

# DoctorAgent-RL: A Multi-Agent Collaborative Reinforcement Learning System for Multi-Turn Clinical Dialogue

Yichun Feng<sup>†1,2</sup>, Jiawei Wang<sup>†3</sup>, Lu Zhou<sup>2</sup>, and Yixue Li<sup>\*2,4</sup>

<sup>1</sup>School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences

<sup>2</sup>Guangzhou National Laboratory

<sup>3</sup>Department of EEIS, University of Science and Technology of China

<sup>4</sup>Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences

<sup>†</sup><https://github.com/JarvisUSTC/DoctorAgent-RL>

## Abstract

Large language models (LLMs) have demonstrated excellent capabilities in the field of biomedical question answering, but their application in real-world clinical consultations still faces core challenges. Existing systems rely on a one-way information transmission mode where patients must fully describe their symptoms in a single round, leading to nonspecific diagnostic recommendations when complaints are vague. Traditional multi-turn dialogue methods based on supervised learning are constrained by static data-driven paradigms, lacking generalizability and struggling to intelligently extract key clinical information. To address these limitations, we propose DoctorAgent-RL, a reinforcement learning (RL)-based multi-agent collaborative framework that models medical consultations as a dynamic decision-making process under uncertainty. The doctor agent continuously optimizes its questioning strategy within the RL framework through multi-turn interactions with the patient agent, dynamically adjusting its information-gathering path based on comprehensive rewards from the Consultation Evaluator. This RL fine-tuning mechanism enables LLMs to autonomously develop interaction strategies aligned with clinical reasoning logic, rather than superficially imitating patterns in existing dialogue data. Notably, we constructed MTMedDialog, the first English multi-turn medical consultation dataset capable of simulating patient interactions. Experiments demonstrate that DoctorAgent-RL outperforms existing models in both multi-turn reasoning capability and final diagnostic performance, demonstrating practical value in assisting clinical consultations.

## 1 Introduction

Large language models (LLMs) such as ChatGPT [1], LLaMA [2], and ChatGLM [3] have demonstrated remarkable capabilities in various natural language processing tasks, including open-domain question answering, dialogue generation, and code synthesis [4]. With their strong generalization abilities, these models are increasingly applied to healthcare domains, where they exhibit potential for providing preliminary medical advice and assisting clinical decision-making [5].

The core flaw of current medical LLMs lies in the inherent contradiction between their passive single-turn interaction paradigm and the active information-gathering requirements of clinical diagnosis.

<sup>†</sup>These authors contributed equally to this work.

<sup>\*</sup>Corresponding author: li\_yixue@gzlab.ac.cn

Although existing systems demonstrate a high level of knowledge accuracy in structured medical question-answering tasks, their "answer-only" characteristic forces patients to fully describe their symptoms in a single round of conversation [6]. However, in reality, complex medical conditions often require multiple targeted inquiries to gradually clarify the situation [7]. This misalignment in interaction patterns leads to two systemic limitations: On the one hand, users tend to provide vague or fragmented initial descriptions, and the diagnostic suggestions generated by the models based on incomplete information often lack specificity and may even pose potential risks [8]. On the other hand, existing methods that attempt to introduce multi-turn dialogue capabilities through supervised learning, limited by the pattern imitation of static dialogue datasets, fail to establish a dynamic decision-making mechanism to weigh the value of information against dialogue efficiency [9].

To address these challenges, we propose DoctorAgent-RL—a multi-agent collaborative reinforcement learning framework that reformulates clinical reasoning as a Markov Decision Process (MDP). Within this framework: (1) A high-fidelity patient agent based on LLMs, which generates pathologically consistent responses while mimicking the diversity of real-world communication; (2) A clinician agent initialized by cloning medical actions from real consultation records and refined through reinforcement learning (RL) to master effective questioning strategies; and (3) A consultation evaluator that provides multi-dimensional rewards based on diagnostic accuracy, patient information responsiveness, and the standardization of questions. Notably, we constructed MTMedDialog, the first English multi-turn medical consultation dataset capable of simulating realistic patient interactions. This dataset enables the training and evaluation of multi-agent systems in dynamic clinical reasoning scenarios. DoctorAgent-RL redefines the diagnostic process through reward-based strategic policy optimization: The doctor agent continuously optimizes its questioning strategy within a RL framework, obtaining immediate feedback through multi-turn interactions with the patient agent, and dynamically adjusting the information-gathering path based on the comprehensive rewards provided by the consultation evaluator. This end-to-end RL fine-tuning mechanism enables large language models to autonomously evolve interaction strategies that align with clinical reasoning logic, rather than simply imitating surface patterns in existing dialogue data. Experiments demonstrate that DoctorAgent-RL outperforms existing models in both multi-turn reasoning capability and final diagnostic performance, as shown in Figure 1, demonstrating practical value in assisting clinical consultations. Through such multi-agent collaborative training, DoctorAgent-RL allows the model to essentially learn how to ask context-relevant questions and when to confidently terminate the consultation, thereby achieving proactive and strategic patient interaction. Our contributions are three-fold:

- We propose DoctorAgent-RL, a multi-agent collaborative RL framework where the doctor agent autonomously develops clinically-aligned questioning strategies through interactions with the patient agent, guided by the consultation evaluator's comprehensive reward mechanism. DoctorAgent-RL dynamically adjusts information-gathering pathways, overcoming the limitations of conventional static data-driven paradigms.
- We construct MTMedDialog, the first English multi-turn medical dialogue dataset capable of simulating patient interactions.
- DoctorAgent-RL achieves state-of-the-art performance in multi-turn medical consultations, demonstrating exceptional capabilities in generating high-quality questions and delivering more accurate clinical diagnoses.

## 2 Related Work

### 2.1 Medical QA Systems

Despite achieving high accuracy in standardized benchmark tests, current medical LLMs are still restricted by their passive single-turn QA pattern [10]. Systems like MedAlpaca [11] and MedDialog [12] concentrate on optimizing single-response generation via supervised fine-tuning, with BioMistral [13], a medical adaptation of Mistral [14], demonstrating particularly strong few-shot performance on clinical benchmarks, which implicitly presumes that users can offer comprehensive symptom descriptions. Nevertheless, this assumption contradicts real-world scenarios where patients typically initial with vague complaints. Although models such as HuaTuo [15] and UltraMedical [16] enhance

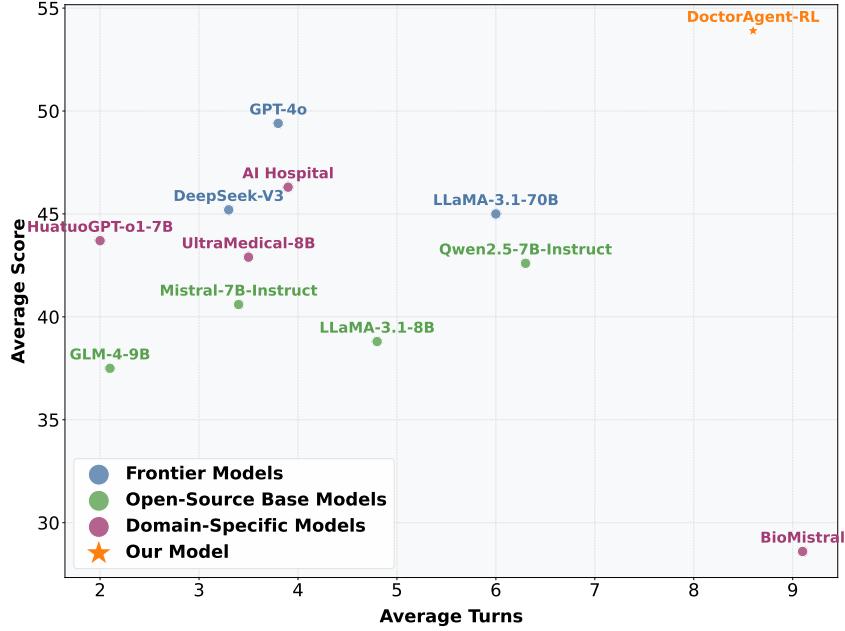


Figure 1: Comparison of the average accuracy scores for diagnosis and recommendation, as well as the average interaction rounds, between DoctorAgent-RL and other models.

answer quality by incorporating medical knowledge graphs, they still function reactively, compelling patients to serve as self-diagnosticians by interpreting general advice like “consult a doctor”. This limitation arises from framing medical QA as a language generation task rather than a sequential decision process. Current approaches prioritize maximizing token-level prediction accuracy but fail to model the clinician’s key challenge of strategically eliciting critical information through adaptive questioning [17]. Thus, even state-of-the-art systems struggle with ambiguous cases where diagnosis relies on iterative hypothesis refinement, a capability beyond single-turn interaction design.

