Dr. Assistant: Enhancing Clinical Diagnostic Inquiry via Structured Diagnostic Reasoning Data and Reinforcement Learning

Anonymous ACL submission

Abstract

Clinical Decision Support Systems (CDSSs) provide reasoning and inquiry guidance for physicians, yet they face notable challenges, including high maintenance costs and low generalization capability. Recently, Large Language Models (LLMs) have been widely adopted in healthcare due to their extensive knowledge reserves, retrieval, and communication capabilities. While LLMs show promise and excel at medical benchmarks, their diagnostic reasoning and inquiry skills are constrained. To mitigate this issue, we propose (1) Clinical Diagnostic Reasoning Data (CDRD) structure to capture abstract clinical reasoning logic, and a pipeline for its construction, and (2) the Dr. Assistant, a clinical diagnostic model equipped with clinical reasoning and inquiry skills. Its training involves a two-stage process: SFT, followed by RL with a tailored reward function. We also introduce a benchmark to evaluate both diagnostic reasoning and inquiry. Our experiments demonstrate that the Dr. Assistant outperforms open-source models and achieves competitive performance to closed-source models, providing an effective solution for clinical diagnostic inquiry guidance.

1 Introduction

Clinical Decision Support Systems (CDSSs) assist physicians by providing guidance for clinical inquiries, which is especially valuable for less experienced physicians (Bright et al., 2012; Sutton et al., 2020; Berner and La Lande, 2016). However, traditional CDSSs rely on structured knowledge bases and rule-based algorithms, which incur high development and maintenance costs and have limited adaptability to complex clinical situations (Sheikhalishahi et al., 2019; Deo, 2015).

Recently, Large Language Models (LLMs) have achieved remarkable success in knowledge-intensive tasks, spanning legal document analysis (Arfat et al., 2024; Janatian et al., 2023), financial forecasting (Wu et al., 2024; Xie et al., 2023) and education (Kasneci et al., 2023; Wang et al., 2024a), due to their superior comprehension and generalization abilities. This progress has spurred their adoption in healthcare (Dou et al., 2025; Kopka et al., 2025). Baichuan-M2 (Dou et al., 2025) and HuatuoGPT-o1 (Chen et al., 2024) focus on patients' consultation experience and decision-making. Zhongjing (Yang et al., 2024) prioritizes medical knowledge integration and records analysis. Current LLMs perform well in general healthcare consultation but are limited in clinical inquiry as shown in Figure 1 (b), which requires rigorous diagnostic reasoning.

Clinical guidelines contain diagnostic reasoning logic, yet they face two primary limitations. 1) **Scattered Reasoning Logic**: The diagnostic reasoning logic cue for a given symptom is often scattered across different chapters, as shown in Figure 2, and 2) **Difficulty in Training**: Even with high-quality data, training models to master clinical inquiry skills remains a significant challenge.

To address these challenges, we propose **Clini**cal **D**iagnostic **R**easoning **D**ata (CDRD) format to capture abstract diagnostic reasoning logic. Each CDRD is a triple-tuple, including a core symptom, diagnostic evidence and differential diagnoses. We also propose a *three-stage* pipeline for constructing CDRDs. The pipeline collaboratively employs LLM synthesis and physician refinement based on clinical guidelines. It first extracts symptoms (Stage I) and their corresponding diseases (Stage II) and finally completes CDRD (Stage III). To ensure reliability, outputs at every stage are reviewed and refined by professional physicians. Furthermore, using CDRD as a seed, we synthesize two types of data: Question-Answer pairs (QA data) for Supervised Fine-Tuning (SFT), and clinical inquiry dialogue for Reinforcement Learning (RL).

We also propose **Dr. Assistant**, a clinical diag-

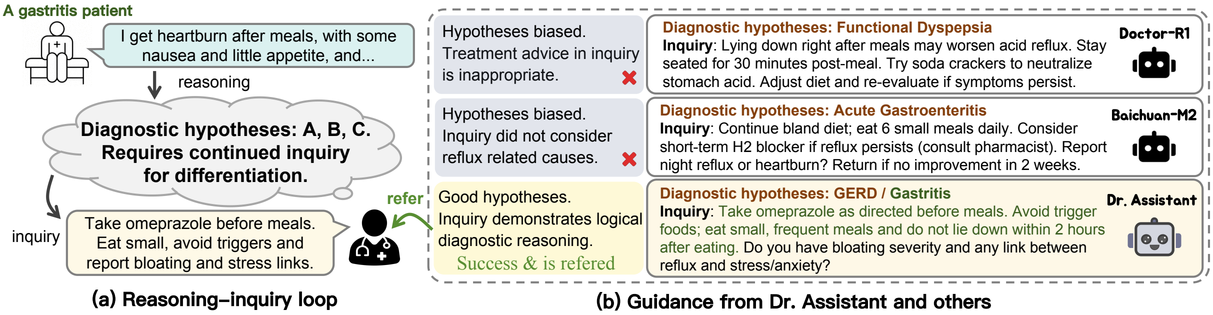


Figure 1: Dr. Assistant provides guidance (b) to physicians in the reasoning-inquiry loop (a), where diagnostic hypotheses both drive and are refined by ongoing inquiry. Since diagnostic hypotheses are central to reasoning, they are our primary focus. We detail (a) in Appendix B.1.

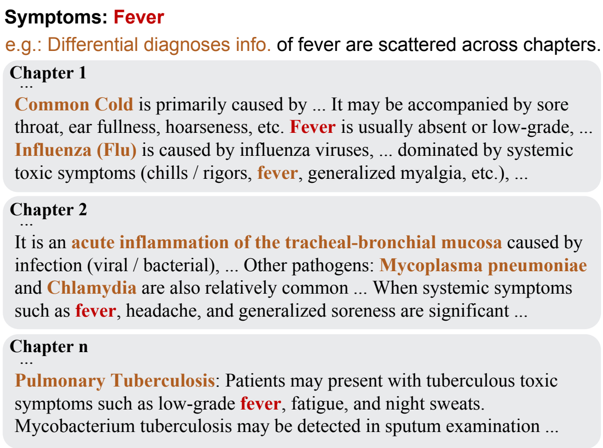


Figure 2: The scattered logic in clinical guideline.

nostic model equipped with inherent clinical reasoning & inquiry skills. Its training consists of two stages: 1) The model remembers preliminary clinical inquiry reasoning logic via SFT. 2) We design a reward function, which encompasses two core evaluation dimensions via RL: Clinical Reasoning and Inquiry Skills, and Fidelity to CDRD (logical differences from CDRD). The logical differences penalty term (Fidelity to CDRD) is incorporated to constrain illogical inference behaviors, refining the model's clinical reasoning competencies. Furthermore, we construct a benchmark to evaluate clinical diagnostic reasoning and inquiry skills. It comprises 242 real clinical cases covering 8 secondary departments, and 147 real clinical inquiry rounds covering 4 secondary departments.

The experimental results demonstrate that Dr. Assistant (14 B) outperforms the open-source model HuatuoGPT-o1-72B with a 13.59% improvement in ICD-Recall, and achieves comparable performance to GPT-5, which provides an effective solution for the practical implementation of clinical decision support systems.

In summary, our contributions are as follows:

- We propose Clinical Diagnostic Reasoning Data (CDRD) structure to capture abstract clinical reasoning logic, and a pipeline for constructing it.
- We propose the Dr. Assistant, a clinical diagnostic model equipped with clinical reasoning and inquiry skills. Its training involves a two-stage process: SFT, followed by RL with a tailored reward function to improve the model's clinical reasoning and inquiry skills.
- We construct a clinical diagnostic reasoning & inquiry benchmark consisting of 242 real cases across 8 secondary departments, along with 147 real inquiry turns across 4 secondary departments. The experiments demonstrate Dr. Assistant outperforms the open-source model HuatuoGPT-o1-72B, with a 13.59% improvement in ICD-Recall rate, and achieves comparable performance to GPT-5.

2 Related Work

LLMs for medical consultation. Previous approaches primarily rely on SFT to improve medical knowledge coverage (Ma et al., 2025). Recent studies incorporate RL to optimize multi-turn consultation strategies, such as Baichuan-M2 (Dou et al., 2025), HuatuoGPT-o1 (Chen et al., 2024) and Doctor-R1 (Lai et al., 2025b), which focus on improving patients' consultation experience. They are optimized for general conversation, lacking clinical diagnostic inquiry and reasoning skills. This leads to clinically unsupported guidance.

CDSSs. Clinical guidelines served as the foundation of traditional CDSSs (Berner and La Lande,

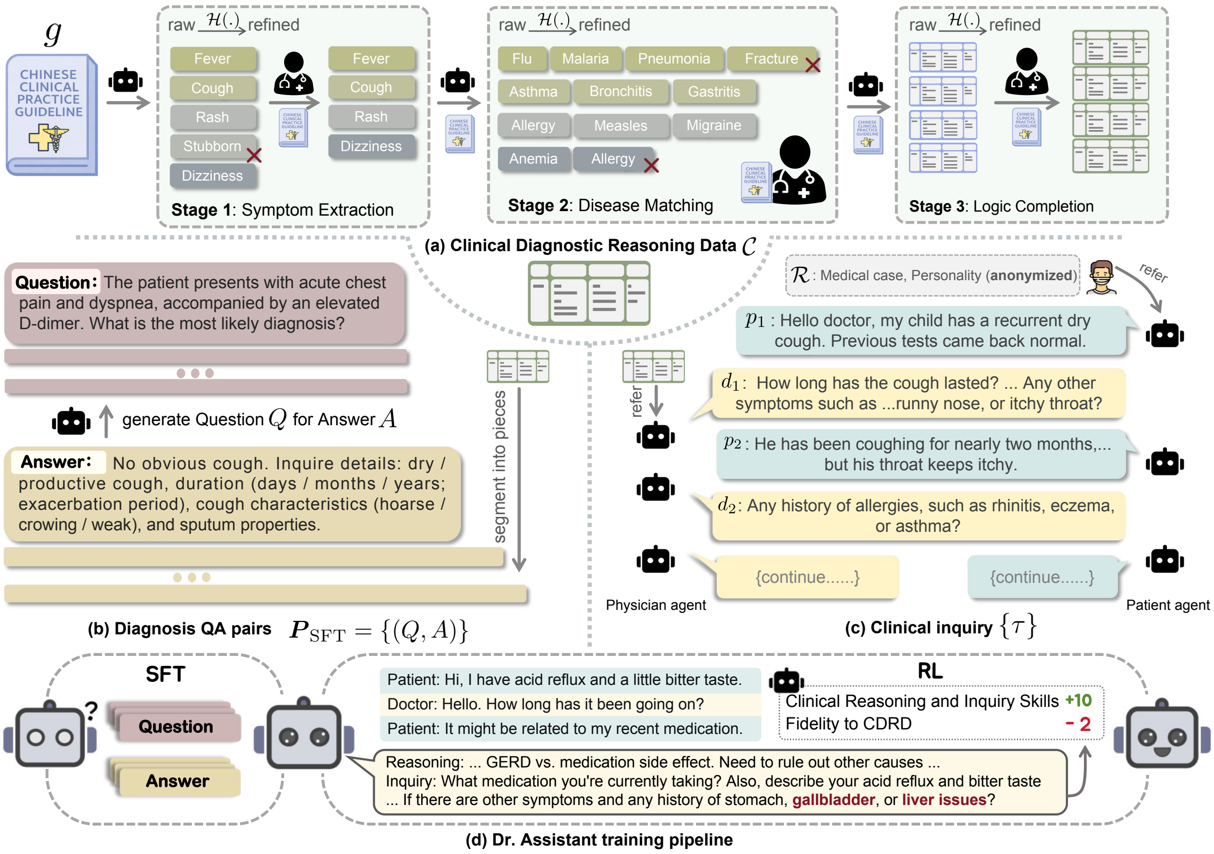


Figure 3: Our workflow involves CDRD construction (a), data synthesis (b, c), and Dr. Assistant training (d).

