: *You are a doctor. The patient's chief complaint is as follows: {chief complaints}. You need to gather more information through conversation with the patient. Physical examination includes: general examination (including height, weight, temperature, blood pressure, pulse, etc.), head, eyes, ears, nose, and throat examination, neck examination (including thyroid, cervical lymph nodes), chest examination (including lungs, heart), abdominal examination, spine and limb examination, skin examination, neurological examination, and genitourinary system examination. Auxiliary examinations include: X-ray, MRI, CT, ultrasound, nuclear medicine imaging, hematological tests, urine tests, stool tests, endoscopic examination, and pathological examination. After obtaining certain information (such as past history and present illness history), please select the physical examinations and auxiliary examinations to inquire about based on the patient's condition. Ask about only one or two items per conversation round. Please inquire about as many examination items as possible (ask at least one item each for physical examination and auxiliary examination) until you can determine the condition. At the end of the conversation, please say "Wishing you a speedy recovery."*

Evaluation metrics is the IoU between the predicted examination items and the ground truth set. This patient-agent design simulates a realistic clinical history-taking process, allowing the doctor-LLM to demonstrate its ability to ask relevant follow-up questions, interpret patient responses, and determine appropriate examinations. The use of a local large model as the Patient Agent ensures consistency in evaluations and improves the reproducibility of the benchmark.

Systematic Evaluation: Furthermore, we conducted a systematic evaluation with a senior physician (10+ years clinical experience) across 10 simulated cases to validate the patient agent's effectiveness. The evaluation focused on three key dimensions: medical history accuracy (match rate between simulated and original EHR data), symptom consistency (temporal coherence of symptom pro-

Evaluation Prompt

Extract Reference Claims

Next, I will give you an imaging report. Please divide it into several statements, ensuring there are no omissions or repetitions between the statements. Output the resulting statements in the form of a list. The output format should be: ["Statement 1", "Statement 2", "Statement 3", ...]

Imaging report:

Entailment Evaluation

Please evaluate whether the following AI assistant-provided imaging report can fully include each of the following statements. Also, explain your answer. For each statement, output "1" or "0", where "1" means the statement can be fully included in the imaging report, and "0" means the statement contains information not included in the imaging report. Output the answers as a list composed of dictionaries in text form, not JSON. The output format should be: [{"Statement": "Original Statement 1", "Explanation": "Reason for the prediction", "Inclusion Prediction": 1 or 0 (whether the statement is included in the imaging report)}, {"Statement": "Original Statement 2", "Explanation": "Reason for the prediction", "Inclusion Prediction": 1 or 0 (whether the statement is included in the imaging report)}, ...]

Statements:

Imaging report:

Figure 7: The prompt for Evaluation in Task3 Examination.

gression), and treatment response fidelity (accuracy of responses to medications and interventions). Each dimension was rated on a 5-point Likert scale (1=poor, 5=excellent). Results demonstrated strong performance across all metrics, with mean scores of 4.0 (variance=1.0) for match rate, 3.9 (variance=1.49) for temporal coherence, and 3.7 (variance=0.81) for treatment response accuracy.

A.2.3 Task 3: Examination

The examination task assesses the LLM's ability to analyze medical images and generate corresponding image reports. The input consists of the medical images from each case, and the output is a free-text image report. This task tests the LLM's capability to interpret visual medical data and articulate findings in a clear, professional manner.

Evaluation metrics: Given the complexity of medical reports, conventional NLP metrics such as BLEU [37] and BERTScore [69] are inadequate for evaluating this task, as they do not capture whether the generated report aligns with the ground truth in terms of medical claims. Instead, we adopt the Claim Recall metric proposed in DocLens [63] to evaluate the completeness of the generated report. First, GPT-4o-mini is used to extract a list of claims from the original ground truth report. Then, GPT-4o-mini assesses whether the generated report entails these reference claims, and the recall score measures the proportion of claims that are correctly reflected in the generated report. The prompts used for this evaluation can be found in Figure 7.

A.2.4 Task 4: Diagnosis

The diagnosis task evaluates the LLM's ability to identify the patient's illness based on the provided symptoms and examination results. The input includes the patient's chief complaint, medical history, examination results, and the department visited. The output is a diagnosis in free-text format.

Evaluation metrics: We use a five-level grading system to evaluate the accuracy of the diagnosis, ranging from 1 to 5. The prompt we used in LLM-based evaluation is followed the instruction of People's Medical Publishing House⁵. The system accounts for the complex nature of medical diagnoses, where multiple factors may contribute to the patient's condition. The prompt is shown in Figure 8.

A.2.5 Task 5: Treatment

The treatment task involves formulating a treatment plan based on the patient's diagnosis and examination results. The input includes the patient's chief complaint, medical history, examination results, and diagnosis. The output is a set of proposed treatment items. This task tests the LLM's ability to synthesize the gathered information and apply medical knowledge to develop an appropriate treatment plan, considering factors such as drug interactions, treatment protocols, and patient-specific considerations.

⁵Diagnostics. 9th Ed. Beijing: People's Medical Publishing House; 2018

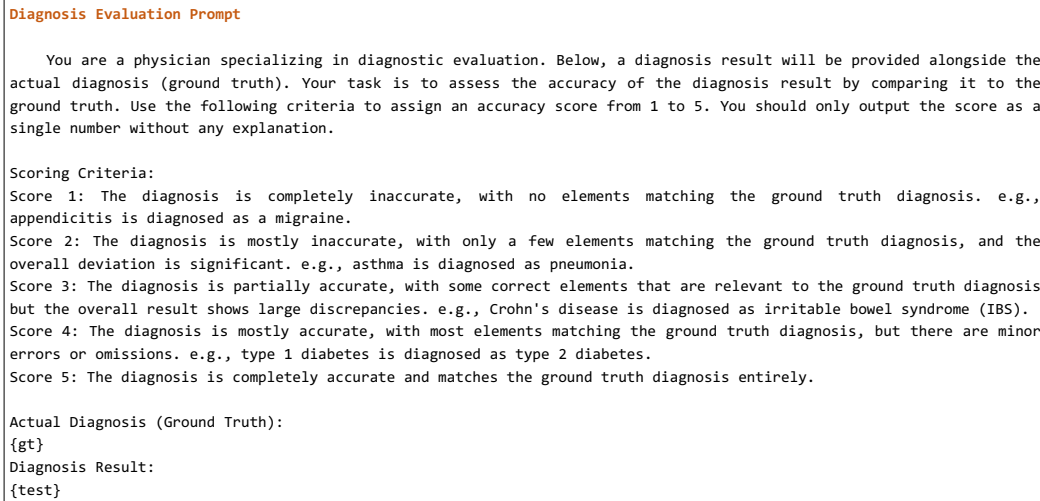


Figure 8: The prompt for Evaluation in Task4 Diagnosis.

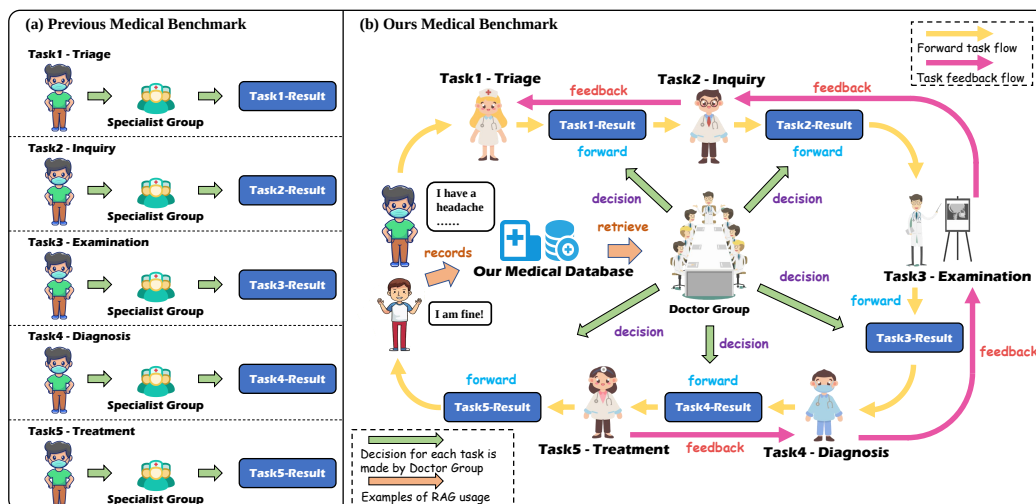


Figure 9: **Comparison between our benchmark method and previous benchmarks.** In Figure (a), the benchmark methods used in previous work are shown, which are based on specific domain problems and cannot achieve a complete end-to-end medical task workflow. In Figure (b), our proposed benchmark is presented, where we effectively accomplish the full workflow of medical tasks through our proposed MedCase-RAG method and the ACFM mechanism.

Evaluation metrics is the IoU between the proposed treatments and the ground truth treatment set. Since multiple treatment options may be appropriate, IoU allows for partial credit when the LLM suggests a subset of the recommended treatments or proposes additional reasonable treatments that are not part of the ground truth.

A.3 Examination Items in Task 2

Physical Examination: General examination (including height, weight, temperature, blood pressure, pulse, etc.), head, eyes, ears, nose and throat examination, neck examination (including thyroid, cervical lymph nodes), chest examination (including lungs, heart), abdominal examination, spine and limb examination, skin examination, neurological examination, urogenital system examination.

Auxiliary Examinations: X-ray, MRI, CT, ultrasound, nuclear medicine imaging, blood tests, urine tests, stool tests, endoscopy, pathological examination.

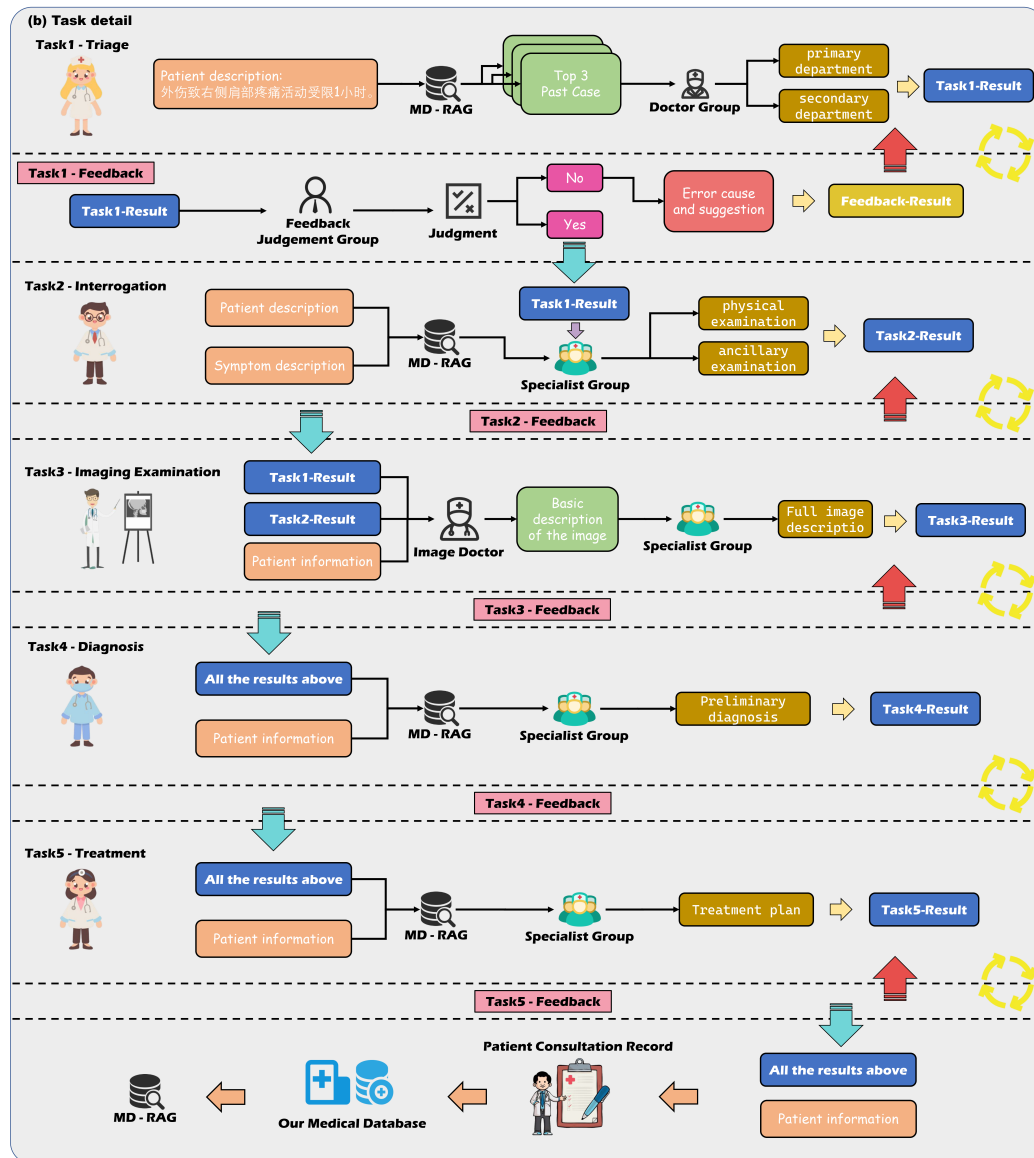


Figure 10: The specific agent design for each task in our MedChain-Agent and the ACFM of the entire process are presented.

A.4 Treatment Items in Task 5

Surgery, interventional therapy, medication, chemotherapy, antibiotic therapy, radiation therapy, physical therapy, immunotherapy, psychological therapy, traditional Chinese medicine, gene therapy.

B MedChain-Agent

B.1 Implementation Details

The overall process and output flow of our framework are illustrated in Figure 10. The patient first describes their condition, after which Task 1 performs specialty referral, assigning the patient to the appropriate primary and secondary departments. Once Task 1 is completed, the doctor group consists of specialists from the corresponding primary department, enabling better analytical decision-making. Task 2 involves these specialists prescribing suitable medical examinations based on the patient's

current situation. Task 3 focuses on analyzing the patient's imaging data to generate a detailed and comprehensive imaging report. Task 4 consists of the specialists conducting a comprehensive analysis based on the results from Task 2 and the imaging report from Task 3 to provide an initial diagnosis for the patient. Finally, Task 5 involves the specialists formulating an appropriate treatment plan based on all the previous tasks' results and analyses.

Additionally, in the aforementioned description, the decision-making and analytical processes for all tasks are conducted through discussions within the Doctor Group. Specifically, three specialists from the same department first engage in discussions and analysis, after which they summarize their findings and results to the final medical generalist (primary care physician) for the final decision. At each step, relevant similar cases are retrieved from our medical database to assist in decision-making. For Task 1, we will divide the tasks of identifying primary and secondary departments into two subtasks based on the Chain of Thought (CoT) approach: first identifying the primary department, followed by identifying the secondary department. Similarly, for Task 2, we first prescribe specific tests before additional examinations, optimizing the decision-making process.

Once the patient completes the entire medical process, a comprehensive treatment follow-up record will be compiled, including treatment outcomes and reports from each stage, along with the patient's information. This data will then be reintroduced into our medical database. Since these data have not been fully validated (we do not know how reasonable the decisions made for the current cases are), we will treat this data as pseudo-data based on the principles of semi-supervised learning and assign it a lower priority. This means that we will first retrieve data from completely validated sources, and if retrieval fails or data is insufficient, we will resort to retrieving from the pseudo-data.

B.2 Feedback Mechanism Details

To date, no work has proposed a multi-agent framework for simulating the entire medical process. Although existing frameworks demonstrate efficiency in specific medical tasks, their performance is suboptimal when linking various medical stages together. In sequential task scenarios, simply concatenating individual tasks is not feasible. Therefore, our full-process task can be regarded as a multi-sequence task, which necessitates a deeper exploration of the issue of error propagation. When a problem arises in the first task, subsequent tasks will analyze and make decisions based on erroneous results, potentially leading to severe impacts on the entire process.

To address this issue, we have introduced a feedback mechanism within the full-process framework, tightly connecting the current task with all subsequent tasks. After the current task is completed, the output results are evaluated by the physicians of the subsequent tasks (Feedback Judgement Group). Only when the results are confirmed to be accurate will the process advance to the next task; if issues are identified, the reasons for the errors and improvement suggestions will be output and fed back to the current task for re-discussion and decision-making. This process will continue iteratively until consensus is reached on the results or the maximum number of discussion rounds is achieved.

By implementing a feedback mechanism between each task, we can effectively manage the propagation of errors while significantly enhancing collaboration and communication among agent groups. This mechanism encourages agents to share information in real-time, improving mutual understanding and allowing for rapid strategic adjustments when issues are identified, thereby enhancing the overall flexibility and adaptability of the system. Each agent can better respond to changes in a dynamic environment, facilitating more efficient medical services. This feedback-driven collaborative model lays the foundation for the efficient operation of multi-agent systems, aiding in the provision of precise and reliable medical decision support in complex clinical environments.

B.3 RAG Details

We performed data restructuring on the original dataset, where each case is mapped into feature vectors from 12 dimensions. Among these, "Symptom Description" is identified as the most representative feature of the current patient and is processed through a Text Embedding model for quantification, which is stored in the database for subsequent dense retrieval tasks.

