Med-PMC: A Personalized Multi-modal Framework for Dynamic Clinical Interaction and Assessment of Large Language Models

Anonymous ACL submission

Abstract

The application of the Multi-modal Large Language Models (MLLMs) in medical clinical scenarios remains underexplored. Previous benchmarks only focus on the capacity of the MLLMs in medical visual question-answering (VQA) or report generation and fail to assess the performance of the MLLMs on complex clinical multi-modal tasks. In this paper, we propose a novel **Med**ical **P**ersonalized **M**ulti-modal **C**onsultation (Med-PMC) paradigm to evaluate the clinical capacity of the MLLMs. Med-PMC builds a simulated clinical environment where the MLLMs are required to interact with a patient simulator to complete the multi-modal information-gathering and decision-making task. Specifically, the patient simulator is decorated with personalized actors to simulate diverse patients in real scenarios. We conduct extensive experiments to access 12 types of MLLMs, providing a comprehensive view of the MLLMs' clinical performance. We found that current MLLMs fail to gather multimodal information and show potential bias in the decision-making task when consulted with the personalized patient simulators. Further analysis demonstrates the effectiveness of Med-PMC, showing the potential to guide the development of robust and reliable clinical MLLMs. Code and data will be released upon acceptance.

1 Introduction

The application of large language models (LLMs) in the medical field has garnered significant attention, with models such as GPT-4 (OpenAI, 2023a) and Med-Palm (Singhal et al., 2022) being deployed for a range of clinical tasks, including medical examinations (Jin et al., 2021; Pal et al., 2022), medical consultation (Liao et al., 2023; Tu et al., 2024), and decision support (Benary et al., 2023; Hager et al., 2024). Specifically, previous research has leveraged the role-playing capability of LLMs to function as patient simulators (Tu et al., 2024; Liao et al., 2024), creating simulated clinical environments to assess models' clinical performance. However, real-world clinical scenarios often involve multi-modal data, such as medical images (e.g., X-ray, CT) and biological signals (e.g., temperature, ECG), which are crucial for a comprehensive patient evaluation. The current limitation of LLMs, which predominantly process only textual information, restricts their ability to fully address the complexities of clinical settings.

While Multi-modal Large Language Models (MLLMs) are designed to handle multi-modal information, their use in medical clinical scenarios remains insufficiently explored. Existing benchmarks mainly focus on gathering information from medical images, such as visual question-answering (VQA)(Lau et al., 2018; Liu et al., 2021; He et al., 2021) and report generation(Thawkar et al., 2023; Hamamci et al., 2024). Other studies address more complex tasks, such as reasoning with multi-modal data (Li et al., 2024) or integrating multiple images (Wu et al., 2023). However, these tasks are typically conducted in static environments that fail to simulate the interactive dynamics inherent in real-world clinical interactions. As a result, these approaches do not provide a complete evaluation of MLLMs' performance and potential in actual clinical contexts.

In this paper, we propose a novel framework, **Med**ical **P**ersonalized **M**ulti-modal **C**onsultation (Med-PMC), designed to simulate a dynamic clinical environment and assess the performance of MLLMs in more realistic clinical scenarios. Med-PMC requires MLLMs to perform multi-turn decision-making based on initial patient information. Within a limited number of interactions, the model must efficiently gather multi-modal symptom data from the patient and ultimately provide diagnostic insights and treatment recommendations. To enhance the reliability of the clinical simulation, we introduce the *patient-actor* agent, which detects the type of actions taken by the doctor and generates corresponding patient responses. Basic information is directly extracted and responded to after being imbued with a unique identity and personality through the actor module. For examination results, relevant departments provide reports. The *patient-actor* agent records the decision-making process, ensuring the simulation's authenticity, while also reflecting the diversity of patient conditions. By incorporating *patient-actor* agents, Med-PMC provides a more accurate evaluation of MLLMs' ability to handle dynamic, real-world clinical interactions, ultimately improving their applicability and effectiveness in medical practice.

In summary, the contributions of this paper are as follows:

- **A Multi-modal Clinical Interactive Evaluation Framework**: We propose Med-PMC, a new evaluation framework that simulates a clinical environment to comprehensively assess MLLMs' performance in real-world medical scenarios. This framework overcomes the limitations of existing approaches by integrating multi-turn decision-making and multi-modal data.

- **Development of the Reliable *patient-actor* Agent**: We introduce the *patient-actor* agent to enhance the reliability of the clinical simulation. This agent interacts dynamically with MLLMs, detects the doctor's actions, and generates relevant responses based on the case. It also simulates patient diversity, providing more realistic and varied settings for evaluation.

- **Comprehensive Assessment of Clinical Capability**: By incorporating multi-turn reasoning, multi-modal data, and patient diversity, Med-PMC enables a more thorough assessment of MLLMs' ability to manage complex and dynamic clinical interactions. Extensive experiments with 12 MLLMs offer a comprehensive view of their clinical performance. Our results demonstrate the effectiveness of Med-PMC and guide the development of more robust and reliable MLLMs for clinical use.