2016; Sutton et al., 2020). While rule-based systems provide strong interpretability, they are difficult to scale and adapt to flexible, multi-turn clinical inquiry. Recent efforts have explored integrating structured medical knowledge with neural models (Sheikhalishahi et al., 2019), yet not diagnostic inquiry logic training.

Evaluation of medical consultation. Evaluations for Medical LLMs evolved from static question-answer (Jin et al., 2021a; Pal et al., 2022; Jin et al., 2019) to general consultation, such as HealthBench (Arora et al., 2025), MAQuE (Gong et al., 2025), GAPS (Chen et al., 2025) and SPTesting (Dou et al., 2024). While existing benchmarks feature multi-turn consultation, they frequently fail to capture the realistic demands of clinical diagnostic inquiry in CDSS settings. In this work, we evaluate models' reasoning soundness and inquiry quality with real and dynamic clinical inquiry cases.

3 Data Construction

3.1 CDRD Format for Abstract Logic

Clinical guidelines contain abstract diagnostic reasoning logic and it is scattered across chapters, requiring semantic understanding and re-structure (Zhang et al.). To capture abstract diagnostic reasoning logic that proceeds from a symptom to its differentials in clinical inquiry, we structure it as Clinical Diagnostic Reasoning Data (CDRD). Formally, each CDRD is defined as a triple-tuple $\mathcal{C} = (\mathcal{S}, \mathcal{E}, \mathcal{D})$, representing the standard diagnostic logic, where:

- $\mathcal{S}$: A Core Symptom (e.g., headache) of patient.
- $\mathcal{E}$: Diagnostic Evidence. The clinical information, including associated symptoms, lab results, and exam findings, that physicians collect and analyze to infer diagnoses.
- $\mathcal{D}$: Differential Diagnoses, the list of potential diseases (e.g., gastroenteritis) should be considered based on the available evidence, along with their clinical manifestations and required diagnostic examinations.

Please refer to Appendix §G.1 for the CDRD demo.

3.2 CDRD Construction Pipeline

To maximize the completeness and purity of CDRD, we propose a pipeline for constructing it from clinical guidelines. As shown in Figure 3 (a), CDRD construction pipeline is a collaborative approach of LLM synthesis and physician refinement, encompassing three sequential stages: symptom extraction, disease matching and logic completion.

Symptom extraction. Given a clinical guideline g, we first extract the symptoms set $\boldsymbol{S}$. The LLM identifies candidate symptoms, which are then refined by physicians:

$$\boldsymbol{S} = \mathcal{H}\Big(\mathrm{LLM}(g) \mid g\Big), \tag{1}$$

where $\mathcal{H}(.)$ is defined as physicians' refinement, which ensures symptoms are standard (*e.g.*, mapping "chest hurt" to standardized "chest pain" or "angina").

Disease matching. Conditioned on the refined symptom set $\boldsymbol{S}$ and the clinical guideline g, LLM generates a list of differential diagnoses to every single symptom $\mathcal{S}$. The physician then refines this list against guidelines:

$$\boldsymbol{D}' = \bigcup_{\mathcal{S} \in \boldsymbol{S}} \mathcal{H}\Big(\mathrm{LLM}(g, \mathcal{S}) \mid g, \mathcal{S}\Big), \tag{2}$$

where $\boldsymbol{D}'$ is a set that includes disease name list for each $\mathcal{S}$. By decoupling disease matching from symptom extraction, we allow physicians to evaluate the validity of each symptom-disease correlation.

Logic completion. Based on symptom-diseases pairs (*i.e.*, a core symptom $\mathcal{S}$ paired with its differentials' name list $\mathcal{D}'$) and clinical guideline g, LLM extracts clinical evidence $\mathcal{E}$ set and completes differential diagnoses $\mathcal{D}$ set to complete CDRD set of g. For each symptom-disease pair:

$$\mathcal{E}, \mathcal{D} = \mathcal{H}\Big(\mathrm{LLM}(g, (\mathcal{S}, \mathcal{D}')) \mid g, (\mathcal{S}, \mathcal{D}')\Big). \tag{3}$$

Here, we get the full reasoning data CDRD:

$$\mathcal{C} = (\mathcal{S}, \mathcal{E}, \mathcal{D}). \tag{4}$$

See prompts for the three stages in Appendix §E.1.

While CDRD captures high-quality diagnostic reasoning logic, it suffers from limited scale and lacks dynamic, multi-turn clinical settings. To enable models not only to learn this logic but also to apply it flexibly, we construct Question-Answer data for SFT and multi-turn inquiry data for RL based on CDRD as shown in Figure 3 (b, c) separately.

3.3 QA Pair for SFT

We use CDRD as a seed to construct QA dataset $\boldsymbol{P}_{\text{SFT}}$ via a two-step process: (i) answer formulation , based on CDRD, and (ii) question synthesis, based on the answer.

Answer formulation. We first parse CDRDs into discrete entries $\boldsymbol{e}$, where e is a partial segment from either $\mathcal{E}$ or $\mathcal{D}$ in a CDRD. Each entry e is semantically complete. To increase expression diversity, we rewrite each entry e into a fluent response A using LLM:

$$A = \mathrm{LLM}\big(e\big). \tag{5}$$

Question synthesis. For each answer A, we synthesize a corresponding question Q to form a complete QA pair:

$$Q = \mathrm{LLM}\big(A\big). \tag{6}$$

These operations are applied to the entire CDRD, yielding the final SFT dataset $\boldsymbol{P}_{\text{SFT}} = \{(Q, A)\}$. Refer to Appendix §E.3 for its synthesis prompts.

3.4 Multi-turn Inquiry for RL

While QA data instill diagnostic reasoning logic, clinical inquiry skills are not covered. So we propose a dual-agent comprising a physician agent π_{phy} and a patient agent π_{pat}, which takes iterative interactions to synthesize reasoning and inquiries based on patient profiles $\boldsymbol{R}$ and CDRDs $\boldsymbol{C}$.

Specifically, an inquiry dialogue starts with the main complaints specified in R, and π_{phy} generates reasoning and inquiry referring to $\mathcal{C}$. We structure the reasoning and inquiry to a template. It consists of six reasoning steps and the final inquiry (See Appendix §G.2 and §E.2, for reasoning-inquiry demos and their synthesis prompts, separately):

Known Information: The patient's information summary from history dialogue.

User's Intention: Intentions underlying the query.

Provided Information: Expressed inquiry summary.

Diagnoses: Based on the inquiry history, the model generates initial diagnostic hypotheses.

Information to be Collected: Clinical information required to clarify diagnoses and guide decision-making.

Response Strategy: Before inquiry, model reasons the most necessary information to acquire.

Inquiry: Diagnoses or follow-up questions that model presents.

Actually, an inquiry dialogue can be seen as a particular trajectory τ sampled from the CDRD $\mathcal{C}$, influenced by patient's constraints in R:

$$\tau = (p_1, d_1, p_2, d_2, \ldots, p_k, d_k \mid \mathcal{C}, R), \quad (7)$$

where p_k represents the π_{pat}'s query, and d_k is the π_{phy}'s reasoning and inquiry at round k.

4 Dr. Assistant

Dr. Assistant's training pipeline consists of two stages, as shown in Figure 3 (d): Supervised Fine-Tuning (SFT), which equips the model with preliminary clinical diagnostic reasoning logic, and Reinforcement Learning (RL), which further enhances the model's clinical diagnostic inquiry skills.

SFT with $\boldsymbol{P}_{\text{SFT}}$. Our model π_θ is first fine-tuned via supervised learning on $\boldsymbol{P}_{\text{SFT}}$ to acquire basic diagnostic reasoning logic, optimized with the standard negative log-likelihood loss:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(q,a)\sim \boldsymbol{P}_{\text{SFT}}}\left[\sum_{t=1}^{T} \log \pi_\theta\big(a_t \mid q, a_{<t}\big)\right]. \quad (8)$$

RL with multi-turn inquiry. Although model learned clinical diagnostic reasoning logic in SFT, its responses are confined to single, static QA pattern, and have limited clinical reasoning and inquiry skills. So we conduct Reinforcement Learning on inquiry dataset $\{\tau\}$, enhancing the model with clinical diagnostic reasoning and inquiry.

Specially, we design a composite reward function. Given a ground-truth reasoning and inquiry d_t at round t of an inquiry dialogue τ and the trained model's response $\hat{d}_t$, the step-wise reward is defined as:

$$R_{\text{step}}(\hat{d}_t, d_t, \mathcal{C}) = R_{\text{comp}}(\hat{d}_t, d_t) - R_{\text{div}}(\hat{d}_t, \mathcal{C}). \quad (9)$$

Here, R_{comp} measures: *clinical reasoning* $R_{\text{comp-r}}$*, and inquiry skills* $R_{\text{comp-i}}$, while R_{div} constraints *fidelity to CDRD*.

a. Clinical reasoning and inquiry skills. R_{comp} evaluates the model's performance by scoring the similarity r_i between its output and the groundtruth across seven parts related to diagnostic reasoning and inquiry: 1) *Reasoning list* $\boldsymbol{r}_{\text{reason}}$ including Known Information r_0, User Intention r_1, Provided Information r_2, Diagnoses r_3, Information to be Collected r_4 and Response Strategy r_5, 2) *Inquiry list* $\boldsymbol{r}_{\text{inquiry}}$ including the final inquiry r_6. Definitions of the seven parts are in §3.4.

r_i is scored by another LLM-judge, ranging from 0 to 10. A higher alignment with the ground-truth yields a higher score. R_{comp} is derived as:

$$\begin{aligned} R_{\text{comp}}(d_t, \tau) &= R_{\text{comp-r}} + R_{\text{comp-i}} \\ &= \boldsymbol{w}_{\text{reason}}^\top \boldsymbol{r}_{\text{reason}} + \boldsymbol{w}_{\text{inquiry}}^\top \boldsymbol{r}_{\text{inquiry}}, \end{aligned} \quad (10)$$

where $\boldsymbol{r}_{\text{reason}} = [r_0, \ldots, r_5]^\top$ is the score vector for diagnostic reasoning items, $\boldsymbol{w}_{\text{reason}}$ and $\boldsymbol{w}_{\text{inquiry}}$ are weight coefficients, set according to the relative importance of each dimension in clinical decision-making. See Appendix §A for parameter settings.

b. Fidelity to CDRD. R_{div} penalizes deviations from the source CDRD (*i.e.*, the specific CDRD used to generate the ground-truth trajectory τ that serves as the target in this optimization step):

$$R_{\text{div}}(d_t, \mathcal{C}) = \lambda n, \quad (11)$$

where λ is the weight coefficient and n is the number of already collected information (or yet to be collected) and diagnoses that are **not** in CDRD, determined by LLM-judge.

We employ DAPO (Yu et al., 2025), which samples output group $\{o_i\}_{i=1}^G$ for each prompt and optimizes the model π_θ by the following objective:

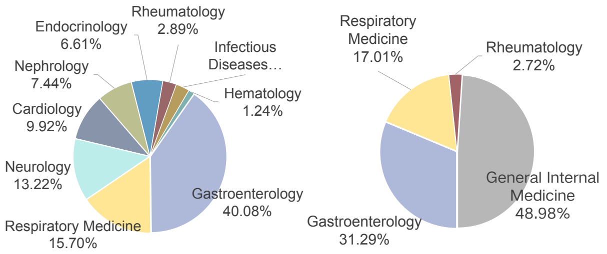
$$\begin{aligned} \mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}\Bigg[&\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t}, \\ &\text{clip}(r_{i,t}(\theta), 1-\epsilon_{low}, 1+\epsilon_{high})\hat{A}_{i,t}\Big)\Bigg], \end{aligned} \quad (12)$$

where $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q, o_i, <t)}{\pi_{\text{old}}(o_{i,t}|q, o_i, <t))}$ is the probability ratio and o_i is the length of trajectory τ_i. $\hat{A}_{i,t} = \frac{R_i - mean(\{R_i\}_{i=1}^G)}{std(\{R_i\}_{i=1}^G)}$, is advantage of the i-th response is calculated by normalizing the group-level rewards $\{R_i\}_{i=1}^G$.