Previous Medical-RAG methods relied on medical question-answering (QA) databases and predominantly used chunked indexing for retrieval. Our approach differs in several ways. Firstly, the content of our foundational medical database is distinct; while most methods have built their databases

using medical QA data, ours utilizes a tree structure. We initially categorize patient information according to primary medical departments (in the experimental section, we categorized into 19 primary departments) and then extract and map patient information into a two-dimensional feature representation. This storage design allows for better retention of the patient’s crucial clinical information to assist in decision-making. Secondly, our retrieval method also diverges from theirs. In our retrieval process, we simulate the everyday practice of physicians. When faced with challenging cases that require decision-making, doctors often refer to past cases as references to enhance current decisions and judgments. We treat these two-dimensional features as the minimal unit of a case, using “Symptom description” as the basis for retrieval. When a new patient arrives, we extract and refine their “Symptom description” feature (approximately 70 characters). We employ direct quantitative matching without tokenizing this feature, using a Text Embedding model for quantification. The resulting feature vector is then compared to each case in our database corresponding to the relevant department through cosine similarity calculations, selecting the top three cases with the highest symptom similarity as the current retrieval results. These results, combined with the current case features, are forwarded to subsequent agents for comprehensive decision-making and judgment, enhancing the agents’ output. Since our retrieval method does not utilize document chunking but rather employs direct matching, it retains more matching information, leading to improved matching accuracy.

B.4 Task Prompts

Figure 11 ~ page 30 showcase prompts for each task and stage within our MedChain-Agent framework.

Table 6: **Comparison of Benchmarks for LLM-based Agent.** The MedChain composed of 12,163 case with five sequential stages of the clinical workflow. It provide interaction environment for LLM-based agent evaluation.

Benchmark	#Num	Modality	Personal Information	Sequential Evaluation	Interaction
MedQA [70]	1,273	Text	✓	✗	✗
PubMedQA [20]	500	Text	✗	✗	✗
MedMCQA [36]	193,155	Text	✓	✗	✗
MMedBench [39]	53,566	Text	✗	✗	✗
MedBench [29]	300,901	Text	✓	✗	✗
Asclepius [54]	3,232	Text, Imaging	✗	✗	✗
GMAI-MMBench [5]	26,000	Text, Imaging	✗	✗	✗
MedChain	12,163	Text, Imaging	✓	✓	✓

C Supplementary for Experiment

Tasks 1 and 2 are subdivided based on the Chain of Thought (CoT) reasoning approach for decision-making. Consequently, Task 1 is split into "specialty referral (Level 1)" and "specialty referral (Level 2)" to evaluate performance at both primary and secondary department levels. Task 2 is divided into "Physical" and "Ancillary," representing performance in physical examinations and ancillary tests. or both ablation and comparative experiments, we split our dataset into training and test sets in a 7:3 ratio. We extracted 12-dimensional features from each training set case and stored them in our Medical Dataset for subsequent retrieval tasks.

D Related Work Discussion

We compare MedChain with most of related work in Table 6 and Figure 9. Especially, we analyze MedChain with AI Hospital and CoD as following:

AI Hospital [11]: While both AI Hospital and our work aim to evaluate LLMs in clinical scenarios through multi-agent interactions, there are several key distinctions. Firstly, MedChain significantly

Task1 Triage Prompt

Mission1: Classification of Primary Medical Departments

sys_prompt: You are a professional triage doctor with extensive experience in triage. Based on the patient's condition, you need to accurately assign them to the appropriate medical department.

user_input: You need to assign the patient to the appropriate department for consultation based on the "patient condition" described below. You must select only one department from the 19 "available departments" listed below. Refer to the "past cases" below to assist in making your decision. Patient condition: "[condition]" Past cases: "\${[case1],[case2],[case3]}\$Available departments: ['Nursing Department', 'Pharmacy Department', 'Dentistry', 'Pediatrics', 'Medical Imaging', 'Ophthalmology', 'Laboratory Medicine', 'Surgery', 'Dermatology and Venereology', 'Psychiatry', 'General Medicine', 'Otolaryngology', 'Internal Medicine', 'Emergency Medicine', 'Oncology', 'Traditional Chinese Medicine', 'Rehabilitation', 'Obstetrics and Gynecology', 'Psychology'] Output example (output only your selection, no extra text): "Surgery"

Mission2: Mission1-Feedback

sys_prompt: You are a triage result evaluator with extensive experience in triage. Please review the triage doctor's "triage result" to determine whether the primary department assignment is correct.

user_input: Please review the triage doctor's "triage result" to determine whether the primary department is correct. Refer to the "past cases" department assignment to assist in decision-making. Output requirements: (You must follow these instructions carefully) 1、If you think it is correct, output "Correct," if it is incorrect, output "Incorrect," and explain the reason!! 2、You must answer first, then give a reason! Triage result: "[result]" Past cases: "\${[case1],[case2],[case3]}\$Output example: For correct result (no need to explain): "Correct" For incorrect result (with reason): "IncorrectReason: {your opinion on the primary department assignment}"

Feedback_result_example: Here is the triage result for the "primary department" based on the patient's condition: "[result]". Here is the feedback from other doctors regarding the current "primary department" triage: "[feedback]"

Mission3: Feedback-Answer

sys_prompt: You are a professional triage doctor with extensive experience in triage. You need to assign the patient to the appropriate department based on their condition with precision.

user_input: Please reasonably consider the opinions of the clinic doctors and assign the patient to the appropriate department for treatment. You need to select only one department from the following 19 "available departments." The original triage result and other doctors' suggestions: "[feedback]" Past cases: "\${[past_case]}\$Available departments: ['Nursing Department', 'Pharmacy Department', 'Dental Department', 'Pediatrics', 'Medical Imaging Department', 'Ophthalmology', 'Laboratory Department', 'Surgery', 'Dermatology', 'Psychiatry', 'General Medicine', 'Otorhinolaryngology', 'Internal Medicine', 'Emergency Department', 'Oncology', 'Traditional Chinese Medicine', 'Rehabilitation Department', 'Obstetrics and Gynecology', 'Psychology'] Output example: (Directly output your choice, without extra content) "Surgery"

Mission4: Secondary Department Classification

sys_prompt: You are a professional triage doctor with extensive experience in triage, and you need to accurately allocate patients to the appropriate department based on their conditions.

user_input: You need to allocate the patient to the appropriate department for treatment based on the described "patient condition." You can choose from the following "available departments," and you may select multiple suitable departments. Refer to the "previous cases" for guidance in your decision-making. Patient condition: "[condition]" Previous cases: "\${[case1],[case2],[case3]}\$Available departments: [list] Output example: (Directly output your options without any extra content) For a single department: "Surgery" For multiple departments: (separated by commas) "Joint Orthopedics, Orthopedics"

Mission5: Mission4-Feedback

sys_prompt: You are a triage result evaluator with extensive triage experience. Please check whether the "secondary department" in the triage result provided by the triage doctor is correct.

user_input: Please check whether the "secondary department" in the triage result provided by the triage doctor is correct. Refer to the "past case" department allocation to assist in your decision-making. Output requirements: (You must think according to the following requirements) If you believe there are no errors, output "Correct." If there is an error, output "Incorrect" and explain the reason! Always respond first, then provide the reason! Triage result: "[result]" Past cases: "\${[case1],[case2],[case3]}\$Output example: If you believe it is correct (no need to provide a reason): "Correct" If you believe it is incorrect (explain the reason): "IncorrectReason: {Your opinion on the secondary department triage result}"

Feedback_result_example: This is the triage result for the "secondary department" based on the patient's condition: "[result]" This is the triage opinion from other doctors regarding the current "secondary department": "[feedback]"

Mission6: Feedback-Answer

sys_prompt: You are a professional triage doctor with extensive experience in triage, and you need to accurately allocate patients to the appropriate department based on their conditions.

user_input: Please reasonably consider the opinions of the clinic doctors and assign the patient to the appropriate department for treatment. You need to choose the suitable departments from the list below and can select multiple appropriate departments. Directly output your selection without additional content!!! Original triage result and other doctors' suggestions: "[feedback]" Available departments: [list] Output example: Single department output: "Surgery" Multiple departments output: (separated by commas) "Joint Orthopedics, Orthopedics"

Figure 11: The prompt for Task 1 in the MedChain-Agent .

Task1 Triage Prompt

Mission1: Classification of Primary Medical Departments

sys_prompt: 你是一个专业的分诊医生，有着丰富分诊经验，你需要根据病人的情况，进行精确的科室分诊。

user_input: 你需要根据下面描述的“病人情况”，将病人分到合适的科室进行就诊，你需要从以下19个“可选择的科室”中仅选择1个科室进行分配。参考下面的“过往病例”的科室分配情况，辅助做出决策。病人情况：“[condition]” 过往病例：“\${[case1]、[case2]、[case3]}”。可选择的科室：['护理科', '药剂科', '口腔科', '儿科', '医学影像科', '眼科', '检验科', '外科', '皮肤性病科', '精神科', '全科', '耳鼻咽喉科', '内科', '急诊科', '肿瘤科', '中医科', '康复科', '妇产科', '心理科']。输出示例：（直接输出你的选项，不要输出多余的内容）“外科”。

Mission2: Mission1-Feedback

sys_prompt: 你是一名分诊结果判读者，拥有着丰富的分诊经验，请你检查分诊医生的“分诊结果”中一级科室是否正确。

user_input: 请你检查分诊医生的“分诊结果”中一级科室是否正确。参考下面的“过往病例”的科室分配情况，辅助做出决策。输出要求：（你一定要按下面的要求进行思考）1、认为没有错误则输出“正确”，有错误则输出“错误”，并且说明原因！！ 2、一定要先回答，然后再回答原因！分诊结果：“[result]” 过往病例：“\${[case1] [case2] [case3]}”。输出示例：认为正确的输出示例：（认为正确则无需输出原因）“正确”。认为错误的输出示例：（认为错误要说明原因）“错误原因：{对一级科室分诊结果的意见}”

Feedback_result_example: 这是基于病人情况做出的“一级科室”的分诊结果：“[result]”。这是其他医生对当前“一级科室”的分诊意见：“[feedback]”

Mission3: Feedback-Answer

sys_prompt: 你是一个专业的分诊医生，有着丰富分诊经验，你需要根据病人的情况，进行精确的科室分诊。

user_input: 请你合理参考诊室医生的意见，将病人分到合适的科室进行就诊，你需要从以下19个“可选择的科室”中仅选择1个科室进行分配。诊室医生的意见不一定正确，请你合理对待。参考下面的“过往病例”的科室分配情况，辅助做出决策。原来的分诊结果和其他医生的建议：“[feedback]”。过往病例：“\${[past_case]}”。可选择的科室：['护理科', '药剂科', '口腔科', '儿科', '医学影像科', '眼科', '检验科', '外科', '皮肤性病科', '精神科', '全科', '耳鼻咽喉科', '内科', '急诊科', '肿瘤科', '中医科', '康复科', '妇产科', '心理科']。输出示例：（直接输出你的选项，不要输出多余的内容）“外科”

Mission4: Secondary Department Classification

sys_prompt: 你是一个专业的分诊医生，有着丰富分诊经验，你需要根据病人的情况，进行精确的科室分诊。

user_input: 你需要根据下面描述的“病人情况”，将病人分到合适的科室进行就诊，你需要从下面“可选择的科室”中选择合适的科室进行分配，可以选择多个合适的科室。参考下面的“过往病例”的科室分配情况，辅助做出决策。病人情况：“[condition]”。过往病例：“\${[case1]、[case2]、[case3]}”。可选择的科室：[list]。输出示例：（直接输出你的选项，不要输出多余的内容），输出为单个科室：“外科”。输出为多个科室：（用逗号分隔开）“关节骨科,骨科”

Mission5: Mission4-Feedback

sys_prompt: 你是一名分诊结果判读者，拥有着丰富的分诊经验，请你检查分诊医生的“分诊结果”中二级科室是否正确。

user_input: 请你检查分诊医生的“分诊结果”中二级科室是否正确。参考下面的“过往病例”的科室分配情况，辅助做出决策。输出要求：（你一定要按下面的要求进行思考）：1、认为没有错误则输出“正确”，有错误则输出“错误”，并且说明原因！！ 2、一定要先回答，然后再回答原因！分诊结果：“[result]”。过往病例：“\${[case1]、[case2]、[case3]}”。输出示例：认为正确的输出示例：（认为正确则无需输出原因）“正确”。认为错误的输出示例：（认为错误要说明原因）：“错误，原因：{对二级科室分诊结果的意见}”

Feedback_result_example: 这是基于病人情况做出的“二级科室”的分诊结果：“[result]”。这是其他医生对当前“二级科室”的分诊意见：“[feedback]”

Mission6: Feedback-Answer

sys_prompt: 你是一名专业的分诊医生，在分诊方面有丰富的经验，你需要根据患者的情况，准确地将患者分配到合适的科室。

user_input: 请你合理参考诊室医生的意见，将病人分到合适的科室进行就诊，你需要从下面“可选择的科室”中选择合适的科室进行分配，可以选择多个合适的科室。直接输出你的选项，不要输出多余的内容！！ 原来的分诊结果和其他医生的建议：“[feedback]”。可选择的科室：[list]。输出示例：输出为单个科室：“外科”。输出为多个科室：（用逗号分隔开）“关节骨科,骨科”。

Figure 12: The prompt for Task 1 in the MedChain-Agent in Chinese.

expands the scale and diversity of medical cases, comprising 12,163 cases across 19 specialties and 156 subspecialties, including 7,338 medical images. This represents a substantial advancement over AI Hospital’s dataset of 506 cases with limited specialty coverage, enabling more comprehensive evaluation of LLMs’ medical capabilities across diverse clinical scenarios. Secondly, MedChain introduces a more sophisticated multi-agent collaboration mechanism. While AI Hospital employs a basic agent interaction model, our framework incorporates a feedback-driven multi-agent system enhanced by the MedCase-RAG module. This module enables dynamic knowledge base expansion and supports case-based reasoning through structured feature representation.

CoD [4] introduces an interactive approach to enhance the interpretability of medical diagnosis. While both CoD and our work incorporate interactive components, they differ substantially in both design objectives and implementation mechanisms. CoD primarily focuses on improving the interpretability of the diagnostic phase through confidence-driven interactions that demonstrate the reasoning process. In contrast, MedChain takes a more comprehensive approach by simulating the complete clinical workflow through a multi-agent collaborative framework. Our framework encompasses five sequential stages from triage to treatment, with each stage’s decisions being evaluated and guided by subsequent stages through a novel cross-stage feedback mechanism. This design enables MedChain to capture the interdependent nature of clinical decision-making, where decisions at each stage influence and are influenced by other stages in the workflow.

Task2 Inquiry Prompt

Mission1: Physical examination
sys_prompt: You are a professional [first_room] physician with extensive clinical experience, and you need to prescribe appropriate medical examinations for the patient.
user_input: Task Description:Based on the current case situation and with appropriate reference to past cases, prescribe suitable "[task_obj]" for the current patient. If no physical examination is needed, you can output "[task_obj]: None." Please adhere strictly to the output format.Only output the examinations that need to be performed; no need to provide examination results!!!Past Cases:"\${case1}[case2][case3]\$"Current Case:"Patient's complaint: [case now]"Output Format:"[task_obj]: {.....}"

Mission2: Mission1-Feedback
sys_prompt: You are a medical physical examination result evaluator with extensive experience in medical examinations. Please review the physical examination items in the physician's "physical examination" to determine whether they are correct.
user_input: Please review the physical examination items in the physician's "physical examination" to determine whether they are correct. Refer to the "past cases" below to assist in your decision-making.Output requirements: (You must follow the requirements below)Output "Correct" if there are no errors, or "Incorrect" if there are errors, and provide reasons!!!Always answer first, then explain the reasons!Physical Examination: " [result] "Past Cases: "\${past_case} \$"Output Example: If correct, the output example is: (No reason needed if correct) " Correct " If incorrect, the output example is: (Must explain the reasons for being incorrect!!!) " Incorrect Reason: {Opinion on the items in the "physical examination" result} "
Feedback_result_example: Here are the "physical examination" items ordered by the medical examination physician:"[result]"Here are the opinions of other doctors regarding the ordered "physical examination" items:"[feedback]"

Mission3: Feedback-Answer
sys_prompt: You are a specialized [first_room] physician with extensive clinical experience in patient consultations, and you need to prescribe appropriate medical examinations for the patient.
user_input: Please reasonably consider the opinions and suggestions of other doctors regarding the "physical examination" items. Based on the patient's condition and the opinions of the consulting physician, reissue the appropriate examination items for the patient's "physical examination" according to the original format. The opinions of other doctors may not be correct, so please evaluate them critically. Only output the modified "physical examination" items without providing reasons or additional explanations.Reference the "past cases" for assistance in decision-making.Original physical examination items and other doctors' suggestions: " [feedback] "Past cases: "\${past_case} \$"Output format: " Physical Examination: {.....} "

Mission4: Auxiliary Examination
sys_prompt: You are a professional [first_room] doctor with extensive clinical interview experience, and you need to prescribe appropriate medical examinations for the patient.
user_input: Task Description:You need to prescribe appropriate "[task_obj]" for the current patient based on the current case situation, with suitable reference to past cases. If no physical examination is required, you can output "[task_obj]: None." Please adhere strictly to the output format.Only output the required examinations, without providing the examination results.Past Cases: "\${case1}[case2][case3] \$"Current Case: " Patient's Chief Complaint: [case now]"Output Format: " [task_obj]: {.....} "

Mission5: Mission4-Feedback
sys_prompt: You are a medical auxiliary examination result evaluator with extensive experience in medical testing. Please review the auxiliary examination items listed by the medical examination doctor to determine their accuracy.
user_input: Please review the auxiliary examination items provided by the medical examination doctor to determine their accuracy. Use the following "past cases" distribution of departments to assist in your decision-making.Output requirements: (You must follow the requirements below for your reasoning)If you believe there are no errors, output "Correct." If there are errors, output "Incorrect" and explain the reason!!!Always respond first, then provide the reason!Auxiliary examination: " [result] "Past cases: "\${past_case} \$"Output examples: If you believe it's correct (no reason needed): " Correct " If you believe it's incorrect (you must explain the reason!!!): " Incorrect Reason: {Your opinion on the "auxiliary examination" items} "
Feedback_result_example: Here are the "auxiliary examination" items prescribed by the medical examination doctor:"[result]"Here are the opinions of other doctors regarding the prescribed "auxiliary examination" items:"[feedback]"

Mission6: Feedback-Answer
sys_prompt: You are a specialized [first_room] physician with extensive clinical examination experience, and you need to prescribe appropriate medical tests for the patient.
user_input: Please reasonably consider the opinions and suggestions of other doctors regarding the "auxiliary examination" items. Based on the patient's condition and the opinions of the consulting physician, please rewrite the appropriate examination items for the patient's "auxiliary examination" according to the original format. The opinions of other doctors may not be correct, so please evaluate them judiciously. Only output the modified "auxiliary examination" items without providing reasons or additional explanations. Refer to the following "past cases" to assist in decision-making.Original auxiliary examination items and other doctors' suggestions: " [feedback] "Past cases: "\${past_case} \$"Output format: " Auxiliary examination: {.....} "

Figure 13: The prompt for Task 2 in the MedChain-Agent .