2 Related Works

In the realm of medical language model development, significant strides have been made by prior works such as HuatuoGPT (Zhang et al., 2023) and Disc-medllm (Bao et al., 2023). HuatuoGPT employs a method that uses multi-turn consultation conversations generated by ChatGPT-like models for training data, whereas Disc-medllm simplifies the dialogue process into three distinct phases: information inquiry, preliminary diagnosis, and treatment suggestion. Similarly, other studies have proposed interaction frameworks to assess the clinical consultation performance of models by leveraging the role-playing capabilities of large language models (LLMs). Johri et al. (2024) and Liao et al. (2023) have demonstrated the effectiveness of such frameworks in evaluating model performance in clinical settings. Furthermore, Liao et al. (2024) employed human evaluation to validate the rationality of the interaction framework and proposed a set of action categories for doctor models during the consultation process, enhancing the interpretability of model behavior. In addition, Schmidgall et al. (2024) considered the integration of multimodal information during the consultation and diagnosis process. However, their approach does not fully evaluate the ability of multimodal large language models (MLLMs) to interpret medical images, which limits the comprehensiveness of their assessment.

State	Meanings
Begin	The start of the consultation
Effective Inquiry	Specific and relevant patient questions
Ineffective Inquiry	Specific but non-relevant patient questions
Ambiguous Inquiry	Board patient questions
Effective Advice	Specific and relevant examination questions
Ineffective Advice	Specific but non-relevant examination questions
Ambiguous Advice	Board examination questions
Other Topic	Other questions

Table 1: States definitions in the state detection stage.

3 Med-PMC Evaluation Framework

To close the gap between the evaluation methods and actual clinical scenarios, we proposed Med-PMC evaluation framework to build a simulated clinical environment. As shown in Figure 1, the MLLMs are required to interact with the personalized *patient-actor* agent and the multi-modal technician agent iteratively until it gathers enough information to make the final decision. Finally, MLLMs generate the diagnosis results and recommendations based on the gathered information.

3.1 Problem Formulation

The proposed Med-PMC evaluation framework builds a simulated clinical environment upon N

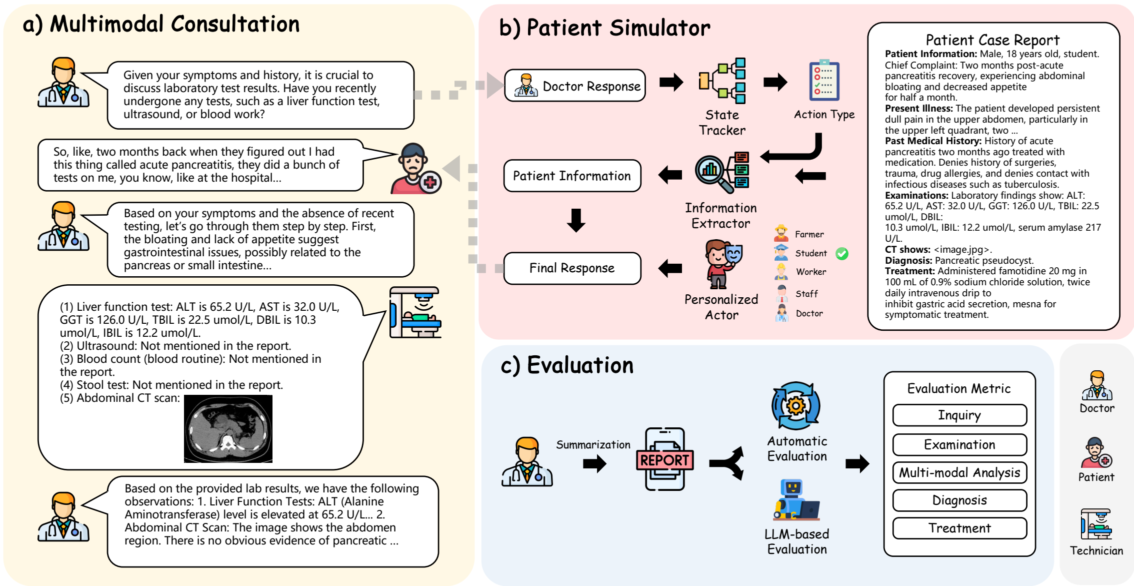


Figure 1: Overview of the Med-PMC evaluation framework. The whole framework can be divided into three parts, including a) Multi-modal consultation, b) Patient Simulator, and c) Evaluation.

realistic patient cases, which can be noted as $\mathcal{M} = \{\mathcal{I}, \boldsymbol{d}, \boldsymbol{t}\}^N$. $\mathcal{I}$ is the patient information, $\boldsymbol{d}$ is the diagnosis results, and $\boldsymbol{t}$ is the treatment recommendations. The patient information $\mathcal{I}$ contains three different types of multi-modal messages: he basic patient information, examination results, and medical image.

For each case, MLLMs are required to interact with the *patient-actor* agent for up to M rounds to gather patient information. In the k-th round interaction, the MLLMs θ_d output the sequence $\boldsymbol{y}_{d,k}$ based on the interaction history $\mathcal{H}_k$ (contain the multimodal information):

$$\mathcal{P}_{\theta_d}(\boldsymbol{y}_{d,k}|\mathcal{H}_k) = \prod_{j=1}^{|\boldsymbol{y}_{d,k}|} \mathcal{P}_{\theta_d}(y_{d,k}^j|y_{d,k}^{<j}, \mathcal{H}_k) \quad (1)$$

where the j denotes the jth token in the response. The *patient-actor* agent or technician θ_p responses $\boldsymbol{y}_{p,k}$ according to the patient information in an autoregressive manner:

$$\mathcal{P}_{\theta_p}(\boldsymbol{y}_{p,k}|\mathcal{I}, \mathcal{H}_k, \boldsymbol{y}_{d,k}) = \prod_{j=1}^{|\boldsymbol{y}_{p,k}|} \mathcal{P}_{\theta_p}(y_{p,k}^j|y_{p,k}^{<j}, \mathcal{I}, \mathcal{H}_k, \boldsymbol{y}_{d,k}) \quad (2)$$