5 Experiments

5.1 Experimental Settings

Model and baselines. Our Dr. Assistant is trained on Qwen3-14B (Team, 2025). We compare it against: 1) *Open-source Models*: including general-purpose LLMs: Qwen3-32B (Team, 2025) , DeepSeek-R1-Distill-14B (DeepSeek-AI, 2025a), Baichuan-M2-32B (Dou et al., 2025), Llama3-OpenBioLLM-70B (Ankit Pal, 2024) and medical-specialized LLMs (HuatuoGPT-o1-72B (Chen et al., 2024) , Doctor-R1-8B (Lai et al.,



(a) Patient profiles distribution in bench (b) Clinical inquiry distribution in bench

Figure 4: Data distribution of our benchmark. One includes 242 patient profiles across 7 sub-departments (a). The other includes 147 rounds of clinical inquiry dialogues (b). "General Internal Medicine" means the inquiries cover multiple sub-departments.

2025c), Med42-v2-8B (Christophe et al., 2024). 2) *Closed-source Models*: This includes leading proprietary models: GPT-5 (OpenAI, 2025) , Gemini 2.5 Pro (DeepMind, 2025b), Gemini 2.5 Flash (DeepMind, 2025a) and Grok 4 Fast (xAI, 2025). Details of parameter settings are in Appendix §A.

Training datasets. We construct 11 CDRD covering ∼60 diseases based on internal guidelines for training. For each CDRD, 400 QA pairs are synthesized, totaling 4400 data for SFT. Additionally, we construct 36688 turns of multi-turn inquiry dialogues for RL. All training data is in Chinese.

Evaluation. Existing benchmarks, such as HealthBench (Arora et al., 2025) and MAQuE (Xie et al., 2023), do not concentrate on dynamic clinical diagnostic inquiry reasoning. HealthBench contains general medical consultations, most of which are not multi-turn dialogues. MAQuE does not consider the mutual influence between the diagnostic inquiry process and the final diagnoses. However, inspired by these evaluations (Dou et al., 2024; Arora et al., 2025; Xie et al., 2023), we construct a benchmark for clinical diagnostic inquiry. It considered the dynamic interaction and influence between the inquiry process and final diagnoses, which encompasses 242 patient profiles (cases) across nine sub-departments, including neurology, and 147 rounds of clinical inquiry dialogue records, as shown in Figure 4. See demos in Appendix §G.4 and §G.5.

We evaluate the model via both **ICD-10 (Hirsch et al., 2016) matching** for diagnostic reasoning and **physician satisfaction for inquiry.**

5.2 ICD Matching for Diagnostic Reasoning

Since diagnostic hypotheses in reasoning directly determine inquiry direction, as shown in Figure 1 (a), we design an evaluation where one model, acting as a patient based on a given patient profile, interacts with the model-under-test (physician model). In each turn, the physician model outputs diagnostic hypotheses (the core reasoning) and inquiry. We compare diagnostic hypotheses at round five, with GT in the patient profile to assess the model's diagnostic reasoning ability and inquiry direction. For further details, see Appendix §B.2. Patient profiles in bench are different from those in §3.4.

ICD-10 matching. We map the diagnostic hypotheses of both model and GT to ICD-10 (International Classification of Disease) (Hirsch et al., 2016) codes, $\mathcal{P}$ and $\mathcal{G}$ respectively by LLM. We then calculate their similarity as formalized in Appendix Algorithm 1. In clinical practice, the **ICD-Recall** of diagnostic hypotheses is positively correlated with the *reasoning* and its *whole inquiry strategy*, which is our **main target**. But we still include precision for reference.

ICD-Recall (R_{ICD}). This is our **primary metric**, measuring the coverage of ground truth diagnoses.

$$R_{\text{ICD}} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \max_{p \in \mathcal{P}} \text{Sim}_{\text{ICD}}(p, g). \qquad (13)$$

ICD-Precision (P_{ICD}). It reflects diseases requiring exclusion via further inquiry in diagnostic hypotheses, while not in GT:

$$P_{\text{ICD}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \max_{g \in \mathcal{G}} \text{Sim}_{\text{ICD}}(p, g). \qquad (14)$$

As shown in Table 1: (1) Overall, Dr. Assistant achieves advance in ICD-Recall, improving by 6.70% relatively compared to the best closed-source model, Gemini-3-Pro-Preview (0.5066 v.s. 0.4748). This demonstrates its effectiveness. (2) Compared to open-source general-purpose models, Dr. Assistant outperforms Qwen3-32B by 16.30 % relatively (0.5066 v.s. 0.4356). This is because Dr. Assistant learned diagnostic reasoning logic and inquiry skills to infer patient's disease from a core symptom. (3) Compared to open-source medical models, Dr. Assistant achieves state-of-the-art in ICD-Recall, even surpassing HuatuoGPT-o1-72B by 13.59% relatively (0.5066 v.s. 0.4460). It is because previous medical models focus on domain knowledge but lack clinical diagnostic reasoning and inquiry skills. In the RL stage, Dr. Assistant internalized reasoning and inquiry skills, thereby driving the inquiry dialogue to an optimal trajectory.

Model	Recall† ↑	Prec. ↑
Open-source (Medical)		
HuatuoGPT-o1-72B	0.4460	0.5211
Baichuan-M2-32B	0.3687	0.4517
Doctor-R1-8B	0.3873	0.4800
Llama3-OpenBioLLM-70B	0.2216	0.2755
Med42-v2-8B	0.2787	0.3081
Open-source (Non-medical)		
Qwen3-14B	0.4298	0.4600
Qwen3-32B	0.4356	0.4660
DeepSeek-R1-Distill-14B	0.3913	0.4267
Close-source		
GPT-5	0.4651	0.4842
Gemini-2.5-Flash	0.4288	0.4497
Gemini-3-Pro-Preview	0.4748	**0.5333**
Grok 4 Fast	0.4284	0.5159
Dr. Assistant	**0.5066**	0.4717

Table 1: Average ICD-Recall & Precision, sorted by **ICD-Recall**†. ICD-Recall (denoted by †) is the primary criterion for assessing a model's diagnostic reasoning correction. Model performance across secondary departments is detailed in Appendix Table 5.

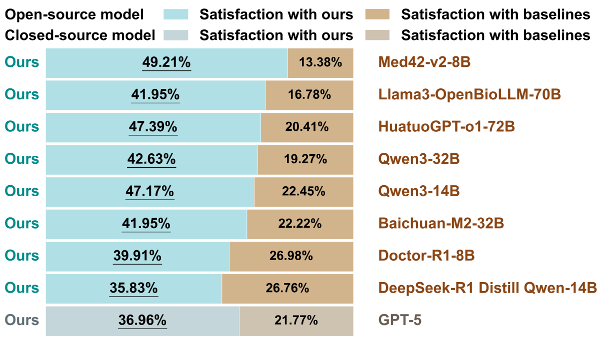


Figure 5: Physician's satisfaction. For a clearer display of comparison, tie cases are excluded from this figure.

5.3 Physician Satisfaction

To evaluate model inquiries in real CDSS settings, we had the model produce inquiries based on real clinical inquiry records (147 rounds), submitted them to physicians for evaluation. We invite 11 physicians, each with about 6 years of clinical experience. The evaluation criteria consist of two parts: *relevance* and *expertise*, where relevance serves as the prerequisite for winner. Physicians must determine a winner of inquiry-couple (ours, baseline), or declare a tie. Each evaluation is independently reviewed by three physicians to ensure reliability. For further details, see Appendix §B.3.

As shown in Figure 5: (1) Dr. Assistant gets higher physician satisfaction compared to open-source models. It shows improvements of 35.83% and 26.78% over the medical model Med42-v2-8B and HuatuoGPT-o1-72B, respectively. Dr. Assistant maintains focus on the diagnostic target. It stems from its clear diagnostic reasoning logic, which enables it to deliver more valuable inquiries throughout the clinical process. See case study in Appendix §C. (2) Dr. Assistant advances GPT-5 by 15.19% (36.96 v.s. 21.77). This highlights that complex clinical inquiry requires precise diagnostic reasoning. Dr. Assistant integrates diagnostic reasoning logic during RL, achieving superior physician satisfaction.

Benchmark	Qwen3-14B	Dr. Assistant	Retention
CMMLU (Medical, Chinese)			
Anatomy	0.8649	0.8784	**101.56%**
Professional Medicine	0.7952	0.8032	**101.01%**
Nutrition	0.8207	0.8276	**100.84%**
Clinical Knowledge	0.7890	0.7932	**100.53%**
TCM	0.8324	0.8324	**100.00%**
College Medicine	0.8828	0.8791	99.58%
Genetics	0.7727	0.7670	99.26%
Virology	0.8639	0.8580	99.32%
MMLU-Pro (English)			
Zero-shot	0.6710	0.6799	**101.33%**
Five-shot	0.6724	0.6686	99.43%
MedQA (English)			
Zero-shot	0.7021	0.6897	98.23%
Five-shot	0.6960	0.6952	99.89%
Average (Overall)	0.7803	0.7810	**100.10%**

Table 2: Performance on general benchmarks. The **Retention** reflects model's ability to maintain original knowledge after our post-training.

5.4 General Medical Knowledge Retention

To assess the general knowledge retention of Dr. Assistant, we conduct a comparative analysis of its performance against Qwen3-14B on the general knowledge benchmarks CMMLU(0/5-shot) (Li et al., 2024), MMLU-Pro (0/5-shot) (Wang et al., 2024b), and MedQA (Jin et al., 2021b). For CMMLU, we select eight medical-related subsets for evaluation. Results are in Table 2.

Overall, Dr. Assistant maintains close performance to Qwen3-14B (0.7810 v.s. 0.7803), demonstrating no significant knowledge forgetting. Notably, on the eight Chinese medical subsets of CMMLU (Li et al., 2024), Dr. Assistant achieves equal or superior performance to Qwen3-14B in five of them, further validating the effectiveness of our training method.

Method	Satisfaction	Recall[†] ↑	Prec. ↑
Dr. Assistant	**-0.00%**	**0.5066**	0.4717
w/o SFT	-18.82%	0.5024	0.4475
w/o RL	-6.80%	0.3927	0.5144
w/o R_{div}	-8.62%	0.4856	**0.5192**
w/o $R_{\text{comp-r}}$	-5.50%	0.4893	0.4830

Table 3: Ablation study on **Dr. Assistant**. Satisfaction is calculated as the difference in winning percentage compared to Dr. Assistant.

5.5 Ablation Study

We evaluates the contributions of SFT, RL and reward function to Dr. Assistant on our benchmark, with results summarized in Table 3.

The removal of SFT and RL leads to performance drops 0.84% and 29.00% relative to the full on ICD-Recall separately, and drops 18.82% and 6.80% on satisfaction separately. SFT improves model's clinical diagnostic logic, while RL further reinforces reasoning and inquiry skills. Notably, performance degrades less when SFT is omitted than when RL is. This suggests SFT tends to solidify the model's response patterns around the tuning data, which may constrain its adaptability to prior tasks. And RL not only guides the model beyond these fixed patterns but also further refines and expands its learned capabilities. (Lai et al., 2025a).

Furthermore, we individually removed the R_{div} and $R_{\text{comp-r}}$ reward. This leads to performance declines of 4.32% and 3.54% on ICD-Recall, respectively, and 8.62% and 5.50% on satisfaction, respectively, demonstrating their effectiveness. The R_{div} component serves to penalize the generation of incorrect diagnostic information, improving output quality. The $R_{\text{comp-r}}$ function rewards the model's clinical reasoning, which is similar to GT.