Task2 Inquiry Prompt

Mission1: Physical examination

sys_prompt: 你是一名专业的[first_room]医生, 拥有着丰富的临床问诊经验, 您需要为患者开具合适的医学检查。

user_input: 任务说明: 1、你需要根据当前病例的情况, 适当参考过往病例, 为当前患者开具合适的"[task_obj]", 若无需开局体格检查则可以输出"[task_obj]: None", 请你严格按照输出格式进行输出。2、只输出需要做的检查即可, 无需输出检查结果!!! 过往病例: "\${case1}[case2][case3]"\$当前病例: "病人主诉: [case now]"输出格式: "[task_obj]: {...}"

Mission2: Mission1-Feedback

sys_prompt: 你是一名医学体格检查结果判别者, 拥有着丰富的医学检查经验, 请你检查医学检查医生的"体格检查"中的这些体格检查项目是否正确。

user_input: 请你检查医学检查医生的"体格检查"中的这些体格检查项目是否正确。参考下面的"过往病例"的科室分配情况, 辅助做出决策。输出要求: (你一定要按下面的要求进行思考) 1、认为没有错误则输出"正确", 有错误则输出"错误", 并且说明原因!!! 2、一定要先回答, 然后再回答原因! 体格检查: "[result]"过往病例: "\${past_case}"\$输出示例: 认为正确的输出示例: (认为正确则无需输出原因)"正确"认为错误的输出示例: (认为错误要说明原因, 一定要说明原因!!!)"错误原因: {对"体格检查"项目结果的意见}"

Feedback_result_example: 这是医学检查医生开具的"体格检查"项目: "[result]"这是其他医生对开具的"体格检查"项目的意见: "[feedback]"

Mission3: Feedback-Answer

sys_prompt: 你是一名专业的[first_room]医生, 拥有着丰富的临床问诊经验, 您需要为患者开具合适的医学检查。

user_input: 请你合理参考其他医生的关于"体格检查"项目的看法和建议, 基于病人情况和诊室医生的意见按照原来的格式重新对病人的"体格检查"开具合适的检查项目。其他医生的意见不一定正确, 请你合理对待。只输出修改后的"体格检查"项目即可, 无需输出原因和其他说明。参考下面的"过往病例"的科室分配情况, 辅助做出决策。原来的体格检查项目和其他医生的建议: "[feedback]"过往病例: "\${past_case}"\$输出格式: "体格检查: {...}"

Mission4: Auxiliary Examination

sys_prompt: 你是一名专业的[first_room]医生, 拥有着丰富的临床问诊经验, 您需要为患者开具合适的医学检查。

user_input: 任务说明: 1、你需要根据当前病例的情况, 适当参考过往病例, 为当前患者开具合适的"[task_obj]", 若无需开局体格检查则可以输出"[task_obj]: None", 请你严格按照输出格式进行输出。2、只输出需要做的检查即可, 无需输出检查结果!!! 过往病例: "\${case1}[case2][case3]"\$当前病例: "病人主诉: [case now]"输出格式: "[task_obj]: {...}"

Mission5: Mission4-Feedback

sys_prompt: 你是一名医学辅助检查结果判别者, 拥有着丰富的医学检查经验, 请你检查医学检查医生的"辅助检查"中的这些辅助检查项目是否正确。

user_input: 请你检查医学检查医生的辅助检查"中的这些辅助检查项目是否正确。参考下面的"过往病例"的科室分配情况, 辅助做出决策。输出要求: (你一定要按下面的要求进行思考) 1、认为没有错误则输出"正确", 有错误则输出"错误", 并且说明原因!!! 2、一定要先回答, 然后再回答原因! 体格检查: "[result]"过往病例: "\${past_case}"\$输出示例: 认为正确的输出示例: (认为正确则无需输出原因)"正确"认为错误的输出示例: (认为错误要说明原因, 一定要说明原因!!!)"错误原因: {对"辅助检查"项目结果的意见}"

Feedback_result_example: 这是医学检查医生开具的"辅助检查"项目: "[result]"这是其他医生对开具的"辅助检查"项目的意见: "[feedback]"

Mission6: Feedback-Answer

sys_prompt: 你是一名专业的[first_room]医生, 拥有着丰富的临床问诊经验, 您需要为患者开具合适的医学检查。

user_input: 请你合理参考其他医生的关于"辅助检查"项目的看法和建议, 基于病人情况和诊室医生的意见按照原来的格式重新对病人的"辅助检查"开具合适的检查项目。其他医生的意见不一定正确, 请你合理对待。只输出修改后的"辅助检查"项目即可, 无需输出原因和其他说明。参考下面的"过往病例"的科室分配情况, 辅助做出决策。原来的辅助检查项目和其他医生的建议: "[feedback]"过往病例: "\${past_case}"\$输出格式: "辅助检查: {...}"

Figure 14: The prompt for Task 2 in the MedChain-Agent in Chinese.

Task3 Examination Prompt

Mission1: Description of imaging by a professional radiologist
sys_prompt: You are a professional medical imaging physician, and you need to provide a feature description based on the input medical images.
user_input: The following are the "basic patient information" and the patient's [img_class] imaging. Please provide a concise and comprehensive description of the imaging.

Basic Patient Information:
 "
 [symptom]
 "

Output Example: (The descriptions in both the imaging and conclusion sections should be as detailed and comprehensive as possible.)
 "
 Chest CT: (Description should be as detailed and comprehensive as possible)
 1. Multiple abnormal density shadows in both lungs, mostly located subpleurally, with a large lesion in the right upper lobe accompanied by multiple cavitary formations, raising the possibility of infection;
 2. Small amount of pericardial effusion and a small amount of pleural effusion on the right side;
 3. Slight thickening of both pleurae.

****Imaging Report Conclusion:**** (Description should be as detailed and comprehensive as possible)
 1. No clear abnormal density shadows were seen in both lungs, bronchial structures appear normal, and no significant dilation was observed.
 2. No obvious enlarged lymph nodes were seen in the mediastinum or at both lung hilum.
 3. An approximately 8.3 cm soft tissue density mass was observed in the anterior mediastinum.
 4. Other structures within the mediastinum showed no significant displacement, and the heart size and shape were normal.

Recommendations:
 1.
 2.
 "

Mission2: Image report generation
sys_prompt: You are a professional {first_room} radiologist with extensive clinical experience, capable of generating medical imaging reports effectively.
user_input: Below is the imaging report generated by the previous radiologist based on the 'patient condition' and the patient's imaging. Please re-evaluate the 'imaging report' below to check for any omissions, incompleteness, or errors, and if any are found, provide the output as per the example. Patient condition: ' [now_case] 'Imaging report: ' [img_report] 'Output instructions: If there are no suggestions and you consider it correct, output: ' None 'If there are supplementary suggestions, output: ' It is recommended to supplement, improve, or modify the following parts of the "imaging report"... '

Mission3: Mission2-Feedback
sys_prompt: You are a specialized {first_room} radiologist with extensive clinical experience in imaging. You need to evaluate whether the previously analyzed report by the expert physician is reasonable based on the patient's imaging findings.
user_input: Task Instructions: Below is the imaging report provided by the radiologist. Based on the report and the patient's imaging, do you think there are issues with this imaging report? If you think the result is correct, output "Correct"; if you think it is incorrect, output "Incorrect" and explain the reason! You must answer first and then provide your reasoning. You only need to output your reason or opinion; do not output the original report result or information. Imaging Doctor's Report Result: "[result]" Output Example: [Correct/Incorrect] Reason:

Mission4: Feedback-Answer
sys_prompt: You are a specialized {first_room} radiologist with extensive clinical experience in imaging. You need to evaluate whether the previously analyzed report by the expert physician is reasonable based on the patient's imaging findings.
user_input: Task Instructions: Please reasonably consider the clinic doctor's opinions and suggestions regarding the imaging results for the above patient. Based on the patient's imaging and the opinions of other expert doctors, modify the "current imaging report" according to the original format. Patient Situation: "[now_case]" Current Imaging Report: "[img_report]" Expert Doctor's Opinion: "[feedback]" Output Example: "Imaging Report: {.....}"

Figure 15: The prompt for Task 3 in the MedChain-Agent .

Task3 Examination Prompt

Mission1: Description of imaging by a professional radiologist

sys_prompt: 你是一名专业的医学影像医生，你需要根据输入的医学影像对其进行特征描述。

user_input: 下面传进来的是“病人的基本情况”和病人的[img_class]影像，请你进行精简、全面的影像描述。病人的基本情况：“[symptom]”输出示例：（影像描述和结论部分的描述尽量详细！）“胸部CT：（描述尽可能详细、全面）1、两肺内多发异常密度影，大部位于胸膜下，右肺上叶病变范围较大伴多发空洞形成，考虑感染可能大；2、心包少量积液，右侧胸腔少量积液；3、两侧胸膜稍厚影像报告结论：（描述尽可能详细、全面）1、两肺内未见明确异常密度影，支气管正常，未见明显扩张。2、纵隔内及两肺门未见明显肿大淋巴结影。3、前纵隔内见一直径约8.3cm的软组织密度肿块。4、余纵隔内诸结构未见明显移位，心影大小形态正常。建议：1、.....2、.....”

Mission2: Image report generation

sys_prompt: 你是一名专业的{first_room}影像学医生，拥有丰富的临床经验，能够很好地生成医学影像报告。

user_input: 下面是之前的影像学医生基于“病人情况”和病人的影像生成的影像报告。请你基于“病人情况”和病人的影像再次检查一下下面的“影像报告”是否有遗漏、不完善或者错误的部分，若有则按照输出示例进行输出。仅做补充说明即可，无需完整回答！！病人情况：“[now_case]”影像报告：“[img_report]”输出说明：1、若无建议，认为正确则输出：“无”2、若有补充建议则输出：“认为“影像报告”中需要补充、完善、修改的部分.....”

Mission3: Mission2-Feedback

sys_prompt: 你是一名专业的{first_room}影像医生，拥有丰富的影像学临床经验，您需要基于患者的影像情况，判断之前专家医生分析的报告是否合理。

user_input: 任务说明：1、下面是影像医生开具影像报告结果。2、基于下面的报告和病人的影像，你认为这个影像报告结果有问题吗？3、你认为这个结果正确则输出-“正确”，认为不正确则输出-“错误”，并且说明原因！！4、你一定要先回答，然后再回答原因。5、你只需要输出你的原因或观点即可，不要输出原来的报告结果或者信息。影像医生的报告结果：“[result]”输出示例：[正确/错误]原因：.....

Mission4: Feedback-Answer

sys_prompt: 你是一名专业的放射科医生，在影像方面有丰富的临床经验。您需要根据患者的影像学检查结果评估专家医师之前分析报告是否合理。

user_input: 任务说明：请你合理参考诊室医生关于上述病人影像结果的看法和建议，基于病人影像图片和其他专家医生的意见按照原来的格式修改“当前影像报告”。病人情况：“[now_case]”当前影像报告：“[img_report]”专家医生的意见：“[feedback]”输出示例：“影像报告：{.....}”

Figure 16: The prompt for Task 3 in the MedChain-Agent in Chinese.

Task4 Diagnosis Prompt

Mission1: Medical Diagnosis
sys_prompt: You are a professional {first_room} physician with extensive clinical diagnostic experience. You need to output the diagnosis based on the patient's condition.
user_input: You need to provide an appropriate "diagnosis" for the current patient based on the described "patient condition" and consider past cases for reference. Only reply with one or more diagnosis names, without any additional explanations, and strictly follow the output format. Patient condition: "[case now]" Past cases: "\$[case1][case2][case3]\$" Output example: (directly output your options without additional content) For a single diagnosis: "Ankylosing Spondylitis" For multiple diagnoses: (separated by commas) "Rheumatoid Arthritis", "Fracture"

Mission2: Mission1-Feedback
sys_prompt: You are a professional {first_room} physician with extensive clinical diagnostic experience, and you need to analyze whether the diagnosis made by the clinic doctor is reasonable based on the patient's condition.
user_input: Please check whether the diagnosis in the "current patient condition and diagnosis results" is correct. Refer to the "past cases" below to assist in making your decision. Output requirements: (You must follow the requirements below in your reasoning) If you believe there are no errors, output "Correct." If there are errors, output "Incorrect," and explain the reason!!! You must answer first, and then provide your reasoning! Current patient condition and diagnosis results: "[result]" Past cases: " [case1][case2][case3] " Output example: If you believe it is correct, output example: (No reason is needed if you think it is correct) " Correct " If you believe it is incorrect, output example: (If you think it is incorrect, provide the reason; you must explain the reason!!!) " Incorrect Reason: {Your opinion on the "diagnosis result"} "
Feedback_result_example: Here is the diagnosis result made by the doctor based on the current patient condition: "[result]" Here are the opinions of other doctors regarding the diagnosis result: "[feedback]"

Mission3: Feedback-Answer
sys_prompt: You are a specialized {first_room} physician with extensive clinical diagnostic experience, and you need to provide a diagnosis based on the patient's condition.
user_input: Please reasonably consider other doctors' opinions and suggestions regarding the current case's "diagnosis." Based on the patient's condition and the opinions of the clinic doctor, revise the patient's "diagnosis" according to the original format. Other doctors' opinions may not be correct, so please assess them critically. Only output the modified "diagnosis" without any reasons or additional explanations. You can refer to the "past cases" diagnosis situation to assist in decision-making. Original diagnosis results and other doctors' suggestions: " [feedback] " Past cases: "\$ [past_case] \$" Output example: (directly output your options without additional content) For a single diagnosis: " 'Spondyloarthritis' " For multiple diagnoses: (separated by commas) " 'Rheumatoid arthritis', 'Fracture' "

Task5 Treatment Prompt

Mission1: Medical treatment
sys_prompt: You are a specialized {first_room} doctor with extensive clinical treatment experience, and you need to provide a treatment plan based on the patient's condition.
user_input: You need to prescribe appropriate treatment options for the current patient based on the described "patient condition." You should choose suitable options from the "available treatment options" list below, and you can select multiple options. Refer to the "past cases" treatment options for assistance in making your decision. Patient condition: " [case now] " Past cases: "\$ [case1][case2][case3] \$" Available treatment options: ["Surgery", "Interventional Therapy", "Medication Therapy", "Chemotherapy", "Antibiotic Therapy", "Radiation Therapy", "Physical Therapy", "Immunotherapy", "Psychotherapy", "Traditional Chinese Medicine", "Gene Therapy"] Output example: (directly output your options without additional content) For a single treatment option: " Surgery " For multiple treatment options: (separate with commas) " Medication Therapy, Antibiotic Therapy, Physical Therapy "

Mission2: Mission1-Feedback
sys_prompt: You are a professional {first_room} doctor with extensive clinical treatment experience. You need to analyze whether the treatment results provided by the clinic doctor are reasonable based on the patient's condition.
user_input: Please check whether the "treatment items" prescribed by the doctor are correct. Refer to the department allocation situation of the "past cases" below to assist in making your decision. Output requirements: (You must think according to the requirements below) If you believe there are no errors, output "Correct." If there are errors, output "Incorrect," and provide a reason! You must answer first, then provide the reason! Treatment items prescribed by the doctor based on the patient's condition: " [result] " Past medical records: " [case1][case2][case3] " Output example: If you believe it is correct, the output example is as follows: (If you believe it is correct, no reason is needed) " Correct " If you believe it is incorrect, the output example is as follows: (If you believe it is incorrect, you must explain the reason!!!) " Incorrect Reason: {Opinion on the treatment item results} "
Feedback_result_example: The following are the "diagnosis results" you provided: "[Result]." Below are the opinions from other doctors regarding your "diagnosis results": "[Feedback]."