The interaction history is updated as follows:

$$\mathcal{H}_k \cup (\boldsymbol{y}_{d,k}, \boldsymbol{y}_{p,k}) \rightarrow \mathcal{H}_{k+1} \quad (3)$$

Once the doctor MLLMs have gathered sufficient patient information with K turns consultation, they are required to generate a summarized report $\mathcal{R}$ and make the final decision on diagnosis and treatment:

$$\mathcal{R} = \theta_d(\mathcal{H}_K) \quad (4)$$

We will evaluate the performance of the MLLMs on clinical tasks by comparing the content of their reports to the ground truth diagnosis results $\boldsymbol{d}$ and treatment plans $\boldsymbol{t}$.

3.2 *Patient-Actor* Agent

To create a realistic and reliable clinical simulation environment, the *patient-actor* agent needs to exhibit appropriate behaviors and responses. In practice, patients generally engage in two types of behaviors: answering questions posed by the doctor, which typically involve providing basic personal information and describing symptoms, and completing specific examinations or tests as requested by the doctor. Therefore, the *patient-actor* agent should first classify the action of the doctor and then exhibit the expected behaviors, generating contextually relevant responses. Moreover, the patient agent also needs to have personalized characteristics, as clinical patients are diverse. Based on these requirements, we have designed the *patient-actor* agent to consist of three main components: a state tracker, a response generator, and a personalized actor.

3.2.1 State Tracker & State Definition

As previously discussed, the state tracker categorizes doctors' actions to guide subsequent behaviors from three perspective. (i) The state tracker first determines whether the action is **Specific** or **Broad**. If the doctor's language is clear and specific, the patient-actor can respond to the doctor's action. On the contrary, if the language is vague and broad, the patient-actor should require the doctor to ask a more specific question. (ii) Following the real-world scenarios, we classify the specific actions into two main categories: **Basic Information Inquiry** and **Medical Examination Recommendations**. This classification determines whether the patient agent directly responds to questions or undergoes the relevant examinations. (iii) Finally, we categorize actions based on whether they retrieve relevant patient information into **Relevant** and **Irrelevant** types. If the information or the suggested examination/test results requested by the doctor are present in the patient's information, the action is considered Relevant; otherwise, the action is Irrelevant. In summary, we have defined a total of eight types of doctor actions to enable the *patient-actor* agent to respond more reliably. The types of actions and their meanings are detailed in Table 1.

3.2.2 Response Generator

The response generator can produce corresponding replies based on the state tracker's classification of the doctor's actions. During experiments, we observed that some doctor models tend to find shortcuts to gather patient information by repeatedly asking vague questions like 'Is there any other information?'. For such a situation, the generator should respond with requiring more specific action. The response generator can directly respond with relevant patient information or obtain the examination results from the technician agent for Basic Information Inquiries and Medical Examination Recommendations, respectively. Finally, For Irrelevant actions that request unavailable information, the *patient-actor* agent should respond 'I don't know' to avoid fabricating false information.

Technician Agent The technician agent aims to respond to questions related to examinations, including blood tests and X-ray images. To provide a more appropriate result rather than the external result without asking, we divided this process into two steps, one for examination detection and another for result provision, which is shown in Figure 1. The result provision step needs two agents to complete the initial response based on the detected state and the final response based on the initial response. Similarly, we only provide the examination results for the technician agent and it can be activated only when the detected state is advice.

3.2.3 Personalized Actor

After obtaining the response information from the response generator, the Personalized Actor will rewrite the response, imbuing the patient agent with specific tones and personalities to simulate the diversity in how different patients express themselves in clinical settings. Specifically, we have set up 10 personas by combining two genders with five distinct professions: farmer, student, worker, office worker, and doctor. This allows us to model patients of varying ages, educational levels, socioeconomic statuses, living environments, and professional backgrounds, thereby capturing a wide range of patient expressions and behaviors.

4 Experiments

4.1 Experimental Settings

We conducted 1,200 evaluation experiments on the medical website[1], focusing mainly on the Department of General Surgery. For each case, patient information is divided into seven parts: Chief Complaint, Present Illness, Medical History, Examination, Diagnosis, Treatment, and Image Report. The ages of the patients range from 15 to 81 years, with a gender ratio of 2 to 3. Each case includes 1-2 radio medical images, including MRI, ultrasound, and X-ray. We evaluate 12 MLLMs as the doctor models, which are widely used and can complete the multi-modal multi-turn dialogue (Chen et al., 2024b; OpenAI, 2023b; Chen et al., 2024a). We adopt the Qwen-Max (Qwen, 2024b) as the backbone for the *patient-actor* agent, and we use GPT-4o (OpenAI, 2024b) as the evaluation LLMs. In practice, we set the seed as 0 to avoid randomness. As the *patient-actor* agent consultation, we both conduct the consultation from scratch and evaluate these two through the standard patient response to avoid the influence of the expression of the actor patient in the evaluation. More information can be found in the Appendix.

4.2 Evaluation Metrics

We primarily assess MLLMs' consultation capabilities from two perspectives: information gathering and final decision-making. Information gathering evaluates whether MLLMs can collect sufficient patient information during the consultation process, while final decision-making assesses whether MLLMs can ultimately provide accurate diagnoses

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

and recommendations. To validate this capability, we use both automatic evaluation and LLM-based evaluation methods.