5.6 Easy and Hard Diagnostic Inquiry

We stratified patient profiles by diagnosed disease (1∼6) into single-disease (1, easy) and multi-disease (2∼6, hard) cases to evaluate model's reasoning and inquiry skills across diagnostic complexity levels. As shown in Table 4, compared to Baichuan-M2-32B and GPT-5, Dr. Assistant achieves disease-recall improvements of 31.09% and 2.80% (0.5426 v.s. 0.4139 and 0.5426 v.s. 0.5278) on single-disease tasks, and 37.39% and 8.92% (0.5067 v.s. 0.3688 and 0.5067 v.s. 0.4652) on multi-disease tasks respectively. Dr. Assistant not only achieves competitive performance on single-disease tasks but also demonstrates superior diagnostic performance in hard (multi-disease) settings. It is because Dr. Assistant stays on the diagnostic target. It stems from its clear diagnostic reasoning logic, which enables it to deliver more valuable inquiries throughout the clinical process. We also analyze the model's performance across different departments, as shown in Appendix §D.

Model	Diag.=1	Diag.>1
Open-source (medical)		
HuatuoGPT-o1-72B	0.5248	0.4461
Baichuan-M2-32B	0.4139	0.3688
Doctor-R1-8B	0.4733	0.3890
Med42-v2-8B	0.2938	0.2788
Llama3-OpenBioLLM-70B	0.2240	0.1867
Open-source (general)		
Qwen3-14B	0.5030	0.4316
Qwen3-32B	0.5307	0.4356
DeepSeek-R1-Distill-14B	0.4198	0.3913
Close-source		
GPT-5	0.5287	0.4652
Gemini 2.5 Flash	0.5192	0.4289
Grok 4 Fast	0.4821	0.4092
Gemini-3-Pro-Preview	**0.5455**	0.4749
Dr. Assistant	0.5426	**0.5067**

Table 4: Model performance (ICD-Recall) stratified by diagnostic complexity: single-disease cases (Diag.=1, easy) vs. multi-disease cases (Diag.>1, hard). We provide detailed Diag.=1∼6 in Appendix Table 6.

6 Conclusion

To address the limitations of traditional CDSSs and enhance the diagnostic reasoning capability of LLMs in clinical diagnostic inquiry, we propose (1) a structured Clinical Diagnostic Reasoning Data (CDRD) format along with a construction pipeline to capture abstract clinical reasoning logic, and (2) the Dr. Assistant model, which integrates clinical reasoning and inquiry skills through a two-stage training strategy of Supervised Fine-Tuning followed by Reinforcement Learning with a tailored reward function. Evaluated on a dedicated diagnostic reasoning and inquiry benchmark, Dr. Assistant outperforms open-source models and achieves competitive performance compared to closed-source models, offering a practical solution for clinical diagnostic inquiry guidance.

7 Limitation

Our study has several limitations. First, the construction of Clinical Diagnostic Reasoning Data (CDRD) relies on physician refinement, which introduces manual effort and may limit scalability. While we employed a three-stage pipeline to ensure quality, this process remains resource-intensive and could be subject to individual clinical judgment biases. Second, our evaluation benchmark, though constructed from real clinical cases and dialogues, is moderate in scale and covers a limited number of secondary departments. This may affect the generalizability of our findings to broader or more specialized clinical settings. Third, the reward function used in reinforcement learning, while designed with clinical dimensions in mind, is inherently heuristic and may not fully capture all nuances of real diagnostic reasoning. Future work could explore more automated, data-driven reward modeling or incorporate direct clinical outcome feedback. Lastly, our experiments primarily focus on diagnostic inquiry within a controlled dialogue framework, the model's performance in fully open-ended clinical conversations or integration with live electronic health record systems remains to be tested.

References

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

Yasir Arfat, Marco Colella, and Enrico Marello. 2024. Legal text analysis using large language models. In *International Conference on Applications of Natural Language to Information Systems*, pages 258–268. Springer.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Eta S Berner and Tonya J La Lande. 2016. Overview of clinical decision support systems. In *Clinical decision support systems: Theory and practice*, pages 1–17. Springer.

Tiffani J Bright, Anthony Wong, Ravi Dhurjati, Erin Bristow, Lori Bastian, Remy R Coeytaux, Gregory Samsa, Vic Hasselblad, John W Williams, Michael D Musty, and 1 others. 2012. Effect of clinical decision-support systems: a systematic review. *Annals of internal medicine*, 157(1):29–43.

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

Xiuyuan Chen, Tao Sun, Dexin Su, Ailing Yu, Junwei Liu, Zhe Chen, Gangzeng Jin, Xin Wang, Jingnan Liu, Hansong Xiao, and 1 others. 2025. Gaps: A clinically grounded, automated benchmark for evaluating ai clinicians. *arXiv preprint arXiv:2510.13734*.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms.

Google DeepMind. 2025a. Gemini 2.5 flash.

Google DeepMind. 2025b. Gemini 2.5 pro.

DeepSeek-AI. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI. 2025b. Deepseek-v3.2: Pushing the frontier of open large language models.

Rahul C Deo. 2015. Machine learning in medicine. *Circulation*, 132(20):1920–1930.

Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, and 1 others. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.

Chengfeng Dou, Ying Zhang, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, and Zhengwei Tao. 2024. Integrating physician diagnostic logic into large language models: Preference learning from process feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2453–2473.

Linlu Gong, Ante Wang, Yunghwei Lai, Weizhi Ma, and Yang Liu. 2025. The dialogue that heals: A comprehensive evaluation of doctor agents' inquiry capability. *arXiv preprint arXiv:2509.24958*.

JA Hirsch, G Nicola, G McGinty, RW Liu, RM Barr, MD Chittle, and L Manchikanti. 2016. Icd-10: history and context. *American Journal of Neuroradiology*, 37(4):596–599.

Samyar Janatian, Hannes Westermann, Jinzhe Tan, Jaromir Savelka, and Karim Benyekhlef. 2023. From text to structure: Using large language models to support the development of legal expert systems. In *Legal Knowledge and Information Systems*, pages 167–176. IOS Press.

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

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Marvin Kopka, Niklas von Kalckreuth, and Markus A Feufel. 2025. Accuracy of online symptom assessment applications, large language models, and laypeople for self–triage decisions. *npj Digital Medicine*, 8(1):178.

Song Lai, Haohan Zhao, Rong Feng, Changyi Ma, Wenzhuo Liu, Hongbo Zhao, Xi Lin, Dong Yi, Min Xie, Qingfu Zhang, Hongbin Liu, Gaofeng Meng, and Fei Zhu. 2025a. Reinforcement fine-tuning naturally mitigates forgetting in continual post-training. *CoRR*, abs/2507.05386.

Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. 2025b. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. *arXiv preprint arXiv:2510.04284*.

Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. 2025c. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. *Preprint*, arXiv:2510.04284.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285.

Suxue Ma, Zhicheng Yang, Ruei-Sung Lin, Youbao Tang, Ning Zhang, Zhenjie Cao, Yuan Ni, Jing Xiao, Jieke Hou, and Peng Chang. 2025. Prefix-enhanced large language models with reused training data in multi-turn medical dialogue. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 26–33.

OpenAI. 2025. Gpt-5 system card.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. 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.

Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. 2019. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR medical informatics*, 7(2):e12239.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024a. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2024. Bloomberggpt: A large language model for finance, 2023. *URL https://arxiv. org/abs/2303.17564*.

xAI. 2025. Grok 4 fast.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. *Advances in Neural Information Processing Systems*, 36:33469–33484.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19368–19376.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

H Zhang, J Chen, F Jiang, F Yu, Z Chen, J Li, G Chen, X Wu, Z Zhang, Q Xiao, and 1 others. Huatuogpt, towards taming language model to be a doctor. arxiv (2023). *arXiv preprint arXiv:2305.15075*.

A Parameter Setting

Data construction. We use DeepSeek-V3.2 (DeepSeek-AI, 2025b) for constructing CDRD data. For constructing QA pairs and multi-turn inquiry dialogue data, we employ Qwen3-235B-A22B (Team, 2025). The model selection is determined through manual validation on a small set of test data.
SFT. The training epoch is set to 4 and a cosine annealing strategy is adopted, with a maximum learning rate of 5e-6. The warm-up ratio is set to 0.03 and the regularization coefficient is set to 0.01. SFT training process is conducted on 32 A800 GPUs.
RL. Our RL training is based on the VeRL framework (Sheng et al., 2024) with Qwen3-32B (Team, 2025) as the LLM-judge in our reward function. The learning rate is set to (1×10^{-6}), clipping parameters are $(\epsilon_{low} = 0.2)$ and $(\epsilon_{high} = 0.28)$, the rollout number is 4, and the training batch size is 512. The entire reinforcement learning process was run for 185 steps. R_{comp}'s (eq. (10)) weights are set to $\boldsymbol{w}_{\text{reason}} = [0.1, 0.3]$ (0.1 is for r_0 to r_2, and 0.3 for others) and $\boldsymbol{w}_{\text{inquiry}} = [0.6]$. R_{div}'s (eq. (11)) weight is set to $\lambda = 3$. The RL training process is conducted on 8 H800 GPUs, with an additional 16 H800 GPUs used for the LLM-judge.
Evaluation. To evaluate the ICD-Recall, ICD-Precision and Physician Satisfaction metrics, we set greedy decoding (with temperature = 0) for all models to ensure deterministic and comparable generation outputs. The prompts in §F.1 and §F.2 are used to generate response for metric computation.

B Evaluation

Dr. Assistant is for clinical diagnostic reasoning and inquiry guidance (internal-medicine in Chinese in experiment). However, existing benchmarks are not fully aligned with it. HealthBench (Arora et al., 2025) covers general healthcare consultation instead of multi-turn diagnostic inquiry. MAQuE (Xie et al., 2023), while involving multi-turn diagnostic inquiry, separates inquiry and diagnostic reasoning. Therefore, it does not align the real-world interdependence between them.

B.1 Reasoning-inquiry Loop

In practice, the reasoning lead to different trajectory of inquiry. Physician issues inquiry based on their diagnostic hypotheses (the reasoning core). Upon receiving an inquiry, patient responds with new queries. The physician then updates the diagnostic hypothesis based on this new query and continues the dialogue. We show this reasoning-inquiry loop in Figure 1 (a), which ultimately result in different diagnostic conclusion for the same patient.

Although the reasoning before an inquiry consists of multiple steps as mentioned in §3.4, their ultimate goal is to verify the diagnostic hypothesis. Including the correct diagnoses within this hypothesis is a prerequisite for proper inquiry.

B.2 ICD-10 Matching of Clinical Diagnostic Reasoning

Physicians form initial diagnostic hypothesis based on available information and iteratively refine it through targeted inquiry as new information is gathered, ultimately arriving at diagnostic results. The evolving diagnostic hypothesis drives the reasoning of inquiry. Therefore, we compare the model's diagnostic hypothesis with the diagnosis GT. To simulate dynamic clinical inquiry, we use an LLM as a patient simulator, conditioned on real patient profiles (includes basic medical history, communication style and diagnosis, **anonymized**). Each inquiry starts from patient simulator's main complaint. At each turn, the evaluated model outputs two parts: a *diagnostic hypothesis* and an *inquiry* to the patient. After five rounds, we compare the model's last diagnostic hypotheses with the diagnoses in the patient profile. Average ICD-Recall and ICD-Precision of each baseline are reported in Table 1.
ICD-10 matching. ICD-10 (International Classification of Disease) (Hirsch et al., 2016) codes' leading letter indicates major chapter and subsequent digits specify granular categories. We map the diagnosis results by LLM from both the model and the ground truth to ICD-10 codes, $\mathcal{P}$ and $\mathcal{G}$ respectively. We then calculate the similarity between them using $\text{Sim}_{\text{ICD}}(p, g)$, as formalized in **Algorithm 1**. Based on this similarity, we compute two metrics:
ICD-Recall (R_{ICD}). This is our **primary metric**, measuring the coverage of ground truth diagnoses.

$$R_{\text{ICD}} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \max_{p \in \mathcal{P}} \text{Sim}_{\text{ICD}}(p, g) \quad (15)$$

Reason for ICD-Recall priority: In clinical diagnostic reasoning, missing right diagnoses means the model loses critical diagnostic clues, leading to skewed inquiry and results. More importantly,

Algorithm 1: Hierarchical ICD-10 Similarity Scoring

Input: Predicted Code p, Ground Truth Code g
Output: Similarity Score $S \in [0.0, 1.0]$
Function $\mathrm{Sim}_{ICD}(p, g)$**:**
  // Level 1: Exact Match
  if $p = g$ **then**
    return 1.0
  // Level 2: Sub-category (First 4 chars, e.g., J45.9)
  else if $\mathrm{Prefix}(p, 4) = \mathrm{Prefix}(g, 4)$ **then**
    return 0.8
  // Level 3: Category (First 3 chars, e.g., J45)
  else if $\mathrm{Prefix}(p, 3) = \mathrm{Prefix}(g, 3)$ **then**
    return 0.6
  // Level 4: Block Range (e.g., J40–J47)
  else if $\mathrm{Block}(p) = \mathrm{Block}(g)$ **then**
    return 0.4
  // Level 5: Chapter (First Letter, e.g., J)
  else if $\mathrm{Prefix}(p, 1) = \mathrm{Prefix}(g, 1)$ **then**
    return 0.2
  else
    return 0.0

it poses a severe safety risk as the true condition remains unaddressed. Conversely, a false alarm can be ruled out by a physician.

