MSDiagnosis: A Chinese Benchmark for Evaluating Large Language Models in Multi-Step Clinical Diagnosis

Anonymous ACL submission

Abstract

Clinical diagnosis is critical in medical practice, typically requiring a continuous and evolving process that includes primary diagnosis, differential diagnosis, and final diagnosis. However, most existing clinical diagnostic tasks are single-step processes, which does not align with the complex multi-step diagnostic procedures found in real-world clinical settings. In this paper, we propose a Chinese clinical diagnostic benchmark, called MSDiagnosis. This benchmark consists of 2,225 cases from 12 departments, covering tasks such as primary diagnosis, differential diagnosis, and final diagnosis. Additionally, we propose a novel and effective framework. This framework combines forward inference, backward inference, reflection, and refinement, enabling the large language model to self-evaluate and adjust its diagnostic results. To this end, we evaluate medical language models, general language models, and our proposed framework. The experimental results demonstrate the effectiveness of the proposed method. We also provide a comprehensive experimental analysis and suggest future research directions for this task. The dataset and codes are available at the anonymous URL https://anonymous.4open.science/r/MDQA-6EC0.

1 Introduction

Clinical diagnosis is a central element of clinical decision-making, involving the integration of chief complaint, present history, and physical examinations, along with clinical experience and medical knowledge, to make a scientifically grounded judgment regarding the nature and cause of a disease (Ball et al., 2015). Accurate diagnosis helps to provide the most appropriate treatment for patients. In recent years, research in data-driven clinical diagnosis has advanced rapidly, particularly with the widespread use of electronic medical records (EMRs), which enable diagnostic models to integrate multi-stage patient data for more precise analysis (Herrero-Zazo et al., 2021).

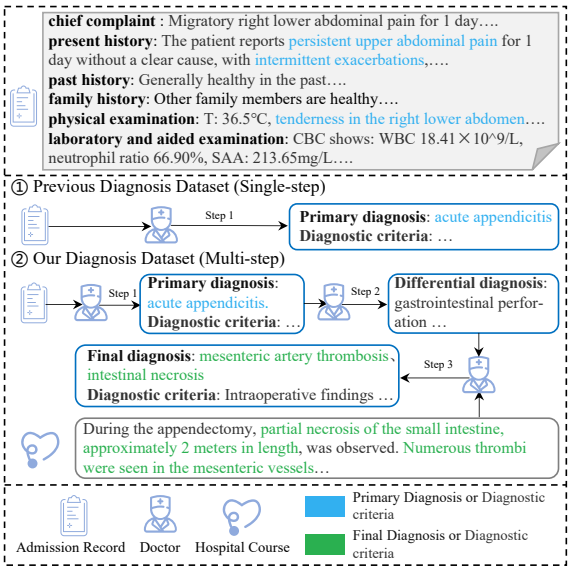

Figure 1: An example of our diagnostic benchmark and its differences from the previous diagnostic benchmark.

The clinical diagnosis process is a continuous and evolving process (Tiffen et al., 2014). Specifically, doctors first make a primary diagnosis based on the patient's chief complaint, present history, and laboratory and aided examination. Then they narrow down the diagnostic options through differential diagnosis. Finally, they determine the final diagnosis by considering clinical changes throughout the diagnostic and treatment process. However, a significant gap remains between the existing diagnostic benchmark and actual clinical practice. Most existing diagnostic tasks are single-step processes (Wang et al., 2024; Lyu et al., 2023; Liu et al., 2024b), where a diagnosis is made directly based on the patient's medical history, chief complaint, and examination results. This single-step approach does not align with the multi-step process typically used in clinical practice. In addition, current evaluations of single-step diagnostic tasks typically use BLEU and ROUGE metrics. However, these metrics, which rely solely on lexical

Table 1: Overview of the clinical diagnosis benchmark. "*" indicates that the dataset contains data for multiple tasks, with the diagnosis being just one of them. In the "Size" column, the value outside the brackets indicates the number of samples related to diagnosis, and the value inside the brackets indicates the total number of samples in the dataset. The terms "Criteria", "Multi-Step", "Q.Type", and "MCQA" stand for "diagnostic criteria", "multi-step diagnosis", "question type", and "multiple-choice QA" respectively.

Dataset	Data Source	Q.Type	Criteria	Multi-Step	Key Points	Size
DDx-basic (Liu et al., 2024c)	Examination Paper	MCQA	✗	✗	✗	150
DDx-advanced (Liu et al., 2024c)	Examination Paper	MCQA	✗	✗	✗	30
CMExam* (Liu et al., 2024b)	Medical Examination	MCQA	✗	✗	✗	1,630(60K+)
AgentClinic-MedQA (Schmidgall et al., 2024)	GPT4+MedQA	OpenQA	✗	✗	✗	107
AI Hospital (Fan et al., 2024)	Medical Website	OpenQA	✓	✗	✗	506
CMB-Clin (Wang et al., 2024)	Medical Textbook	OpenQA	✓	✗	✗	74
RJUA-QA (Lyu et al., 2023)	Synthetic	OpenQA	✓	✗	✗	2,132
MIMIC-CDM (Hager et al., 2024)	MIMIC	OpenQA	✓	✗	✗	2,400
Ours	Medical Website	OpenQA	✓	✓	✓	2,225

overlap, cannot effectively assess the rationale behind the diagnosis or the coverage of key evidence.

Hence, in this paper, we propose a **M**ulti-**S**tep clinical **Diagnosis** benchmark (MSDiagnosis). In this benchmark, a primary diagnosis and diagnostic criteria are generated based on the patient's chief complaint, present history, and examination results. Then, a differential diagnosis is made to narrow down the possible diseases. Finally, the final diagnosis and diagnostic criteria are made by combining the hospital course, primary diagnosis, and differential diagnosis. The specific process is shown in Fig. 1. The MSDiagnosis benchmark is designed to evaluate the diagnostic capabilities of large language models (LLMs). The benchmark comprises 2,225 medical records collected from medical websites, covering 12 departments. Each medical record contains five related diagnostic questions. For ease of comparison, we list the most relevant works in Table 1. Additionally, to accurately assess diagnostic criteria, we annotate each answer with corresponding key points and evaluate the quality of the answers by calculating the macro-recall for each key point.

For the multi-step clinical diagnostic benchmark, we propose a simple and effective pipeline framework. This framework consists of two stages. The first stage is forward inference. Specifically, we retrieve similar EMRs to serve as in-context learning (ICL), allowing the LLM to diagnose the patient. The second stage involves backward inference, reflection, and refinement. Specifically, we have designed corresponding rules for each of these processes and combined them with ICL to enable the LLM to review and refine its diagnostic results.

Our contributions are summarized as follows:

- We propose a multi-step clinical diagnosis benchmark that includes three tasks: primary diagnosis, differential diagnosis, and final diagnosis.
- We propose a simple and efficient framework that combines forward and backward inference, enabling LLMs to validate and refine diagnostic results.
- We evaluate both medical and general LLMs on MSDiagnosis and conduct extensive experiments on our framework. The experimental results demonstrate the effectiveness of the proposed method.

2 Problem Formulation

In this paper, we consider the multi-step diagnosis task as a multi-round dialogue problem. For each complex EMR, we simulate the interaction between an examiner and a candidate, where the candidate is required to provide an answer to a specific diagnostic question. Specifically, given a patient's admission record E, which includes the chief complaint, present history, past history, physical examination, and laboratory and aided examination, the candidate answers two questions $R_1 = [Q_1, Q_2]$ in the first round of dialogue. These questions mainly involve the patient's primary diagnosis and its diagnostic criteria. As illustrated in Fig. 3, Q_1 and Q_2 are "What is the patient's primary diagnosis?" and "What is the criterion for the primary diagnosis?" respectively. In the second round of dialogue, the candidate needs to answer the question $R_2 = [Q_3]$. This question refers to asking the patient about their differential diagnosis. In the third round of dialogue, the candidate receives the patient's hospital course T, the dialogue history H, and two additional questions $R_3 = [Q_4, Q_5]$.

These questions focus on the patient's final diagnosis and its diagnostic criteria. Therefore, the input X for the current dialogue can be formally expressed as $X = E+T+H+Q_i$, where $Q_i \in R_3$. The reference answer for this input is A_i.

3 MSDiagnosis

In this section, we detail the construction process of the dataset, including the data collection, selection, and data annotation.