4.2.1 Auto-Metric Evaluation

Information Gathering Metrics We categorize patient information into three main types: basic information, examination information, and multi-modal (medical imaging) information. Basic information includes Chief Complaint, Present Illness, and Medical History. We assess the information-gathering capability of doctor MLLMs by calculating the recall score of these three types of information in the final report generated during the consultation. For the intermedia consultation process, we defined the Info-Metric to analyze the consultation dynamics between patient information and responses. In practice, we use RaTEScore (Zhao et al., 2024), which is specifically designed to evaluate radiology reports, to calculate it.

Decision-Making Metrics Decision-making primarily encompasses two aspects: diagnosis and treatment plan. Similar to Information Gathering Metrics, we evaluate the model's decision-making capability by calculating the recall rate of diagnostic and treatment plan information from the patient data. However, unlike Information Gathering, diagnosis and treatment plans require the model to make predictions based on the patient information collected during the consultation, thus placing a greater emphasis on the model's reasoning ability.

4.2.2 LLM-Based Evaluation

We use the different prompts to set up various standards for inquiry, examination, multi-modal analysis, diagnosis, and treatment. The prompt we used in LLM-based evaluation is inspired by assessment plans from Peking University and National Health Commission of China (PMPH, 2018; NHC, 2023; BJMU, 2024), and the score is on a scale of 1-5.

4.3 Consultation with Standard Patient

4.3.1 Direct Prompt

We evaluated the consultation performance of 12 MLLMs with standard patients with the result shown in Table 2. Among these results, the GPT-4v achieve the best performance in both information-gathering and decision-making evaluations, demonstrating their potential for application in medical consultations. The capabilities of the models across different dimensions are not consistent. Specifically, Gemini1.5-Flash excels in medical image interpretation but is weaker in gathering information from the examination. Conversely, Qwen-VL-Max performs exceptionally well in consultations (nearly the best), but its ability to analyze medical images is quite poor. On the other hand, the models' decision-making capabilities are relatively consistent. If a model can provide a good diagnosis, it is also likely to offer an effective treatment plan. In summary, the strongest existing MLLMs can collect a substantial amount of basic patient information and some examination data. However, even the most advanced medical MLLM, HuatuoGPT-Vision-34B, shows significant shortcomings in handling multimodal medical information. This highlights a new direction for improving future MLLMs in medical applications.

4.3.2 Performance with CoT

To comprehensively assess the capacity of the MLLMs, we further adopt the zero-shot CoT and one-shot CoT to enhance the models' consultation ability. For zero-shot CoT, we use the specific prompt 'Let's think step by step' to encourage the model to think through each step before generating a response. For the one-shot CoT, we select an outside medical case and use GPT-4o to generate a CoT consultation dialogue between the doctor and the patient based on the full information in the report. We use the example as the one-shot CoT prompt and add it to the original doctor prompt to make the doctor models complete one-shot CoT. The CoT results are shown in Table 2. It can be observed that the performance improvement with zero-shot CoT is limited, as models rely solely on their knowledge without task-specific guidance. In contrast, one-shot CoT leads to an inconsistent performance enhancement. While there is an improvement for the treatment, GLM-4V and GPT-4o, other models do not show significant gains, demonstrating a stronger ability to process and reason with task-specific examples.

4.4 Consultation with *patient-actor* agent

For the consultation with *patient-actor* agent (personalized actor agent generates the response), we select 4 categories of MLLMs with 3 inference methods to validate the impact of the personalized patient in the consultation process. As shown in Table 3, the performance of most MLLMs and inference methods declined under the influence of a personalized actor. In terms of information gathering, the collection of inquiries suffered the most sig-

Model	Information Gathering				Decision-Making			Lens	Turns
	Inquiry	Exam.	MMA.	Avg.	Diagn.	Treat.	Avg.		
Direct Prompt									
InternVL-1.5 (Chen et al., 2024b)	40.47	9.28	14.58	21.44	33.40	28.61	31.00	36.47	8.63
Qwen-VL-Chat (Bai et al., 2023)	46.07	4.64	11.68	20.79	32.82	29.94	31.38	32.62	10.00
Mini-InternVL-1.5 (Chen et al., 2024c)	43.66	2.59	15.07	20.44	31.92	30.60	31.26	41.79	8.00
HuatuoGPT-Vision-7B (Chen et al., 2024a)	41.67	10.58	4.37	18.87	28.57	30.00	29.28	44.44	9.76
HuatuoGPT-Vision-34B (Chen et al., 2024a)	41.44	15.65	15.92	24.33	33.18	28.84	31.01	49.89	8.70
GLM-4V (GLM et al., 2024)	37.68	10.54	18.29	22.17	29.34	26.66	28.00	65.51	7.90
Qwen-VL-Max (Qwen, 2024a)	43.47	**22.17**	21.85	29.16	34.12	30.45	32.28	35.56	9.33
Gemini1.5-Flash (Reid et al., 2024)	44.78	16.30	**25.59**	28.89	37.04	**32.26**	34.65	27.39	9.53
Gemini1.5-Pro (Reid et al., 2024)	44.59	17.97	23.32	28.62	36.91	31.18	34.04	25.10	9.70
GPT-4V (OpenAI, 2023b)	**46.91**	18.95	25.47	**30.44**	**38.01**	31.81	**34.91**	48.46	9.70
GPT-4o (OpenAI, 2024b)	44.51	20.83	19.25	28.19	36.61	29.53	33.07	39.54	9.43
GPT-4o-mini (OpenAI, 2024a)	45.41	13.36	15.44	24.73	35.30	30.30	32.80	30.04	9.00
Zero-Shot CoT									
GLM-4V (GLM et al., 2024)	35.90	8.42	19.60	21.30	28.86	26.91	27.88	63.89	7.60
Qwen-VL-Max (Qwen, 2024a)	40.76	13.66	23.50	25.97	30.45	28.93	29.69	31.15	8.90
Gemini1.5-Pro (Reid et al., 2024)	**44.59**	18.16	15.54	26.09	**36.08**	31.58	**33.83**	23.59	8.90
GPT-4o (OpenAI, 2024b)	42.96	**23.66**	**20.47**	**29.03**	35.80	**31.75**	33.78	43.92	9.16
One-Shot CoT									
GLM-4V (GLM et al., 2024)	36.50	13.96	26.29	25.58	32.70	28.95	30.82	130.36	7.80
Qwen-VL-Max (Qwen, 2024a)	42.61	19.02	15.47	25.70	31.85	29.42	30.63	60.71	8.23
Gemini1.5-Pro (Reid et al., 2024)	38.55	18.52	15.39	24.15	36.21	30.25	33.23	63.98	7.80
GPT-4o (OpenAI, 2024b)	**44.88**	**22.66**	**27.92**	**31.82**	**37.93**	**30.84**	**34.38**	64.78	8.66