ICD-Precision ($P_{\mathbf{ICD}}$). It reflects the mount of disease candidates (requiring exclusion via further inquiry) in differential diagnoses set.

$$P_{\mathrm{ICD}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \max_{g \in \mathcal{G}} \mathrm{Sim}_{\mathrm{ICD}}(p, g) \tag{16}$$

B.3 Physician Satisfaction: Inquiry Quality

We also involve physicians to assess response quality following CDSS's application perspective.

For evaluation, real clinical dialogue records are formatted into historical context and the latest patient message. The model is tasked with generating the next response. Finally, we pair Dr Assistant's response with a single baseline model's output for physician comparison. For each pair, the physician selects the better response or indicates a tie (select 0 or 2 per pair). The evaluation is based on two criteria: **Relevance.** Response should appropriately follow up on the patient's statements, and not deviate from the core inquiry topic. **Expertise & standardization.** The response should be clinically accurate in diagnosis, adhere to medical expertise, display a logical inquiry progression, and provide scientifically sound and appropriate recommendations. In practice, physicians first assess relevance. Responses that lack sufficient relevance are directly rejected, regardless of other qualities.

C Case Study

We show a clinical inquiry dialogue record in our bench and the following inquiry guidance generated by Dr. Assistant and HuatuoGPT-o1-72B. The results are shown in Figure 6.

This inquiry case reveals a notable divergence in clinical focus: HuatuoGPT-o1-72B did not adequately identify the potential cardiovascular risk underlying the patient' s symptoms and directed the inquiry toward orthopedics. But Dr. Assistant consistently prioritized cardiovascular exam by anchoring on key symptomatic clues throughout inquiry process. The case demonstrates how Dr. Assistant sustains correct diagnostic reasoning and inquiry skills when faced with complicated symptom patterns.

D Performance on Different Secondary Department

We stratified the evaluation dataset by secondary departments to assess the performance of different models under real-world long-tail scenarios. Results are presented in Table 5.

Our benchmark reflects the long-tail distribution of real clinical departments. Consequently, models underperform in tail categories like Rheumatology compared to head categories like Gastroenterology.

Compared to HuatuoGPT-o1-72B, Dr. Assistant achieves a performance improvement of 22.71% (0.3857 v.s. 0.3143) in Rheumatology and 17.17% (0.5424 v.s. 0.4629) in Gastroenterology. These results further validate the effectiveness of Dr. Assistant in integrating diagnostic reasoning logic within clinical inquiry scenarios.

Model	Metric	Avg.	Gastro.	Resp.	Neuro.	Card.	Neph.	Endo.	Infec.	Rheum.	Hema.
Open-source LLMs											
HuatuoGPT-o1-72B	Recall	0.4460	0.4629	**0.5061**	0.3526	0.3875	**0.5278**	0.5094	0.3178	0.3143	0.3444
	Prec.	0.5211	0.5265	0.6000	0.4260	0.4064	0.6130	0.6375	0.5333	0.3000	0.6000
Qwen3-14B	Recall	0.4298	0.5294	0.3899	0.2438	0.3792	0.4722	0.5135	0.2571	0.1714	0.4111
	Prec.	0.4600	0.5442	0.5053	0.2583	0.3672	0.4907	0.4771	0.3857	0.2000	0.5667
Qwen3-32B	Recall	0.4356	0.5194	0.4061	0.3854	0.3840	0.4000	0.4260	0.2486	0.2429	0.2000
	Prec.	0.4660	0.5540	0.5149	0.3875	0.3558	0.3963	0.4250	0.4000	0.1714	0.2000
DeepSeek-R1-Distill-14B	Recall	0.3913	0.4526	0.4285	0.2354	0.3645	0.4000	0.4844	0.2600	0.1000	0.2444
	Prec.	0.4267	0.4766	0.4914	0.2458	0.3523	0.4111	0.5604	0.5238	0.1071	0.4000
Baichuan-M2-32B	Recall	0.3687	0.3844	0.4031	0.1969	0.3826	0.4833	0.6094	0.2143	0.1000	0.1667
	Prec.	0.4517	0.4471	0.5724	0.2385	0.4486	0.5519	0.6667	0.4000	0.2000	0.3333
Doctor-R1-8B	Recall	0.3873	0.4433	0.4026	0.2521	0.3806	0.4778	0.3604	0.2771	0.0571	**0.5111**
	Prec.	0.4800	0.5237	0.5368	0.3359	0.4222	0.6111	0.4000	0.4714	0.0571	1.0000
Med42-v2-8B	Recall	0.2787	0.1575	0.4532	0.1813	0.4674	0.3056	0.5578	0.4111	0.1000	0.1333
	Prec.	0.3081	0.1857	0.5435	0.2396	0.4638	0.2648	0.5311	0.4333	0.1286	0.1333
Llama3-OpenBioLLM-70B	Recall	0.2216	0.1227	0.2412	0.1563	0.2708	0.2000	0.4042	0.2238	0.0857	0.1667
	Prec.	0.2755	0.1619	0.3593	0.2042	0.3083	0.2167	0.4625	0.3714	0.1143	0.3333
Close-source Models											
GPT-5	Recall	0.4651	0.4923	0.4961	0.3641	0.4021	0.3944	0.6115	0.3610	0.5571	0.4556
	Prec.	0.4842	0.5068	0.5816	0.3869	0.4426	0.3537	0.5896	0.4810	0.3898	0.3444
Gemini 2.5 Flash	Recall	0.4288	0.5181	0.4325	0.2724	0.3877	0.4278	0.4156	0.2771	0.2286	0.4111
	Prec.	0.4487	0.5222	0.5114	0.2833	0.3783	0.3972	0.4958	0.4714	0.1524	0.3889
Grok-4-fast-non-reasoning	Recall	0.4284	0.4672	0.4162	0.3051	0.5017	0.4500	0.6136	0.2722	0.0857	0.2444
	Prec.	0.5159	0.5316	0.5886	0.3615	0.5352	0.5370	0.7636	0.4667	0.1143	0.3111
Gemini-3-Pro-Preview	Recall	0.4748	0.4974	0.4491	0.4247	0.4529	0.5111	0.5200	0.3514	**0.6000**	0.3444
	Prec.	0.5333	0.5391	0.6036	0.4667	0.4964	0.5093	0.5422	0.5571	0.5167	0.5333
Dr. Assistant	**Recall**	**0.5066**	**0.5424**	0.4368	**0.4370**	**0.5313**	0.4722	**0.7219**	**0.3981**	0.3857	0.3778
	Prec.	0.4717	0.5146	0.4882	0.3938	0.4124	0.4444	0.6113	0.3381	0.2796	0.3667

Table 5: Comprehensive comparison of ICD-Recall and ICD-Precision across all 9 secondary departments.

Model	Metric	Number of Disease in Case's Diagnosis					
		Diag.=1	**Diag.=2**	**Diag.=3**	**Diag.=4**	**Diag.=5**	**Diag.=6**
Open-source LLMs							
HuatuoGPT-o1-72B	Recall	0.5248	0.3949	0.3920	0.3286	0.4400	**0.5667**
	Prec.	0.4807	0.5259	0.6040	0.5619	<u>1.0000</u>	<u>1.0000</u>
Qwen3-14B	Recall	0.5030	0.3960	0.3387	0.3679	0.0000	0.4000
	Prec.	0.4475	0.4675	0.4853	0.5036	0.0000	0.6667
Qwen3-32B	Recall	0.5307	0.3810	0.3360	0.3321	0.4400	0.2333
	Prec.	0.4621	0.4647	0.4360	0.4952	<u>1.0000</u>	0.8000
DeepSeek-R1-Distill-14B	Recall	0.4198	0.3820	0.3194	0.3893	0.1200	0.4667
	Prec.	0.3745	0.4484	0.4938	0.5200	0.4667	0.6000
Baichuan-M2-32B	Recall	0.4139	0.3310	0.3627	0.3571	0.0000	0.2667
	Prec.	0.4030	0.4580	0.5753	0.5500	0.0000	0.7333
Doctor-R1-8B	Recall	0.4733	0.3434	0.3227	0.2286	0.4400	0.2333
	Prec.	0.4521	0.4939	0.5287	0.4786	<u>1.0000</u>	0.7000
Med42-v2-8B	Recall	0.2938	0.2639	0.2806	0.2821	0.0000	0.4667
	Prec.	0.2656	0.3131	0.3583	0.4571	0.0000	0.9333
Llama3-OpenBioLLM-70B	Recall	0.2240	0.1808	0.1333	0.0679	0.0000	0.2333
	Prec.	0.2160	0.2700	0.2533	0.1857	0.0000	<u>1.0000</u>
Close-source Models							
GPT-5	Recall	0.5287	0.4340	0.4000	0.3750	0.3600	0.1667
	Prec.	0.4491	0.5111	0.5224	0.4690	0.6000	0.5000
Gemini 2.5 Flash	Recall	0.5192	0.3910	0.2667	0.3679	0.4400	0.1667
	Prec.	0.4414	0.4578	0.4120	0.5048	0.5000	0.6000
Grok-4-fast-non-reasoning	Recall	0.4821	0.4032	0.3545	0.3875	0.0000	0.2333
	Prec.	0.4592	<u>0.5588</u>	0.5174	0.6333	0.0000	<u>1.0000</u>
Gemini-3-Pro-Preview	Recall	**0.5455**	0.4330	0.3920	0.4321	0.3600	0.3333
	Prec.	<u>0.5003</u>	0.5498	0.5140	0.6495	0.6000	<u>1.0000</u>
Grok-3-Mini	Recall	0.4787	0.3629	0.3884	0.2821	0.4400	0.2333
	Prec.	0.4779	0.5449	<u>0.6797</u>	0.5714	<u>1.0000</u>	<u>1.0000</u>
Dr. Assistant	**Recall**	0.5426	**0.4750**	**0.4613**	**0.5500**	**0.5200**	**0.5667**
	Prec.	0.4173	0.4874	0.5165	<u>0.6536</u>	0.4667	0.7500

Table 6: Performance comparison of ICD-Recall and ICD-Precision across different numbers of ground truth ICD-10 codes (Diag.=1 to Diag.=6).

Clinical inquiry dialogue history:

Pain in the left and right shoulders and upper back for one month, occasional chest pain, with a history of hypertension. No ECG or cardiac-related tests have been performed.

How well is your hypertension controlled? Have you monitored your blood pressure recently? Do you usually have issues with obesity or high blood lipids?

I am overweight, but I don't know about blood lipids. But haven't checked, probably not high.

Not necessarily.

When high, blood pressure is around 150 systolic, diastolic 90 or 100.

You need to check your blood lipids and heart.

Usually it's around 120 systolic, diastolic 70–80. Should I check if my blood lipids are high?

Yes.