3.1 Data Collection and Selection

In this study, we select a Chinese medical website[1] as the source of EMRs. The raw data are anonymized, and we apply additional anonymization to ensure no protected health information (PHI) remains. After de-identifying the data, we obtain a total of 11,900 EMRs. To ensure high-quality EMR, we select the data through four steps. First, we remove redundant information unrelated to medical content, such as web elements, copyright statements, and advertisements. Second, we eliminate EMRs with duplicate field values and missing fields, such as those lacking primary diagnosis, differential diagnosis, final diagnosis, or hospital course fields. After this step, we retain 5096 EMRs. Third, we deduplicate EMRs from the same department. If two records have identical chief complaints, present histories, and physical examinations, one of them is removed. After this step, we retain 4179 EMRs. Fourth, to avoid duplicate patients in the dataset, we conduct a meticulous matching process using patients' demographic information (such as gender, occupation, etc.) and filter out individuals with identical profiles. After rigorous screening, we ensure the dataset contains no patients with fully identical records. After four steps, we ultimately obtain 3,501 high-quality, complex, and authentic EMRs.

3.2 Data Annotation

3.2.1 Question Construction

We first manually construct five seed questions, including the patient's primary diagnosis and its criteria, the differential diagnosis, and the final diagnosis and its criteria. The definition of each question is shown in Appendix D. Then, we use GPT-4 to expand each constructed question with ten similar questions. Finally, we manually check and filter out unreasonable question expressions.

[1] https://www.iiyi.com/

3.2.2 Answer Annotation

To ensure the quality of the dataset construction, we form a professional team comprising three inspectors and one reviewer. All team members undergo specialized medical training to understand primary diagnosis, differential diagnosis, and final diagnosis. The construction process includes three stages: first-round annotation, second-round checking, and third-round review.

First-round annotation. For questions related to primary diagnosis, differential diagnosis, and final diagnosis, if the original medical records contain relevant answers, the standard answers are directly extracted from the EMR. However, statistical analysis reveals that 1,989 cases lack diagnostic results. Additionally, since the original data does not provide criteria for the primary and final diagnoses, we initially use GPT-4 to generate diagnostic criteria for 3,501 cases.

Second-round checking. We engage three college students to simultaneously check the rationality of all question-answer pairs. Samples that all three inspectors considered unreasonable are directly discarded. If one or two inspectors find a sample unreasonable, it is manually re-annotated and retained only if all three inspectors subsequently agree on its reasonableness. After rigorous check, we identify 1,222 cases that are unanimously deemed unreasonable.

Third-round review. A verified batch is given to a medical expert for double review. The medical expert randomly inspects 20% of the batch samples, totaling 700 samples. Any unqualified annotations are returned to the check team with explanations, which can further refine the standards. This process is repeated until the batch accuracy reaches 95%. After this stage, we remove 54 cases, ultimately retaining 2,225 high-quality cases.

3.2.3 Key Points Annotation

To more accurately evaluate open-ended questions such as primary diagnostic criteria and final diagnostic criteria, we annotate the key points of the answers to these questions. Specifically, for each primary and final diagnostic criteria, we first construct key point extraction prompts and employ GPT-4 to extract the information from the standard answers. The key point for the diagnostic criteria mainly includes four categories: medical history, symptoms, physical signs, and examination results. Then, we invite two students with specialized medical training to validate the extracted information.

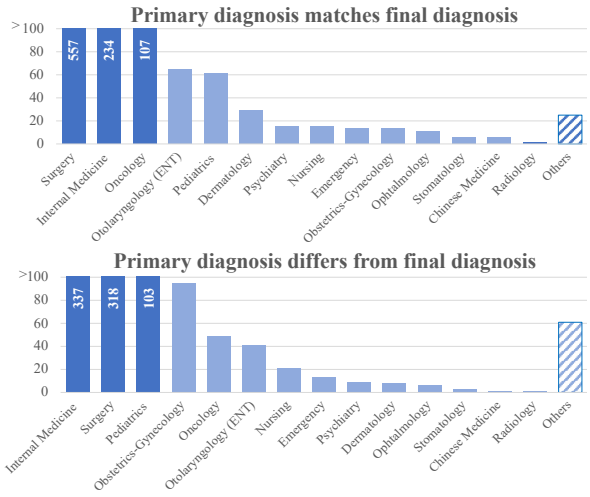

Figure 2: Distribution of "Primary diagnosis matches final diagnosis" and "Primary diagnosis differs from final diagnosis" across departments in MSDiagnosis.

Table 2: Statistics of our constructed dataset. "Diag-M" and "Diag-D" refer to EMRs where the primary diagnosis matches or differs from the final diagnosis, respectively. "AvgP" and "AvgF" represent the average number of diseases in the primary and final diagnoses.

Type	Size	Diag-M	Diag-D	AvgP	AvgF
Train	1,557	808	749	2.53	2.78
Test	445	234	211	2.44	2.77
Dev	223	118	105	2.44	2.88
Total	2,225	1,160	1,065	2.51	2.79

If any inconsistencies are found in the EMRs, the data is re-annotated. Finally, we obtain 2,225 high-quality samples to construct MSDiagnosis. We calculate the Cohen's Kappa (Banerjee et al., 1999) score for the key points of the primary diagnostic criteria and the final diagnostic criteria. The Cohen's Kappa for the primary diagnostic criteria is 0.81, indicating a high level of agreement, while the Cohen's Kappa for the final diagnostic criteria is 0.79, indicating a moderate level of agreement.

4 Dataset Analysis

In this section, we introduce the statistics and the characteristics of the MSDiagnosis in detail.

4.1 Data Statistics

As reported in Table 2, the dataset consists of 2,225 EMRs across 12 departments, which are divided into training, validation, and test sets according to a 7:1:2 ratio. Additionally, we conduct a detailed statistical analysis of patient diagnoses from three main perspectives. 1) *Department Distribution*. Fig. 2 presents the distribution of EMRs across different departments, categorized by whether the primary diagnosis is consistent with the final diagnosis. In the category where the primary and final diagnoses are the same, surgery has the highest proportion (48.02%). In the category where the primary and final diagnoses are different, internal medicine has the highest proportion (31.64%). 2) *Number of Diseases*. The average number of diseases diagnosed in the primary diagnosis stage is 2, with a maximum of 21 diseases. At the final diagnosis stage, the average number of diseases is 3, with a maximum of 21 diseases. 3) *Types of Diagnostic Changes*. Statistical analysis reveals that diagnosis changes primarily fall into three categories: addition, deletion, and modification.

4.2 Data Characteristics

The MSDiagnosis has several significant advantages compared to previous medical diagnostic datasets: 1) *Reliability and Authenticity of Data*. The MSDiagnosis is entirely based on EMRs, including detailed treatment plans and course records. 2) *Rich Variety of Departments and Diseases*. The MSDiagnosis covers EMRs from 12 different departments, encompassing a wide range of diseases, including common diseases, rare diseases, acute diseases, and chronic diseases. This broad diversity ensures the data's wide applicability. 3) *Inclusion of Multi-step Diagnostic Processes*. The dataset reflects real-world clinical scenarios by incorporating multi-step diagnostic processes. By recording diagnostic changes, it captures the complex and dynamic nature of medical diagnostics.

5 Method

In this section, we first provide an overview of our framework. Then, we describe each module of the framework in detail.

5.1 Framework

As illustrated in Fig. 3, our proposed framework mainly consists of two stages. The first stage involves forward inference. In this stage, we retrieve similar EMRs to serve as ICL, enabling the LLM to diagnose the patient. The second stage is backward inference, reflection, and refinement. In this stage, we first validate the diagnostic criteria against the facts derived from the diagnostic results. Then, the LLM uses designed reflection rules to evaluate the diagnostic outcomes. Finally, the diagnosis is refined by integrating all the previous results.