## 2.2 Multi-turn Dialogue Systems

Although systems like Bianque [18] and DialoGPT [19] attempt to introduce multi-turn capabilities through supervised fine-tuning on synthetic or annotated dialogue datasets, they remain constrained by static training paradigms. The standardized multiple-choice evaluation framework [20] aims to mitigate LLMs’ hallucination risks through structured testing, yet its scripted key-value dialogue protocol fails to capture authentic doctor-patient interactions. Recent approaches like the APP system [21] propose entropy-minimized diagnostic optimization, achieving higher symptom recall than conventional methods through medical guideline grounding. MDDial [22] contributes a specialized dataset of 1,200 expert-annotated diagnostic dialogues with turn-level reliability scores, though its template-based generation limits linguistic diversity and complexity, failing to fully reflect the richness of real clinical conversations. These methods still rely on predefined dialogue paths rather than learning optimal questioning strategies from clinical outcomes. While systems incorporating turn reliability scoring and iterative differential diagnosis surpass single-turn response capabilities, they fundamentally suffer from static training data limitations and lack mechanisms to optimize dialogue strategies based on interaction results [23]. Consequently, their ability to ask the right questions at the right time remains inadequate.

## 2.3 Multi-Agent Systems in Medicine

Recent advances in medical multi-agent systems have demonstrated significant potential in clinical reasoning simulation and diagnostic accuracy improvement, yet fundamental limitations persist in achieving human-centric adaptability. AMIE [24] optimize diagnostic dialogues through self-play simulation environments, outperforming primary care doctors in structured evaluations, but their reliance on static training data restricts adaptability to novel clinical scenarios. The MAC

framework [25] enhances rare disease diagnosis by simulating multidisciplinary team discussions, yet its template-driven dialogue patterns lack linguistic diversity and contextual flexibility. Agent Hospital [26] incorporate dynamic medical record libraries and LLM-generated disease progression models to evolve agent behaviors, though the generated experiences carry risks of factual inconsistencies. While the AI Hospital framework [27] improves diagnostic accuracy through multi-role collaboration protocols, it fails to adequately account for the diversity of treatment options and the effectiveness of alternative strategies in real clinical settings. Current systems neither achieve dynamic integration of real-time medical evidence nor adjust questioning depth based on patient comprehension levels, resulting in suboptimal information delivery efficiency.

## 2.4 Reinforcement Learning in Medicine

The application of RL in the medical field is transitioning from static decision-making to dynamic interaction paradigms, demonstrating unique value in optimizing strategies for multi-turn consultation systems. MedVLM-R1 [28] and Med-R1 [29] employ the Group Relative Policy Optimization (GRPO) [30] framework, which integrates RL reward mechanisms with radiological imaging feature analysis to incentivize vision-language models in generating interpretable reasoning pathways. This approach significantly enhances diagnostic accuracy and reduces hallucinatory reasoning, yet exhibits limited adaptability to emerging modalities. HuatuoGPT-01 [31] enhances clinical reasoning through verifiable question generation and medical validation feedback mechanisms, but its reliance on multiple-choice question data conversion limits adaptability to unstructured symptom descriptions. MedRIA [32] employs actor-critic frameworks to optimize inquiry efficiency in emergency scenarios, yet requires manual feature engineering for complex diagnoses. PPME [33] leverages clinical experience replay to prioritize high-value diagnostic pathways, achieving domain-specific accuracy gains at the cost of poor generalization. These studies collectively indicate that RL requires further refinement in dynamic interaction objective modeling and clinically oriented evaluation frameworks to address the complexity and uncertainty inherent in real-world medical practice.

## 3 Method

### 3.1 Dataset and Evaluation Metrics

#### 3.1.1 Dataset creation process

We present MTMedDialog, the first English multi-turn medical consultation dataset capable of simulating patient interactions. The dataset comprises 8,086 training samples and 2,082 test samples derived from three Chinese benchmark datasets: IMCS21 [34], CHIP-MDCFNPC [35], and MedDG [36]. The test set covers 8 major disease categories, with detailed distribution statistics provided Appendix B.

As the source datasets were collected from real doctor-patient conversations, we implemented a two-stage denoising strategy: (1) filtering shallow dialogues with less than three turns through exact turn-count matching, and (2) removing noisy segments containing consecutive meaningless responses using DeepSeek-V3 [37]. We strictly preserved the original data partitioning protocol during this cleaning process to ensure evaluation reliability. The retained dialogues were then translated into English using DeepSeek-V3.

The test set serves dual evaluation purposes: 1) assessing doctor agent’s questioning and diagnostic capabilities using complete dialogue trajectories with gold-standard diagnosis labels; 2) evaluating patient agent’s response quality through a randomly selected subset of 500 samples.

#### 3.1.2 Evaluation metrics

The doctor agent evaluation focuses on two core dimensions: (1) Diagnosis and Recommendation Accuracy: This dimension evaluates eight disease categories using a 6-level quantitative matching scale (0=completely incorrect, 5=exact match). Semantic consistency between the agent’s diagnosis/recommendations and gold-standard labels is assessed via Qwen2.5-32B-Instruct [38], with specific prompts shown in Appendix C. The final score is converted to a 100-point scale by multiplying the raw score by 20; (2) Interaction Turns: The average number of dialogue turns required for the doctor agent and the patient agent to reach a final diagnosis.

The patient agent assessment employs a three-dimensional DeepSeek-V3-based scoring system: (1) Information Control: Quantifies unsolicited information disclosure through a baseline score of 1.0 (perfect compliance), with 0.2-point deductions per unauthorized disclosure instance; (2) Response Completeness: Evaluates critical information omission against doctor queries using an initial score of 1.0, penalized by 0.2 points per essential detail omission; (3) Factual Conflict: Detects contradictions with medical records through a violation counter starting at 0.0, accumulating 0.2 points per identified inconsistency. Patient agent’s rating prompts shown in Appendix C.

### 3.2 Task Formulation

We model the multi-round medical consultation process as a multi-agent collaborative RL system, where the doctor agent serves as the main agent for strategy optimization, the patient agent acts as a collaborative counterpart to form a dynamic game relationship with it, and the consultation evaluator functions as a neutral arbiter that coordinates doctor-patient interactions and guides the optimization process of the doctor agent’s strategy through well-designed reward mechanisms.

#### 3.2.1 Doctor Agent

As the decision-making doctor agent, its state space  $s_t \in \mathcal{S}$  encompasses the dialogue history  $H_t$ , providing a comprehensive record of the consultation. The agent’s actions are drawn from the action space  $\mathcal{A} = \{a_{query}, a_{diagnose}\}$ , which includes two distinct behaviors: generating medical inquiries and executing diagnostic decisions.