Mission3: Feedback-Answer
sys_prompt: You are a specialized {first_room} doctor with extensive clinical treatment experience, and you need to provide a treatment plan based on the patient's condition.
user_input: Please reasonably consider the opinions and suggestions of other doctors regarding the "treatment options" for the current case, and based on the patient's condition and the doctor's opinions, re-prescribe appropriate "treatment options" for the patient according to the original format. You need to choose appropriate options from the following "available treatment options," and you can select multiple suitable options. Only output the modified "treatment options," without providing reasons or additional explanations. You may refer to the "past cases" treatment options to assist in your decision-making. Original treatment options and other doctors' suggestions: " [feedback] " Past cases: "\$ [past_case] \$" Available treatment options: ["Surgery", "Interventional Therapy", "Medication", "Chemotherapy", "Antibiotic Therapy", "Radiation Therapy", "Physical Therapy", "Immunotherapy", "Psychotherapy", "Traditional Chinese Medicine", "Gene Therapy"] Output example: (directly output your options, do not output extra content) For a single treatment option: " Surgery " For multiple options: (separated by commas) " Medication, Antibiotic Therapy, Physical Therapy "

Figure 17: The prompt for Task 4 and Task5 in the MedChain-Agent .

Task4 Diagnosis Prompt

Mission1: Medical Diagnosis
sys_prompt: 你是一名专业的{first_room}医生, 有着丰富的临床诊断经验, 你需要根据病人的情况, 输出诊断结果。
user_input: 你需要根据下面描述的"病人情况", 适当参考过往病例, 为当前患者进行合适的"诊断"。仅回复一个或多个诊断结果的名称, 不要包含其他任何解释, 请严格按照输出格式进行输出。参考下面的"过往病例"的诊断情况, 辅助做出决策。病人情况: "[case now]" 过往病例: "\$[case1][case2][case3]"\$输出示例: (直接输出你的选项, 不要输出多余的内容)输出为单个诊断: "'强直性脊柱炎'" 输出为多个诊断: (用逗号分隔开)" "'风湿性关节炎', '骨折'"

Mission2: Mission1-Feedback
sys_prompt: 你是一名专业的{first_room}医生, 拥有着丰富的临床诊断经验, 您需要从患者的情况分析诊室医生的诊断结果是否合理。
user_input: 请你检查下面"当前病人情况和诊断结果"中的诊断结果是否正确。参考下面的"过往病例"的科室分配情况, 辅助做出决策。输出要求: (你一定要按下面的要求进行思考) 1、认为没有错误则输出"正确", 有错误则输出"错误", 并且说明原因!!! 2、一定要先回答, 然后再回答原因! 当前病人情况和诊断结果: "[result]" 过往病历: "[case1][case2][case3]" 输出示例: 认为正确的输出示例: (认为正确则无需输出原因)" 正确" 认为错误的输出示例: (认为错误要说明原因, 一定要说明原因!!!) "错误原因: {对"体格检查"项目结果的意见}"
Feedback_result_example: 这是医生基于当前病人情况的诊断结果: "[result]"。这是其他医生对诊断结果的意见: "[feedback]"

Mission3: Feedback-Answer
sys_prompt: 你是一名专业的{first_room}医生, 有着丰富的临床诊断经验, 你需要根据病人的情况, 输出诊断结果。
user_input: 请你合理参考其他医生的关于当前病例"诊断"的看法和建议, 基于病人情况和诊室医生的意见按照原来的格式重新对病人进行合适的"诊断"。其他医生的意见不一定正确, 请你合理对待。只输出修改后的"诊断"即可, 无需输出原因和其他说明。可以参考下面的"过往病例"的诊断情况, 辅助做出决策。原来的诊断结果和其他医生的建议: "[feedback]" 过往病例: "\$[past_case]"\$ 输出示例: (直接输出你的选项, 不要输出多余的内容) 输出为单个诊断: "'强直性脊柱炎'" 输出为多个诊断: (用逗号分隔开)" "'风湿性关节炎', '骨折'"

Task5 Treatment Prompt

Mission1: Medical treatment
sys_prompt: 你是一名专业的{first_room}医生, 有着丰富的临床治疗经验, 你需要根据病人的情况, 输出治疗方案。
user_input: 你需要根据下面描述的"病人情况", 为当前患者开具合适的治疗项目。你需要从下面"可选择的治疗项目"中选择合适的项目, 可以选择多个合适的项目。参考下面的"过往病例"的治疗项目情况, 辅助做出决策。病人情况: "[case now]" 过往病例: "\$[case1][case2][case3]"\$ 可选择的治疗项目: ["手术", "介入治疗", "药物治疗", "化学治疗", "抗生素治疗", "放射治疗", "物理疗法", "免疫疗法", "心理治疗", "中医治疗", "基因治疗"] 输出示例: (直接输出你的选项, 不要输出多余的内容) 输出为单个治疗项目: "手术" 输出为多个科室: (用逗号分隔开)" 药物治疗, 抗生素治疗, 物理疗法"

Mission2: Mission1-Feedback
sys_prompt: 你是一名专业的{first_room}医生, 拥有着丰富的临床治疗经验, 您需要从患者的情况分析诊室医生的治疗结果是否合理。
user_input: 请你检查医生开具的"治疗项目"是否正确。参考下面的"过往病例"的科室分配情况, 辅助做出决策。输出要求: (你一定要按下面的要求进行思考) 1、认为没有错误则输出"正确", 有错误则输出"错误", 并且说明原因!!! 2、一定要先回答, 然后再回答原因! 医生基于病人情况开具的治疗项目: "[result]" 过往病历: "[case1][case2][case3]" 输出示例: 认为正确的输出示例: (认为正确则无需输出原因)" 正确" 认为错误的输出示例: (认为错误要说明原因, 一定要说明原因!!!) "错误原因: {对治疗项目结果的意见}"
Feedback_result_example: 以下是您提供的"诊断结果": "[结果]"。"以下是其他医生对您的"诊断结果"的意见: "[反馈]"。

Mission3: Feedback-Answer
sys_prompt: 您是一名专业的第一诊室医生, 有丰富的临床治疗经验, 需要根据患者的情况提供治疗方案。
user_input: 您是一名专业的第一诊室医生, 有丰富的临床治疗经验, 需要根据患者的情况提供治疗方案。请合理考虑其他医生对当前病例"治疗方案"的意见和建议, 并根据患者的病情和医生的意见, 按照原格式重新为患者开出合适的"治疗方案"。您需要从以下"可用的治疗选项"中选择合适的选项, 并且您可以选择多个合适的选项。只输出修改后的"治疗选项", 不提供原因或其他解释。您可以参考"过去的病例"治疗方案来帮助您做出决策。原治疗方案及其他医生建议: "[反馈]" 既往病例: "\$ [past_case] \$" 可用治疗方案: ["手术", "介入治疗", "药物", "化疗", "抗生素治疗", "放射治疗", "物理治疗", "免疫治疗", "心理治疗", "中药", "基因治疗"] 输出示例: (直接输出您的选项, 不输出额外内容) 对于单一治疗方案: "手术" 对于多个选项: (以逗号分隔) "药物治疗, 抗生素治疗, 物理治疗"

Figure 18: The prompt for Task 4 and Task5 in the MedChain-Agent in Chinese.

Task 1 - Input: Patient Information: *Chief Complaint: Right shoulder pain and limited mobility due to trauma, lasting 1 hour. Detailed Situation: Patient fell to the ground while working 1 hour ago, landing on their back. They experienced severe pain in the right shoulder immediately, with limited mobility in the right shoulder and right upper limb. At the time, there was no headache, chest pain, or abdominal pain, and no loss of consciousness. Patient came to our hospital for examination. X-ray shows: right scapular fracture. Proposed admission for right scapular fracture treatment. Since admission, patient appears weak, has not had bowel or bladder movements." Output Example: (If there are multiple options for Secondary Department, separate them with commas) *Primary Department: Obstetrics and Gynecology Secondary Department: Gynecologic Endocrinology, Gynecologic Oncology Reasoning: Because....."
AutoGen Role1: Experienced Triage Nurse *Primary Department: Surgery Secondary Department: Trauma Orthopedics Reasoning: Based on the patient's chief complaint of 'right shoulder pain and limited mobility due to trauma for 1 hour,' and the detailed description of 'right scapular fracture,' we can determine that the patient needs further treatment in the Surgery Department. This is because Surgery is the main department for treating trauma and fractures." Role1: Experienced Triage Nurse *Primary Department: Surgery Secondary Department: Trauma Orthopedics Reasoning: Based on the patient's chief complaint of 'right shoulder pain and limited mobility due to trauma for 1 hour,' and the detailed description of 'right scapular fracture,' we can determine that the patient needs further treatment in the Surgery Department. This is because Surgery is the main department for treating trauma and fractures, and Trauma Orthopedics specifically specializes in treating shoulder fractures." Role1: Experienced Triage Nurse *Primary Department: Surgery Secondary Department: Trauma Orthopedics Reasoning: Based on the patient's chief complaint of 'right shoulder pain and limited mobility due to trauma for 1 hour,' and the detailed description of 'right scapular fracture,' we can determine that the patient needs further treatment in the Surgery Department. This is because Surgery is the main department for treating trauma and fractures, and Trauma Orthopedics specifically specializes in treating shoulder fractures."
DyLAN Agent1: Question: Patient's Chief Complaint: Right shoulder pain and limited mobility due to trauma, lasting 1 hour. *Detailed Situation: Patient fell to the ground while working 1 hour ago, landing on their back. They experienced severe pain in the right shoulder immediately, with limited mobility in the right shoulder and right upper limb. At the time, there was no headache, chest pain, or abdominal pain, and no loss of consciousness. Patient came to our hospital for examination. X-ray shows: right scapular fracture. Since admission, patient appears weak, has not had bowel or bladder movements. *Primary Department: Surgery *Secondary Department: Trauma Orthopedics Agent2: Question: Patient's Chief Complaint: Right shoulder pain and limited mobility due to trauma, lasting 1 hour. *Detailed Situation: Patient fell to the ground while working 1 hour ago, landing on their back. They experienced severe pain in the right shoulder immediately, with limited mobility in the right shoulder and right upper limb. At the time, there was no headache, chest pain, or abdominal pain, and no loss of consciousness. Patient came to our hospital for examination. X-ray shows: right scapular fracture. Since admission, patient appears weak, has not had bowel or bladder movements. *Primary Department: Surgery *Secondary Department: Trauma Orthopedics, Vascular Surgery Agent Importance Scores: [2.0, 1.0] Final Answer: "Primary Department": "Surgery", "Secondary Department": ["Orthopedics", "Vascular Surgery"]
MedAgent Expert Gathering: Trauma Medicine, Pain Management, Orthopedics, Rehabilitation Medicine, Emergency Medicine. Analysis Proposition: Trauma Medicine: 'Right shoulder pain and limited mobility due to trauma for 1 hour. This description indicates the patient experienced right shoulder trauma within the past hour, resulting in pain and limited mobility. This condition may be caused by various factors, including but not limited to shoulder dislocation, rotator cuff tear... Therefore, it is recommended that the patient seek medical attention promptly for detailed physical examination and imaging studies to determine specific diagnosis and treatment plan.' Pain Management: 'Right shoulder pain and limited mobility due to trauma for 1 hour. This description indicates the patient experienced right shoulder trauma within the past hour, resulting in pain and limited mobility. This condition may be caused by various factors, including but not limited to muscle strain, ligament tear, bone... If the patient experiences other symptoms such as numbness, tingling, or sensory abnormalities, immediate medical attention should be sought as these symptoms may indicate nerve damage.' Orthopedics: 'Right shoulder pain and limited mobility due to trauma for 1 hour.' Rehabilitation Medicine: 'Right shoulder pain and limited mobility due to trauma for 1 hour. This description indicates the patient experienced right shoulder trauma within the past hour, resulting in pain and limited mobility. This condition may be caused by various factors, including but not limited to muscle strain, ligament tear, bone... If the patient experiences other symptoms such as numbness, tingling, or sensory abnormalities, immediate medical attention should be sought as these symptoms may indicate nerve damage.' Emergency Medicine: 'Right shoulder pain and limited mobility due to trauma for 1 hour. This description suggests possible shoulder injury, specifically shoulder dislocation, rotator cuff tear, subacromial impingement syndrome, etc. Among these, shoulder dislocation is the most severe injury requiring immediate medical attention. Rotator cuff... Additionally, if the patient has other underlying conditions such as heart disease, hypertension, etc., prompt medical attention is needed to avoid complications.' Report Summarization: 'Key Points: Right shoulder pain and limited mobility due to trauma for 1 hour. The most concerning aspect is shoulder dislocation, as delayed treatment may lead to permanent joint dysfunction. Additionally, if the patient has other underlying conditions such as heart disease, hypertension, etc., prompt medical attention is needed to avoid complications. Overall Analysis: Right shoulder pain and limited mobility due to trauma for 1 hour. The most concerning aspect is shoulder dislocation, as delayed treatment may lead to permanent joint dysfunction. Additionally, if the patient has other underlying conditions such as heart disease, hypertension, etc., prompt medical attention is needed to avoid complications.' Collaborative Consultation: Trauma Medicine: yes Pain Management: yes Orthopedics: yes Rehabilitation Medicine: yes Emergency Medicine: yes Decision Making: Prompt: Based on the discussion results from each expert above, choose the most appropriate department for patient triage from the given primary and secondary departments. Final Answer: '-Primary Department: Emergency Department\n-Secondary Department: Trauma Orthopedics'
MDAgent Complexity Check: basic Final Answer: 'Recommended Department Classification:\n-Primary Department: Orthopedics\n-Secondary Department: Trauma Orthopedics'
MedChain-Agent Primary Department Classification by doctor (Agent): Surgery Feedback: Correct Secondary Department Classification by doctor (Agent): (Subdivision under Surgery as the primary department) Orthopedics Feedback: Correct Final Answer: Primary Department: Surgery Secondary Department: Orthopedics
Ground Truth: Primary Department: "Surgery" Secondary Department: "Orthopedics"

Figure 19: The case study among different multi-agent for task1.