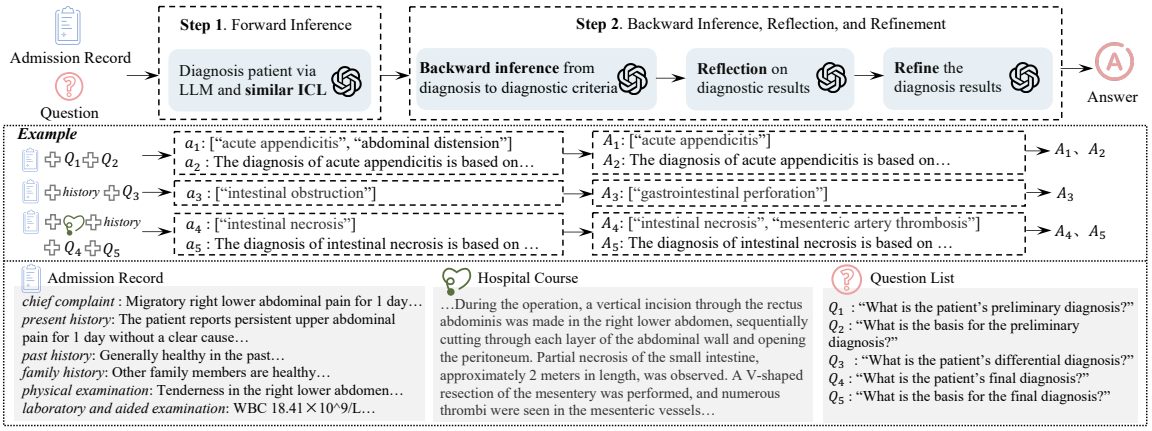

Figure 3: Our framework for the multi-step clinical diagnosis. The top portion of the figure illustrates the flow of the framework, comprising two stages. The first stage involves the forward inference diagnosis. The second stage focuses on backward inference, reflection, and refinement.

5.2 Forward Inference

This part aims to make diagnoses for patients based on admission records. In this paper, we utilize an LLM (e.g., GPT4o-mini) with the ICL method to achieve this purpose. Its core idea is to select similar EMRs from the training set, guiding the model in making accurate diagnoses for patients.

Specifically, given the admission record E and question Q_o, where o is the total number of questions, we select ICL examples with similar semantics to E through the following steps. First, we use the BGE (Xiao et al., 2023) model to obtain the representations of E and each training sample Y_i $(1 \leq i \leq u)$, where u is the size of the training set. Second, we calculate their cosine similarity and select the top K samples with the highest similarity as ICL examples. Along with E and the top K samples, we provide the LLM with a role definition ("You are a professional doctor, and you need to complete the diagnosis task") and predefined output format rules ("The output can be loaded directly using the JSON.load() function"). Finally, the LLM would generate the answer a_o for question Q_o. The detailed prompt is shown in Appendix I.

5.3 Backward Inference and Reflection

After obtaining the forward inference results, this part aims to conduct backward inference, reflection, and refine the patient's diagnosis. In this section, we define some rules to achieve this goal.

Specifically, given the forward inference diagnostic results a_o, we first perform backward inference from the diagnosis to diagnostic criteria. In this step, we define backward inference rules to guide the LLM in generating outputs that comply with pre-defined content and format constraints. For content constraints, we have "For each diagnosis, recall the representative medical history, symptoms, physical signs, and examination results" (Rule 1). For format constraints, we consider "The recalled content should follow the format: Medical History: Recall the representative medical history for the disease; delete this item if not applicable" (Rule 2). Then, we design reflection rules to review the diagnoses. The specific reflection rule is: "If a diagnosis's characteristics don't align with the medical record, delete or revise it, and provide the rationale". Finally, we combine the aforementioned backward inference, reflection results, and their criteria to let the model optimize the diagnostic results, with specific prompts shown in Appendix I.

6 Experiments

In this section, we conduct a series of experiments to evaluate the performance of LLMs on the MS-Diagnosis and the effectiveness of our framework.

6.1 Experimental Setup

6.1.1 Baseline

In this paper, we mainly describe several types of baseline methods, including medical LLMs, general LLMs, and other methods. In the medical LLMs, we employ MMedLM (Qiu et al., 2024), PULSE (Xiaofan Zhang, 2023), Llama3-OpenBioLLM (Ankit Pal, 2024), and Apollo2-7B (Zheng et al., 2024) for comparison. Based on these models, we manually construct an example to serve as ICL. In general LLMs, we categorize them into two types: small LLMs (<20B) and large LLMs (>20B). For small LLMs, we compare

Table 3: Method comparisons on MSDiagnosis (%). "Pri" refers to the primary diagnosis. "Fin" stands for the final diagnosis. "DD" means differential diagnosis. All results are the average of three runs. The best and second results, excluding the human method, are highlighted in **bold** and underline, respectively.

Model	Settings	F1			Macro-Recall		Rouge-L		BLEU-1		BERTScore	
		Pre	DD	Fin	Pre	Fin	Pre	Fin	Pre	Fin	Pre	Fin
Medical LLMs												
Llama3-OpenBioLLM-8B	1-Shot	00.41	9.90	0.14	18.62	10.19	11.56	7.30	8.57	5.50	50.19	49.43
MMed-Llama-3-8B	1-Shot	22.03	8.87	18.19	40.73	32.68	22.29	14.77	12.83	6.02	60.90	60.13
PULSE-20bv5	1-Shot	26.40	9.55	13.18	27.57	24.37	45.93	42.20	36.11	32.44	80.46	78.46
Apollo2-7B	1-Shot	33.61	8.47	26.78	37.01	33.15	51.25	44.91	37.00	33.10	80.70	78.50
General LLMs												
Llama-3.1-8B	1-Shot	32.72	5.54	19.99	37.88	22.11	51.91	32.10	34.42	27.86	82.10	77.13
	LoRA	20.59	7.86	16.32	17.04	14.06	25.07	19.76	18.69	15.00	70.97	70.20
glm4-chat-9b	1-Shot	34.71	9.97	31.76	38.53	40.48	50.41	51.35	38.65	39.37	85.13	85.53
	LoRA	38.78	10.66	34.00	42.36	42.89	59.23	58.51	44.50	45.89	84.96	85.73
Baichuan2-13B	1-Shot	24.77	2.05	15.83	29.67	15.49	38.54	31.02	29.88	23.22	77.16	73.50
	LoRA	32.51	13.84	28.25	51.15	52.58	60.69	58.07	45.46	44.08	74.60	74.56
ChatGPT3.5-turbo	1-Shot	30.93	8.08	20.99	39.29	31.46	51.54	46.10	47.36	40.82	82.16	81.46
	CoT	30.32	3.63	22.29	26.51	19.10	38.65	34.18	29.47	25.91	72.36	70.89
	DAC	24.74	8.07	25.05	28.82	35.34	40.40	49.64	29.88	42.37	81.17	82.35
	Ours	34.60	8.23	23.29	39.38	39.02	51.54	47.19	47.04	45.05	82.27	83.31
GPT4o-mini	1-Shot	31.00	10.10	28.92	34.61	35.82	53.28	53.61	44.91	45.45	84.39	84.56
	CoT	33.79	7.06	31.28	27.63	26.31	48.95	48.67	38.66	37.55	78.63	79.03
	DAC	26.95	10.17	24.90	23.18	24.26	52.33	51.96	40.42	39.86	80.83	81.13
	Ours	34.78	11.13	34.32	42.67	43.28	55.95	55.97	**49.15**	**47.94**	86.13	86.06
DeepSeek-V3	1-Shot	37.28	14.17	35.17	51.75	53.37	62.08	62.13	41.32	40.35	86.04	86.27
	CoT	36.17	11.99	34.78	42.57	41.06	54.24	60.19	38.87	36.33	83.44	83.52
	DAC	37.53	18.76	34.99	46.19	41.58	60.93	60.50	39.84	37.16	85.77	85.99
	Ours	**38.82**	**19.89**	**36.59**	**55.96**	**56.74**	**62.85**	**62.79**	45.22	41.54	**88.25**	**87.37**
Other Method												
Human	-	94.11	84.31	95.31	96.08	97.16	97.64	97.44	97.27	97.41	98.03	98.19

Llama (Touvron et al., 2023), glm (GLM et al., 2024), and Baichuan (Yang et al., 2023) under 1-shot and LoRA settings. For large LLMs, we compare the 1-shot method, CoT (Wei et al., 2022), and DAC (Zhang et al., 2024) methods. In other methods, we mainly consider manual answering methods. Specifically, we randomly select 100 samples from MSDiagnosis and invite one college student (different from the annotation team in Section 3.2.2) to answer the questions. The specific settings are shown in the Appendix A.