But my arms don't hurt at night when sleeping; sometimes it feels like pain in the gaps.

That could be periarthritis of the shoulder. It is related to orthopedics.

What about the occasional chest pain? Sometimes it feels like something is coming up, but once I was checked and told I had pharyngitis.

Any acid reflux or heartburn?

No heartburn, no acid reflux. Just feels like something is coming up, very uncomfortable.

You could try omeprazole + domperidone to improve the symptoms.

Patient's query: **Okay. Sometimes there is pain around the collarbone.**

Inquiry guidance by assistants:

HuatuoGPT-o1-72B:

Your symptoms do sound concerning, but please rest assured that we will work through this step by step. The shoulder and back pain could be due to conditions like frozen shoulder or cervical spondylosis. **I recommend that you consult an orthopedic specialist for further evaluation.** ✗

Dr. Assistant:

The collarbone pain, together with your previous shoulder and chest pain, requires evaluation for shoulder or heart concerns. **Your blood pressure control remains suboptimal. Prompt reassessment of your lipid levels and an ECG are advised to exclude cardiac risks.**

Figure 6: Case study. Assistants generate final inquiry guidance based on the dialogue history and the latest query. Please note that the data in our study is in Chinese. English is used here solely for illustrative purposes.

E Prompt for Data Synthesis

E.1 Prompt for CDRD Construction

Stage I | Symptom Extraction Prompt

以下是《临床诊疗指南》某一章的核心内容：
» Below is the core content of a chapter in “Clinical Diagnosis and Treatment Guidelines”:
{指南}
» {Guideline}

基于上述指南，请针对症状 “{症状}” 筛选出与其相关的 10 个【疾病】
» Based on the above guidelines, please identify 10 [Diseases] related to the symptom “{Symptom}”.

要求：提取的【疾病】需要是具体的疾病，不可是大类疾病。不可以捏造，只能提取！
» Requirements: The extracted [Diseases] need to be specific diseases, not broad categories. Do not fabricate, only extract!

输出为 List 格式：
» Output in List format:
[“AAA”, “BBB”, “CCCC”]

Stage II | Disease Matching Prompt

以下是《临床诊疗指南》某一章的核心内容：
» Below is the core content of a chapter in “Clinical Diagnosis and Treatment Guidelines”:
{指南}
» {Guideline}

已知症状 “{症状}” 对应的疾病列表为：{疾病列表}
» The disease list corresponding to the known symptom “{Symptom}” is: {Disease_List}

请基于上述指南，补全诊断部分（疾病必须严格对应已知列表，不可新增或遗漏，内容严格依据指南提取）：
» Based on the above guidelines, please complete the diagnosis section (diseases must strictly correspond to the known list, no additions or omissions, with content strictly extracted from the guidelines):
要求：
» Requirements:
1. 每个疾病需包含 “疾病特点” 和 “辅助检查” 两部分
» 1. Each disease must include two parts: “Disease Features” and “Auxiliary Examinations”.
2. 疾病特点：突出该疾病与当前症状相关的核心特征，明确鉴别要点
» 2. Disease Features: Highlight the core features of the disease related to the current symptom, clarifying differential points.
3. 辅助检查：列出指南推荐的、用于明确诊断的检查项目
» 3. Auxiliary Examinations: List the examination items recommended by the guidelines for confirming the diagnosis.
4. 格式严格遵循示例：
» 4. Strictly follow the example format:
    一、xxx 疾病

» 1. xxx Disease
疾病特点
» Disease Features
* 核心症状：xxxxxxx
» * Core Symptoms: xxxxxxx
* 伴随症状：xxxxxxx
» * Accompanying Symptoms: xxxxxxx
辅助检查
» Auxiliary Examinations
* 检查项目 1：xxxxxxx
» * Exam Item 1: xxxxxxx
* 检查项目 2：xxxxxxx
» * Exam Item 2: xxxxxxx

返回格式严格为 JSON，仅包含诊断字段，不可添加额外内容：
» Return format strictly as JSON, containing only the diagnosis field, no extra content:

{“诊断（疾病特点+进一步检查）”: “上述格式的诊断内容”}
» {“Diagnosis (Disease Features + Further Checks)”: “Diagnostic content in the above format”}

Stage III | Logic Completion Prompt

以下是《临床诊疗指南》某一章的核心内容：
» Below is the core content of a chapter in “Clinical Diagnosis and Treatment Guidelines”:
{指南}
» {Guideline}

已知信息：
» Known information:
- 症状：{症状}
» - Symptom: {Symptom}
- 对应疾病列表：{疾病列表}
» - Corresponding Disease List: {Disease_List}
- 已补全的诊断内容：{诊断内容}
» - Completed Diagnosis Content: {Diagnosis_Content}
请基于上述指南和诊断内容，补全病情收集部分（需覆盖所有疾病的鉴别要点，具有临床实用性）：
» Based on the above guidelines and diagnosis content, please complete the medical condition collection section (must cover differential points for all diseases, with clinical practicality):
要求：
» Requirements:
1. 包含但不限于 发作频率、诱发因素、伴随症状、发病程度、既往史、用药史、个人史 等维度
» 1. Including but not limited to dimensions such as attack frequency, inducing factors, accompanying symptoms, severity, past history, medication history, personal history, etc.
2. 每个维度需明确收集方向，需能支撑诊断中各疾病的鉴别
» 2. Each dimension requires a clear collection direction to support the differentiation of diseases in the diagnosis.
3. 格式严格遵循示例：

» 3. Strictly follow the example format:

1. 发作频率：是否突发或反复发作，发作间隔时间，持续时长
» 1. Attack Frequency: Sudden or recurrent, interval time, duration
2. 发病程度：症状严重程度（如水肿范围、疼痛等级），是否影响视力或日常生活
» 2. Severity: Severity of symptoms (e.g., range of edema, pain level), whether it affects vision or daily life
3. 诱发因素：是否接触过敏原、外伤、感染史、劳累、情绪波动等
» 3. Inducing Factors: Contact with allergens, trauma, infection history, fatigue, mood swings, etc.
4. 伴随症状：是否有眼红、分泌物增多、发热、淋巴结肿大、视力下降等
» 4. Accompanying Symptoms: Red eyes, increased secretions, fever, lymph node enlargement, vision loss, etc.
5. 既往疾病史：是否有眼部疾病、皮肤病、传染病、慢性病等病史
» 5. Past Medical History: Eye diseases, skin diseases, infectious diseases, chronic diseases, etc.
6. 用药史：是否使用过抗生素、激素类药物、抗过敏药物等，用药时间及疗效
» 6. Medication History: Use of antibiotics, hormones, anti-allergy drugs, etc., duration and efficacy
7. 个人史：是否有过敏史、职业暴露史、家族遗传病史等
» 7. Personal History: Allergy history, occupational exposure, family genetic history, etc.

返回格式严格为 JSON，仅包含病情收集字段，不可添加额外内容：
» Return format strictly as JSON, containing only the condition collection field, no extra content:
{“病情收集”: “上述格式的病情收集内容”}
» {“Condition Collection”: “Condition collection content in the above format”}

E.2 Prompt for Inquiry Dialogue Synthesis

医生模拟器提示词 (Doctor Simulator Prompt for Inquiry Dialogue Synthesis)

你是一名医生。目标是：根据患者所述，在确认或高度怀疑特定疾病后，给出清晰、可执行的治疗建议与就医建议。
» You are a doctor. The goal is: based on the patient’s statement, after confirming or highly suspecting a specific disease, provide clear and actionable treatment advice and medical advice.
重要规则

1. 严格按照模板格式输出
» 1. Strictly output according to the template format

2. 【诊断】字段仅填写本模板对应的疾病名称，不添加额外前缀
» 2. The [Diagnosis] field should only contain the disease name corresponding to this template, without adding extra prefixes

3. 优先提供治疗方案、用药建议、生活方式与随访计划，并明确何时需要急诊/就医
» 3. Prioritize providing treatment plans, medication advice, lifestyle and follow-up plans, and clearly state when emergency/medical attention is needed

4. 若信息不足，先用 1-2 个关键问题快速确认，再给出安全的过渡性建议
» 4. If information is insufficient, first use 1-2 key questions to quickly confirm, then give safe interim advice

5. **每次回答只提问 1-2 个关键问题**
» **5. Ask only 1-2 key questions per answer**

症状需要收集的病情
» **### Symptom Conditions to Collect ###**
{症状收集清单}
» {Symptom_Checklist}

症状的鉴别诊断（疾病特点+进一步检查）###
» **### Differential Diagnosis of Symptoms (Disease Features + Further Tests) ###**
{鉴别诊断}
» {Differential_Diagnosis}

示例格式（严格遵循此格式）###
» **### Example Format (Strictly Follow This Format) ###**
【】包裹每个字段的符号必须完全相同
» The symbols [] wrapping each field must be exactly the same

```
【已知信息】
>> [Known Information]

【待解决用户需求】
>> [User Needs to Address]

【已提供给用户信息】
>> [Information Provided to User]

【诊断】
>> [Diagnosis]

【待收集信息】
>> [Information to Collect]

【回复策略】
>> [Response Strategy]

【回复】
>> [Response]

例如:
>> For example:
【已知信息】
>> [Known Information]
小儿持续打喷嚏、流涕1个月，晨起及接触宠物后加重，家长疑过敏。
>> Child continuously sneezing and runny nose for 1 month, worse in the morning and after
contact with pets, parents suspect allergy.

【待解决用户需求】
>> [User Needs to Address]
了解可能病因、是否与过敏相关、是否需检查/用药。
>> Understand possible causes, whether related to allergies, whether examination/medication is
needed.

【已提供给用户信息】
>> [Information Provided to User]
提及接触宠物后症状加重及过敏怀疑，无其他症状、病史。
>> Mentioned symptoms worsening after pet contact and allergy suspicion, no other symptoms or
medical history.

【诊断】
>> [Diagnosis]
初步诊断：过敏性鼻炎。
>> Preliminary diagnosis: Allergic rhinitis.
```

```
【待收集信息】
>> [Information to Collect]
鼻痒、眼痒、皮疹情况；症状是否季节性；既往过敏史。
>> Itchy nose, itchy eyes, rash condition; whether symptoms are seasonal; past history of
allergies.

【回复策略】
>> [Response Strategy]
给出阶段性建议，同步补充收集关键信息。
>> Give phased advice while simultaneously collecting key information.

【回复】
>> [Response]
初步考虑过敏性鼻炎。请问孩子有鼻痒、眼痒或皮疹吗？症状是否春天加重？可进一步咨询医师明确过敏原，
制定治疗方案。
>> Preliminary consideration is allergic rhinitis. Does the child have itchy nose, itchy eyes or
 rash? Do symptoms worsen in spring? You can further consult a physician to clarify allergens
and formulate a treatment plan.
```

这是你与患者的对话历史：
» This is the history of your dialogue with the patient:
{历史对话}
» {Dialogue_History}

这是患者的最新消息：
» This is the latest message from the patient:
{最新消息}
» {Latest_Message}

这是患者的症状：
» This is the patient's symptom:
{主诉症状}
» {Symptom}

生成要求
» **### Generation Requirements ###**

1. 严格遵循模板格式
   » 1. Strictly follow the template format
2. 【诊断】仅考虑本模板疾病，不加前缀
   » 2. The [Diagnosis] field must only contain the disease corresponding to this template, without any prefix.
3. 每次聚焦 1-2 个关键确认问题 **(即回复时只提问 1-2 个关键问题)**，其后给出可执行的治疗与就医建议
   » 3. Focus on 1-2 key confirmation questions each time **(i.e., ask only 1-2 key questions when replying)**, followed by actionable treatment and medical advice
4. 语言简洁，剂量清晰，避免过度专业术语；强调安全性与随访
   » 4. Language should be concise, dosage should be clear, avoid excessive professional terminology; emphasize safety and follow-up
5. 明确何时需急诊或尽快就医
   » 5. Clearly state when emergency or immediate medical attention is needed
6. 回答时，如果涉及到药品信息，仅回复药品名和使用方式即可，禁止说明药品使用的剂量、频率、周期等

» 6. When answering, if drug information is involved, only reply with the drug name and usage method, and it is forbidden to explain the dosage, frequency, cycle, etc. of the drug use

患者模拟器提示词 (Patient Simulator Prompt for Inquiry Dialogue Synthesis)

你是一个患者，请根据以下患者画像信息和历史对话，生成一个自然、真实的患者回复。
» You are a patient. Please generate a natural and realistic patient response based on the following patient profile information and dialogue history.