Table 2: The consultation performance with the **standard *patient-actor* agent**. We evaluate different doctor models with three types of inference modes, including *Direct Prompt*, *Zero-CoT*, and *One-shot CoT*. We assess the model's performance from two perspectives. 'MMA.' indicates Multi-modal Analysis, 'Diagn.' indicates Diagnosis, and 'Treat' indicates Treatment. The best results of each dimension are **Bold**.

nificant drop, as the information obtained through questioning is directly affected by the patient's expression. Consequently, the decision-making performance also suffered significantly.

A surprising finding is that Zero-shot CoT significantly mitigates the impact of the patient-actor, with performance declining by only -4.24 and -2.65 points in total, much less than Direct Prompt (-8.82 and -5.91) and One-Shot CoT (-11.28 and -5.81). This may be because the model's step-by-step reasoning before making a decision allows it to better handle information expressed in various ways. On the other hand, the failure of One-Shot CoT is likely due to the given example not including similar situations.

More LLM-based evaluation results of the standard patient and actor patient are shown in Figure 2. The performance of the MLLMs consulting with standard patients is better than that of the patient actor, which is consistent with the automatic metrics. Besides, it is obvious that the MLLMs fall short of gathering the examination information.

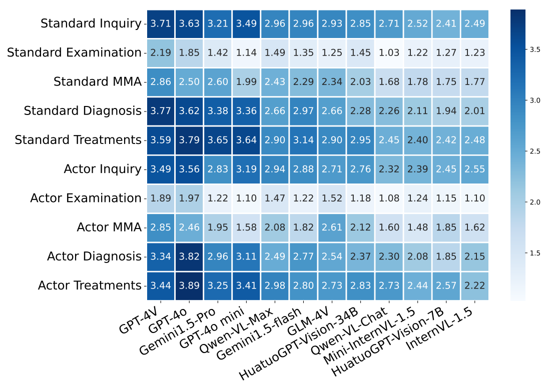


Figure 2: Results of LLM-based evaluation on consultation with both standard patient and actor patient.

4.5 Patient Evaluation

To demonstrate the effectiveness of our evaluation framework, We evaluated the performance of the standard patient agent and actor agent in response generation, and the result is shown in Table 4 and Table 5. In detail, we randomly select 3 different

Model	Information Gathering				Decision		
	Inquiry	Examination	MMA	Average	Diagnosis	Treatment	Average
Direct Prompt							
GLM-4V	36.76 (-0.92)	13.20 (+2.66)	18.30 (0.01)	22.75 (+0.58)	27.15 (-2.10)	26.64 (-0.02)	26.85(-1.10)
Qwen-VL-Max	39.34 (-4.13)	14.96(-7.21)	20.15 (-1.70)	24.81 (-4.34)	30.91(–3,23)	29.03 (-1.42)	29.97 (-2.31)
Gemini1.5-Pro	41.91 (-2.68)	16.40 (-1.57)	20.18(-3.14)	26.16(-2.35)	35.33 (-1.58)	31.04 (-0.14)	33.18 (-0.85)
GPT-4o	41.61 (-1.90)	15.47 (-5.36)	18.36(-0.89)	25.48 (-2.71)	35.08 (-1.53)	29.21(-0.32)	32.14(-0.93)
Zero-Shot CoT							
GLM-4V	38.47 (+2,57)	7.04 (-1.38)	20.3 (+0.70)	21.93 (+0.63)	27.32(-1.54)	26.25(-0.66)	26.78 (-1.09)
Qwen-VL-Max	40.02(-0.74)	20.06(+6.40)	19.16(-4.34)	26.41 (+0.44)	31.41(+0.96)	29.40(+0.47)	30.40 (+0.71)
Gemini1.5-Pro	43.14(-1.45)	14.39(-3.77)	21.06(+5.52)	26.19 (+0.11)	35.61(-0.47)	30.62(-0.96)	33.11 (-0.71)
GPT-4o	42.49(-0.47)	17.96 (-5.70)	10.36(-10.11)	23.60(-5.42)	35.37(-0.43)	29.06(-2.69)	32.21 (-1.56)
One-Shot CoT							
GLM-4V	34.97 (-1.53)	7.43 (-6.53)	24.02 (-2.27)	22.14 (-3.44)	31.65 (-1.05)	27.70 (-1.25)	29.68 (-1.15)
Qwen-VL-Max	38.06 (-4.55)	13.26 (-5.76)	13.94 (-1.53)	21.75 (-3.95)	31.11 (-0.74)	30.09 (+0.67)	30.60 (-0.03)
Gemini1.5-Pro	35.68 (-2.87)	20.07 (+1.55)	19.22 (+3.83)	24.99 (+0.84)	30.49 (-5.72)	30.84 (+0.59)	30.67 (-2.57)
GPT-4o	42.05 (-2.83)	18.50 (-4.16)	25.75 (-2.17)	28.77 (-3.05)	34.81 (-3.12)	29.84 (-1.00)	32.33 (-2.06)