6.1.2 Evaluation Details

For different types of problems, we introduce various evaluation metrics. Specifically, for primary diagnosis, differential diagnosis, and final diagnosis, we introduce entity $F1$ (Liu et al., 2022) as the evaluation metric. For primary and final diagnosis criteria, we employ two types of evaluation metrics. First, we adapt $Rouge-L$ (Lin, 2004), $BERTScore$ (Zhang* et al., 2020), and $BLEU$-1 (Papineni et al., 2002) to measure the similarity between the generated text and the reference text. Second, we introduce the $Macro-Recall$ metric, which is calculated based on the key points in the answers. More details are shown in Appendix B.

6.1.3 Implementation Details

For all the open-source models mentioned above, we use their default hyperparameters. All experiments are conducted three times, and the average performance across three runs is computed. Further implementation details are listed in Appendix C.

6.2 Main Results

To verify the effectiveness of our proposed method, we compare it with all baselines on the MSDiagnosis test set. The results are reported in Table 3.

From the table, we conclude that: 1) All LLMs perform poorly on MSDiagnosis, with a significant gap from the human final diagnosis $F1$ of 95.31%. 2) Our framework outperforms all baseline models across multiple metrics, demonstrating the effectiveness of the proposed approach. Specifically, under the DeepSeek-V3 model, our method outperforms the 1-shot method by 1.42% on the final $F1$ metric, the CoT method by 1.81%, and the DAC method by 1.6%. 3) In general small LLMs, the glm4 and Baichuan2 models outperform those without instruction fine-tuning in diagnostic reasoning after fine-tuning. However, the LlaMA's performance declines after fine-tuning, with primary $F1$ and final $F1$ metrics decreasing by 12.13% and 3.67%, respectively. This decline likely results

Table 4: The ablation results of our framework (%). *w/o* indicates the removal of the corresponding module.

Method	F1			Macro-Recall		Rouge-L		BLEU-1		BERTScore	
	Pre	DD	Fin	Pre	Fin	Pre	Fin	Pre	Fin	Pre	Fin
w/o Backward inference	33.59	10.96	32.04	41.32	41.01	55.48	55.48	48.93	47.32	85.83	84.76
w/o Reflection	33.68	10.04	32.19	41.08	41.60	55.53	55.89	49.16	47.87	85.93	85.93
w/o Refinement	34.03	10.81	32.92	41.95	42.15	55.63	55.39	49.06	47.72	85.16	84.83
Ours	**34.78**	**11.13**	**34.32**	**42.67**	**43.28**	**55.95**	**55.97**	**49.15**	**47.94**	**86.13**	**86.06**

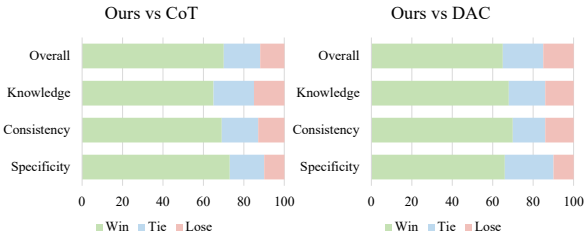

Figure 4: Human evaluation results of baseline methods. Red is for Win, yellow for Tie, and green for Lose.

from the relatively small proportion of Chinese in LlaMA's training corpus. Using Chinese data for instruction fine-tuning could further degrade its performance. 4) The model performs worse in differential diagnosis than in primary and final diagnoses. This is due to the increased complexity of the differential diagnosis task, which requires distinguishing between similar diseases and demands more specialized medical knowledge.

6.3 Human Evaluation

We conduct human evaluations to compare our framework with two baselines: CoT and DAC. We randomly select 100 samples and ask three medical students to assess them based on Knowledge, Consistency, and Specificity. Additionally, we compare the overall quality, and our framework outperforms both baselines in all aspects, as shown in Fig. 4.

6.4 Detailed Analysis

6.4.1 Ablation Study

In this section, we analyze the effectiveness of each module in the framework through experiments on the GPT4o-mini. Specifically, we sequentially remove backward inference, reflection, and refinement, assessing the effectiveness of the remaining components. The results are shown in Table 4.

From the Table 4, it can be observed that when backward reasoning is removed, the $F1$ and $Macro - Recall$ scores for the final diagnosis decrease significantly, specifically by 2.28% and 2.27%, respectively. These findings suggest that backward reasoning effectively enhances both the accuracy and interpretability of the final diagnosis.

6.4.2 The Impact of Different Parameters of the Model on the Framework

To evaluate the impact of models with varying parameter scales on the framework's performance, we evaluate Qwen2.5 models with 7B, 14B, 32B, and 72B parameters, and also compare with the 1-shot method for analysis convenience. The detailed experimental results are presented in Table 5. From the table, the following conclusions emerge: as the number of parameters increases, the framework's performance improves progressively. Specifically, the 32B model outperforms the 7B model in the F1 and Macro-Recall metrics for final diagnosis by 2.18% and 6.96%, respectively. In comparison to the 14B model, the 32B model achieves improvements of 1.26% and 1.12% in these metrics.

6.4.3 Error Analysis

To further investigate the limitations of our method, we analyze 100 error samples. After manual classification, error types are divided into three types and other errors: (a) Lack of domain knowledge (52%): The model lacks the specialized medical knowledge for accurate clinical diagnosis. (b) Deficiencies in medical record extraction (30%): The model fails to correctly identify or extract key information from the patient's medical history. (c) Inconsistent diagnostic criteria (18%): In the reasoning process, facts that contradict the patient's condition may occur. We observe that lack of domain knowledge errors are the most frequent. To illustrate these error types more intuitively, we provide examples as shown in Fig. 5.

7 Related Work

7.1 Clinical Diagnosis on EMRs

Clinic diagnosis based on EMRs is crucial for improving healthcare quality and patient outcomes. However, existing datasets for EMR-based diagnosis primarily focus on single-step diagnosis, with diagnostic types generally divided into two cate-

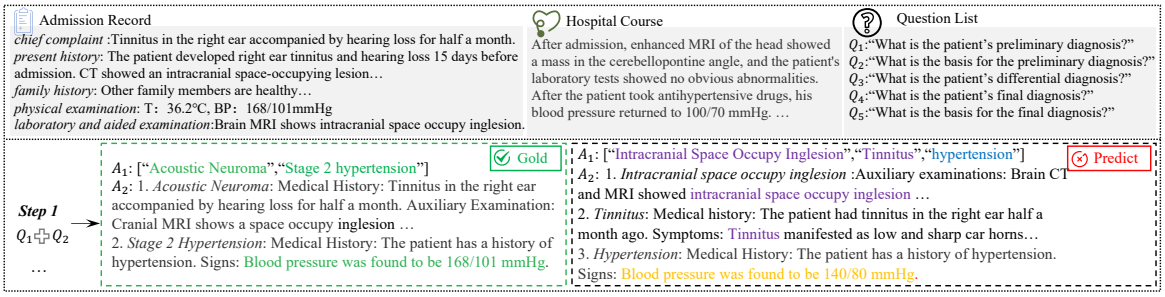

Figure 5: Error Case. The green(red) highlight indicates correct(incorrect) results. Purple marks lack of domain knowledge, blue marks deficiencies in medical record extraction, and yellow marks inconsistent diagnostic criteria.

Table 5: Comparison of models with different parameter sizes on our framework.

Model	Method	F1			Macro-Recall		Rouge-L		BLEU-1		BERTScore	
		Pre	DD	Fin	Pre	Fin	Pre	Fin	Pre	Fin	Pre	Fin
Qwen2.5-7B-Instruct	1-Shot	34.65	10.56	29.97	34.90	36.07	53.19	54.41	41.73	42.26	86.67	87.21
	Ours	36.97	9.01	33.35	39.39	40.23	58.60	59.81	50.32	49.25	87.61	87.57
Qwen2.5-14B-Instruct	1-Shot	36.88	11.72	33.82	43.12	43.18	60.27	60.61	53.06	50.48	88.68	89.84
	Ours	36.95	12.41	34.27	44.31	46.07	63.07	61.04	53.17	52.01	89.03	89.56
Qwen2.5-32B-Instruct	1-Shot	37.43	11.92	33.40	43.46	45.82	52.99	51.16	38.07	36.45	85.72	85.61
	Ours	38.27	11.79	35.53	45.24	47.19	62.83	63.38	55.43	52.64	88.37	88.48
Qwen2.5-72B-Instruct	1-Shot	36.87	10.56	34.11	47.41	50.01	59.34	60.11	52.51	52.24	87.27	87.83
	Ours	38.76	10.56	34.97	49.24	51.09	64.71	65.10	54.14	53.64	87.52	88.83

gories: primary diagnosis and differential diagnosis. Primary diagnosis involves determining the likely disease based on the patient's history and symptoms, or in combination with examination results. Most current research focuses on extracting abnormal features from medical records or integrating external medical knowledge to make a diagnosis (Xu et al., 2024; Jia et al., 2024; Zhu et al., 2024). Differential diagnosis is a list of potential diseases that could cause the patient's symptoms (Adler-Milstein et al., 2021). It enables a more comprehensive evaluation of clinical EMRs, allowing for the identification of less obvious but critical conditions. Current studies often employ deep learning methods to extract features from medical records for differential diagnosis (Zhou et al., 2024; Wu et al., 2023). In summary, previous research on EMR-based diagnosis has primarily focused on single-step diagnostic processes. However, this approach does not align with actual clinical diagnostic workflows.