患者画像信息：
» Patient Profile Information:
{患者画像}
» {Patient_Profile}

历史对话：
» Dialogue History:
{历史对话}
» {Dialogue_History}

请基于以上信息，以患者的身份回复医生。要求：
» Please respond to the doctor as a patient based on the above information. Requirements:

1. 回复要符合患者的年龄、性别、症状特征和语气风格
   » 1. The response should match the patient's age, gender, symptom characteristics and tone style
2. 回复要自然口语化，不要过于正式或书面化
   » 2. The response should be natural and colloquial, not too formal or literary
3. 回复要针对医生上一轮的询问或建议
   » 3. The response should address the doctor's inquiry or suggestion in the previous round
4. 不要包含任何思考过程或分析，直接输出患者的回复
   » 4. Do not include any thinking process or analysis, directly output the patient's response
5. 禁止将无关联的症状强行关联
   » 5. Do not forcibly associate unrelated symptoms
6. 如果涉及到药品信息，仅回复药品名和使用方式即可，禁止说明药品使用的剂量、频率、周期等
   » 6. If drug information is involved, only reply with the drug name and usage method, and it is forbidden to explain the dosage, frequency, cycle, etc. of the drug use

注意： 你所说的话必须符合客观事实
» **Note:** What you say must be consistent with objective facts
患者回复：
» **Patient Response:**

E.3 Prompt for QA-pair Data Synthesis

QA 数据生成提示词 (QA-pair Data Synthesis Prompt)

基于以下【参考答案】，生成 1 道临床考试问题：
» Based on the following [Reference Answer], generate 1 clinical examination question:

【参考答案】:
» **[Reference Answer]:**
{抽取片段}
» {Extracted_Content}

要求：
» **Requirements:**

1. 【问题】需基于上述其他信息以及【参考答案】设计，考查临床知识。【问题】不需要有实际场景，但每个答案都需要有对应问题。
   » 1. The [Question] should be designed based on the above information and [Reference Answer] to test clinical knowledge. The [Question] does not need a real scenario, but each answer must have a corresponding question.
2. 【问题】只能是主观题，不能是选择题和判断题。不可生成没有对应【参考答案】的问题。不可出现 "{症状}" 字样，因为这需要考察学生对这个症状的敏感度。
   » 2. The [Question] can only be a subjective question, not multiple-choice or true/false. Do not generate questions without corresponding [Reference Answer]. Do not include "{Symptom}" text, as this is intended to test the student's sensitivity to the symptom.
3. 【参考答案】需要根据【问题】添加 "胶水词" 连接成完整的一段话，可以调换片段顺序，但必须紧密贴合原始片段的所有语意。
   » 3. The [Reference Answer] needs to add "connecting words" according to the [Question] to form a complete paragraph. The order of segments can be rearranged, but must closely adhere to all the semantics of the original segments.
4. 【参考答案】需要包含 "{症状}" 的信息。
   » 4. The [Reference Answer] must contain information about "{Symptom}".
5. 【参考答案】中的医学信息只能从原始片段中来，不可以擅自补全。
   » 5. Medical information in the [Reference Answer] can only come from the original segments and cannot be arbitrarily supplemented.
6. 确保参考答案准确反映诊疗要点。总字数控制在 800 字以内。
   » 6. Ensure the reference answer accurately reflects diagnostic and treatment key points. Total word count should be within 800 characters.

格式：
» **Format:**

```
【参考答案】
>> [Reference Answer]

【问题】
>> [Question]
```

F Prompt for Evaluation

F.1 Data Generation Prompt for ICD-Recall and ICD-Precision Metric

ICD 测评数据 患者模型提示词 (Patient Model Prompt for ICD Metric)

你是一位患者，以下是你的个人信息：
» You are a patient. Below is your personal information:
{患者画像}
» {Patient_Profile}

以下是之前的对话历史：
» Below is the previous dialogue history:
{历史对话}
» {Dialogue_History}

你的角色要求：
» Your role requirements:

1. 严格遵循画像中的信息，不可无中生有捏造
   » 1. Strictly follow the information in the profile; do not fabricate information.
2. 基于历史对话，自然回应医生的提问/建议，或主动询问关心的问题
   » 2. Based on the dialogue history, respond naturally to the doctor's questions/suggestions, or proactively ask about concerns.
3. 不提前透露未在画像中提及的信息
   » 3. Do not reveal information not mentioned in the profile in advance.
4. 只输出语言信息，不输出神态/动作信息
   » 4. Output only verbal information; do not describe expressions or actions.
5. 请直接输出回复内容，不要输出思考过程。
   » 5. Please output the response content directly; do not output the thought process.

输出格式：
» Output format:
患者：
» Patient:

ICD 测评数据 被测试模型 (医生) 提示词 (Tested-model (Physician) Prompt for ICD Metric)

你是一位内科医生，需要为患者提供专业、耐心的医疗指导。
» You are an internal medicine physician, expected to provide professional and compassionate medical guidance.

以下是之前的对话历史：
» Below is the previous dialogue history:
{历史对话}

» {Dialogue_History}

你的角色要求:
» Your role requirements:

1. 提问需明确，帮助进一步了解病情
   » 1. Questions must be clear to facilitate further understanding of the condition.

2. 建议具体可行，符合诊疗常识
   » 2. Recommendations must be specific, feasible, and consistent with clinical practice.

3. 只输出语言信息，不输出神态/动作信息
   » 3. Output only verbal information; do not describe expressions or actions.

输出格式要求:
» Output format requirements:
医生:
» Doctor:
【诊断】（仅写病名，如不确定可写疑似）
» [Diagnosis] (Only write the disease name; if uncertain, write "suspected")
【回复】 xxxx
» [Response] xxxx

953

F.2 Data Generation Prompt in Satisfaction Metric 954

满意度测评数据生成提示词 (Data Generation Prompt for Satisfication)

你是专业临床医生，请根据你与患者的历史对话、诊疗模板（症状需要收集的病情、症状的鉴别诊断），生成 3 个回复
You are a professional clinician. Please generate 3 responses based on your dialogue history with the patient and the diagnostic template (conditions to collect for symptoms, differential diagnosis of symptoms)

患者信息: {性别}, {年龄}
» Patient Information: {Sex}, {Age}

这是你与患者的历史对话:
» This is your dialogue history with the patient:
{历史对话}
» {Dialogue_History}

这是患者的最新消息:
» This is the latest message from the patient:
{最新消息}
» {Latest_Message}

要求:
» **## Requirements:**

1. 对于每个**【回复】**，优先回答患者的需求，照顾患者害怕担忧等情绪。再进行诊疗。 955

» 1. For each [Response], prioritize addressing the patient's needs and caring for their emotions such as fear and worry. Then proceed with diagnosis and treatment.

2. 其次，若能给出诊断，则需要优先给出。
   » 2. Secondly, if a diagnosis can be made, it should be provided first.

3. 若患者回复仅为 "好的""谢谢" 等无诊疗信息陈述句，则【回复 1】**优先**祝福患者早日康复！
   » 3. If the patient's reply is merely a declarative sentence without diagnostic information such as "OK" or "Thank you", [Response 1] should **prioritize** wishing the patient a speedy recovery!

4. 输出格式为：每个回复不超过 100 字，且每个【回复】独立存在
   【回复 1】
   【回复 2】
   【回复 3】
   » 4. Output format: each response should not exceed 100 characters, and each [Response] should be independent.
   » [Response 1]
   » [Response 2]
   » [Response 3]

5. 若有诊疗需要的提问，每个【回复】提问不超过两个问号！
   » 5. If there are questions needed for diagnosis and treatment, each [Response] should contain no more than two question marks!

6. 若有用药，不可给出明确剂量！
   » 6. If medication is mentioned, do not specify exact dosage!

7. 不可重复医生说过的话！
   » 7. Do not repeat what the doctor has already said!

例如：
» **For Example:**

```
【回复1】
>> [Response 1]
aaaaa

【回复2】
>> [Response 2]
bbbbbbbbbb

【回复3】
>> [Response 3]
ccccccccc
```

请输出 3 个回复：
» **## Please output 3 responses:**

956

G Data Demos

G.1 CDRD

结构化的临床诊断推理数据 (CDRD)

主症状
» **Chief Symptom**
咳嗽
» Cough

病情收集
» **Symptom Collection**

1. **咳嗽的性质**: 干咳、有痰
   » 1. **Nature of Cough**: Dry cough, productive cough (with sputum)

2. **咳嗽的时间**: 几天、几个月、几年加重几天
   » 2. **Duration of Cough**: Days, months, years; aggravation for days

3. **咳嗽的音色**: 咳嗽嘶哑、鸡鸣样咳嗽、咳嗽声音低微
   » 3. **Timbre of Cough**: Hoarse cough, whooping cough, weak cough sound

4. **痰的性质**: 粘液性痰、泡沫样痰、脓性痰、痰中带血
   » 4. **Nature of Sputum**: Mucous sputum, frothy sputum, purulent sputum, blood-stained sputum

5. **痰量**: 少、多
   » 5. **Sputum Amount**: Scant, copious

6. **伴随症状**: 伴发热、伴胸痛、伴咳血、伴脓痰、伴哮喘、伴呼吸困难、咽干，咽痒，咽痛，鼻塞，喷嚏，流清涕
   » 6. **Accompanying Symptoms**: With fever, with chest pain, with hemoptysis, with purulent sputum, with asthma, with dyspnea, dry throat, itchy throat, sore throat, nasal congestion, sneezing, clear runny nose

---

诊断考虑
» **Diagnosis Considerations**
1. 上呼吸道感染
» **1. Upper Respiratory Tract Infection**

疾病特点: 起病较急，几天。表现为咳嗽，咽干，咽痒，甚至咽痛，伴有鼻部症状如鼻塞，喷嚏，流清涕。病原体多为病毒，少数为细菌。
» **Disease Features**: Acute onset, days. Manifested as cough, dry throat, itchy throat, even sore throat, accompanied by nasal symptoms such as nasal congestion, sneezing, clear runny nose. Pathogens are mostly viruses, occasionally bacteria.

实验室检查: 白细胞计数正常或偏低，伴淋巴细胞比例偏高。
» **Laboratory Tests**: WBC count normal or low, with elevated lymphocyte percentage.

治疗反应: 多饮水、保持室内空气流通和防治继发性细菌感染。一般 5 ~ 7 天痊愈，伴发并发症者可致病程迁延。
» **Treatment Response**: Drink plenty of water, maintain indoor air circulation, and prevent secondary bacterial infection. Generally recovers in 5-7 days; complications may prolong the course.

2. 急性支气管炎
» **2. Acute Bronchitis**

疾病特点: 常起病较急，几天，发生于寒冷季节或气候突变时，也可由急性上呼吸道感染迁延不愈所致。全身症状较轻，可有发热。初为干咳或少量黏痰。
» **Disease Features**: Often acute onset, days, occurring in cold seasons or during weather changes, or may result from unresolved acute upper respiratory infection. Mild systemic symptoms, possible fever. Initially dry cough or scant mucous sputum.

实验室检查: 周围血白细胞计数可正常，但由细菌感染引起者，可伴白细胞总数和中性粒细胞百分比升高。血沉加快，痰培养可见致病菌。
» **Laboratory Tests**: Peripheral WBC count may be normal, but in bacterial infection, total WBC and neutrophil percentage may be elevated. ESR increased, sputum culture may show pathogenic bacteria.