Table 3: The consultation performance with the ***patient-actor* agent** (personalized responses). We evaluate different doctor models with three types of inference modes, including *Direct Prompt*, *Zero-CoT*, and *One-shot CoT*. The results in '(·)' represent the delta performance compared to the standard patient-actor agent shown in Table 2. 'MMA.' indicates Multi-modal Analysis.

S-T	D	I	C	U	CM	Total
w	2.00	1.85	1.96	1.81	1.85	9.48
w/o	1.00	0.96	1.36	1.10	1.10	5.52

Table 4: The LLM-based evaluation result of standard patient agent performance. Each aspect is worth 2 points. S-T: State Tracker, D: Description clarity, I: Information completeness, C: Cooperation, U: Understanding, CM: Communication Ability.

Setting	Semantic Consistency	Character Feature
w	4.85	4.59
w/o	-	0.77

Table 5: The LLM-based evaluation result of actor agent performance. Each aspect is worth 5 points.

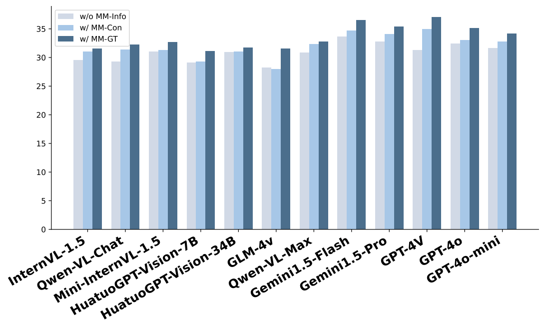


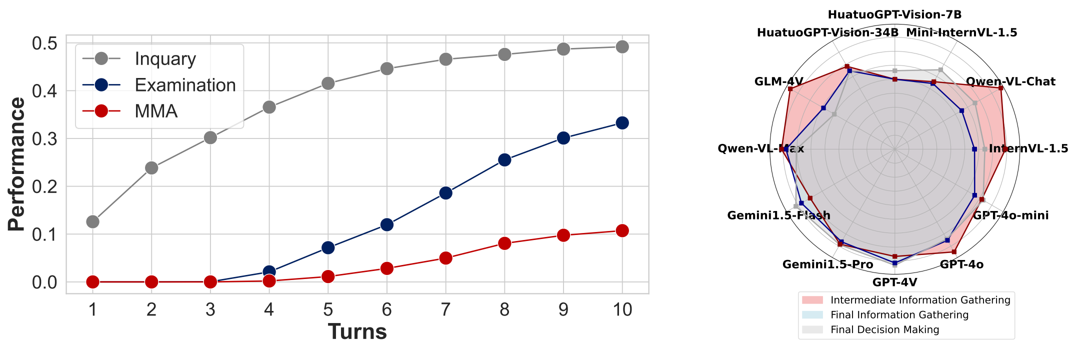
Figure 3: Ablation study of multi-modal information on the average of the diagnosis performance and treatment performance.

dialogues in each consultation case and use GPT-4o to evaluate the performance. As the standard patient, we consider five aspects: description, information completeness, cooperation, understanding, and communication ability. We assess the semantic consistency and character features between the standard and the actor output. The results in the two tables show that the patient agent can understand the doctor's purpose and generate an appropriate response, making the framework evaluate the doctor effectively. Furthermore, the ablation study of the state tracker and personalized actor also suppose the effectiveness of these two modules.

5 Discussion

5.1 Effectiveness of Multi-modal Information

In this section, we primarily analyze the importance of multimodal information in decision-making. We set up three scenarios: 1) ***w/o* MM-Info**: MLLMs have no access to the patient's multimodal information. 2) ***w/* MM-Con**: MLLMs can acquire and analyze the patient's multimodal information through consultation, representing the standard scenario. 3) ***w/* MM-GT**: MLLMs have direct access to the patient's medical imaging reports. By comparing the performance differences of MLLMs in decision-making across these three scenarios, we can evaluate the effectiveness of multimodal information. The results are shown in Figure 3. On one hand, multimodal information significantly impacts



(a) Different consultation turns.

(b) Different consultation models.

Figure 4: The illustration of information gathering performance.

diagnosis but has a lesser effect on treatment. The 'w/ MM-GT' scenario notably enhances MLLMs' decision-making performance. On the other hand, the 'w/ MM-Con' scenario is often on par with the 'w/o MM-Info' scenario, or even performs worse. This indicates that MLLMs are currently unable to effectively analyze medical images, falling short of clinical standards. In summary, while multimodal information is crucial for clinical diagnosis, current MLLMs cannot fully utilize medical imaging information to make accurate diagnoses.