7.2 Prompting Strategies of LLM

With the development of LLMs, many researchers have applied these models to medical tasks. These methods can be categorized into three types: IO prompting, CoT prompting, and the DAC paradigm. IO prompting (Yao et al., 2024) is a standard prompting strategy where input is combined with instructions and/or a few ICLs to generate a response. CoT prompting (Wei et al., 2022) aims to emulate the step-by-step thought process humans use to tackle complex tasks, such as combinatorial reasoning and mathematical calculations. There are also various CoT variants, such as CoT with self-consistency (CoT-SC) prompting (Wang et al., 2022), designed to address the limitations of CoT in exploration. The DAC paradigm (Zhang et al., 2024) mainly refers to simply breaking down the input sequence into multiple sub-inputs to enhance LLM performance on certain specific tasks. While these methods show promise in reasoning tasks, they are challenging to apply directly to medical diagnostic reasoning.

8 Conclusion

This paper introduces MSDiagnosis, a multi-step clinical diagnostic benchmark that is collected and annotated from open source medical websites. MSDiagnosis addresses the limitations of existing datasets by constructing multi-step diagnostic tasks that better align with actual clinical diagnostic scenarios. For this benchmark, we propose a simple and effective framework. We implement both medical and general LLMs and conduct extensive experiments. The results show that tasks in our benchmark effectively measure the multi-step clinical diagnostic abilities, and the framework proposed in this paper shows effectiveness on this benchmark.

Limitations and Future Work

While MSDiagnosis offers a valuable benchmark for evaluating current LLMs, it has two main limitations: Firstly, due to limited data sources, our medical records dataset exhibits an uneven distribution across different departments. This issue can be addressed through machine learning methods (Cao et al., 2019) and data sampling strategies (Chawla et al., 2002). Secondly, the proposed framework requires multiple uses of the LLM, resulting in a longer inference time for multi-step diagnostic reasoning compared to direct reasoning.

In future work, we will first address the imbalance across departments by acquiring additional medical records from the relevant departments to ensure a more balanced distribution. Secondly, while the framework proposed in this paper has improved diagnostic performance to some extent, there remains a gap when compared to real-world clinical applications. Future research can further enhance this by focusing on the following areas: 1) Training the model with multi-step diagnostic data to enable it to accurately differentiate the details of various diagnostic tasks, thereby improving clinical diagnostic capabilities; 2) Incorporating a more comprehensive medical corpus and utilizing retrieval techniques to enhance the model's reasoning ability.

Ethical Statement

The raw data used in this study comes from an open-source medical platform, which includes a large volume of EMRs. The platform permits data for research and education, as confirmed by prior studies (Yim et al., 2024a; Li et al., 2024; Yim et al., 2024b). Throughout the data collection process, we strictly ensured the avoidance of copyright and privacy issues. Furthermore, we conducted a thorough review of the dataset to ensure it contains no harmful content, such as gender bias, racial discrimination, or other inappropriate materials.

References

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

Julia Adler-Milstein, Jonathan H Chen, and Gurpreet Dhaliwal. 2021. Next-generation artificial intelligence for diagnosis: from predicting diagnostic labels to "wayfinding". *Jama*, 326(24):2467–2468.

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

John R Ball, Bryan T Miller, and Erin P Balogh. 2015. Improving diagnosis in health care.

Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. 1999. Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27(1):3–23.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Song Dingjie, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. HuatuoGPT-II, one-stage training for medical adaption of LLMs. In *First Conference on Language Modeling*.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, Xiangmin Xu, et al. 2023. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:2310.15896*.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

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

Maria Herrero-Zazo, Tomas Fitzgerald, Vince Taylor, Helen Street, Afzal N Chaudhry, John Bradley, Ewan

Birney, and Victoria L Keevil. 2021. Big data analysis of electronic health records: Clinically interpretable representations of older adult inpatient trajectories using time-series numerical data and hidden markov models. *medRxiv*, pages 2021–06.

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

Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. 2024. medikal: Integrating knowledge graphs as assistants of llms for enhanced clinical diagnosis on emrs. *arXiv preprint arXiv:2406.14326*.

Lei Li, Xiangxu Zhang, Xiao Zhou, and Zheng Liu. 2024. Automir: Effective zero-shot medical information retrieval without relevance labels. *arXiv preprint arXiv:2410.20050*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024b. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36.

Mianxin Liu, Weiguo Hu, Jinru Ding, Jie Xu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, Pengfei Liu, Xiaofan Zhang, Shanshan Wang, Kang Li, Haofen Wang, Tong Ruan, Xuanjing Huang, Xin Sun, and Shaoting Zhang. 2024c. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *Big Data Mining and Analytics*.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. Meddg: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *Natural Language Processing and Chinese Computing*, pages 447–459, Cham. Springer International Publishing.

Shiwei Lyu, Chenfei Chi, Hongbo Cai, Lei Shi, Xiaoyan Yang, Lei Liu, Xiang Chen, Deng Zhao, Zhiqiang Zhang, Xianguo Lyu, Ming Zhang, Fangzhou Li, Xiaowei Ma, Yue Shen, Jinjie Gu, Wei Xue, and Yiran Huang. 2023. Rjua-qa: A comprehensive qa dataset for urology. *Preprint*, arXiv:2312.09785.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *arXiv preprint arXiv:2402.13963*.

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

Jennifer Tiffen, Susan J. Corbridge, and Lynda Slimmer. 2014. Enhancing clinical decision making: Development of a contiguous definition and conceptual framework. *Journal of Professional Nursing*, 30(5):399–405.

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

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024. CMB: A comprehensive medical benchmark in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205, Mexico City, Mexico. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Lin Wu, Liying Huang, Mei Li, Zhaojun Xiong, Dinghui Liu, Yong Liu, Suzhen Liang, Hua Liang, Zifeng Liu, Xiaoxian Qian, et al. 2023. Differential diagnosis of secondary hypertension based on deep learning. *Artificial Intelligence in Medicine*, 141:102554.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Shaoting Zhang Xiaofan Zhang, Kui Xue. 2023. Pulse: Pretrained and unified language service engine.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May D Wang, Joyce C Ho, and Carl Yang. 2024. Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records. *arXiv preprint arXiv:2403.00815*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen-Yildiz, and Martin Krallinger. 2024a. Overview of the mediqa-m3g 2024 shared task on multilingual multimodal medical answer generation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 581–589.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2024b. Dermavqa: A multilingual visual question answering dataset for dermatology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 209–219. Springer.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhou Zhang, Lun Du, Defu Cao, Qiang Fu, and Yan Liu. 2024. An examination on the effectiveness of divide-and-conquer prompting in large language models. *Preprint*, arXiv:2402.05359.

Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. 2024. Efficiently democratizing medical llms for 50 languages via a mixture of language family experts. *arXiv preprint arXiv:2410.10626*.

Shuang Zhou, Sirui Ding, Jiashuo Wang, Mingquan Lin, Genevieve B Melton, and Rui Zhang. 2024. Interpretable differential diagnosis with dual-inference large language models. *arXiv preprint arXiv:2407.07330*.

Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, et al. 2024. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models. *arXiv preprint arXiv:2402.07016*.

Appendix

Table of Contents

A Details of the Baseline

In this paper, we mainly describe several types of baseline methods, including medical LLMs, general LLMs, closed-source LLMs, and other method.