Task 1 - Input: 病人主诉： "病人主诉： 外伤致右侧肩部疼痛活动受限1小时。 详细情况： 患者于1小时前干活时不慎摔倒在地，背部着地，当时感右侧肩部疼痛难忍，右侧肩部及右上肢活动受限，当时无头痛、胸痛、腹痛，无昏迷，入院就诊行x片检查示：右侧肩胛骨骨折，拟右侧肩胛骨骨折收入住院。入院来，患者精神软，大小便未解。" 输出示例：（二级科室若有多个选项则用逗号分开） "一级科室：妇产科 二级科室：妇科内分泌,妇科肿瘤 判断理由：因为....."
AutoGen role1：经验丰富的分诊护士 "一级科室：外科 二级科室：创伤骨科 判断理由：根据病人的主诉，“外伤致右侧肩部疼痛活动受限1小时”，以及详细情况中描述的“右侧肩胛骨骨折”，可以判断病人需要到外科进行进一步治疗因为外科是处理创伤和骨折等外伤的主要科室。” role1：经验丰富的分诊护士 "一级科室：外科 二级科室：创伤骨科 判断理由：根据病人的主诉，“外伤致右侧肩部疼痛活动受限1小时”，以及详细情况中描述的“右侧肩胛骨骨折”，可以判断病人需要到外科进行进一步治疗因为外科是处理创伤和骨折等外伤的主要科室，而创伤骨科是处理肩部骨折的专业科室。” role1：经验丰富的分诊护士 "一级科室：外科 二级科室：创伤骨科 判断理由：根据病人的主诉，“外伤致右侧肩部疼痛活动受限1小时”，以及详细情况中描述的“右侧肩胛骨骨折”，可以判断病人需要到外科进行进一步治疗因为外科是处理创伤和骨折等外伤的主要科室，而创伤骨科是处理肩部骨折的专业科室。”
DyLAN Agent1：问题：病人主诉：外伤致右侧肩部疼痛活动受限1小时。 *详细情况：患者于1小时前干活时不慎摔倒在地，背部着地，当时感右侧肩部疼痛难忍，右侧肩部及右上肢活动受限，当时无头痛、胸痛、腹痛，无昏迷，入院就诊。行X片检查示：右侧肩胛骨骨折。入院来，患者精神软，大小便未解。 *一级科室：外科 *二级科室：创伤骨科 Agent2：问题：病人主诉：外伤致右侧肩部疼痛活动受限1小时。 *详细情况：患者于1小时前干活时不慎摔倒在地，背部着地，当时感右侧肩部疼痛难忍，右侧肩部及右上肢活动受限，当时无头痛、胸痛、腹痛，无昏迷，入院就诊。行X片检查示：右侧肩胛骨骨折。入院来，患者精神软，大小便未解。 *一级科室：外科 *二级科室：创伤骨科、血管外科 Agent Importance Scores：智能体重要性评分：[2.0, 1.0] Final Answer： "一级科室": "外科", "二级科室": ["骨科", "血管外科"]
MedAgent expert gathering: 创伤医学、疼痛管理、骨科、康复医学、急诊医学 analysis proposition: 创伤医学：'外伤致右侧肩部疼痛活动受限1小时。该描述表明患者在过去1小时内经历了右侧肩部的外伤，导致疼痛和活动受限。这种情况可能由多种原因引起，包括但不限于肩关节脱位、肩袖撕裂、...因此，建议患者尽快就医，进行详细的体格检查和影像学检查，以确定具体的诊断和治疗方案。' 疼痛管理：'外伤致右侧肩部疼痛活动受限1小时。该描述表明患者在过去1小时内经历了右侧肩部的外伤，导致疼痛和活动受限。这种情况可能由多种原因引起，包括但不限于肌肉拉伤、韧带撕裂、骨...患者出现其他症状，如麻木、刺痛或感觉异常，应立即就医，因为这些症状可能表明神经损伤。' 骨科：'外伤致右侧肩部疼痛活动受限1小时。' 康复医学：'外伤致右侧肩部疼痛活动受限1小时。该描述表明患者在过去1小时内经历了右侧肩部的外伤，导致疼痛和活动受限。这种情况可能由多种原因引起，包括但不限于肌肉拉伤、韧带撕裂、骨...患者出现其他症状，如麻木、刺痛或感觉异常，应立即就医，因为这些症状可能表明神经损伤。' 急诊医学：'外伤致右侧肩部疼痛活动受限1小时。该描述提示患者可能发生了肩部损伤，具体可能是肩关节脱位、肩袖撕裂、肩峰下撞击症等。其中，肩关节脱位是最严重的损伤，需要立即就医。肩袖...外，如果患者有其他基础疾病，如心脏病、高血压等，也需要及时就医，以避免并发症的发生。' report summarization: '关键点：外伤致右侧肩部疼痛活动受限1小时。最令人担忧的方面是肩关节脱位，因为如果不及治疗，可能会导致永久性的关节功能丧失。此外，如果患者有其他基础疾病，如心脏病、高血压等，也需要及时就医，以避免并发症的发生。总分析：外伤致右侧肩部疼痛活动受限1小时。最令人担忧的方面是肩关节脱位，因为如果不及治疗，可能会导致永久性的关节功能丧失。此外，如果患者有其他基础疾病，如心脏病、高血压等，也需要及时就医，以避免并发症的发生。' Collaborative consultation: 创伤医学: yes 疼痛管理: yes 骨科: yes 康复医学: yes 急诊医学: yes Decision Making: prompt: 基于上面各个专家给出的讨论结果，在下述给定的一级科室和二级科室中选择最合适的科室进行病患分诊。 final answer: '一级科室: 急诊科\n二级科室: 创伤骨科'
MDAgent Complexity Check: basic Final Answer: '建议科室分级: \n一级科室: 骨科\n二级科室: 创伤骨科'
MedChain-Agent 一级科室分类doctor (Agent) : 外科 Feedback: 正确 二级科室分类doctor (Agent) : （外科为一级科室的前提下在进行细分）骨科 Feedback: 正确 Final Answer: 一级科室: 外科 二级科室: 骨科
Ground Truth: 一级科室: "外科" 二级科室: "骨科"

Figure 20: The case study among different multi-agent for task 1in Chinese.

Task 2 - Input: Patient Information: Chief Complaint: Right shoulder pain and limited mobility due to trauma, lasting 1 hour. Detailed Situation: Patient fell to the ground while working 1 hour ago, landing on their back. They experienced severe pain in the right shoulder immediately, with limited mobility in the right shoulder and right upper limb. At the time, there was no headache, chest pain, or abdominal pain, and no loss of consciousness. Patient came to our hospital for examination. X-ray shows: right scapular fracture. Proposed admission for right scapular fracture treatment. Since admission, patient appears weak, has not had bowel or bladder movements. Output Example: Physical Examination: General examination, Chest examination Auxiliary Examination: X-ray, Colonoscopy Reasoning: Because..... AutoGen Role1: Experienced General Practitioner 1 Physical Examination: General examination, Head examination, Chest examination, Abdominal examination, Spine, limb joints and motor function examination Auxiliary Examination: X-ray, Chest X-ray, Abdominal X-ray, Magnetic Resonance Imaging (MRI), Complete Blood Count, Urinalysis Reasoning: Physical examination: Including general examination, head examination, chest examination, abdominal examination, spine, limb joints and motor function examination. Due to the patient's history of trauma, it is necessary to perform general examination, head examination, chest examination, abdominal examination, spine, limb joints and motor function examination. Auxiliary Examination: X-ray, Chest X-ray, Abdominal X-ray, Magnetic Resonance Imaging (MRI), Complete Blood Count, Urinalysis. Due to the patient's history of trauma, X-ray, chest X-ray, abdominal X-ray, and MRI examinations are needed to understand the fracture situation. Additionally, complete blood count and urinalysis are needed to check for other complications." Role2: Experienced General Practitioner 2 Physical Examination: General examination, Head examination, Chest examination, Abdominal examination, Spine, limb joints and motor function examination Auxiliary Examination: X-ray, Chest X-ray, Abdominal X-ray, Magnetic Resonance Imaging (MRI), Complete Blood Count, Urinalysis Reasoning: Due to the patient's history of trauma, it is necessary to perform general examination, head examination, chest examination, abdominal examination, spine, limb joints and motor function examination. Furthermore, due to the fracture, X-ray, chest X-ray, abdominal X-ray, and MRI examinations are needed to understand the fracture situation. Additionally, complete blood count and urinalysis are needed to check for other complications." Role2: Experienced General Practitioner 2 Physical Examination: General examination, Head examination, Chest examination, Abdominal examination, Spine, limb joints and motor function examination Auxiliary Examination: X-ray, Chest X-ray, Abdominal X-ray, Magnetic Resonance Imaging (MRI), Complete Blood Count, Urinalysis Reasoning: Due to the patient's history of trauma, it is necessary to perform general examination, head examination, chest examination, abdominal examination, spine, limb joints and motor function examination. Furthermore, due to the fracture, X-ray, chest X-ray, abdominal X-ray, and MRI examinations are needed to understand the fracture situation. Additionally, complete blood count and urinalysis are needed to check for other complications." DyLAN #### Agent1: - --Routine examination--: Needed to assess overall condition. - --Right shoulder and arm examination--: Critical for evaluating injury extent. - --Chest examination--: Related due to back injury, but lower priority for shoulder issues. - --Neck examination--: Important step in evaluating potential cervical spine issues. - --Advantages--: Includes key examination information, such as routine checks, shoulder/arm and neck assessments. - --Disadvantages--: Not detailed enough, may miss critical neck examination details or more precise shoulder and arm checks. - --Score--: 2.5/5 #### Agent2: - --Routine examination--: Necessary for understanding patient's overall condition. - --Neck range of motion examination--: Important given shoulder area trauma. - --X-ray and CT scan--: Necessary for detailed imaging. - --Right shoulder MRI--: Recommended for better soft tissue injury observation. - --Advantages--: Detailed explanation of various imaging methods, helpful for structural assessment. - --Disadvantages--: Over-emphasizes imaging tests, neglects detailed physical examination. - --Score--: 4.0/5 ### Detailed Updated Answer: 1. --Physical Examination--: - --Routine examination--: Covers vital signs, overall physical condition, and awareness of any emergencies. - Upper limb examination: Especially right shoulder and arm. - --Neck examination--: Consider possible cervical injury given back injury history. - --Neck range of motion examination (ROM)--: Ensure stability and check for any trauma-induced movement restrictions or injuries. 2. --Imaging Studies--: - --Right shoulder X-ray--: For initial bone integrity assessment. - --Right shoulder CT--: For detailed examination of bones, joints, and soft tissue. - --Right shoulder MRI--: Provides comprehensive soft tissue imaging, including rotator cuff, bursa, and ligaments. Particularly important for ligament tears or severe soft tissue injuries not visible on X-ray or CT. 3. --None-- ### Consensus Answer: - --Routine examination-- - --Right shoulder and arm examination-- - --Neck examination and range of motion assessment-- - --Right shoulder and axilla X-rays-- (including acromioclavicular view, sagittal lateral and glenohumeral joint) - --Right shoulder CT scan-- - --Right shoulder MRI-- ### Agent Importance Scores: --Agent Importance Scores--: [2.0, 1.0] ### Final Answer: Auxiliary Examination: "MRI, CT" Physical Examination: "General examination", "Neck examination" MedAgent Expert Gathering: Orthopedics, Pain Management, Rehabilitation Medicine, Emergency Medicine, Surgery Analysis Proposition: Orthopedics: Based on the description, the patient has right shoulder pain and limited mobility due to trauma, which may involve multiple injuries, including but not limited to shoulder dislocation, rotator cuff tear, clavicle fracture, or proximal humeral fracture.\n\nThe most concerning aspect is shoulder dislocation, due to... examination (like X-ray or MRI) for diagnosis, and take appropriate treatment measures based on specific conditions, such as reduction, fixation, or surgery. Pain Management: Right shoulder pain and limited mobility due to trauma for 1 hour.\n\nThis case describes a patient who developed right shoulder pain and limited mobility within 1 hour. This condition may be caused by multiple factors, including but not limited to shoulder dislocation, rotator cuff injury, subacromial... conduct detailed physical examination and imaging studies to determine if there is shoulder dislocation and take appropriate treatment measures promptly. Rehabilitation Medicine: Based on the description, the patient has right shoulder pain and limited mobility due to trauma, which may involve multiple injuries, including but not limited to shoulder dislocation, rotator cuff injury, subacromial impingement syndrome, acromioclavicular joint injury.\n\nThe most concerning aspect is shoulder dislocation, perform detailed physical examination and imaging studies (such as X-ray, MRI, etc.) to clarify diagnosis and develop corresponding treatment plan. Emergency Medicine: This case describes a patient who developed right shoulder pain and limited mobility after trauma. This condition may involve multiple possible diagnoses, including but not limited to shoulder dislocation, rotator cuff injury, subacromial bursitis, acromioclavicular joint injury.\n\nMost... symptoms such as swelling, bruising or sensory abnormalities. This information can help doctors diagnose and treat patients more accurately. Surgery: Right shoulder pain and limited mobility due to trauma for 1 hour.\n\nThis case describes a patient who developed right shoulder pain and limited mobility within 1 hour. This condition may be caused by multiple factors, including but not limited to shoulder dislocation, rotator cuff injury, subacromial... conduct detailed physical examination and imaging studies to determine if there is shoulder dislocation and take appropriate treatment measures promptly. Report Summarization: Key points: Right shoulder pain and limited mobility due to trauma for 1 hour. The most concerning aspect is shoulder dislocation. It is recommended to conduct detailed physical examination and imaging studies, such as X-ray or MRI, as soon as possible to determine whether there is shoulder dislocation or other related injuries.\n\nOverall analysis: This case describes a patient who developed right shoulder pain and limited mobility within 1 hour. This condition may be caused by multiple factors, including but not limited to shoulder dislocation, rotator cuff injury, subacromial bursitis, acromioclavicular arthritis. The most concerning or noteworthy aspect is shoulder dislocation. Shoulder dislocation is a serious injury that can cause shoulder pain, limited mobility and shoulder deformity. If not diagnosed and treated promptly, it may lead to permanent disability. Therefore, for this patient, detailed physical examination and imaging studies, such as X-ray or MRI, should be conducted as soon as possible to determine whether there is shoulder dislocation and take appropriate treatment measures promptly. Collaborative Consultation: Orthopedics: yes Pain Management: yes Rehabilitation Medicine: yes Emergency Medicine: yes Surgery: yes Decision Making: prompt: Based on the discussion results from each expert above, determine appropriate initial physical examination and auxiliary examination. Final answer: "\nPhysical examination: General examination, Visual inspection, Palpation, Movement examination, Neurological examination\nAuxiliary examination: X-ray, MRI" MDAgent Complexity Check: basic Final Answer: "\nPhysical examination: Oral examination\nAuxiliary examination: CT, Complete blood count\n" MedChain-Agent Physical examination doctor group: doctor1: Physical examination: ["General examination", "Head-eyes-ears-nose-throat examination", "Neck examination", "Chest examination", "Abdominal examination", "Spine and extremity examination", "Skin examination"] doctor2: Physical examination: ["General examination", "Head-eyes-ears-nose-throat examination", "Neck examination", "Chest examination", "Abdominal examination", "Spine and extremity examination", "Skin examination"] doctor3: Physical examination: ["General examination", "Head-eyes-ears-nose-throat examination", "Neck examination", "Chest examination", "Abdominal examination", "Spine and extremity examination", "Skin examination"] Summary: Physical examination: ["General examination", "Head-eyes-ears-nose-throat examination", "Neck examination", "Chest examination", "Abdominal examination", "Spine and extremity examination", "Skin examination"] Feedback: Comments on physical examination items (Physical examination items are incomplete, lacking neurological examination) Refine: Physical examination: ["General examination", "Head-eyes-ears-nose-throat examination", "Neck examination", "Chest examination", "Abdominal examination", "Spine and extremity examination", "Skin examination", "Neurological examination"] Auxiliary examination doctor group: doctor1: Auxiliary examination: ["X-ray", "CT", "X-ray film"] doctor2: Auxiliary examination: ["X-ray", "CT"] doctor3: Auxiliary examination: ["X-ray", "CT"] Summary: Auxiliary examination: ["X-ray", "CT"] Feedback: Comments on auxiliary examination (Auxiliary examination items should be selected based on the patient's specific condition, not generalized) Refine: Auxiliary examination: ["X-ray", "CT"] Final Answer: Physical examination: ["General examination", "Head-eyes-ears-nose-throat examination", "Neck examination", "Chest examination", "Abdominal examination", "Spine and extremity examination", "Skin examination", "Neurological examination"] Auxiliary examination: ["X-ray", "CT"] Ground Truth: Physical examination: General examination, Spine and extremity examination Auxiliary examination: X-ray
--

Figure 21: The case study among different multi-agent for task2.

Figure 22: The case study among different multi-agent for task2 in Chinese.