影像学检查: X 线胸片大多为肺纹理增强，少数无异常发现。
» **Imaging**: Chest X-ray mostly shows increased lung markings; few show no abnormalities.

3. 肺部感染性疾病
» **3. Pulmonary Infectious Diseases**

疾病特点: 常见症状为咳嗽、咳痰，或原有呼吸道症状加重，并出现脓性痰或血痰，伴或不伴胸痛。病变范围大者可有呼吸困难、呼吸窘迫。大多数病人有发热。
» **Disease Features**: Common symptoms include cough, sputum production, or worsening of existing respiratory symptoms, with purulent or bloody sputum, with or without chest pain. Extensive lesions may cause dyspnea or respiratory distress. Most patients have fever.

确定病原体: 血培养和痰培养分离到相同细菌，可确定为肺炎的病原菌。测定特异性 IgM 抗体滴度，可判断支原体、衣原体、嗜肺军团菌和病毒感染等。
» **Pathogen Identification**: Isolation of the same bacteria from blood and sputum cultures can confirm the pathogen of pneumonia. Measuring specific IgM antibody titers can identify Mycoplasma, Chlamydia, Legionella pneumophila, and viral infections.

影像学检查: X 线胸片多为肺实质浸润表现。
» **Imaging**: Chest X-ray mostly shows pulmonary parenchymal infiltration.

治疗反应: 抗感染治疗：选择可能覆盖病原体的抗菌药物，抗感染治疗有效。
» **Treatment Response**: Anti-infective therapy: select antibiotics that may cover the pathogens; anti-infective treatment is effective.

G.2 Inquiry Example with Reasoning

带推理的对话示例 (Inquiry Example with Six Reasoning Steps)

输入/提示词
» **Input/Prompt**
历史:
» **History:**

患者: 咳嗽、喉咙痒，术后伤口长肉芽发炎
» **Patient**: Cough, itchy throat, granulation tissue inflammation in postoperative wound.

医生: 请问咳嗽持续了几天？是否有痰液或发热？
建议：多饮温水、保持空气湿润，术后伤口需每日消毒（如碘伏），若伤口红肿加重或出现脓液，请立即就医。
» **Doctor**: How long has the cough lasted? Any sputum or fever?
» Advice: Drink warm water, keep air humid, disinfect wound daily with iodine; if redness worsens or pus appears, seek medical help immediately.

患者: 咳嗽三四天了，没痰也不发烧。
» **Patient**: Coughing for 3-4 days, no sputum, no fever.

医生: 请问咳嗽时是否伴有咽干、咽痒或鼻塞？术后伤口是否有胸痛或呼吸困难？
建议：多饮温水、保持空气湿润，可尝试使用右美沙芬止咳糖浆（具体用法遵说明书）。若咳嗽持续加重、出现痰液或发热，请及时就医。
» **Doctor**: Is the cough accompanied by dry/itchy throat or nasal congestion? Any chest pain or difficulty breathing at the wound site?
» Advice: Drink warm water, keep air humid, try Dextromethorphan cough syrup (follow instructions). If cough worsens, sputum or fever appears, seek medical help.

患者: 有咽痒，伤口没胸痛或呼吸困难。
» **Patient**: Have itchy throat, no chest pain or breathing difficulty at the wound.

医生: 请问咳嗽时是否伴有鼻塞或流清涕？近期是否接触过花粉、宠物等过敏原？
建议：继续使用右美沙芬止咳糖浆（具体用法遵说明书），保持空气湿润，避免刺激性气体。若咳嗽持续超过 *1* 周或出现发热、痰液，请及时就医。术后伤口需每日碘伏消毒，若红肿加重或有脓液，请立即就诊。
» **Doctor**: Is the cough accompanied by nasal congestion or clear runny nose? Have you been in contact with allergens such as pollen or pets recently?
» Advice: Continue using Dextromethorphan cough syrup (follow instructions), keep air humid, avoid irritating gases. If cough persists >1 week or fever/sputum appears, seek medical help. Disinfect wound daily with iodine; if redness worsens or pus appears, seek medical help immediately.

最新消息/患者: 没鼻塞流涕，没接触过敏原。
» **Latest Message - Patient:** No nasal congestion or runny nose, no contact with allergens.

模型输出/推理与回复
» **Model Output - Reasoning & Response**

```
【已知信息】
>> [Known Information]
咳嗽3-4天，干咳伴咽痒，无痰、无发热、无鼻塞流涕，术后伤口无胸痛或呼吸困难，未接触过敏原。
>> Cough for 3-4 days, dry cough with itchy throat, no sputum, no fever, no nasal congestion or
runny nose, no chest pain or dyspnea at wound site, no allergen contact.

【待解决用户需求】
>> [User Needs to Address]
明确咳嗽病因，排除感染或术后并发症，指导用药及伤口护理。
>> Identify the cause of cough, exclude infection or postoperative complications, provide
medication and wound care guidance.

【已提供给用户信息】
>> [Information Provided to User]
已排除过敏性鼻炎典型症状（鼻塞流涕），建议使用右美沙芬止咳，伤口需碘伏消毒。
>> Excluded typical allergic rhinitis symptoms (nasal congestion and runny nose), advised
Dextromethorphan for cough, wound needs iodine disinfection.

【诊断】
>> [Diagnosis]
急性支气管炎
```

```
>> Acute Bronchitis

【待收集信息】
>> [Information to Collect]
咳嗽是否伴有胸痛或呼吸困难？术后伤口是否有红肿加重或脓液？
>> Is the cough accompanied by chest pain or difficulty breathing? Is there worsening redness/
swelling or pus at the wound site?

【回复策略】
>> [Response Strategy]
聚焦关键点（胸痛/呼吸困难提示肺部感染，伤口感染需急诊），同步强化居家护理建议。
>> Focus on key points (chest pain/dyspnea suggests lung infection; wound infection needs ER),
reinforce home care advice.

【回复】
>> [Response]
请问咳嗽时是否伴有胸痛或呼吸困难？术后伤口目前是否有红肿加重或脓液？
>> Does the cough come with chest pain or difficulty breathing? Is the postoperative wound
currently showing increased redness, swelling, or pus?

建议继续使用右美沙芬止咳糖浆（具体用法遵说明书），保持空气湿润，多饮温水。术后伤口需每日碘伏消
毒，若出现红肿加重、脓液或发热，请立即就医。
>> I recommend continuing Dextromethorphan cough syrup (follow instructions), keeping the air
humid, and drinking warm water. Disinfect the wound daily with iodine; if redness worsens, pus
appears, or fever develops, seek medical attention immediately.
```

G.3 QA Pair

QA 数据示例 (QA Data Examples)

【问题 1】
» **[Question 1]**
请简述在对某一呼吸系统症状进行问诊时，应如何全面评估患者的呼吸道分泌物及相关表现？
» Please briefly describe how to comprehensively evaluate a patient's respiratory secretions and related manifestations when taking a history for a respiratory symptom.

【参考答案 1】
» **[Reference Answer 1]**
患者无咳嗽，但若出现咳嗽，应注意其音色是否嘶哑或呈鸡鸣样，声音是否低微；同时应观察咳出物的性质，如是否为粘液性、泡沫样、脓性或痰中带血；痰量的多少也是重要的评估内容，需判断为少量或大量；此外，还需注意是否伴有其他相关症状，以帮助明确病因和诊断。
» The patient has no cough, but if cough occurs, attention should be paid to whether the sound is hoarse or whooping, and whether the voice is weak. At the same time, the nature of the sputum should be observed, such as whether it is mucous, frothy, purulent, or blood-stained sputum. The amount of sputum is also an important aspect of the assessment, and it should be determined as scant or copious. In addition, attention should be paid to whether there are other related symptoms to help clarify the etiology and diagnosis.

【问题 2】
» **[Question 2]**
请简述肺炎的常见临床表现，并说明在临床诊断中为何不能仅依赖 "咳嗽" 这一症状进行判断。同时，请指出实验室检查在肺炎诊疗中的作用。
» Please briefly describe the common clinical manifestations of pneumonia, and explain why clinical diagnosis cannot rely solely on the symptom of "cough". Also, please indicate the role of laboratory

tests in the diagnosis and treatment of pneumonia.

【参考答案 2】
» **[Reference Answer 2]**
肺炎是一种常见的呼吸系统感染性疾病，其典型临床表现包括咳嗽、咳痰，部分患者痰中可带血，常伴有发热、寒战和乏力等症状。在病情较重时，患者可能出现胸痛或呼吸困难。值得注意的是，并非所有患者都会出现咳嗽症状，因此在临床诊断中应综合考虑其他表现。实验室检查是明确诊断的重要手段，有助于判断感染类型及严重程度，从而指导后续治疗。
» Pneumonia is a common respiratory infectious disease. Its typical clinical manifestations include cough and sputum production; some patients may have blood-stained sputum, often accompanied by fever, chills, and fatigue. In severe cases, patients may experience chest pain or dyspnea. It is worth noting that not all patients will present with cough symptoms, so clinical diagnosis should comprehensively consider other manifestations. Laboratory tests are an important means of confirming the diagnosis, helping to determine the type and severity of infection, thereby guiding subsequent treatment.

G.4 Patient Profile

患者画像示例 (Patient Profile Example)

年龄: 45 岁
» **Age**: 45 years old
性别: 男
» **Gender**: Male
性格特征: 较为顺从，对医生建议接受度高，但因病情反复而略显沮丧，经常询问病情进展。
» **Personality Traits**: Relatively compliant, high acceptance of doctor's advice, but slightly frustrated due to recurring illness, frequently asks about disease progression.
语气特征: 语气平和但带有担忧，对治疗效果和未来预后表示关心。
» **Tone Characteristics**: Peaceful tone but with worry, concerned about treatment effect and future prognosis.
发言字数: 每次发言字数中等，约 25-45 字。
» **Speaking Word Count**: Medium length per speech, about 25-45 words.
症状: 右肺中叶见磨玻璃样小结节，伴有长期干咳，夜间加重，有过敏性鼻炎史，偶有咽痛，吸烟。
» **Symptoms**: Ground-glass nodule in the right middle lobe, accompanied by chronic dry cough, worse at night, history of allergic rhinitis, occasional sore throat, smoker.
诊断: 咳嗽变异性哮喘，肺部磨玻璃结节（考虑良性）。
» **Diagnosis**: Cough variant asthma, pulmonary ground-glass nodule (considered benign).
其他信息: 使用吸入性糖皮质激素后症状减轻，但停药后易复发，有 10 年吸烟史，过敏性鼻炎定期发作。
» **Other Information**: Symptoms alleviated after using inhaled corticosteroids, but prone to relapse after stopping medication; 10-year smoking history; periodic attacks of allergic rhinitis.
第一句话: "最近三个月一直干咳乏力。"
» **First Sentence**: "I've had a dry cough and fatigue for the last three months."

G.5 Real Clinical Inquiry Dialogue in Bench

测评集中的真实临床问诊记录 (Real Clinical Inquiry Dialogue Example in Bench)

这是你与患者的历史对话:
» This is the history of your dialogue with the patient:
患者: 左右肩膀后背疼痛一个月，偶尔胸口疼，有高血压病史，未进行心电图或心脏相关检查。(匿名，女，35 岁)
» Patient: Pain in both shoulders and back for a month, occasional chest pain, history of hypertension, no ECG or heart-related examinations. (Anonymous, Female, 35 years old)
医生: 高血压控制得如何? 最近有没有监测过血压?
» Doctor: How is the hypertension controlled? Have you monitored your blood pressure recently?
医生: 平时有肥胖 血脂高吗
» Doctor: Do you have obesity or high blood lipids?
患者: 胖，但是血脂高不高不知道，没查过，应该是不高
» Patient: I am overweight, but I don't know if my blood lipids are high, I haven't checked, presumably not high.
医生: 不一定
» Doctor: Not necessarily.

这是患者的最新消息:
» This is the latest message from the patient:
患者: 血压就高时候一百五左右，低压九十或者一百
» Patient: When blood pressure is high, it is around 150, diastolic 90 or 100.