Environmental dynamics are governed by the patient agent through state transitions:

$$s_{t+1} \sim P(s_{t+1}|s_t, a_t)$$

Here,  $P$  represents the transition probability function, which quantifies the likelihood of moving to a new state  $s_{t+1}$  given the current state  $s_t$  and the action  $a_t$  taken by the doctor agent.

The consultation continues until either a predefined limit on conversation turns is reached or the doctor agent provides a final diagnosis, at which point a complete dialogue history  $H_T$  forms the consultation trajectory.

The reward signal  $R \in \mathbb{R}$  for each trajectory is provided by an independent consultation evaluator (detailed in Section 3.2.3), considering diagnostic accuracy, information acquisition efficiency, and protocol compliance as metrics.

To enhance the stability of policy optimization and eliminate the requirement for an additional value function approximation, we propose Group Relative Policy Optimization (GRPO) [30] as our policy gradient algorithm. Unlike Proximal Policy Optimization (PPO) [39], GRPO employs the average reward of multiple sampled outputs as a baseline instead of relying on a learned value function. Specifically, for each patient’s consultation  $x$ , GRPO samples a set of trajectories  $y_1, y_2, \dots, y_G$  through the interaction between the doctor agent  $\pi_D$  and the patient agent  $\pi_p$ . The doctor agent, as the policy model, is then optimized by maximizing the following objective function:

$$\begin{aligned} \mathcal{J}_{GRPO}(\theta) = & \mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^G \sim \pi_{D_{old}}(\cdot|x; \pi_p)} \left[ \frac{1}{G} \sum_{i=1}^G \frac{1}{\sum_{t=1}^{|y_i|} I(y_{i,t})} \sum_{t=1: I(y_{i,t})=1}^{|y_i|} \min \left( \frac{\pi_D(y_{i,t}|x, y_{i,<t})}{\pi_{D_{old}}(y_{i,t}|x, y_{i,<t})} \hat{A}_{i,t}, \right. \right. \\ & \left. \left. \text{clip} \left( \frac{\pi_D(y_{i,t}|x, y_{i,<t})}{\pi_{D_{old}}(y_{i,t}|x, y_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right) - \beta \mathbb{D}_{KL} [\pi_D || \pi_{D_{ref}}] \right], \end{aligned} \quad (1)$$

where  $\epsilon$  and  $\beta$  are hyperparameters, and  $\hat{A}_{i,t}$  represents the advantage, computed based on the relative rewards of outputs within each group. Here,  $I(y_{i,t}) = 1$  indicates that the token  $y_{i,t}$  is generated by  $\pi_D$ . Since doctors cannot predict patients’ symptoms in advance, during training, the responses of the patient agent are masked.

#### 3.2.2 Patient Agent

The patient agent is implemented using Qwen2.5-7B-Instruct [38] and is incorporated into a two-phase dialogue simulation framework via carefully crafted prompt engineering.

In the first phase, the system combines patient self-reports and multi-turn dialogue content to create case descriptions. Additionally, it augments potential symptom features using standard diagnostic results, thereby forming a more comprehensive hidden medical profile. This design effectively mitigates the symptom coverage issues stemming from incomplete doctor inquiries in traditional datasets.

In the second phase, the patient agent utilizes dynamic symptom release strategies in response to the real-time queries of the doctor agent. It maintains strict pathological consistency while mimicking the natural variability in patients' symptom description granularity and the order of their complaints. The detailed prompt designs for both phases are presented in Appendix E. By retaining complete hidden case data, the patient agent ensures that its natural language responses adhere to clinical standards and are generated solely based on the dialogue history.

### 3.2.3 Consultation Evaluator

In our RL framework, guiding the doctor agent to master essential clinical diagnostic skills is paramount. To achieve this, we've designed a sophisticated Consultation Evaluator, acting as a multi-faceted reward system that assesses the agent's performance across critical dimensions of a medical consultation. This evaluator comprises three core components, each contributing to a comprehensive assessment of the agent's diagnostic capabilities and consultation conduct.

**Diagnostic Accuracy Reward.** The first pillar of our Consultation Evaluator focuses on the agent's diagnostic precision and treatment recommendations. To ensure the reliability of this evaluation and prevent any potential reward hacking, we employ a rule-based reward mechanism. This mechanism meticulously calculates the word-level F1 score between the doctor agent's predicted diagnosis and the gold-standard diagnosis, as well as for the recommended treatments. The formulation for this reward is as follows:

$$R_{\text{accuracy}} = 5 \times (\text{F1}_{\text{diagnosis}} + \text{F1}_{\text{recommendation}}) \quad (2)$$

The coefficient 5 serves to adjust the relative weight of this crucial reward signal within the overall evaluation. This design ensures a balanced and robust assessment of both the diagnostic output and the quality of suggested interventions, providing a stable foundation for learning.

**Information Acquisition Efficiency Reward.** A proficient clinician knows how to ask the right questions efficiently. To foster this skill in our doctor agent, the Consultation Evaluator incorporates a dynamic reward mechanism that promotes valuable questioning while discouraging repetitive or unhelpful queries. This reward is directly tied to the patient agent's feedback after each dialogue turn and accumulates throughout the interaction:

$$R_{\text{information}}^t = \begin{cases} 1, & \text{if the patient agent answers normally} \\ -2, & \text{if the patient agent refuses to answer} \end{cases} \quad (3)$$

$$R_{\text{information}} = \sum_t R_{\text{information}}^t \quad (4)$$

Through this feedback loop, the model learns to optimize its questioning strategy, prioritizing the acquisition of diagnostically relevant information and refining its inquiry process.

**Protocol Compliance Reward.** Adherence to established clinical interview protocols is a hallmark of professional medical practice. To instill this discipline, our Consultation Evaluator includes a compliance reward. This component penalizes deviations from predefined norms and ensures that the agent completes the diagnostic process within specified limits:

$$R_{\text{compliance}}^t = \begin{cases} -2, & \text{if the question format violates predefined norms} \\ -5, & \text{if no diagnosis is provided within the allowed turns} \\ 0, & \text{otherwise (i.e., protocol is followed)} \end{cases} \quad (5)$$

$$R_{\text{compliance}} = \sum_t R_{\text{compliance}}^t \quad (6)$$

This mechanism reinforces the learning of structured interview processes and ensures the timely completion of a diagnosis, closely mirroring the practical constraints and expectations of real-world clinical environments.

By combining these three critical components, the total consultation evaluation score, or reward, received by the doctor agent at each time step  $t$  is calculated as:

$$R = R_{\text{accuracy}} + R_{\text{information}} + R_{\text{compliance}} \quad (7)$$

This sophisticated, multi-dimensional Consultation Evaluator not only guides the model toward superior diagnostic accuracy but also actively encourages the development of efficient information-gathering strategies and professional, compliant clinical behavior, ultimately aiming to achieve diagnostic capabilities that closely resemble those of expert human clinicians.

### 3.3 Training Procedure

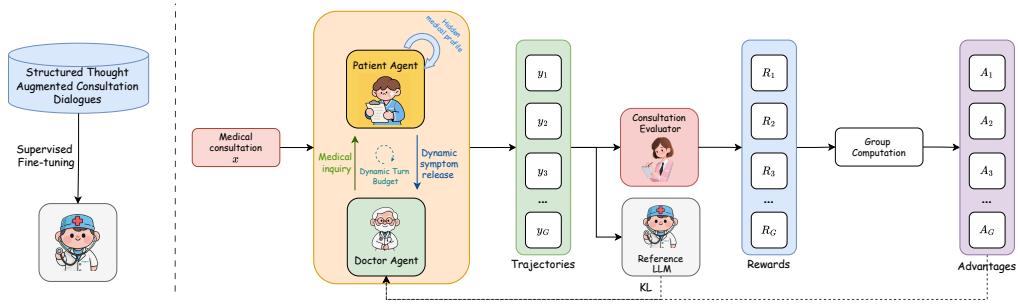


Figure 2: The multi-agent collaborative reinforcement learning framework for DoctorAgent-RL. During the rollout stage, multi-turn interactions are conducted between the doctor agent and the patient agent.