Task 3 - Input:
I need you to analyze these medical images through discussion, and ultimately provide a concise Chinese comprehensive imaging report. Image path: [/datasets/Medimg/20807.png]
AutoGen
Role 1: Experienced radiologist 1 Based on the provided medical images, the patient may have a certain type of tumor. The images show clear abnormal tissue, with irregular borders that contrast significantly with the surrounding tissue. The imaging report recommends further examinations, such as MRI or CT scans, as well as possible pathological examination, to determine the type and stage of the tumor, and further formulate a treatment plan. Role 2: Experienced radiologist 2 The patient may have a certain type of tumor. The images show clear abnormal tissue, with irregular borders that contrast significantly with the surrounding tissue. It is recommended to conduct further examinations, such as MRI or CT scans, as well as possible pathological examination, to determine the type and stage of the tumor, and further formulate a treatment plan. Role 1: Experienced radiologist 1 Based on the provided medical images, the patient may have a certain type of tumor. The images show clear abnormal tissue, with irregular borders that contrast significantly with the surrounding tissue. The imaging report recommends further examinations, such as MRI or CT scans, as well as possible pathological examination, to determine the type and stage of the tumor, and further formulate a treatment plan. Role 2: Experienced radiologist 2 Based on the provided medical images, the patient may have a certain type of tumor. The images show clear abnormal tissue, with irregular borders that contrast significantly with the surrounding tissue. It is recommended to conduct further examinations, such as MRI or CT scans, as well as possible pathological examination, to determine the type and stage of the tumor, and further formulate a treatment plan.
Dylan
Agent1: #### Patient Information - **Name**+: [Patient Name] - **Gender**+: [Patient Gender] - **Age**+: [Patient Age] - **Examination Date**+: [Examination Date] #### Imaging Description This imaging examination includes anteroposterior (AP) and lateral X-ray films. The imaging shows the following: 1. **Skeletal Structure**+: - The bilateral shoulder joints, clavicles, ribs, spine, pelvis, and bilateral femurs are all clearly displayed. - No obvious fractures or abnormalities are seen in the clavicles and ribs. - The spine shows normal physiological curvature, with no significant scoliosis or abnormalities. 2. **Soft Tissue**+: - No obvious abnormalities are seen in the soft tissues around the bilateral shoulder and hip joints. - The bilateral femoral heads and acetabula appear normal. 3. **Internal Fixation**+: - Metallic internal fixation devices, including screws and plates, are visible in the proximal right femur. - Metallic internal fixation devices are also visible in the proximal left femur, including screws and plates. #### Diagnosis and Analysis - **Proximal Right Femur Internal Fixation**+: - The position of the internal fixation in the proximal right femur is good, with no obvious displacement or loosening. - The screws and plates are in normal position, with no signs of foreign body reaction. - **Proximal Left Femur Internal Fixation**+: - The position of the internal fixation in the proximal left femur is good, with no obvious displacement or loosening. - The screws and plates are in normal position, with no signs of foreign body reaction. #### Conclusion - This imaging examination shows that the patient's bilateral proximal femur internal fixation devices are in good position, with no obvious displacement or loosening. - The skeletal structure and soft tissues show no significant abnormalities. #### Recommendations - It is recommended that the patient continue follow-up examinations as instructed by the physician to monitor the condition of the internal fixation devices. - If the patient experiences any discomfort or changes in symptoms, they should seek medical attention promptly. #### Reporting Physician - **Signature**+: [Physician Signature] - **Date**+: [Report Date]
MedAgent
expert gathering: orthopedics, radiology, imaging, thoracic medicine, spine medicine analysis proposition: orthopedics, radiology, imaging, thoracic medicine, spine medicine: 'This X-ray shows an image of the patient's left shoulder and left upper arm. The following points can be observed from the image.\n\n1. Fracture: The humerus (i.e. upper arm bone) of the left upper arm has an obvious fracture near the shoulder joint. The fracture line is clearly visible, indicating that the bone is broken here.\n\n2. Internal fixation device: Metal internal fixation devices, including screws and plates, can be seen at the fracture site. This indicates that the patient may have undergone surgery to stabilize the fracture and promote healing.\n\n3. Soft tissue condition: Around the fracture area, the soft tissue density is high, and there may be swelling or hematoma.\n\n4. Shoulder joint: No obvious abnormality is found in the shoulder joint area, but further examination is needed to rule out possible shoulder joint injury.\n\nThe most worrying aspects:\n- Stability of the fracture: Despite the presence of an internal fixation device, the stability of the fracture needs to be confirmed to prevent displacement or re-fracture.\n- Soft tissue condition: Swelling and hematoma may indicate the risk of infection or other complications, which require close observation and treatment.\n- Rehabilitation progress: Patients need to follow the doctor's rehabilitation instructions to ensure that the fracture heals properly and avoid secondary injury.\n\nIt is recommended that patients seek medical attention as soon as possible for detailed examinations and follow-up treatment to ensure the stability of the fracture and the progress of rehabilitation.'
report summarization:Key points: fracture, internal fixation device, soft tissue condition, shoulder joint, fracture stability, infection risk, rehabilitation progress, detailed examination, follow-up treatment.\n\nTotal analysis: This X-ray shows the image of the patient's left shoulder and left upper arm, in which the humerus of the left upper arm has an obvious fracture near the shoulder joint and has undergone surgery for internal fixation. The soft tissue condition needs further observation to rule out possible infection or other complications. It is recommended that patients seek medical attention as soon as possible for detailed examinations and follow-up treatment to ensure the stability of the fracture and the progress of rehabilitation.'
Collaborative consultation: Orthopedics: yes Radiology: yes Imaging: yes Thoracic medicine: no, Recommendations: 'Revisions: In the report, it is recommended that the patient seek medical attention as soon as possible for detailed examinations and follow-up treatment to ensure the stability of the fracture and the progress of rehabilitation. It is recommended that the specific hospital or clinic that the patient needs to see a doctor can be further clarified, and the patient is advised to perform initial self-care before seeing a doctor, such as ice compresses, elevating the affected limb, etc., to reduce pain and swelling. In addition, it is recommended that the report add recommendations for rehabilitation guidance for patients, such as appropriate physical therapy and rehabilitation training to promote fracture healing and restore function.'
Spinal medicine: no, Recommendations: 'Revisions: In the report, it is recommended that the patient seek medical attention as soon as possible for detailed examinations and follow-up treatment to ensure the stability of the fracture and the progress of rehabilitation. It is recommended that the specific examinations and treatment plans that the patient needs to receive can be further clarified, such as whether physical therapy or rehabilitation training is needed, and whether regular X-rays are needed to monitor fracture healing. In addition, it is recommended to pay attention to the patient's pain management to ensure that the patient can comfortably carry out necessary activities during the rehabilitation process.'
Decision Making: prompt: Based on the discussion results given by the experts above, start with an appropriate imaging report. final answer: Overall analysis: This X-ray shows the patient's left shoulder and left upper arm. The humerus of the left upper arm has an obvious fracture near the shoulder joint and has undergone internal fixation. The soft tissue condition needs further observation to rule out possible infection or other complications. It is recommended that the patient seek medical attention as soon as possible for detailed examination and follow-up treatment to ensure the stability of the fracture and the progress of rehabilitation.'
MDAgent
Complexity Check: basic Final Answer: 'X-ray' 'Imaging report'\nPatient number: XX\nExamination date: XX/XX/XX\n\nImage description:\n1. Chest X-ray images show that there is a large amount of gas in the chest cavity, the lung tissue is compressed, and the transparency of the lung tissue increases, indicating the possible presence of pneumothorax.\n2. X-ray images of the pelvis and bilateral femurs show that there is a metal foreign body in the proximal end of the right femur, and the foreign body is in contact with the acetabular joint, suggesting a possible surgical internal fixation device.\n3. There are metal implants at the upper end of both femurs, which are in the correct position, and no foreign body displacement or breakage is found.\n4. Spine images show that the spine is neatly aligned, without obvious deformity or dislocation.\n5. No obvious abnormal lesions, such as nodules or masses, are found in the lung field.\n\nIn summary, the images suggest that the patient has pneumothorax and may require further clinical evaluation and treatment. X-ray images of the pelvis and femurs show that there is a surgical internal fixation device, which is in a normal position and no abnormalities are found.'
MedChain-Agent
Description by a professional imaging doctor:\n\nImage description:\n1. The X-ray shows a clear fracture line in the right scapula area, which is oblique and extends from the top to the bottom of the scapula.\n2. The bone density around the fracture line is uneven, and there may be slight callus formation.\n3. The joint space of the shoulder joint is well maintained, and there is no obvious joint effusion or joint destruction.\n4. There is no obvious abnormality in the soft tissue structure of the scapula.\n\nConclusion of the imaging report:\n1. The right scapula is fractured, and the fracture line is oblique, extending from the top to the bottom of the scapula.\n2. The bone density around the fracture line is uneven, and there may be slight callus formation.\n3. The joint space of the shoulder joint is well maintained, and there is no obvious joint effusion or joint destruction.\n4. There is no obvious abnormality in the soft tissue structure of the scapula.\n\nRecommendations:\n1. Patients should avoid strenuous activities to avoid further injury.\n2. It is recommended to further perform CT or MRI examination to evaluate the specific situation of the fracture and the soft tissue injury.\n3. Depending on the specific situation of the fracture, surgical fixation or conservative treatment may be required.\n4. Regular follow-up to observe the fracture healing and functional recovery.\n\nDoctor Group: doctor1: The report content is sufficient, no additional content doctor2: The report content is sufficient, no additional content doctor3: The report content is sufficient, no additional content Feedback: Correct Final Answer: 'Image description:\n1. The X-ray shows a clear fracture line in the right scapula area, which is oblique and extends from the top to the bottom of the scapula.\n2. The bone density around the fracture line is uneven, and there may be slight callus formation.\n3. The joint space of the shoulder joint is well maintained, and there is no obvious joint effusion or joint destruction is found'\n4. There is no obvious abnormality in the soft tissue structure of the scapula.\n\nConclusion of the imaging report:\n1. The right scapula is fractured, and the fracture line is oblique, extending from the top to the bottom of the scapula.\n2. The bone density around the fracture line is uneven, and there may be slight callus formation.\n3. The joint space of the shoulder joint is well maintained, and there is no obvious joint effusion or joint destruction.\n4. There is no obvious abnormality in the soft tissue structure of the scapula.\n\nRecommendations:\n1. Patients should avoid strenuous activities to avoid further injury.\n2. Further CT or MRI examinations are recommended to evaluate the specific situation of the fracture and soft tissue injury.\n3. Depending on the specific situation of the fracture, surgical fixation or conservative treatment may be required.\n4. Regular review to observe the fracture healing and functional recovery.\n\n''
Ground Truth:
Comminuted fracture of the right calcaneus, please combine with clinical findings; no obvious fractures were found in the bones of the right ankle joint.

Figure 23: The case study among different multi-agent for task3.

Task 3 - Input: 我需要你们通过讨论，对这些医学影像进行分析，最终提供一个中文简短的综合性影像报告。 影像路径：[/datasets/Medimg/20807.png]
AutoGen role1: 经验丰富的影像科医生1 根据提供的医学影像，患者可能患有某种类型的肿瘤。影像显示有明显的异常组织，边界不规则，与周围组织对比明显。影像报告建议进行进一步的检查，例如MRI或CT扫描，以及可能的病理学检查，以确定肿瘤的类型和分期，并进一步制定治疗计划。 role2: 经验丰富的影像科医生2 患者可能患有某种类型的肿瘤。影像显示有明显的异常组织，边界不规则，与周围组织对比明显。建议进行进一步的检查，例如MRI或CT扫描，以及可能的病理学检查，以确定肿瘤的类型和分期，并进一步制定治疗计划。 role1: 经验丰富的影像科医生1 根据提供的医学影像，患者可能患有某种类型的肿瘤。影像显示有明显的异常组织，边界不规则，与周围组织对比明显。影像报告建议进行进一步的检查，例如MRI或CT扫描，以及可能的病理学检查，以确定肿瘤的类型和分期，并进一步制定治疗计划。 role2: 经验丰富的影像科医生2 根据提供的医学影像，患者可能患有某种类型的肿瘤。影像显示有明显的异常组织，边界不规则，与周围组织对比明显。建议进行进一步的检查，例如MRI或CT扫描，以及可能的病理学检查，以确定肿瘤的类型和分期，并进一步制定治疗计划。
DyLAN Agent1: ##### 患者信息 --姓名--: 患者姓名患者姓名 --性别--: 患者性别患者性别 --年龄--: 患者年龄患者年龄 --检查日期--: 检查日期检查日期 ##### 影像描述 本次影像检查包括前后位 (AP) 和侧位 (Lateral) X线片。影像显示如下: 1. **骨骼结构**: - 双侧肩关节、锁骨、肋骨、脊柱、骨盆及双侧股骨均显示清晰。 - 锁骨和肋骨未见明显骨折或异常。 - 脊柱显示正常生理弯曲，未见明显侧弯或异常。 2. **软组织**: - 双侧肩关节和腕关节周围软组织未见明显异常。 - 双侧股骨头和髌臼显示正常。 3. **内固定物**: - 右侧股骨近端可见金属内固定物，包括螺钉和钢板。 - 左侧股骨近端亦可见金属内固定物，包括螺钉和钢板。 ##### 诊断与分析 --右- 右侧股骨近端内固定物--: - 右侧股骨近端内固定物位置良好，未见明显移位或松动。 - 螺钉和钢板位置正常，未见明显异物反应。 --左- 左侧股骨近端内固定物--: - 左侧股骨近端内固定物位置良好，未见明显移位或松动。 - 螺钉和钢板位置正常，未见明显异物反应。 ##### 结论 - 本次影像检查显示，患者双侧股骨近端内固定物位置良好，未见明显移位或松动。 - 骨骼结构及软组织未见明显异常。 ##### 建议 - 建议患者继续按医嘱进行复查，观察内固定物情况。 - 如有任何不适或症状变化，请及时就医。 ##### 报告医师 --签名--: 医师签名医师签名 --日期--: 报告日期报告日期 ##### Agent2: 同上 Agent Importance Scores: 智能体重要性评分: [0.5, 0.5] Final Answer: "X-ray": ##### 影像报告\n\n#### 患者信息\n\n--姓名--: [患者姓名]\n\n--性别--: [患者性别]\n\n--年龄--: [患者年龄]\n\n--检查日期--: [检查日期]\n\n#### 影像描述\n\n本次影像检查包括前后位 (AP) 和侧位 (Lateral) X线片。影像显示如下: \n\n1. **骨骼结构**:\n\n- 双侧肩关节、锁骨、肋骨、脊柱、骨盆及双侧股骨均显示清晰。 \n\n- 锁骨和肋骨未见明显骨折或异常。 \n\n- 脊柱显示正常生理弯曲，未见明显侧弯或异常。 \n\n2. **软组织**:\n\n- 双侧肩关节和腕关节周围软组织未见明显异常。 \n\n- 双侧股骨头和髌臼显示正常。 \n\n3. **内固定物**:\n\n- 右- 右侧股骨近端可见金属内固定物，包括螺钉和钢板。 \n\n- 左- 左侧股骨近端亦可见金属内固定物，包括螺钉和钢板。 \n\n#### 诊断与分析\n\n--右- 右侧股骨近端内固定物--: \n\n- 右侧股骨近端内固定物位置良好，未见明显移位或松动。 \n\n- 螺钉和钢板位置正常，未见明显异物反应。 \n\n--左- 左侧股骨近端内固定物--: \n\n- 左侧股骨近端内固定物位置良好，未见明显移位或松动。 \n\n- 螺钉和钢板位置正常，未见明显异物反应。 \n\n#### 结论\n\n- 本次影像检查显示，患者双侧股骨近端内固定物位置良好，未见明显移位或松动。 \n\n- 骨骼结构及软组织未见明显异常。 \n\n#### 建议\n\n- 建议患者继续按医嘱进行复查，观察内固定物情况。 \n\n- 如有任何不适或症状变化，请及时就医。 \n\n#### 报告医师\n\n--签名--: [医师签名]\n\n--日期--: [报告日期]"
MedAgent expert gathering: 骨科、放射学、影像学、胸腔医学、脊柱医学 analysis proposition: 骨科、放射学、影像学、胸腔医学、脊柱医学: '这张X光片显示了患者的左肩和左上臂的影像。从影像中可以观察到以下几点: \n\n1. 骨折: 左上臂的肱骨 (即上臂骨) 在靠近肩关节处有明显的骨折。骨折线清晰可见，表明骨骼在此处断裂。 \n\n2. 内固定装置: 在骨折处可以看到金属内固定装置，包括螺钉和钢板。这表明患者可能已经接受了手术治疗，以稳定骨折并促进愈合。 \n\n3. 软组织情况: 在骨折区域周围，软组织密度较高，可能存在肿胀或血肿。 \n\n4. 肩关节: 肩关节区域未见明显异常，但需要进一步检查以排除可能的肩关节损伤。 \n\n5. 最令人担忧的方面: \n\n- 骨折的稳定性: 尽管有内固定装置，但仍需确认骨折的稳定性，以防止移位或再次断裂。 \n\n- 软组织情况: 肿胀和血肿可能提示有感染或其他并发症的风险，需要密切观察和处理。 \n\n- 康复进展: 患者需要遵循医生的康复指导，确保骨折正确愈合，避免二次损伤。 \n\n建议患者尽快就医，进行详细检查和后续治疗，以确保骨折的稳定性和康复进展。' report summarization: '关键点: 骨折、内固定装置、软组织情况、肩关节、骨折稳定性、感染风险、康复进展、详细检查、后续治疗。 \n\n总结分析: 这张X光片显示了患者的左肩和左上臂的影像，其中左上臂的肱骨在靠近肩关节处有明显的骨折，并已进行了内固定装置的手术。软组织情况需要进一步观察，以排除可能的感染或其他并发症。建议患者尽快就医，进行详细检查和后续治疗，以确保骨折的稳定性和康复进展。' Collaborative consultation: 骨科: yes 放射学: yes 影像学: yes 胸腔医学: no, 建议: '修订: 在报告中，建议患者尽快就医，进行详细检查和后续治疗，以确保骨折的稳定性和康复进展。建议可以进一步明确患者需要就医的具体医院或诊所，并建议患者在就医前先进行初步的自我护理，如冰敷、抬高患肢等，以减轻疼痛和肿胀。此外，建议报告中增加对患者进行康复指导的建议，如适当的物理治疗和康复训练，以促进骨折的愈合和恢复功能。' 脊柱医学: no, 建议: '修订: 在报告中，建议患者尽快就医，进行详细检查和后续治疗，以确保骨折的稳定性和康复进展。建议可以进一步明确患者需要接受的具体检查和治疗方案，例如是否需要物理治疗或康复训练，以及是否需要定期复查X光片以监测骨折愈合情况。此外，建议关注患者的疼痛管理，以确保患者在康复过程中能够舒适地进行必要的活动。' Decision Making: prompt: '基于上面各个专家给出的讨论结果，开局合适的影像报告。 final answer: '总结分析: 这张X光片显示了患者的左肩和左上臂的影像，其中左上臂的肱骨在靠近肩关节处有明显的骨折，并已进行了内固定装置的手术。软组织情况需要进一步观察，以排除可能的感染或其他并发症。建议患者尽快就医，进行详细检查和后续治疗，以确保骨折的稳定性和康复进展。'
MDAgent Complexity Check: basic Final Answer: "X-ray": '影像报告\n\n患者编号: XX\n检查日期: XX年XX月XX日\n\n影像描述: \n\n1. 胸片影像显示胸腔内含有大量气体，肺组织被压缩，肺组织透明度增加，提示可能存在气胸。 \n\n2. 骨盆及双侧股骨X线影像显示右侧股骨近端存在金属异物，异物位置与髌臼关节接触，提示可能的手术内固定装置。 \n\n3. 双侧股骨上端有金属内植入物，位置正确，未见异物移位或断裂。 \n\n4. 脊柱影像显示脊柱排列整齐，未见明显畸形或脱位。 \n\n5. 视野中未见明显异常病变，如结节或肿块。 \n\n综上所述，影像提示患者存在气胸，可能需要进一步的临床评估及治疗。 \n\n骨盆及股骨X线影像提示存在手术内固定装置，位置正常，未见异常。"
MedChain-Agent 专业影像学医生描述: "\n\n影像描述: \n\n1. 该X光片显示右侧肩胛骨区域有明显的骨折线，骨折线呈斜行，从肩胛骨的上方延伸至下方。 \n\n2. 骨折线周围的骨质密度不均，可能存在轻微的骨痂形成。 \n\n3. 肩关节的关节间隙保持良好，未见明显关节积液或关节破坏。 \n\n4. 肩胛骨的软组织结构未见明显异常。 \n\n影像报告结论: \n\n1. 右侧肩胛骨骨折，骨折线呈斜行，从肩胛骨的上方延伸至下方。 \n\n2. 骨折线周围的骨质密度不均，可能存在轻微的骨痂形成。 \n\n3. 肩关节的关节间隙保持良好，未见明显关节积液或关节破坏。 \n\n4. 肩胛骨的软组织结构未见明显异常。 \n\n建议: \n\n1. 患者应避免剧烈活动，避免进一步损伤。 \n\n2. 建议进一步行CT或MRI检查，以评估骨折的具体情况及各软组织损伤情况。 \n\n3. 根据骨折的具体情况，可能需要手术固定或保守治疗。 \n\n4. 定期复查，观察骨折愈合情况及功能恢复情况。 \n\n" Doctor Group: doctor1: 报告内容充分，无补充内容 doctor2: 报告内容充分，无补充内容 doctor3: 报告内容充分，无补充内容 Feedback: 正确 Final Answer: "\n\n影像描述: \n\n1. 该X光片显示右侧肩胛骨区域有明显的骨折线，骨折线呈斜行，从肩胛骨的上方延伸至下方。 \n\n2. 骨折线周围的骨质密度不均，可能存在轻微的骨痂形成。 \n\n3. 肩关节的关节间隙保持良好，未见明显关节积液或关节破坏。 \n\n4. 肩胛骨的软组织结构未见明显异常。 \n\n影像报告结论: \n\n1. 右侧肩胛骨骨折，骨折线呈斜行，从肩胛骨的上方延伸至下方。 \n\n2. 骨折线周围的骨质密度不均，可能存在轻微的骨痂形成。 \n\n3. 肩关节的关节间隙保持良好，未见明显关节积液或关节破坏。 \n\n4. 肩胛骨的软组织结构未见明显异常。 \n\n建议: \n\n1. 患者应避免剧烈活动，避免进一步损伤。 \n\n2. 建议进一步行CT或MRI检查，以评估骨折的具体情况及各软组织损伤情况。 \n\n3. 根据骨折的具体情况，可能需要手术固定或保守治疗。 \n\n4. 定期复查，观察骨折愈合情况及功能恢复情况。 \n\n" Ground Truth: 右跟骨粉碎性骨折，请结合临床；右侧踝关节诸骨未见明显骨折。