In the medical LLMs, we employ MMedLM (Qiu et al., 2024), PULSE (Xiaofan Zhang, 2023), Lmama3-OpenBioLLM (Ankit Pal, 2024), Llama3-OpenBioLLM (Ankit Pal, 2024), and Apollo2-7B (Zheng et al., 2024) for comparison. Based on these models, we manually construct an example to serve as ICL. The complete example is shown in Section J. We previously tested several medical models (HuatuoGPT2-7B (Chen et al., 2024), DoctorGLM (Xiong et al., 2023), and BianQue-2 (Chen et al., 2023)), but they were excluded due to poor performance, as the length of our dataset approached their context window limits. According to our statistics, in the diagnostic tasks of this study, the average number of context tokens for the primary diagnosis is 3245.14, for the differential diagnosis is 3336.75, and for the final diagnosis is 4807.20.

In the general LLMs, we categorize them into two types: small LLMs (<20B) and large LLMs (>20B). For small LLMs, we use Llama-3.1 (Touvron et al., 2023), glm4 (GLM et al., 2024), and Baichuan2 (Yang et al., 2023) for comparison. Based on these models, we design two settings: 1-shot reasoning and instruction tuning. In the first setting, we utilize the previously constructed example as ICL. In the second setting, we use parameter-efficient fine-tuning with LoRA (Hu et al., 2021). The instruction data is generated by transforming the input and output from the training data.

For large LLMs, we use ChatGPT3.5-turbo[2] , DeepSeek-V3 (Liu et al., 2024a), and GPT4o-mini (Achiam et al., 2023) for comparison. We are unable to include this for all MSDiagnosis benchmarks due to the extraordinarily high cost of o1-preview inference. Based on these models, we employ two settings: 1-shot and prompting methods. In the first approach, we utilize the same example as previously used. In the second setting, we consider comparing method COT (Wei et al., 2022) and method DAC (Zhang et al., 2024). COT (Wei et al., 2022) uses "Let's think step by step" to enhance the model's reasoning ability for task-solving. DAC (Zhang et al., 2024) adopts a simple divide-and-conquer prompting strategy, where the input sequence is simply divided into multiple sub-inputs, which can enhance the reasoning performance of the LLM.

In other method, we mainly consider manual answering methods. Specifically, we randomly select 100 samples from MSDiagnosis and invite one college student (different from the annotation team in Section 3.2.2) to answer the questions. Our evaluation method for human results is consistent with other baselines, utilizing the automated assessment approach described in Section B.

B Details of the Evaluation Metrics

For the disease entities in the diagnosis results D and the reference results R in the medical records, we employ a more rigorous evaluation method. This approach aligns better with actual clinical needs and facilitates targeted treatment for patients. Specifically, we compute the edit distance to associate these entities with ICD-10 terms, thereby

[2] https://platform.openai.com/docs/models/gpt-3-5-turbo

Table 6: The impact of different departments on our framework. Green highlights indicate that our framework performs well in processing medical records for the respective department, while red highlights indicate poorer performance in that department.

Department	F1			Macro-Recall		Rouge-L		BLEU-1		BERTScore	
	Pre	DD	Fin	Pre	Fin	Pre	Fin	Pre	Fin	Pre	Fin
Pediatrics	37.58	10.16	35.84	49.54	47.35	52.03	52.27	45.75	46.80	84.56	84.19
Internal Medicine	27.09	14.30	22.10	39.08	36.63	51.10	52.14	47.35	45.95	84.74	84.38
Emergency	32.14	29.64	32.14	44.91	42.88	61.09	57.71	53.15	58.51	87.66	88.06
Nursing Department	54.60	11.11	45.71	29.16	36.66	62.24	63.70	51.34	43.09	85.87	85.76
Surgery	40.68	10.40	39.81	43.21	47.65	60.93	61.17	52.25	50.93	87.40	87.00
Obstetrics-Gynecology	37.14	22.44	23.08	32.00	29.41	49.97	47.51	49.65	40.76	85.34	84.02
Otolaryngology (ENT)	44.82	5.64	41.83	44.42	50.28	63.23	63.60	49.37	47.78	87.53	87.53
Psychiatry	33.33	0.00	22.22	19.97	31.97	69.46	50.77	43.16	46.91	84.30	83.45
Oncology	7.63	2.57	5.89	24.67	25.85	39.83	40.80	40.90	42.48	82.34	83.06
Ophtalmology	20.83	7.14	20.83	44.01	36.46	55.66	53.64	47.56	48.08	86.25	86.92
Dermatology	83.33	0.00	75.00	38.54	48.96	33.93	67.86	52.08	45.76	86.65	88.36
Stomatology	75.00	8.33	83.33	48.87	62.07	74.22	85.90	56.57	64.02	88.50	93.03
Traditional Chinese Medicine	0.00	40.0	0.00	61.88	43.75	83.33	53.33	58.95	42.38	88.63	83.80
Others	38.76	10.97	23.82	37.87	36.34	62.97	61.62	50.66	47.16	87.07	85.68

mapping D and R to two standardized disease sets, S_d and S_r, respectively. We then compute the entity $F1$ score based on S_d and S_r.

For primary and final diagnosis criteria, we employ two types of evaluation metrics. First, we adapt $Rouge-L$ (Lin, 2004), $BERTScore$ (Zhang* et al., 2020), and $BLEU$-1 (Papineni et al., 2002) to measure the longest matching sequence between the generated text and the reference text, capturing the similarity in the overall structure of the two sequences. Second, we introduce the $Macro-Recall$ metric, which is calculated based on the key points in the answers. For our predefined N key point categories, the calculation method for key point $Macro-Recall$ is as follows:

$$Marco-Recall = \frac{1}{N}\sum_{i=1}^{N} Recall_i, \quad (1)$$

where $Recall_i$ represents the recall rate of the i category.

C Details of Implementation

To enhance the stability and reliability of the experimental results and reduce the impact of random factors, we conduct each experiment three times and then calculate the average of three results. For the backbone model, we utilize the OpenAI API, specifying the model as "GPT4o-mini". We set the top_p parameter to 0.01, and all other hyperparameters of the OpenAI API are maintained at default values. When selecting similar examples, we set K to 1. During the LoRA fine-tuning phase, we employ the Adam optimizer with weight decay correction for fine-tuning the model. The initial learning rate is set to 1e-4, and the batch size is set to 1, with Cross-Entropy Loss serving as the loss function. The input during training is: "instruction: task description. Medical Record: patient case. question: diagnosis question". The output is: "the answer to the question". We calculate the loss only for the answer, excluding the instruction and input. For the open-source models, our experiments are conducted on four Nvidia A100 GPUs, each with 40GB of memory, and we use PyTorch[3] in Python[4].

Table 7: The definition of question

Question	Definition
Q_1	Inquire about the patient's primary diagnosis.
Q_2	Inquire about the patient's primary diagnostic criteria.
Q_3	Inquire about the patient's differential diagnosis.
Q_4	Inquire about the patient's final diagnosis.
Q_5	Inquire about the patient's final diagnostic criteria.

D Question Definition

In this section, we introduce in detail the definition of the diagnostic questions corresponding to each EMR. In MSDiagnosis, the questions corresponding to each case are initially constructed manually and then expanded using GPT-4. The definitions of

[3] https://pytorch.org/
[4] https://www.python.org/

each question are shown in Table 7.

E The Impact of Different Departments on the Framework

In this section, to evaluate the impact of different departments on our framework, we conduct an analysis of the department-specific results. The detailed experimental outcomes are presented in Table 6. From this analysis, we can draw the following conclusions: Different departments have different influences on the framework. The Oncology have the greatest impact on our framework. Specifically, the final $F1$ score in the Oncology department is only 5.89%. The best performance is achieved in the Stomatology department. Specifically, the final $F1$ score in the Oncology department is only 83.33%.

F The Impact of the Number of Diagnosed Diseases on the Framework

In this section, we primarily analyze the impact of the number of diseases on the framework using the GPT4o-mini model. In this experiment, we categorize the number of diseases into three levels: 1 to 5 diseases, 5 to 10 diseases, and more than 10 diseases. The specific experimental results are shown in Table 8. From the results, we observe that as the number of diseases increases, the performance of the framework gradually declines. Specifically, when comparing patients with 5 to 10 diseases to those with 1 to 5 diseases, the final $F1$ score decreases by 6.04%. However, when comparing patients with more than 10 diseases to those with 5 to 10 diseases, the final $F1$ score improves by 2.72%. This improvement might be due to the fact that there are fewer patients with more than 10 diseases. Our analysis shows that there are only 5 such patients.