As illustrated in Figure 2, our training framework for the doctor agent is built upon Qwen2.5-7B-Instruct, following the DeepSeek-R1 training paradigm [40]. The approach employs a two-stage training pipeline, integrating SFT and RL to cultivate clinical reasoning capabilities. Detailed experimental settings are illustrated in Appendix H.

Specifically, we randomly select 1,000 multi-turn consultation dialogues from the training corpus. Each doctor’s query in the sampled dialogues is augmented with structured thought processes using DeepSeek-V3. These include hypothesis generation, evidence evaluation, and differential diagnosis steps derived from context. The doctor agent is fine-tuned on this enriched dataset to activate core capabilities in question-asking, diagnostic reasoning, and recommendation generation.

Following SFT, we refine the agent’s decision-making under interaction constraints using the policy optimization algorithm detailed in Section 3.2. To enhance robustness and mimic real-world clinical scenarios, we introduce a **Dynamic Turn Budget Training Strategy**. Each training episode is assigned a random dialogue turn budget (2–10 turns). The model is explicitly reminded of the remaining turns after each interaction step, encouraging efficient information gathering. This two-stage approach ensures the agent first internalizes clinical reasoning patterns via supervised learning, then refines its strategy through interactive reward optimization. Ablation studies comparing alternative training strategies (direct SFT, direct RL, fixed-turn training) are presented in the experimental section.

## 4 Experiments

### 4.1 Patient Agent Behavior Analysis on MTMedDialog

Table 1 presents the evaluation results of the effectiveness of patient agent in simulating real-world medical interactions. In the dimension of information control, Qwen2.5-7B demonstrates optimal compliance performance, precisely constraining output content and effectively avoiding the leakage of irrelevant information. In terms of response completeness, while DeepSeek-V3 performs best,

Model	Information Control	Response Completeness	Factual Conflict
DeepSeek-V3	86.4	<b>86.1</b>	0.0
HuatuoGPT-o1-7B	88.1	81.7	0.0
Qwen2.5-7B-Instruct	<b>88.8</b>	84.4	0.0
LLaMA-3.1-8B	80.3	77.2	0.0
GLM-4-9B	72.6	70.7	0.0

Table 1: Performance comparison of different models in simulating patient agent on MTMedDialog. The best results are highlighted in bold.

Qwen2-5.7B still maintains near-optimal levels of key information density required for doctor-patient dialogue. Notably, all tested models achieve zero error rates in the dimension of factual conflicts, fully validating the reliability of LLMs in medical knowledge. Considering both model performance and cost-effectiveness in training, we ultimately selected Qwen2-5.7B as the implementation solution for the patient agent. For a more detailed description of the dialogue flow, please refer to Appendix G

#### 4.2 Comparative Performance Evaluation of Multi-Model Approaches on MTMedDialog

Experimental results in Table 2 show DoctorAgent-RL achieving a comprehensive average score of 53.9%, demonstrating significant improvements over frontier models, open-source base models, and domain-specific models. More detailed experimental results can be found in Appendix F. The system particularly maintains stable advantages in disease types requiring in-depth consultation, proving its superior ability to simulate doctors' clinical consultation processes.

Analysis of model dialogue processes reveals key limitations in existing methods' interaction quality. While all models receive explicit instructions to "ask only one question at a time," frontier models show strict compliance but lack professional medical consultation knowledge, often missing critical symptom information in their questions. Some open-source models (e.g., GLM-4, Mistral) achieve decent comprehensive average scores but improperly combine multiple questions, reflecting inadequate instruction-following capability that affects complex condition diagnosis. The domain-specific model BioMistral exhibits the highest interaction frequency but frequently repeats similar questions due to ineffective planning, resulting in the lowest diagnostic accuracy. These findings demonstrate that question quality - not quantity - determines diagnostic effectiveness.

DoctorAgent-RL employs a phased training strategy: initial fine-tuning with medical dialogue data containing detailed reasoning processes establishes systematic consultation capabilities, followed by RL that optimizes questioning strategies through simulated doctor-patient dialogues. This approach enables strict compliance with consultation norms while allowing flexible, doctor-like question adjustment based on acquired symptom information. The resulting intelligent questioning strategy significantly improves comprehensive average scores while maintaining reasonable interaction frequency.

#### 4.3 Adaptive Fine-Tuning Strategies for Task-Specific Optimization

Experimental results demonstrate the significant advantages of our proposed two-stage optimization framework (SFT + RL) in medical dialogue tasks. As shown in Table 3, DoctorAgent-RL achieves a superior average score of 53.9, outperforming all baseline models. The key advantage stems from our phased strategy: (1) SFT on doctor-patient dialogues establishes reasonable questioning baselines through behavioral cloning, followed by (2) RL optimization that enables dynamic adjustment of questioning strategies for high-value information acquisition. This combined approach improves the average diagnosis and recommendation score by 25.9% compared to the base model while enhancing proactive questioning efficiency by 36.7%. Ablation studies on three critical components further validate our design:

**w/o Dynamic\_Turn:** When trained with fixed budget of turns during RL, the model shows only a 1.2% performance drop in matched scenarios but reveals strategy rigidity during inference—it mechanically adheres to the training budget of turns regardless of specified variations, proving impractical for real-world deployment.

Model	DSD	RSD	ID	GSD	ND	CSD	ED	SD	Avg. Score	Avg. Turns
<i>Frontier Models</i>										
GPT-4o	<b>49.5</b>	<b>50.7</b>	<b>48.6</b>	<b>46.8</b>	<b>47.9</b>	<b>52.5</b>	<b>46.3</b>	41.5	<b>49.4</b>	3.8
DeepSeek-V3	44.8	46.6	44.2	45.0	45.1	46.3	43.7	43.0	45.2	3.3
LLaMA-3.1-70B	45.4	44.5	45.6	46.1	42.1	44.9	39.5	39.0	45.0	6.0
<i>Open-Source Base Models</i>										
GLM-4-9B	37.5	38.3	39.7	35.2	39.1	38.3	40.0	30.5	37.5	2.1
LLaMA-3.1-8B	38.8	38.7	39.3	37.6	39.0	39.6	36.3	41.5	38.8	4.8
Mistral-7B-Instruct	40.3	41.5	39.8	41.9	44.2	39.8	36.3	35.5	40.6	3.4
Qwen2.5-7B-Instruct	43.2	42.2	40.8	43.6	39.5	45.5	35.0	40.5	42.6	6.3
<i>Domain-Specific Models</i>										
BioMistral	28.8	28.9	29.2	23.9	29.5	25.6	37.0	25.5	28.6	<b>9.1</b>
UltraMedical-8B	42.8	43.8	39.9	41.7	44.7	45.4	42.2	38.5	42.9	3.5
HuatuoGPT-o1-7B	43.4	44.5	44.5	41.8	43.3	49.4	45.6	37.0	43.7	2.0
AI Hospital	46.3	47.5	44.8	46.0	46.1	44.4	45.9	<u>43.5</u>	46.3	3.9
DoctorAgent-RL (Ours)	<b>54.5</b>	<b>52.9</b>	<b>55.1</b>	<b>52.4</b>	<b>50.4</b>	<b>57.0</b>	<b>51.4</b>	<b>48.0</b>	<b>53.9</b>	8.6

Table 2: Main results by disease category on MTMedDialog dataset, showing the average of diagnostic accuracy and recommendation accuracy scores. Abbreviations: DSD (Digestive System Diseases), RSD (Respiratory System Diseases), ID (Infectious Diseases), GSD (Genitourinary System Diseases), ND (Neurological Disorders), CSD (Circulatory System Diseases), ED (Endocrine Disorders), SD (Skin Diseases). Avg. Score represents the average of diagnostic and recommendation accuracy scores across all disease categories. Avg. Turns indicates the average number of interaction turns across all disease categories. Best performing metrics are highlighted in bold. The second best results are indicated with underlines.