Figure 24: The case study among different multi-agent for task3 in Chinese.

Task 4 - Input: The patient's comprehensive report is as follows: "Patient's main complaint: right shoulder pain and limited movement for 1 hour due to trauma. Details: The patient accidentally fell to the ground while working 1 hour ago, and landed on his back. At that time, he felt unbearable pain in the right shoulder, and limited movement of the right shoulder and right upper limb. He had no headache, chest pain, abdominal pain, or coma at the time. He was admitted to our hospital for treatment. X-ray examination showed: right scapula fracture. He was hospitalized for right scapula fracture. Since admission, the patient was weak and had difficulty urinating and defecating." Only reply with the name of one or more diagnosis results, and do not include any other explanations. Output example: "Diagnosis result: fracture, knife wound Judgment reason: because..."
AutoGen role1: Experienced general practitioner 1 "Diagnosis: right scapula fracture Judgment reason: X-ray examination shows: right scapula fracture." role2: Experienced general practitioner 2 "Diagnosis: right scapula fracture Judgment reason: X-ray examination shows: right scapula fracture." role2: Experienced general practitioner 2 "Diagnosis: right scapula fracture Judgment reason: X-ray examination shows: right scapula fracture." role1: Experienced general practitioner 1 "Diagnosis: right scapula fracture Judgment reason: X-ray examination shows: right scapula fracture."
DyLAN Agent 1: Shoulder joint Agent 2: Shoulder dislocation Agent Importance Scores: [2.0, 1.0] Final Answer: Diagnosis: Shoulder dislocation
MedAgent expert gathering: Traumatology, Orthopedics, Pain Management, Imaging, Rehabilitation analysis proposition: Traumatology: The patient was diagnosed with right shoulder pain and limited mobility due to trauma, and was treated at Traumatology and Orthopedics. X-rays showed degenerative lesions of bilateral shoulder joints and a history of surgery on the right shoulder joint. The most worrying aspect is the history of surgery on the right shoulder joint, which may indicate that the patient has...disease, which may further affect the function of the shoulder joint. Therefore, timely treatment is required to avoid further loss of joint function. Orthopedics: The patient was diagnosed with right shoulder pain and limited mobility due to trauma, and was treated at Traumatology and Orthopedics. X-rays showed degenerative lesions of bilateral shoulder joints and a history of surgery on the right shoulder joint. The most worrying aspect is the history of surgery on the right shoulder joint, which may indicate that the patient has...disease, which may further affect the function of the shoulder joint. Therefore, timely treatment is required to avoid further loss of joint function. Pain management: According to the information provided, the patient has right shoulder pain and limited mobility due to trauma, and X-rays show bilateral degenerative lesions of the shoulder joints and a history of surgery on the right shoulder joint. These findings suggest that the patient may have chronic problems with the shoulder joint, such as rotator cuff injury, shoulder degeneration... Treatment options: This may include physical therapy, medication, or surgical intervention, depending on the patient's specific situation and medical history. Imaging: According to the information provided, the patient has right shoulder pain and limited mobility due to trauma, and X-rays show bilateral degenerative lesions of the shoulder joints and a history of surgery on the right shoulder joint. These findings suggest that the patient may have chronic problems with the shoulder joint, such as rotator cuff injury, shoulder degeneration... Further imaging examinations, such as MRI or CT scans, are required to determine the specific pathological changes and develop an appropriate treatment plan. Rehabilitation: The patient was treated by trauma orthopedics and orthopedics due to right shoulder pain and limited mobility caused by trauma. X-rays show bilateral degenerative lesions of the shoulder joints and a history of surgery on the right shoulder joint. The most worrying aspect is the surgical history of the right shoulder joint, which may indicate that the patient has a loss of joint function in the next step. It is recommended that the patient seek medical attention as soon as possible for detailed examination and treatment to relieve symptoms and prevent worsening of the condition. report summarization: Key points: Traumatic right shoulder pain and limited activity for 1 hour. X-rays show degenerative lesions of bilateral shoulder joints and a history of surgery on the right shoulder joint. The most worrying aspect is the surgical history of the right shoulder joint, which may indicate that the patient has previous shoulder surgery, such as shoulder replacement or revision surgery. Although these surgeries can relieve pain and improve function, they may also increase the risk of postoperative complications such as infection, dislocation, or prosthesis loosening. In addition, the degenerative lesions shown in the X-rays may indicate that the patient has osteoarthritis or other joint diseases, which may further affect the function of the shoulder joint. Timely treatment is required to avoid further loss of joint function. Summary analysis: The patient was treated in the trauma orthopedics and orthopedics department due to right shoulder pain and limited activity caused by trauma. X-rays show degenerative lesions of bilateral shoulder joints and a history of surgery on the right shoulder joint. The most worrying aspect is the surgical history of the right shoulder joint, which may indicate that the patient has previous shoulder surgery, such as shoulder replacement or revision surgery. Although these surgeries can relieve pain and improve function, they may also increase the risk of postoperative complications, such as infection, dislocation, or loosening of the prosthesis. In addition, the degenerative changes shown in the X-ray may indicate that the patient has osteoarthritis or other joint diseases, which may further affect the function of the shoulder joint. Therefore, timely treatment is required to avoid further loss of joint function. Collaborative consultation: Traumatology: yes Orthopedics: yes Pain management: yes Imaging: yes Rehabilitation: yes Decision Making: prompt: Based on the discussion results given by the experts above, make a reasonable diagnosis. If no decision can be made, output None. final answer: None
MDAgent Complexity Check: intermediate Expert Recruitment: Agent 1 (👤 1. Orthopedic Surgeon): Specializes in the surgical and non-surgical; Agent 2 (👤 2. Radiologist): Focuses on the interpretation and diagnosis of medical images such as X; Agent 3 (👤 3. Traumatologist): Specializes in the diagnosis, treatment, and rehabilitation of trauma Collaborative Decision Making: Hierarchy Selection Participatory Debate (orthopedic surgeon): Trauma (radiologist): Multiple comminuted fractures of the left tibia and fibula with soft tissue swelling of the right calf and left ankle (traumatologist): Fracture of the left tibia and fibula, fracture of the left patella Analysis and Synthesis: == Round 1 == [Turn 1 Agent 1 (👤 orthopedic surgeon) -> Agent 1 (👤 orthopedic surgeon) : As an orthopedic surgeon who specializes in both surgical and non-surgical treatments of musculoskeletal injuries and disorders, I believe that the patient requires a thorough evaluation of her fractures and potential complications. The opinions from the radiologist and traumatologist are valuable, but an orthopedic surgeon's expertise is crucial in determining the most appropriate management plan, which may include surgical intervention if needed. It's essential to address any potential issues that may arise from the fractures, such as joint instability, infection, or non-union.; Agent 2 (👤 radiologist): No; Agent 3 (👤 traumatologist): No [Turn 2 Agent 1 (👤 orthopedic surgeon): No; Agent 2 (👤 radiologist): No; Agent 3 (👤 traumatologist): No Final Decision: Right scapula fracture
MedChain-Agent Doctor group: doctor1: fracture doctor2: scapula fracture doctor3: fracture Summary: fracture Feedback: correct Final Answer: fracture
Ground Truth: fracture

Figure 25: The case study among different multi-agent for task4.

Task 4 - Input:
患者综合报告如下: "病人主诉: 外伤致右侧肩部疼痛活动受限1小时。 详细情况: 患者于1小时前干活时不慎摔倒在地, 背部着地, 当时感右侧肩部疼痛难忍, 右侧肩部及右上肢活动受限, 当时无头痛、胸痛、腹痛, 无昏迷, 入我院就诊, 行x片检查示: 右侧肩胛骨骨折, 拟右侧肩胛骨骨折收入住院。入院来, 患者精神软, 大小便未解。" 仅回复一个或多个诊断结果的名称, 不要包含其他任何解释。 输出示例: "诊断结果: 骨折,刀伤 判断理由: 因为....."
AutoGen role1: 经验丰富的全科医生1 "诊断结果: 右侧肩胛骨骨折 判断理由: X片检查示: 右侧肩胛骨骨折。" role2: 经验丰富的全科医生2 "诊断结果: 右侧肩胛骨骨折 判断理由: X片检查示: 右侧肩胛骨骨折。" role2: 经验丰富的全科医生2 "诊断结果: 右侧肩胛骨骨折 判断理由: X片检查示: 右侧肩胛骨骨折。" role1: 经验丰富的全科医生1 "诊断结果: 右侧肩胛骨骨折 判断理由: X片检查示: 右侧肩胛骨骨折。"
DyLAN Agent1: 肩关节 Agent2: 肩关节脱位 Agent Importance Scores: [2.0, 1.0] Final Answer: 诊断结果: 肩关节脱位
MedAgent expert gathering: 创伤骨科、骨科、疼痛管理、影像学、康复 analysis proposition: 该患者因外伤导致右侧肩部疼痛和活动受限, 就診于创伤骨科和骨科。X光片显示双侧肩关节的退行性病变和右侧肩关节的手术史。最令人担忧的方面是右侧肩关节的手术史, 这可能表明患者有...疾病, 这可能会进一步影响肩关节的功能。因此, 需要及时处理以避免进一步的关节功能丧失。 骨科: 该患者因外伤导致右侧肩部疼痛和活动受限, 就診于创伤骨科和骨科。X光片显示双侧肩关节的退行性病变和右侧肩关节的手术史。最令人担忧的方面是右侧肩关节的手术史, 这可能表明患者有...疾病, 这可能会进一步影响肩关节的功能。因此, 需要及时处理以避免进一步的关节功能丧失。 疼痛管理: 根据提供的信息, 患者因外伤导致右侧肩部疼痛和活动受限, X光片显示双侧肩关节的退行性病变和右侧肩关节的手术史。这些发现表明患者可能存在肩关节的慢性问题, 如肩袖损伤、肩关节退...的治疗方案。这可能包括物理治疗、药物治疗或手术干预, 具体取决于患者的具体情况和病史。 影像学: 根据提供的信息, 患者因外伤导致右侧肩部疼痛和活动受限, X光片显示双侧肩关节的退行性病变和右侧肩关节的手术史。这些发现表明患者可能存在肩关节的慢性问题, 如肩袖损伤、肩关节退...要进一步的影像学检查, 如MRI或CT扫描, 以确定具体的病理变化和制定适当的治疗计划。 康复: 该患者因外伤导致右侧肩部疼痛和活动受限, 就診于创伤骨科和骨科。X光片显示双侧肩关节的退行性病变和右侧肩关节的手术史。最令人担忧的方面是右侧肩关节的手术史, 这可能表明患者有...步的关节功能丧失。建议患者尽快就医, 进行详细的检查和治疗, 以缓解症状并防止病情恶化。 report summarization: 关键点: 外伤致右侧肩部疼痛活动受限1小时。X光片显示双侧肩关节的退行性病变和右侧肩关节的手术史。最令人担忧的方面是右侧肩关节的手术史, 这可能表明患者有既往的肩部手术, 如肩关节置换或修复手术。这些手术虽然可以缓解疼痛和改善功能, 但也有可能增加术后并发症的风险, 如感染、脱位或假体松动。此外, X光片显示的退行性病变可能表明患者有有关节炎或其他关节疾病, 这可能会进一步影响肩关节的功能。需要及时处理以避免进一步的关节功能丧失。 Collaborative consultation: 创伤骨科: yes 骨科: yes 疼痛管理: yes 影像学: yes 康复: yes Decision Making: prompt: 基于上面各个专家给出的讨论结果, 进行合理的诊断。若无法做出决策, 则输出None。 final answer: None
MDAgent Complexity Check: intermediate Expert Recruitment: Agent 1 (👤 1. Orthopedic Surgeon): Specializes in the surgical and non-; Agent 2 (👤 2. Radiologist): Focuses on the interpretation and diagnosis of medical images such as X-; Agent 3 (👤 3. Traumatologist): Specializes in the diagnosis, treatment, and rehabilitation of trauma Collaborative Decision Making: Hierarchy Selection Participatory Debate (orthopedic surgeon): 创伤 (radiologist): 左胫腓骨多处粉碎性骨折伴右小腿及左踝部软组织肿胀 (traumatologist): 左胫腓骨骨折, 左髌骨骨折 Analysis and Synthesis: == Round 1 == └Turn 1 Agent 1 (👤 orthopedic surgeon) -> Agent 1 (👤 orthopedic surgeon) : As an orthopedic surgeon who specializes in both surgical and non-surgical treatments of musculoskeletal injuries and disorders, I believe that the patient requires a thorough evaluation of her fractures and potential complications. The opinions from the radiologist and traumatologist are valuable, but an orthopedic surgeon's expertise is crucial in determining the most appropriate management plan, which may include surgical intervention if needed. It's essential to address any potential issues that may arise from the fractures, such as joint instability, infection, or non-union; Agent 2 (👤 radiologist): No; Agent 3 (👤 traumatologist): No └Turn 2 Agent 1 (👤 orthopedic surgeon): No; Agent 2 (👤 radiologist): No; Agent 3 (👤 traumatologist): No Final Decision: 右侧肩胛骨骨折
MedChain-Agent Doctor group: doctor1: 骨折 doctor2: 肩胛骨骨折 doctor3: 骨折 Summary: 骨折 Feedback: 正确 Final Answer: 骨折
Ground Truth: 骨折

Figure 26: The case study among different multi-agent for task4 in Chinese.