5.2 Information Gains Capability

We further explored how models collect different types of information during the consultation process. As shown in Figure 4, we compiled the cumulative information collected by all models at each round and found that MLLMs tend to gather general information first, followed by examination information, and finally image information. This indicates that MLLMs possess a basic logical structure for consultations, and their preferences in clinical tasks are similar to those of human doctors. Furthermore, we compare the information-gathering performance across various models using the Info-Metric to assess the validity of intermediate responses during the consultation process. Our findings suggest that the high performance in final information gathering is not strongly correlated with the logits of intermediate responses, indicating that more attention should be paid to the consultation logits.

5.3 Gender Bias of the MLLMs

We further explore the gender bias on Gemini-pro and GPT-4o in medical consultation. Specifically, we set the gender of all patients in the entire test set to either male or female to observe any potential gender bias in the model.[2] The results in Figure 5 show that both models exhibit varying degrees of gender bias. Specifically, GPT-4o performs worse with female patients, while Gemini1.5-Pro performs worse with male patients. Notably, gender has almost no impact on the inquiry performance. As a result, the gender effect can significantly influence model performance in practice. Therefore, addressing and mitigating gender bias should be a key focus for future research and development efforts.

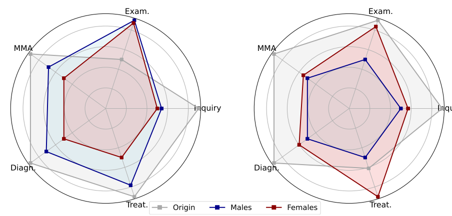


Figure 5: Gender bias of MLLMs. On the left is the GPT-4o. On the right is the Gemini1.5-Pro. Both MLLMs exhibit varying degrees of gender bias in medical consultations. All the scores are normalized.

6 Conclusions

In this paper, we proposed a Multi-modal Consultation Evaluation Framework to assess the clinical performance of the MLLMs and explore their potential application in realistic clinical environments. The results show that current MLLMs fail shot in consulting with personalized actor and can not fully utilize the multi-modal information. Our results demonstrate the effectiveness of `Med-PMC`, thereby guiding the development of more robust and reliable MLLMs for clinical use.

[2]As these diseases are not sensitive to gender, this setting about gender cannot influence the final result. These adjustments are validated by medical professionals.

Limitations

We only evaluate the performance of the general surgery department. Although this department is highly representative, the performance of the model could not be further evaluated due to data limitations. Furthermore, due to the constraints of departmental data, we were unable to assess the model's generalizability in other departments or broader clinical settings. Future research could expand the data set and include a variety of departments to further validate the reliability and adaptability of the model.

Ethical Consideration

The medical cases are collected from the iiyi website, where doctors voluntarily upload and share information. This data is explicitly permitted for use in research and educational purposes. To ensure the protection of patient privacy, our dataset does not include any personally identifiable information such as patient names, hospital details, or any other sensitive information. Consequently, there is no risk of privacy breaches associated with our dataset. Additionally, all data usage complies with ethical standards and regulations governing medical information and research.

References

Jinze Bai, Shuai Bai, Shusheng Yang, and et al. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, et al. 2023. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11):e2343689–e2343689.

BJMU. 2024. *Peking University Health Science Center Clinical Medicine Program Objective Structured Clinical Examination (OSCE) Instructions*. Peking University Health Science Center Office of Education, Peking, China.

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, et al. 2024a. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Zhe Chen, Weiyun Wang, Hao Tian, and et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Gao Zhangwei, Cui Erfei, and et al. 2024c. Mini-internvl 1.5: A powerful pocket multimodal model with 8% parameters for 80% performance.

Team GLM, Aohan Zeng, Bin Xu, and et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

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*, pages 1–10.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. 2024. Ct2rep: Automated radiology report generation for 3d medical imaging. *arXiv preprint arXiv:2403.06801*.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2021. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 708–718, Online. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. CRAFT-MD: A conversational evaluation framework for comprehensive assessment of clinical LLMs. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Yusheng Liao, Yutong Meng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2023. An automatic evaluation framework for multi-turn medical consultations capabilities of large language models. *arXiv preprint arXiv:2309.02077*.

Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

NHC. 2023. *Standardized Training for Residents in Surgery and All Directions of Surgery Clinical Practice Ability Assessment Program (2023 Edition)*. Health and Wellness Commission of China, Peking, China.

OpenAI. 2023a. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2023b. Gptv_system_card.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024b. Hello gpt-4o.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

PMPH. 2018. *Diagnostics, 9th ed*. People's Medical Publishing House, Peking, China.

Qwen. 2024a. The github of qwen-vl-max.

Qwen. 2024b. Introducing qwen1.5.

Machel Reid, Nikolay Savinov, Denis Teplyashin, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, Miami, Florida, USA. Association for Computational Linguistics.

A The details on Doctor models

Model	Size	Precision	Belongs
Mini-InternVL-1.5	4.2B	BF16	OpenGVLab
InternVL-1.5	25.5B	BF16	OpenGVLab
Gemini-flash	API	-	Google
Gemini-pro	API	-	Google
GLM4v	API	-	THUDM
GPT-4o	API	-	OpenAI
GPT-4o mini	API	-	OpenAI
GPT-4v	API	-	OpenAI
Qwen-vl-chat-v1	9.6B	FP16	Ali
Qwen-vl-max	API	-	Ali
HuatuoGPT-Vision-7B	7.94B	BF16	SRIBD
HuatuoGPT-Vision-34B	34.8B	BF16	SRIBD

Table 6: The details about the doctor models.