Method	DSD	RSD	ID	GSD	ND	CSD	ED	SD	Avg. Score	Avg. Turns
DoctorAgent-RL (Ours)	<b>54.5</b>	<b>52.9</b>	<b>55.1</b>	<b>52.4</b>	50.4	<b>57.0</b>	<b>51.4</b>	<b>48.0</b>	<b>53.9</b>	8.6
w/o Dynamic Turn	53.5	52.4	51.4	51.4	<b>50.9</b>	51.8	44.1	45.0	52.7	8.5
w/o SFT	48.4	48.3	48.4	48.8	49.5	50.0	48.2	43.0	48.4	5.8
w/o RL	48.3	46.5	43.6	44.5	49.0	50.2	48.5	36.0	47.4	<b>9.0</b>
Qwen2.5-7B-Instruct	43.2	42.2	40.8	43.6	39.5	45.5	35.0	40.5	42.6	6.3

Table 3: Performance comparison of different fine-tuning methods for Qwen2.5-7B-Instruct on MTMedDialog. The best results are highlighted in bold.

**w/o SFT:** Direct RL training without SFT initialization causes a 5.5% average score degradation with the lowest average turns. While capable of planning effective questions for information gathering, the model demonstrates insufficient initiative in question generation due to the absence of behavioral cloning foundations.

**w/o RL:** SFT-only training results in a 6.5% lower average score despite having the highest turn count. The model memorizes question sequences from training data without truly understanding diagnostic logic, leading to mechanical questioning rather than strategic information acquisition.

The comprehensive performance of DoctorAgent-RL confirms each component’s necessity: SFT establishes reliable behavioral baselines, RL injects dynamic decision-making capability, and adaptive turn mechanisms ensure strategic flexibility—together forming a reproducible paradigm for task-oriented medical dialogue optimization.

#### 4.4 Impact Analysis of Budget of Turns on Diagnosis and Recommendation Performance

The experimental results shown in Figure 3 demonstrate that as the budget of turns increases, the average performance of diagnosis and recommendation exhibits a distinct two-phase characteristic: in the initial phase (low-turn range), performance rises rapidly with additional turns, primarily due to the multi-turn dialogue mechanism enabling the LLM to progressively collect and refine patient information through iterative questioning; in the intermediate phase (medium-to-high turn range), the performance curve’s slope noticeably flattens as the valuable information patients can provide gradually becomes saturated, making it difficult for the LLM to extract more meaningful information through additional questioning. Nevertheless, on the whole, a larger budget of turns still leads to better performance, as additional interaction opportunities can capture potential subtle information differences. It is worth noting that diagnostic performance consistently outperforms recommendation performance, as diagnosis tasks can progressively confirm symptom characteristics through multi-turn

interactions, while recommendation tasks rely more on established medical knowledge bases, leaving relatively limited room for performance improvement.

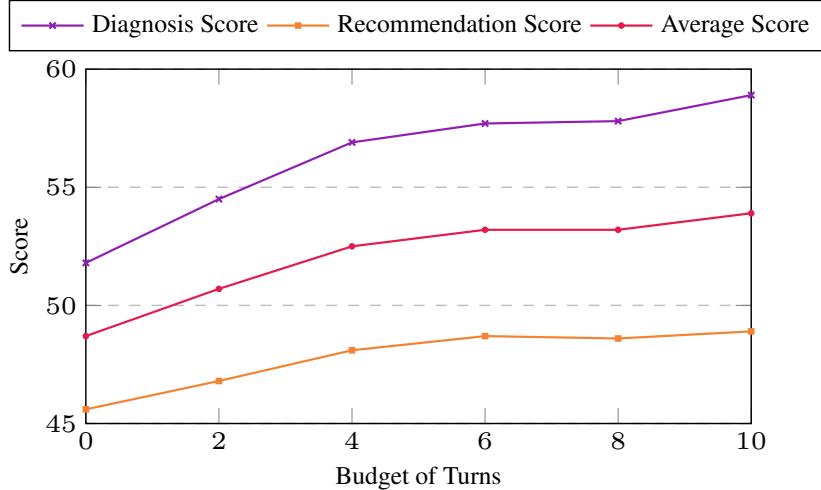


Figure 3: Comparative analysis of Diagnosis Score, Recommendation Score, and the resulting Average Score of DoctorAgent-RL on MTMedDialog at different levels of the Budget of Turns.

## 5 Conclusion and future work

We propose DoctorAgent-RL, a multi-agent collaborative reinforcement learning framework that establishes an innovative paradigm enabling LLMs to progressively refine diagnoses through proactive questioning. The framework effectively addresses the unrealistic requirement of conventional systems for patients to fully describe symptoms in a single interaction by synergistically integrating a high-fidelity patient agent simulating pathologically consistent symptom expressions with a doctor agent optimizing adaptive questioning strategies through reinforcement learning, guided by an evaluator providing multidimensional clinical reward signals. Experimental results demonstrate that DoctorAgent-RL significantly outperforms traditional methods in diagnostic accuracy. These findings not only mark a paradigm shift in medical AI from static Q&A to dynamic reasoning, but also provide clinicians with interpretable decision support through its proactive questioning strategies enabled by multi-turn interactions, while simultaneously pioneering new technical pathways for patient-led symptom screening.

Future research will focus on advancing multimodal medical reasoning capabilities by integrating heterogeneous data sources such as medical imaging, pathological slides, and real-time physiological signals from wearable devices to enhance the model’s comprehensive understanding of complex conditions. Concurrently, systematic efforts are needed to address data bias and decision transparency, establishing a fairness evaluation framework that accounts for demographic characteristics and disease spectrum variations to ensure model robustness and interpretability across diverse patient populations. Throughout technological iterations, an ethical governance framework for medical AI must be developed through interdisciplinary research, addressing critical issues such as diagnostic accountability, privacy and data security, and doctor-patient trust mechanisms, ultimately achieving a dynamic balance between technological innovation and adherence to medical ethics and patient safety standards.

## References

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

- [2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [3] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.
- [4] Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1):3280, 2025.
- [5] Marvin Kopka, Niklas von Kalckreuth, and Markus A Feufel. Accuracy of online symptom assessment applications, large language models, and laypeople for self-triage decisions. *npj Digital Medicine*, 8(1):178, 2025.
- [6] 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.
- [7] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.
- [8] Andrew D Auerbach, Tiffany M Lee, Colin C Hubbard, Sumant R Ranji, Katie Raffel, Gilmer Valdes, John Boscardin, Anuj K Dalal, Alyssa Harris, Ellen Flynn, et al. Diagnostic errors in hospitalized adults who died or were transferred to intensive care. *JAMA Internal Medicine*, 184(2):164–173, 2024.
- [9] Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaotong Zhang. Interactive evaluation for medical llms via task-oriented dialogue system. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4871–4896, 2025.
- [10] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):A1oa2300138, 2024.
- [11] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.
- [12] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9241–9250, 2020.
- [13] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistrail: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.
- [14] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- [15] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023.
- [16] Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081, 2024.
- [17] Harsha Nori, Naoto Usuyama, Nicholas King, Scott Mayer McKinney, Xavier Fernandes, Sheng Zhang, and Eric Horvitz. From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *arXiv preprint arXiv:2411.03590*, 2024.