G The Impact of the Number of ICL on the Framework

In this section, to analyze the impact of the number of ICL examples on the method's performance, we introduced varying numbers of ICL examples into the framework for comparison. In this experiment, when the number of ICL examples reaches 4, the context length exceeds the model's token limit, so we compared experiments with fewer than 4 ICL examples. The specific experimental results are shown in Table 9. As observed from the table, the performance of the framework gradually improves as the number of ICL examples increases.

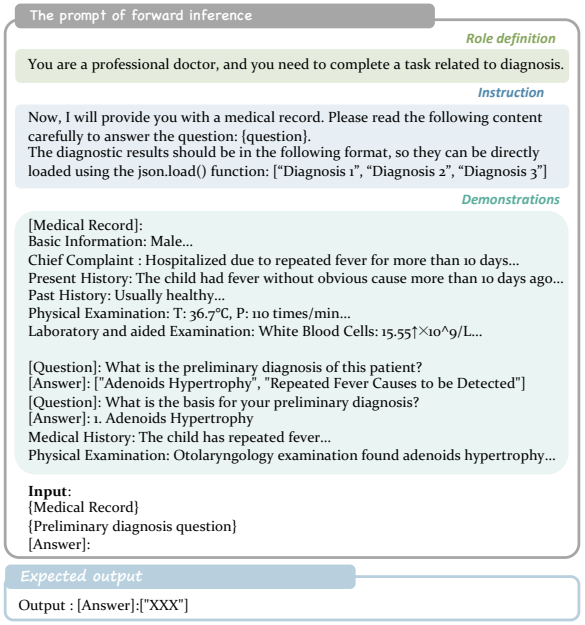

Figure 6: The prompt of forward inference.

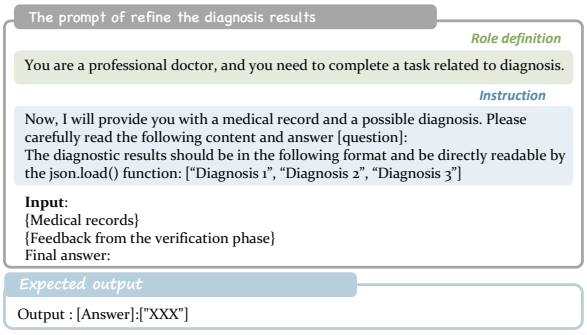

Figure 7: The prompt of refine the diagnosis results.

H Ablation analysis of a multi-step diagnostic process

In this section, we design an experiment to analyze the multi-step diagnostic process of MSDiagnosis. Specifically, for the ablation of the multi-step diagnostic process, we sequentially remove the primary diagnosis, differential diagnosis, and both to evaluate the effectiveness of the final diagnosis. Notably, the LLM is used for direct reasoning in this experiment. The results are shown in Table 10.

From the Table 10, we conclude that: 1) The primary diagnosis significantly influences the final diagnosis's performance, likely due to errors in the primary diagnosis propagating through the multi-step process, reducing the final diagnosis's effectiveness. 2) The multi-step diagnostic process helps improve the interpretability of the final diagnosis. Specifically, by introducing a multi-step diagnosis, the final diagnosis is based on more cri-

Table 8: The impact of the number of diagnosed diseases (final diagnosis) on the framework.

Number of Disease	F1			Macro-Recall		Rouge-L		BLEU-1		BERTScore	
	Pre	DD	Fin	Pre	Fin	Pre	Fin	Pre	Fin	Pre	Fin
1-5	35.93	11.63	32.50	42.52	45.73	57.33	58.4	50.17	48.62	86.32	86.23
5-10	30.03	9.77	26.46	31.73	22.47	49.76	45.92	45.39	45.16	84.58	84.75
>10	34.27	9.35	29.18	17.94	22.57	48.82	50.19	45.95	42.86	84.66	86.17

Table 9: Experimental results with different numbers of ICL examples.

Model	F1			Macro-Recall		Rouge-L		BLEU-1		BERTScore	
	Pre	DD	Fin	Pre	Fin	Pre	Fin	Pre	Fin	Pre	Fin
ICL=0	34.70	10.09	31.91	34.58	31.91	54.25	53.31	47.04	44.66	84.30	84.45
ICL=1	34.78	10.97	31.69	40.67	42.28	55.95	55.97	49.15	47.94	85.19	85.72
ICL=2	34.76	10.72	31.53	41.60	43.88	56.92	55.28	50.42	48.72	86.45	86.04
ICL=3	35.37	10.73	31.69	42.45	43.08	54.16	54.94	50.49	49.31	87.19	87.72

The prompt of reflection on diagnostic result

Role definition

You are a professional doctor, and you need to complete a task related to diagnosis.

Instruction

I will now provide you with a medical record, diagnostic results, and relevant disease characteristics. Your task is to review the diagnostic results against the medical record and relevant disease characteristics according to the following criteria:
Rule 1. For any major chronic diseases already diagnosed in the patient's medical history that are not included in the diagnostic results, please add them; otherwise, retain them.
Rule 2. If you find any significant inconsistencies between the disease characteristics of a diagnosis and the medical record, please remove or correct the diagnosis and provide the basis for its exclusion.
Rule 3. Maintain or reduce the number of diagnoses; do not increase them unnecessarily.
Rule 4. The final returned results should be in the following format, with no additional content:
• Diagnostic Results: ["Diagnosis 1", "Diagnosis 2", "Diagnosis 3"]
• Basis for Diagnosis:
• Excluded Diagnoses: <Names of excluded diseases, delete this item if not applicable>
• Basis for Exclusion: <Your basis for excluding the diagnoses, delete this item if not applicable>

Input:
{Medical history}
{Initial answer}
{Related disease characteristics}
Diagnosis result:
Diagnosis basis:
Excluded diagnosis:
Exclusion basis:

Expected output

Output : [Answer]:"XXX"

Figure 8: The prompt of reflection on diagnostic result.

teria, thereby enhancing its interpretability. This is primarily because the multi-step diagnostic process, by breaking down tasks, conducting step-by-step reasoning, and increasing traceability, allows for a clear demonstration of the reasoning path and basis at each stage, thereby enhancing the interpretability of the diagnostic process.

I The Prompt used in our Framework

In this section, we primarily introduce the prompts corresponding to the four instances where the LLM is used within the framework. It is noteworthy that the prompts in our framework follow a general pattern, including role definition, formatting instructions, and examples, rather than being meticulously designed. The prompt for forward reasoning is shown in Fig. 6. The prompt for backward reasoning is shown in Fig. 9. The prompt for reflecting on the diagnostic results is shown in Fig. 8. The prompt for optimizing the diagnostic results is shown in Fig. 7.

The prompt of backward inference from diagnosis to diagnostic criteria

Role definition

You are a professional doctor, and you need to complete a task related to diagnosis.

Instruction

Now, I will provide you with a diagnostic result that includes several disease. Please complete the following tasks based on this diagnostic result:
Rule 1. For each diagnosis in the diagnostic result, recall the representative medical history, symptoms, physical signs, and auxiliary examination results for the disease.
Rule 2. The recalled content should be in the following format:
Medical History: <Recall the representative medical history for the disease; delete this item if not applicable>
Symptoms: <Recall the representative symptoms for the disease; delete this item if not applicable>
Physical Signs: <Recall the representative physical signs for the disease; delete this item if not applicable>
Auxiliary Examination Results: <Recall the representative auxiliary examination results for the disease; delete this item if not applicable>

Demonstrations

Diagnostic Results: ["Acute Pancreatitis", "Fatty Liver", "Right Renal Cyst"]
Recall:
1. Acute Pancreatitis
• Medical History: Often associated with gallstones, alcohol abuse, or hyperlipidaemia
• Symptoms: Severe upper abdominal pain radiating to the back, nausea, vomiting, fever
• Physical Signs: Abdominal tenderness, muscle guarding, possible jaundice
• Auxiliary Examination Results: Elevated serum amylase and lipase, abdominal ultrasound or CT showing pancreatic enlargement or fluid collection
2. Fatty Liver
• Medical History: Common in patients with obesity, hyperlipidaemia, or diabetes
• Symptoms: Often asymptomatic, some patients may experience fatigue, upper abdominal discomfort
• Physical Signs: Hepatomegaly, palpable enlarged liver
• Auxiliary Examination Results: Liver function tests may be normal or slightly abnormal, abdominal ultrasound showing hepatic steatosis
3. Right Renal Cyst
• Medical History: Typically asymptomatic, may be discovered incidentally during physical exams or imaging studies
• Symptoms: Usually asymptomatic, large cysts may cause flank pain or abdominal discomfort
• Physical Signs: Physical exam usually unremarkable, large cysts may be palpable
• Auxiliary Examination Results: Abdominal ultrasound, CT, or MRI showing cystic structures within the kidney filled with fluid

Input:
{Initial answer}
Disease characteristics:

Expected output

Output : [Answer]:"XXX"

Figure 9: The prompt of backward inference from diagnosis to diagnostic criteria.