Task 5 - Input: The patient's comprehensive report is as follows: *Patient's main complaint: right shoulder pain and limited movement for 1 hour due to trauma. Details: The patient accidentally fell to the ground while working 1 hour ago, and landed on his back. At that time, he felt unbearable pain in the right shoulder, and limited movement of the right shoulder and right upper limb. He had no headache, chest pain, abdominal pain, or coma at the time. He was admitted to our hospital for treatment. X-ray examination showed: right scapula fracture. He was planned to be hospitalized for right scapula fracture. Since admission, the patient was weak and had difficulty urinating and defecating." Only reply with the name of one or more treatment items, do not include any other explanations, and make sure the answer is among the above eleven items!!! If there are multiple options, separate them with commas. Output example: *Treatment items: Traditional Chinese medicine treatment, gene therapy, surgery Judgment reason: because....."
AutoGen role1: Experienced general practitioner 1 *Treatment: Surgery Judgment reason: The patient has a scapula fracture and needs surgical treatment." role2: Experienced general practitioner 2 *Treatment: Surgery Judgment reason: The patient has a scapula fracture and needs surgical treatment." role2: Experienced general practitioner 2 *Treatment: Surgery Judgment reason: The patient has a scapula fracture and needs surgical treatment."
DyLAN Agent1: *Treatment: Surgery Agent2: *Treatment: Surgery Agent Importance Scores: [0.5, 0.5] Final Answer: *Treatment: Surgery
MedAgent Expert gathering: Traumatology, Orthopedics, Pain Management, Imaging, Rehabilitation Medicine analysis proposition: Traumatology: The patient had limited right shoulder pain for 1 hour due to trauma, and X-ray examination showed a right scapula fracture. The patient had no history of hypertension, diabetes, coronary heart disease, infectious diseases such as hepatitis and tuberculosis, no history of drug and food allergies, and no history of trauma surgery. After admission, he suffered from... changes and a history of surgery on the right shoulder joint. These lesions and surgical history may lead to further loss of joint function and need to be treated in time. Orthopedics: The patient had limited right shoulder pain for 1 hour due to trauma, and X-ray examination showed a right scapula fracture. The patient had no history of hypertension, diabetes, coronary heart disease, infectious diseases such as hepatitis and tuberculosis, no history of drug and food allergies, and no history of trauma surgery. After admission, he suffered from... changes and a history of surgery on the right shoulder joint. These lesions and surgical history may lead to further loss of joint function and need to be treated in time. Pain management: The patient had limited right shoulder pain and activity for 1 hour due to trauma, and X-ray examination showed a right scapula fracture. The patient had no history of hypertension, diabetes, coronary heart disease, infectious diseases such as hepatitis and tuberculosis, no history of drug and food allergies, and no history of trauma surgery. Admission... Changes and right shoulder joint surgery history. These lesions and surgical history may lead to further loss of joint function and need to be treated in time. Imaging: The patient had limited right shoulder pain and activity for 1 hour due to trauma, and X-ray examination showed a right scapula fracture. The patient had no history of hypertension, diabetes, coronary heart disease, infectious diseases such as hepatitis and tuberculosis, no history of drug and food allergies, and no history of trauma surgery. Admission... Changes and right shoulder joint surgery history. These lesions and surgical history may lead to further loss of joint function and need to be treated in time. Rehabilitation medicine: The patient had limited right shoulder pain and activity for 1 hour due to trauma, and X-ray examination showed a right scapula fracture. The patient had no history of hypertension, diabetes, coronary heart disease, infectious diseases such as hepatitis and tuberculosis, no history of drug and food allergies, and no history of traumatic surgery. After admission, he suffered from... and a history of surgery on the right shoulder joint. These lesions and surgical history may lead to further loss of joint function and need to be treated in time. report summarization: Key points: The patient had right shoulder pain and limited activity for 1 hour due to trauma, and X-ray examination showed a right scapula fracture. The patient had no history of hypertension, diabetes, coronary heart disease, infectious diseases such as hepatitis and tuberculosis, no history of drug and food allergies, and no history of traumatic surgery. After admission, the patient was weak in spirit and had unresolved urination and defecation. The most worrying aspect is the degenerative lesions of the bilateral shoulder joints and the history of surgery on the right shoulder joint. These lesions and surgical history may lead to further loss of joint function and need to be treated in time. \n\nTotal analysis: The patient had right shoulder pain and limited activity for 1 hour due to trauma, and X-ray examination showed a right scapula fracture. The patient has no history of hypertension, diabetes, coronary heart disease, infectious diseases such as hepatitis and tuberculosis, no history of drug and food allergies, and no history of trauma surgery. After admission, the patient was weak and had difficulty urinating and defecating. The most worrying aspect is the degenerative lesions of the bilateral shoulder joints and the surgical history of the right shoulder joint. These lesions and surgical history may lead to further loss of joint function and need to be treated in time. ' Collaborative consultation: Traumatology: yes Orthopedics: yes Pain management: yes Imaging: yes Rehabilitation medicine: yes Decision Making: prompt: Please propose the treatment items that need to be taken for this patient based on the patient's chief complaint, past history, current medical history, physical examination results, visiting department and diagnosis results provided below. final answer: Surgery
MDAgent Complexity Check: intermediate Expert Recruitment: Agent 1 (👤 1. Pulmonologist): Specializes in the diagnosis and treatment of respiratory system disorders. Agent 2 (👤 2. Trauma Surgeon): Focuses on the diagnosis and treatment of injuries caused by trauma. Agent 3 (👤 3. Orthopedic Surgeon): Specializes in the surgical treatment of musculoskeletal disorders and injuries. Collaborative Decision Making: Hierarchy Selection Participatory Debate (pulmonologist): antibiotic treatment (trauma surgeon): interventional treatment (orthopedic surgeon): surgery, antibiotic treatment, physical therapy Analysis and Synthesis: == Round 1 == Turn 1 Agent 1 (👤 pulmonologist): No Agent 2 (👤 trauma surgeon): No Agent 3 (👤 orthopedic surgeon): No Final Decision: Medication
MedChain-Agent Doctor group: doctor1: surgery, physical therapy doctor2: surgery doctor3: surgery Summary: surgery Feedback: correct Final Answer: surgery Data summary: extract 12-dimensional information from the current case and add it to RAG—DataSet
Ground Truth: Surgery

Figure 27: The case study among different multi-agent for task5.

 Task 5 - Input: 患者综合报告如下: "病人主诉: 外伤致右侧肩部疼痛活动受限1小时。 详细情况: 患者于1小时前干活时不慎摔倒在地, 背部着地, 当时感右侧肩部疼痛难忍, 右侧肩部及右上肢活动受限, 当时无头痛、胸痛、腹痛, 无昏迷, 入我院就诊, 行X片检查示: 右侧肩胛骨骨折。拟右侧肩胛骨骨折收入住院。入院来, 患者精神软, 大小便未解。" 仅回复一个或多个治疗项目的名称, 不要包含其他任何解释, 并确保回答在上述十一项之中!!! 若有多个选项, 则用逗号隔开。 输出示例: "治疗项目: 中医治疗, 基因治疗, 手术 判断理由: 因为....." AutoGen role1: 经验丰富的全科医生1 "治疗项目: 手术 判断理由: 患者有肩胛骨骨折, 需要手术治疗。" role2: 经验丰富的全科医生2 "治疗项目: 手术 判断理由: 患者有肩胛骨骨折, 需要手术治疗。" role2: 经验丰富的全科医生2 "治疗项目: 手术 判断理由: 患者有肩胛骨骨折, 需要手术治疗。" DyLAN Agent1: 治疗项目: 手术 Agent2: 治疗项目: 手术 Agent Importance Scores: [0.5, 0.5] Final Answer: 治疗项目: 手术 MedAgent expert gathering: 创伤骨科、骨科、疼痛管理、影像学、康复医学 analysis proposition:创伤骨科: 患者因外伤致右侧肩部疼痛活动受限1小时, X片检查示右侧肩胛骨骨折。患者无高血压、糖尿病、冠心病等病史, 无肝炎、肺结核等传染病史, 无药物及食物过敏史、无外伤手术史。入院后患_变和右侧肩关节的手术史。这些病变和手术史可能导致进一步的关节功能丧失, 需要及时处理。 骨科: 患者因外伤致右侧肩部疼痛活动受限1小时, X片检查示右侧肩胛骨骨折。患者无高血压、糖尿病、冠心病等病史, 无肝炎、肺结核等传染病史, 无药物及食物过敏史、无外伤手术史。入院后患_变和右侧肩关节的手术史。这些病变和手术史可能导致进一步的关节功能丧失, 需要及时处理。 疼痛管理: 患者因外伤致右侧肩部疼痛活动受限1小时, X片检查示右侧肩胛骨骨折。患者既往无高血压、糖尿病、冠心病等病史, 无肝炎、肺结核等传染病史, 无药物及食物过敏史、无外伤手术史。入院_变和右侧肩关节的手术史。这些病变和手术史可能导致进一步的关节功能丧失, 需要及时处理。 影像学: 患者因外伤致右侧肩部疼痛活动受限1小时, X片检查示右侧肩胛骨骨折。患者无高血压、糖尿病、冠心病等病史, 无肝炎、肺结核等传染病史, 无药物及食物过敏史、无外伤手术史。入院后患_变和右侧肩关节的手术史。这些病变和手术史可能导致进一步的关节功能丧失, 需要及时处理。 康复医学: 患者因外伤致右侧肩部疼痛活动受限1小时, X片检查示右侧肩胛骨骨折。患者无高血压、糖尿病、冠心病等病史, 无肝炎、肺结核等传染病史, 无药物及食物过敏史、无外伤手术史。入院后患_变和右侧肩关节的手术史。这些病变和手术史可能导致进一步的关节功能丧失, 需要及时处理。 report summarization: "关键点: 患者因外伤致右侧肩部疼痛活动受限1小时, X片检查示右侧肩胛骨骨折。患者无高血压、糖尿病、冠心病等病史, 无肝炎、肺结核等传染病史, 无药物及食物过敏史、无外伤手术史。入院后患者精神软, 大小便未解。最令人担忧的方面是双侧肩关节的退行性病变和右侧肩关节的手术史。这些病变和手术史可能导致进一步的关节功能丧失, 需要及时处理。"\n\n总结分析: 患者因外伤致右侧肩部疼痛活动受限1小时, X片检查示右侧肩胛骨骨折。患者无高血压、糖尿病、冠心病等病史, 无肝炎、肺结核等传染病史, 无药物及食物过敏史、无外伤手术史。入院后患者精神软, 大小便未解。最令人担忧的方面是双侧肩关节的退行性病变和右侧肩关节的手术史。这些病变和手术史可能导致进一步的关节功能丧失, 需要及时处理。" Collaborative consultation: 创伤骨科: yes 骨科: yes 疼痛管理: yes 影像学: yes 康复医学: yes Decision Making: prompt: 请根据以下提供的病人的主诉、既往史、现病史、查体结果、就诊科室以及诊断结果, 提出需要对该病人采取的治疗项目。 final answer: 手术 MDAgent Complexity Check: intermediate Expert Recruitment: Agent 1 (👤 1. Pulmonologist): Specializes in the diagnosis and treatment of respiratory system disorders. Agent 2 (👤 2. Trauma Surgeon): Focuses on the diagnosis and treatment of injuries caused by trauma. Agent 3 (👤 3. Orthopedic Surgeon): Specializes in the surgical treatment of musculoskeletal disorders and injuries. Collaborative Decision Making: Hierarchy Selection Participatory Debate (pulmonologist): 抗生素治疗 (trauma surgeon): 介入治疗 (orthopedic surgeon): 手术、. 抗生素治疗, 物理疗法 Analysis and Synthesis: == Round 1 == └┐ Turn 1 Agent 1 (👤 pulmonologist): No Agent 2 (👤 trauma surgeon): No Agent 3 (👤 orthopedic surgeon): No Final Decision: 药物治疗 MedChain-Agent Doctor group: doctor1: 手术、物理疗法 doctor2: 手术 doctor3: 手术 Summary: 手术 Feedback: 正确 Final Answer: 手术 Data summary: 把当前病例的情况提取出12维信息, 加入RAG—DataSet Ground Truth: 手术

Figure 28: The case study among different multi-agent for task5 in Chinese.

Task1

作为一名专业的分诊医生，你需要根据病人的主诉，将其分诊至适当的科室。分诊流程分为两步：

1. 首先将病人分诊至唯一的第一级科室。

2. 然后将病人分诊至一个或多个对应的第二级科室。

请仅回复将病人分诊的第一级科室和第二级科室的名称，不要包含其他任何解释。科室必须从以下列表中选择：

护理科：基础护理、内科护理、外科护理

药剂科：西药房、药理实验室、中药房、调剂科

口腔科：口腔粘膜科、牙周科、儿童口腔科、口腔修复科、牙体牙髓科、种植科、正畸科、口腔急诊科、口腔预防科、颌面外科

儿科：小儿感染科、儿童保健科、小儿耳鼻喉、小儿免疫科、小儿血液科、小儿肾内科、小儿内科、小儿心外科、小儿呼吸科、小儿骨科、小儿内分泌科、小儿整形科、小儿消化科、小儿泌尿科、小儿精神科、儿童康复科、小儿神经外科、小儿外科、小儿心内科、小儿皮肤科、新生儿科、小儿神经内科、小儿急诊科

医学影像科：MRI室、CT室、B超科、X线室、彩超科、放射科、心超科、超声科、核医学科

眼科：眼外伤、青光眼、眼眶及肿瘤、眼视光学、角膜科、白内障、小儿眼科、眼底

检验科：体液检验、临床检验室、血液检验、生化室、病理科

外科：胰腺外科、创伤骨科、脊柱外科、外伤科、麻醉疼痛科、骨肿瘤科、泌尿外科、胃肠外科、神经外科、心脏外科、器官移植、普外科、乳腺外科、关节骨科、骨科、血管外科、肝胆外科、手外科、甲状腺外科、心胸外科、整形科、烧伤科、胸外科、肛肠外科、微创外科

皮肤性病科：激光室、性病科、皮肤美容、皮肤科

精神科：精神科

全科：全科

耳鼻咽喉科：耳鼻咽喉科

内科：肾脏内科、结核病科、肝病科、营养科、消化内科、内分泌科、神经内科、传染科、过敏反应科、干部诊疗科、呼吸科、免疫科、心血管内科、血液科、老年病科、普通内科、艾滋病科、风湿科、感染科

急诊科：急诊科

肿瘤科：肿瘤妇科、肿瘤外科、放疗科、肿瘤内科

中医科：中医肛肠科、中医消化科、中医男科、推拿科、中医肿瘤科、中医肾脏内科、中医外科、中医感染内科、中医皮肤科、中医妇产科、中医内分泌、中医心内科、中医医结合科、中医血液科、中医骨伤科、中医免疫内科、中医肝病科、中医神经内科、中医按摩科、中医儿科、中医老年病科、中医五官科、中医精神科、中医呼吸科、中医内科、针灸科

康复科：理疗科、康复科

妇产科：妇科内分泌、生殖中心、产前检查科、妇科肿瘤、计划生育科、泌尿尿科、高危产科、产科、妇科

心理科：心理科

病人主诉：{pation_contribution}

Task2

你是一个医生。已知病人的主诉如下：

{chief_complaints}

你需要通过与病人对话来获取更多信息。体格检查包括：一般检查（包括身高、体重、体温、血压、脉搏等）、头颅眼耳鼻咽喉检查、颈部检查（包括甲状腺、颈部淋巴结）、胸部检查（包括肺部、心脏）、腹部检查、脊柱和四肢检查、皮肤检查、神经系统检查、泌尿生殖系统检查。辅助检查包括：X-ray、MRI、CT、超声、核医学成像、血液学检查、尿液检查、粪便检查、内镜检查、病理检查。在获取一定信息后（如既往史、现病史），请你根据病人情况选择要询问的体格检查和辅助检查情况。每轮对话只问一项或两项，请尽可能多的询问检查项目（体格检查和辅助检查每个至少问一项），直至可以判定病情。最后结束对话时请说“祝您早日康复”。

Task3

请你扮演一名放射科医生，你将接收到一张或多张同一模态的医学影像，这些影像均来自同一位患者。请你尝试对这些影像进行分析，并提供一个简短的综合性影像报告。请勿包含其他任何解释。影像报告：

Task4

作为一名医生，请根据以下提供的病人的主诉、既往史、现病史、查体结果以及就诊科室，作出对该病人的诊断。如果无法作出诊断，请仅回复“None”。请勿包含其他任何解释。

- 主诉：{zhusu}
- 既往史：{jiwangshi}
- 现病史：{xianbingshi}
- 查体结果：{chati}
- 就诊科室：{keshi}

仅回复一个或多个诊断结果的名称，不要包含其他任何解释。

Task5

作为一名医生，请根据以下提供的病人的主诉、既往史、现病史、查体结果、就诊科室以及诊断结果，提出需要对该病人采取的治疗项目。如果无法提出治疗项目，请仅回复“None”。请勿包含其他任何解释。治疗项目必须从以下十一项中选择：

1. 手术
2. 介入治疗
3. 药物治疗
4. 化学治疗
5. 抗生素治疗
6. 放射治疗
7. 物理疗法
8. 免疫疗法
9. 心理治疗
10. 中医治疗
11. 基因治疗

- 主诉：{zhusu}
- 既往史：{jiwangshi}
- 现病史：{xianbingshi}
- 查体结果：{chati}
- 就诊科室：{keshi}
- 诊断结果：{jiegou}

仅回复一个或多个治疗项目的名称，不要包含其他任何解释，并确保回答在上述十一项之中。

Figure 29: The prompt of task 1 to 5 for other baseline.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: See abstract and Section 1.

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: See section 7.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: See section 5 and Appendix A, B, and C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: Dataset is uploaded in <https://huggingface.co/datasets/ljwztc/MedChain>. Code will be released in <https://github.com/ljwztc/MedChain>.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: See section 5 and Appendix A, B, and C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: All model weights are obtained from their official repositories on Hugging Face to ensure consistency and reliability.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: See section 5.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: The research fully conforms to the NeurIPS Code of Ethics, adhering to all ethical guidelines without deviation.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: See in section 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We obtain the permission from the website.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: **[Yes]**

Justification: See the abstract for a link to the dataset, website, and the code, which include details about our new benchmark

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: **[No]**

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: **[No]**

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We test the performance of LLM in our dataset.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.