- [18] Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, Xiangmin Xu, et al. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:2310.15896*, 2023.
- [19] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.
- [20] Yusheng Liao, Yutong Meng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. An automatic evaluation framework for multi-turn medical consultations capabilities of large language models. *arXiv preprint arXiv:2309.02077*, 2023.
- [21] Jiayuan Zhu and Junde Wu. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. *arXiv preprint arXiv:2502.07143*, 2025.
- [22] Srija Macherla, Man Luo, Mihir Parmar, and Chitta Baral. Mddial: A multi-turn differential diagnosis dialogue dataset with reliability evaluation. *arXiv preprint arXiv:2308.08147*, 2023.
- [23] Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.
- [24] Tao Tu, Mike Schaeckermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9, 2025.
- [25] Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159, 2025.
- [26] Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024.
- [27] Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*, 2024.
- [28] Jiazheng Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvilm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025.
- [29] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.
- [30] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [31] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.
- [32] Xuan Zou, Weijie He, Yu Huang, Yi Ouyang, Zhen Zhang, Yu Wu, Yongsheng Wu, Lili Feng, Sheng Wu, Mengqi Yang, et al. Ai-driven diagnostic assistance in medical inquiry: Reinforcement learning algorithm development and validation. *Journal of Medical Internet Research*, 26:e54616, 2024.
- [33] Zhoujian Sun, Ziyi Liu, Cheng Luo, Jiebin Chu, and Zhengxing Huang. Improving interactive diagnostic ability of a large language model agent through clinical experience learning. *arXiv preprint arXiv:2503.16463*, 2025.
- [34] Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. A Benchmark for Automatic Medical Consultation System: Frameworks, Tasks and Datasets. *Bioinformatics*, 2022.

- [35] Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*, 2021.
- [36] Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer, 2022.
- [37] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [38] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.
- [39] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [40] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [41] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyuan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2024.
- [42] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.
- [43] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.
- [44] Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025.
- [45] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- [46] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.
- [47] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

## Appendix

### A Comparison of Diagnostic Consultation Paradigms

This section introduces the comparison of different diagnostic consultation paradigms, as shown in Figure 4.

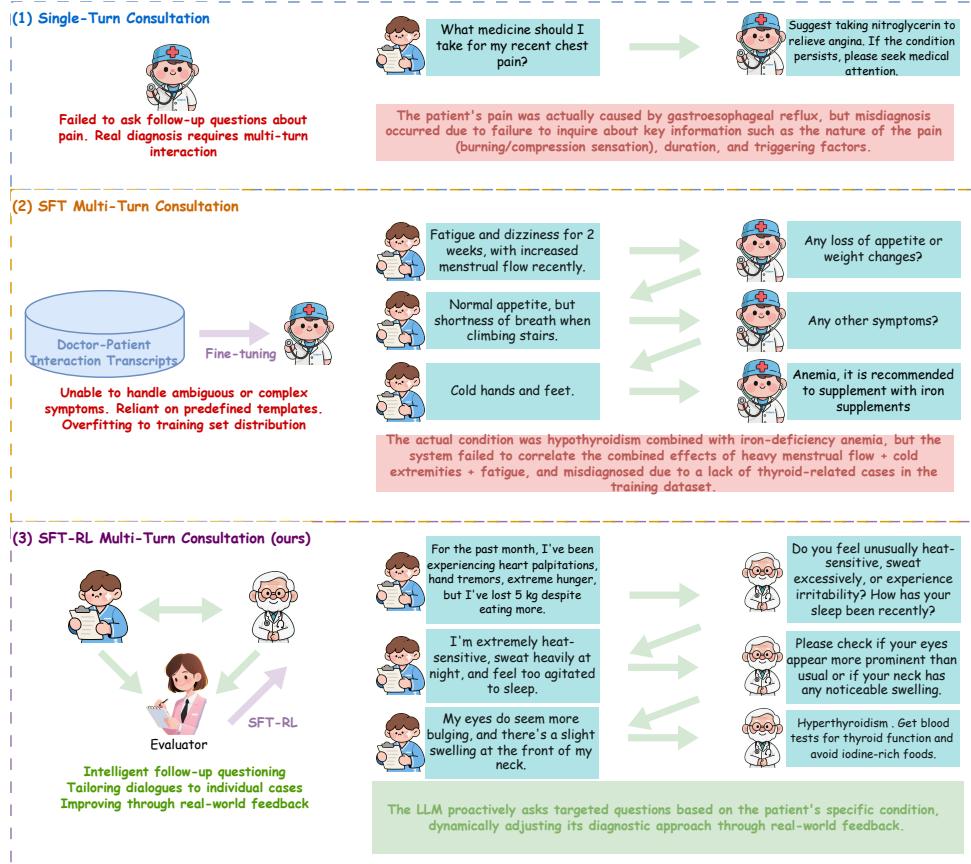


Figure 4: Comparison of diagnostic consultation paradigms. (A) Single-Turn Consultation: Direct diagnosis based on initial query without follow-up interactions, prone to misdiagnosis due to insufficient symptom clarification. (B) SFT Multi-Turn Consultation: Implements basic multi-turn dialogue through supervised fine-tuning but lacks adaptive optimization. (C) SFT-RL Multi-Turn Consultation: Integrates RL with real-world feedback to dynamically optimize questioning strategies, enabling context-aware symptom investigation (e.g., pain nature, duration, triggers) and reducing diagnostic errors through iterative interaction.

### B Details of MTMedDialog

This section presents the statistical distribution of disease categories in the MTMedDialog test set. We employed the DeepSeek-V3 to automatically classify the diagnostic results of each data entry, strictly adhering to a predefined eight-category disease classification system. Samples that did not conform to this classification framework were subsequently removed, resulting in a final collection of 2,082 high-quality test samples. Detailed data are shown in Table 4.

Disease Category	Number of Cases
Digestive System Diseases	1,290
Respiratory System Diseases	402
Infectious Diseases	118
Genitourinary System Diseases	100
Neurological Disorders	77
Circulatory System Diseases	48
Endocrine Disorders	27
Skin Diseases	20

Table 4: Disease Category Distribution in MTMedDialog Test Set

## C Prompts for Evaluation Metrics

This section will provide a detailed explanation of the evaluation prompts for both the doctor agent and patient agent. The evaluation prompt for the doctor agent’s Diagnosis and Recommendation Accuracy is presented in Figure 5, while the evaluation prompt for the patient agent is shown in Figure 6.

**Prompt for Evaluating the Doctor Agent’s Diagnosis and Recommendation Accuracy**

Task: As a medical expert, evaluate the semantic similarity between the model-generated medical text and the ground truth reference. Score on a 0–5 point scale based on meaning alignment (wording differences are acceptable if meaning matches).

Criteria:

- 5: Identical meaning (different wording okay).
- 4: Minor wording/detail differences; overall meaning aligned.
- 3: Partial meaning overlap; important differences exist but core intent is partially shared.
- 2: Limited meaning overlap; key details or context differ significantly.
- 1: Minimal meaning overlap; mostly unrelated or only superficially related.
- 0: Unrelated or completely different meaning (no meaningful semantic connection).

Now evaluate:

Candidate: "{candidate}"

Reference: "{reference}"

OUTPUT FORMAT: <think> [Your Thinking Process] </think><answer> [Your Score] </answer>

Figure 5: Prompt for Evaluating the Doctor Agent’s Diagnosis and Recommendation Accuracy. {candidate} represents the model-generated medical text (doctor agent’s output), while {reference} denotes the ground truth clinical reference standard.

## D Prompt for Doctor Agent

This section elaborates on the prompt design for the doctor agent during both training and inference phases. The complete prompt structure is illustrated in Figure 7.