Table 10: The ablation results of the multi-step diagnostic (%). *w/o* represents deleting the corresponding process.

Method	F1			Macro-Recall		Rouge-L		BLEU-1		BERTScore	
	Pre	DD	Fin	Pre	Fin	Pre	Fin	Pre	Fin	Pre	Fin
Primary diagnosis matches final diagnosis											
w/o primary diagnosis	-	8.05	44.46	-	42.66	-	59.53	-	52.30	-	86.40
w/o differential diagnosis	40.8	-	41.28	35.96	42.10	57.14	58.67	48.08	50.62	85.55	86.08
multi-step	38.11	8.23	39.33	35.94	42.96	55.94	58.62	46.88	49.21	85.14	85.77
Primary diagnosis differs from final diagnosis											
w/o primary diagnosis	-	9.80	21.17	-	29.72	-	49.66	-	42.35	-	83.05
w/o differential diagnosis	25.19	-	18.78	32.70	30.08	50.28	49.04	43.69	41.91	83.62	83.03
multi-step	23.99	11.49	18.12	32.50	30.91	50.74	48.76	42.90	41.88	83.46	83.09

J Complete Multi-step Diagnostic Example

We present a comprehensive multi-step diagnostic example, as illustrated in Fig. 10. In the first stage, the input is the medical record introduction, and Q1 and Q2 are answered sequentially. The responses to these questions are A1 and A2. In the second stage, the input includes the medical record introduction and the Q&A history from the first stage, and Q2 must be addressed. In the third stage, the input consists of the medical record introduction, the Q&A history from the first two stages, and the patient's diagnosis and treatment process. In this stage, Q4 and Q5 must be answered sequentially.

"***Case Description***": {
"Basic Information": "Male, 14 years old, student",
"Chief Complaint": "Left shoulder injury from an accidental fall, lasting over two days",
"Present Illness History": "The patient reported injuring his left shoulder two days ago due to trauma, immediately experiencing localized pain and restricted movement. There was no accompanying limb numbness, headache, history of unconsciousness or vomiting, chest pain, or abdominal pain. No specific treatment was administered. The patient was urgently admitted to the hospital after an X-ray examination at the outpatient clinic, which indicated a left clavicle fracture. During the illness, the patient remained in good mental condition with normal bowel and urinary functions.",
"Past Medical History": "The patient reported good general health, with a history of mango allergy. There was no history of hypertension, diabetes, liver disease, or kidney disease. No drug allergies were noted. The patient denied recent overseas travel, exposure to epidemic areas, confirmed COVID-19 diagnosis, or contact with confirmed cases.",
"Physical Examination": "T: 36.6° C, P: 84 bpm, R: 19 bpm, BP: 106/67 mmHg. The patient was conscious, with no abnormalities in the head or face. Skin and sclera showed no signs of jaundice. Superficial lymph nodes were not swollen. The neck was supple, and tongue protrusion was midline. Pupils were equal, round, and reactive to light. Bilateral lung breath sounds were clear, with no dry or moist rales. Heart sounds were normal with regular rhythm, and no murmurs or abnormal heart sounds were detected in any valve area. The abdomen was soft, with no tenderness or rebound tenderness. \n Specialized examination: Passive posture, transported via wheelchair on admission, with a pained expression. Tenderness and restricted movement were noted in the left shoulder, with palpable crepitus. The spine displayed normal physiological curvature, with no tenderness. Lumbar spine movement was unrestricted. No numbness of the extremities was observed. Muscle strength and tone in other limbs were normal. Pathological signs were negative.",
"Auxiliary Examination": "X-ray indicated a left clavicle fracture; ECG showed sinus arrhythmia; CT scan revealed no significant abnormalities in both lungs; blood tests were essentially normal."}

"***Treatment Process***": "1. Conducted comprehensive examinations. 2. Provided symptomatic treatment. 3. Monitored disease progression. 4. Reported the condition to senior physicians and explained it to the patient and their family. Opted for surgical treatment. \n Anesthesia and emergency care process: Intravenous access was established in the operating room with routine monitoring of NIBP, SPO2, HR, and RR. Oxygen was administered via mask. Under ultrasound guidance, left brachial plexus and cervical plexus blocks were performed using 0.3% ropivacaine hydrochloride (30 ml, 20 ml for the brachial plexus, and 10 ml for the cervical plexus). The anesthesia was effective, with no pain at the needle insertion plane. The patient reported nervousness and desired sleep. The anesthesiologist administered 2 mg of midazolam and dexmedetomidine (4 μg/ml) at 5 ml/h via micro-pump infusion. After 10 minutes, the patient entered light sleep. The circulating nurse observed urticaria on the infusion arm but no other locations. The patient denied chest tightness, discomfort, or respiratory depression. Dexmedetomidine infusion was stopped immediately, and the infusion set was replaced. Dexamethasone (10 mg) and calcium gluconate (10 ml of 10% solution) were administered intravenously. Five minutes later, the urticaria resolved without further issues. Surgery commenced with no recurrence of urticaria. The patient's vital signs remained stable, and after a 10-minute observation period, the patient was transferred back to the ward for further monitoring and instructed to continue oxygen therapy. \n Surgical procedure: Following successful anesthesia, the patient was placed in the supine position. After routine disinfection and draping, an arcuate incision was made over the midsection of the left clavicle, centered at the fracture site. Layers of skin and subcutaneous tissue were incised, with care taken to protect nerves and blood vessels. The periosteum was incised and stripped to expose the fracture ends. Blood clots and soft tissue were cleared from the fracture site. The fracture was reduced and fixed using an 8-hole upper limb reconstruction plate secured with screws. After ensuring stable fixation, the surgical site was irrigated, hemostasis achieved, and the incision closed in layers with sterile dressing applied. Postoperative care included anti-inflammatory treatment (cefmetazole sodium 2 g IV daily), hemostasis (tranexamic acid 1 g IV daily), and symptomatic treatment. The patient was discharged in good condition on postoperative day 7."

"***Step 1: Primary Diagnosis***": {
"**Input**": **Case Description,**
"**QA Pairs**": [{
"Q1": "What is the preliminary diagnosis based on the symptoms and signs on admission?",
"A1": ["Left clavicle fracture"]
}, {
"Q2": "What clinical data supports the preliminary diagnosis?",
"A2": "1. Left clavicle fracture\n History: The patient reported injuring his left shoulder two days ago due to trauma, with immediate localized pain and restricted movement.\n Symptoms: Pain in the left shoulder and restricted movement.\n Signs: Tenderness in the left shoulder, restricted movement, and palpable crepitus.\n Auxiliary Examination: X-ray indicated a left clavicle fracture."
}]},
"***Step 2: Differential Diagnosis*** ": {
"**Input**": **Case Description + Q1 + A1 + Q2 + A2,**
"**QA Pairs**": [{
"Q3": "What are the possible differential diagnoses?",
"A3": ["Pathological fracture"]
}]},
"***Step 3: Final Diagnosis*** ": {
"**Input**": **Case Description + Q1 + A1 + Q2 + A2 + Q3 + A3 + Treatment Process,**
"**QA Pairs**": [{
"Q4": "What is the final diagnosis after a comprehensive examination?",
"A4": ["Left clavicle fracture", "Type I hypersensitivity reaction"]
}, {
"Q5": "What evidence supports the final diagnosis?",
"A5": "1. Left clavicle fracture\n History: The patient reported injuring his left shoulder two days ago due to trauma, with immediate localized pain and restricted movement.\n Symptoms: Pain in the left shoulder and restricted movement.\n Signs: Tenderness in the left shoulder, restricted movement, and palpable crepitus.\n Auxiliary Examination: X-ray indicated a left clavicle fracture.\n\n2. Type I hypersensitivity reaction\n History: Known mango allergy.\n Signs: Urticaria on the infusion arm during surgery.\n Auxiliary Examination: Urticaria resolved following anti-allergy treatment administered during surgery."
}]}

Figure 10: A complete multi-step diagnostic example.