## E Prompt for Patient Agent

This section first details the patient agent’s implicit disease knowledge generation mechanism, with the complete prompt structure shown in Figure 8. Subsequently, it examines the agent’s prompt design methodology for both training and inference phases, as fully illustrated in Figure 9.

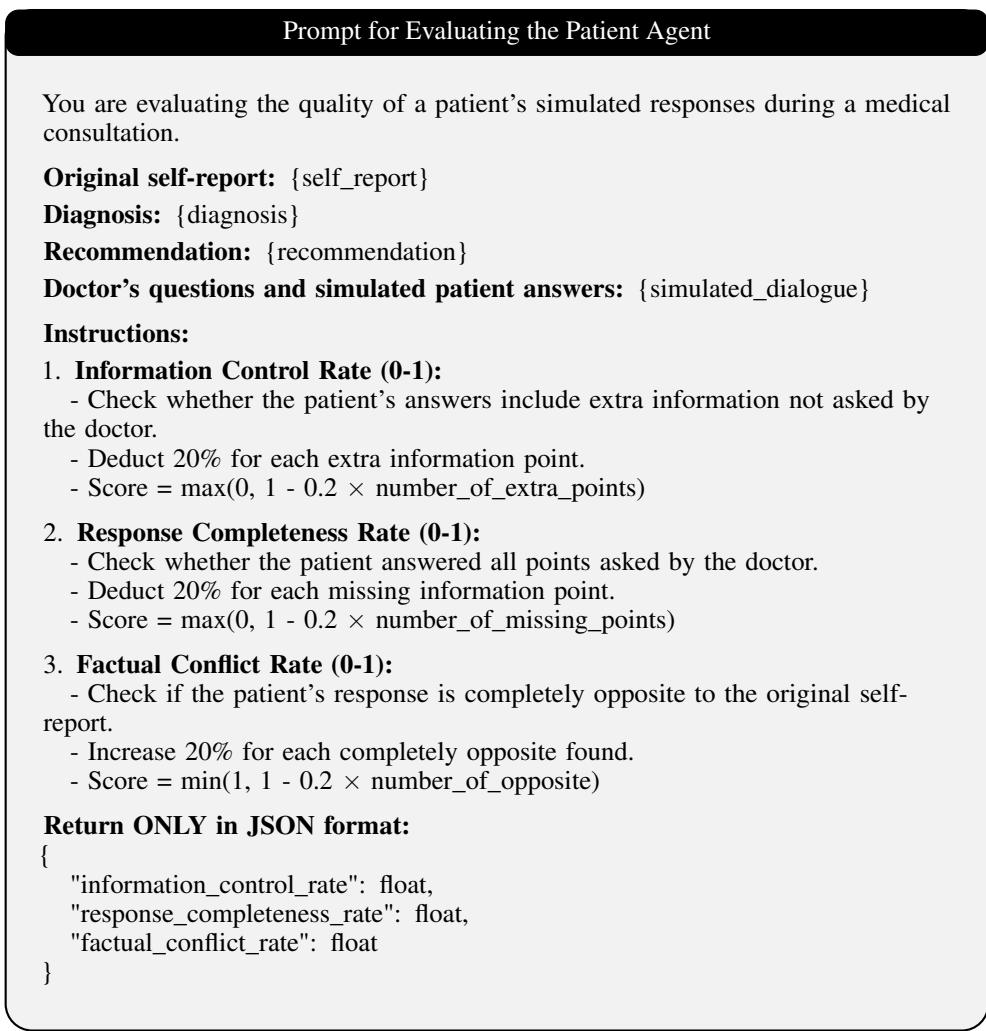


Figure 6: Prompt for Evaluating the Patient Agent.

## F Detailed Comparative Performance Evaluation of Multi-Model Approaches on MTMedDialog

The experimental results of diagnostic accuracy are shown in Table 5. DoctorAgent-RL achieved an average diagnostic accuracy of 58.9% on the MTMedDialog dataset, showing a significant advantage over the comparison models. This result indicates that multi-turn questioning trained with RL can gather more patient information and effectively improve disease identification accuracy.

The experimental results of recommendation accuracy are shown in Table 6. DoctorAgent-RL outperformed the baseline models with an average recommendation accuracy of 48.9%. Notably, the AI Hospital model showed partial advantages in the recommendation tasks for skin diseases and endocrine diseases, possibly due to the smaller test sample sizes for these diseases, which amplified the output randomness of pre-trained models. In contrast, the multi-agent collaborative architecture of AI Hospital, through parallel interactions, more easily obtained effective solutions via probability sampling. Nevertheless, DoctorAgent-RL still maintained the best overall accuracy.

All models exhibited higher diagnostic accuracy than recommendation accuracy. This stems from the core differences between the two tasks. Disease diagnosis, as a symptom-driven reasoning process, has a target space that, while not predefined as a closed set, converges due to pathological constraints. In contrast, treatment recommendation generation requires coordinating multi-dimensional parameters such as drugs, dosages, and treatment durations, leading to an exponentially expanded space of

### Prompt for Doctor Agent

You are an experienced doctor who needs to provide professional diagnosis and advice to patients through consultation. Please listen carefully to the patient's description, ask targeted questions, and collect sufficient information before giving a diagnosis and treatment recommendation.

#### Quick Guide

##### Objectives:

1. Obtain key information through effective questioning, each round of questions should be modified based on the previous round's content, meaning you shouldn't ask similar questions.
2. Comprehensively analyze the patient's condition to provide an accurate diagnosis and appropriate treatment recommendations.

##### Rules:

1. You can only choose one of the options to respond, you cannot both answer questions and provide a diagnosis simultaneously.
2. Absolutely do not repeat or ask questions similar or identical to those previously asked.

##### Response:

<think> [your thinking] </think>

<answer>If you believe there is insufficient information, please only ask one question, in this format:

Question: (your question).

</answer> | <answer>If you believe you have obtained enough information, please only provide diagnosis and recommendations, in this format:

Diagnosis: (the patient's most likely disease or symptoms)

Recommendation: (corresponding treatment plan or advice)

</answer>

##### Rewards:

Incorrect format: -2.0

Effective question (patient can provide an answer and the question is helpful for diagnosis): +1.0

Ineffective questions do not count towards score

Repeated questions: -2.0

The number of conversation turn is limited. Reaching maximum interaction rounds without providing a diagnosis: -5.0

Completely correct diagnosis and recommendations: +10.0

Figure 7: Prompt for Doctor Agent.

possible solutions. Particularly in open scenarios, treatment decisions may have multiple equivalent solutions (e.g., drug substitution therapies), while the evaluation criteria only adopt the optimal path from clinical guidelines, objectively increasing the matching difficulty of the recommendation task. Overall, our method uses RL to teach LLM to dynamically plan questioning paths and infer reasonable answers, achieving state-of-the-art performance in both tasks.

## G Detailed Description of the Dialogue Flow in DoctorAgent-RL

This section provides a detailed explanation of the dialogue flow in DoctorAgent-RL, as illustrated in Figure 10. The framework aims to simulate a realistic diagnostic process by coordinating a structured dialogue flow between two agents: the doctor and the patient. This appendix comprehensively describes the mechanisms for dialogue initiation, progression, and termination.

### Prompt for Patient Agent to Develop a Comprehensive Implicit Health Profile

As a medical assistant, expand the patient's symptom description based on:

- Original self-report: {self\_report}
- Dialogue history: {dialogue\_history}
- Diagnosis: {diagnosis}
- Recommendation: {recommendation}

#### Processing Rules:

1. Summarize the patient's information: Combine the 'Original self-report' and all patient responses from 'dialogue' into a single coherent paragraph. Include only factual patient statements and exclude the doctor's questions. If a patient response only makes sense in the context of the doctor's question, infer its meaning based on the context.
2. Based on diagnosis and recommendations, add medical evidence to clearly support symptoms.
3. Never contradict the patient's original statements.
4. Keep the language natural and clinical.
5. Return ONLY the enhanced description.

Figure 8: Prompt for Patient Agent to Develop a Comprehensive Implicit Health Profile.

### Prompt for Patient Agent Training and Inference

You are interacting with a doctor.

#### Medical Response Instructions:

Answer each medical question concisely in one sentence, strictly describing symptoms while avoiding any mention of diagnoses or recommendations.

If the question is unrelated to your chief complaint, state: "Sorry, I cannot answer this question."

If the question is repetitive, reply: "Sorry, you've already asked this question."

#### Your chief complaint:

{description}

#### doctor's question history:

{history\_questions}

#### Current doctor question:

{question}

#### Output format:

<think>[Your reasoning]</think><answer>[Your response]</answer>

Figure 9: Prompt for Patient Agent Training and Inference.

The diagnostic process unfolds through multiple rounds of interaction between the two agents. The patient's self-report serves as the starting point for the first round of dialogue between the doctor and patient agents. During this phase, the doctor agent actively conducts a comprehensive inquiry, collecting diagnostic information through targeted questioning. The patient agent, functioning as a non-player character (NPC), provides feedback on their condition to the doctor agent in each dialogue round based on carefully designed prompts.

The diagnostic phase concludes when either of the following conditions is met: (1) The doctor agent determines that the collected information is sufficient for diagnosis and directly outputs the diagnostic

Model	DSD	RSD	ID	GSD	ND	CSD	ED	SD	Avg. Diag	Avg. Turns
<i>Frontier Models</i>										
GPT-4o	<u>52.6</u>	<u>54.9</u>	<u>50.5</u>	<u>50.0</u>	<u>53.5</u>	<u>58.3</u>	<u>51.9</u>	46.0	<u>52.6</u>	3.8
DeepSeek-V3	47.3	49.5	49.7	48.6	52.0	52.5	43.0	45.0	48.2	3.3
LLaMA-3.1-70B	46.4	46.6	48.8	48.6	43.4	49.1	38.6	41.0	46.4	6.0
<i>Open-Source Base Models</i>										
GLM-4-9B	43.8	45.1	46.3	41.4	47.8	46.7	43.7	40.0	44.3	2.1
LLaMA-3.1-8B	36.8	38.5	36.8	37.2	36.9	40.0	37.8	<u>47.0</u>	37.5	4.8
Mistral-7B-Instruct	38.2	40.2	37.0	38.0	45.7	39.6	34.1	36.0	38.7	3.4
Qwen2.5-7B-Instruct	47.8	47.2	44.7	50.6	41.3	52.7	40.9	46.0	47.3	6.3
<i>Domain-Specific Models</i>										
BioMistral	29.0	29.8	28.0	23.4	32.5	25.8	40.7	23.0	29.0	<b>9.1</b>
UltraMedical-8B	42.8	45.3	38.5	38.2	47.0	49.2	42.2	43.0	43.1	3.5
HuatuoGPT-01-7B	43.5	46.3	43.2	41.8	48.3	50.8	45.9	40.0	44.4	2.0
AI Hospital	47.0	48.9	45.1	46.2	48.6	46.7	44.4	40.0	47.2	3.9
Ours	<b>59.3</b>	<b>58.3</b>	<b>61.2</b>	<b>55.8</b>	<b>54.3</b>	<b>65.5</b>	<b>58.1</b>	<b>58.0</b>	<b>58.9</b>	<u>8.6</u>

Table 5: Mean diagnostic scores by disease category on MTMedDialog. The best results are highlighted in bold. The second best results are indicated with underlines.

Model	DSD	RSD	ID	GSD	ND	CSD	ED	SD	Avg. Recom	Avg. Turns
<i>Frontier Models</i>										
GPT-4o	<u>46.9</u>	<u>46.5</u>	<u>46.8</u>	43.6	42.3	<u>46.7</u>	40.7	37.0	<u>46.3</u>	3.8
DeepSeek-V3	42.3	43.7	38.8	41.4	38.2	40.0	44.4	41.0	42.1	3.3
LLaMA-3.1-70B	44.5	42.4	42.4	43.6	40.1	40.6	40.5	37.0	43.6	6.0
<i>Open-Source Base Models</i>										
GLM-4-9B	30.5	31.4	33.1	29.0	30.4	30.0	36.3	21.0	30.7	2.1
LLaMA-3.1-8B	40.8	39.0	41.9	38.0	36.9	39.2	34.8	36.0	40.1	4.8
Mistral-7B-Instruct	42.5	42.7	42.7	45.8	42.6	40.0	38.5	35.0	42.5	3.4
Qwen2.5-7B-Instruct	38.6	37.2	36.9	36.6	37.6	38.2	29.0	35.0	37.8	6.3
<i>Domain-Specific Models</i>										
BioMistral	28.5	28.0	30.3	24.4	29.5	25.4	33.3	28.0	28.2	<b>9.1</b>
UltraMedical-8B	42.7	42.3	41.4	45.2	42.3	41.7	42.2	34.0	42.6	3.5
HuatuoGPT-01-7B	43.3	42.7	45.8	38.4	38.2	47.9	45.2	34.0	42.9	2.0
AI Hospital	45.6	46.0	44.6	<u>45.8</u>	<u>43.6</u>	42.1	<b>47.4</b>	<b>47.0</b>	45.5	3.9
Ours	<b>49.7</b>	<b>47.6</b>	<b>49.0</b>	<b>49.0</b>	<b>46.5</b>	<b>48.5</b>	<b>44.8</b>	<b>38.0</b>	<b>48.9</b>	<u>8.6</u>

Table 6: Mean recommendation scores by disease category on MTMedDialog. The best results are highlighted in bold. The second best results are indicated with underlines.

results. (2) The predefined maximum number of interaction rounds is reached, compelling the doctor agent to provide a final diagnosis. This mechanism ensures a structured and finite diagnostic process.

## H Settings of RL training in DoctorAgent-RL

**Hardware & Software.** All training and evaluation were performed on a system featuring eight NVIDIA A100 80GB PCIe GPUs, an Intel Xeon Platinum 8369B 32-Core Processor, and 1.0 TB of RAM. For supervised fine-tuning, we utilized the LLaMA-Factory framework [41] to fine-tune with LoRA [42]. Our DoctorAgent-RL, is built upon the VERL framework (v0.2) [43] for reinforcement learning with language models, with RAGEN [44] providing the multi-turn RL architecture. We employed vLLM (v0.8.5) [45] for efficient LLM inference and evaluation, PyTorch (v2.4.0) with CUDA 12.4 for deep learning, and Ray for distributed training and serving. Flash Attention 2 [46] was integrated to optimize attention computation.

**Hyperparameter.** For completeness and reproducibility, all hyperparameters employed in DoctorAgent-RL are detailed in Table 7. We observed that the decoupled clip approach from DAPO [47] significantly enhances exploration during RL training, and thus, we adopted it for our final training process.



Figure 10: An example of dialogue flow among Doctor and Patient in DoctorAgent-RL.

Supervised Fine-Tuning	learning rate	1e-4
	batch size	64
	epochs	3
	lora_rank	8
DoctorAgent-RL	actor learning rate	1e-6
	state masking	true
	kl loss coef	0.001
	kl penalty	low_var_kl
	entropy coeff	0.001
	clip high	0.28
	clip low	0.2
	batch size	128
	epochs	1
	rollout group size	8
	rollout temperature	0.7

Table 7: Hyperparameters for Supervised Fine-Tuning and Reinforcement Learning experiments.