More details on doctor models are shown in Table 6. Furthermore, we set the temperature as 0.01 and 'do_sample=False' to demonstrate the most primitive capabilities of the model, and we set the seed as 0 to make the results reproducible.

B The Details on Evaluation

B.1 Auto-Metric Formulation

For the sake of convenience, we summarize the symbols used in the formula in Table 7.

B.1.1 Information Gathering

Inquiry The inquiry measures how much patient information is obtained in the consultation and summarizes it in the report.

$$\text{Iquiry} = \frac{1}{N}\sum_{j=1}^{N}(\text{Recall}(\boldsymbol{i}_p^j, \mathcal{R}^j)), \tag{5}$$

Where $\boldsymbol{i}_p^i$ denotes the ground truth of the inquiry, including the patient information, chief, present illness, and past medical history, $\mathcal{R}^j$ is the final report generated by the doctor agent, and the N denotes the total number of cases. We use RaTEScore (Zhao et al., 2024) as the implementation of the recall calculation without special instructions.

Examination The examination measures how many examination items are made in the consultation and summarizes them in the report. We first detect the exam entity in the patient information, which is formulated as e.

$$\mathcal{E} = \{e | e \in \text{NER}(\boldsymbol{i}_e)\}, \tag{6}$$

Symbol	Significance
N	The total number of cases.
K	The total turns of consultation.
$\boldsymbol{i}_p$	The ground truth of the inquiry, including the patient information, chief, present illness, and past medical history.
$\boldsymbol{i}_e$	The ground truth of the examination.
$\boldsymbol{i}_m$	The ground truth of the image analysis.
$\boldsymbol{d}$	The ground truth of the diagnosis.
$\boldsymbol{t}$	The ground truth of the treatment.
$\mathcal{R}$	The final report generated by the doctor agent.
$\boldsymbol{y}_d$	The output generated by the doctor agent.
$\boldsymbol{y}_p$	The response of the patient-actor agent.
$\mathcal{E}$	The exam items.
Info	The amount of information gained

Table 7: The significance of different symbols.

where $\boldsymbol{g}_e$ denotes the ground truth of the examinations and NER denotes the entity extraction tool Spacy (Honnibal et al., 2020). Then, we calculate the examination score:

$$\text{Examination} = \frac{1}{N}\sum_{j=1}^{N}\frac{\sum_{e\in\mathcal{E}^j}\mathbb{I}(e\in\mathcal{R}^j)}{|\mathcal{E}^j|}, \tag{7}$$

Where $\mathbb{I}(\cdot)$ equals 1 only when the equation in $(\cdot)$ is held and $|\mathcal{E}^j|$ denotes the number of entities in $\mathcal{E}^j$.

Multi-modal Analysis This metric measures how much the correlation between the doctor's response and the image report after obtaining the image.

$$\text{MMA} = \frac{1}{N}\sum_{j=1}^{N}(\text{Recall}(\boldsymbol{i}_m^j, \boldsymbol{y}_{d,k}^j)), \tag{8}$$

where the $\boldsymbol{y}_{d,k}^i$ denotes the output after the doctor obtains the image in the ith case, i.e. the k turns.

Info-Metric Information gain mainly measures the logic of the multi-turn consultation and we use the information gained in various dialogue turns to measure the logic.

$$\text{Info} = \frac{1}{\text{K}}\sum_{\text{k}=1}^{\text{K}}(\text{Recall}(\boldsymbol{y}_{\text{p,k}}, \boldsymbol{g}_{\text{p}})), \tag{9}$$

where the Info denotes the amount of information gain, $\boldsymbol{y}_{p,k}$ is the kth response from the patient-actor agent.

B.1.2 Decision-Making

Diagnosis The diagnosis measures the degree of diagnosis of the disease, such as the location of the disease and the type of disease.

$$\text{Diagnosis} = \frac{1}{N}\sum_{j=1}^{N}\text{Recall}_{\text{L}}(\boldsymbol{d}^{j}, \mathcal{R}^{j}), \tag{10}$$

where we use the RaTEScoree as the measure in recall calculation.

Treatment The treatment measures how much the treatment plan coincides with the ground truth.

$$\text{Treatment} = \frac{1}{N}\sum_{j=1}^{N}\text{Recall}(\boldsymbol{t}^{j}, \mathcal{R}^{j}). \tag{11}$$

C Further Evaluation Results

We further employ the Rouge-1, Rouge-L, and the Bert score metric in consultation performance evaluation, and the results are shown on Table 8 - Table 10. The overall performance trends of each model under the Rough-1 and Rough-L metrics are consistent with the results presented in the main text, showing that the GPT series models perform relatively well, and that larger model parameters generally lead to better results. However, since BERTScore is not well-suited for evaluating long texts, its results differ significantly from those of the other metrics.

C.1 LLM-based Prompts

The prompt for LLM-based evaluation is also divided into three parts, including the information gathering, decision making, and other aspects, and the prompt is shown in Table 11 - Table 12.

C.2 Patient Evaluation Prompts

The prompt for patient evaluation is divided into the standard patient agent prompt and actor agent prompt, which is shown in Table 13.

D The Prompts for Agents

D.1 The prompt for doctor agent

The prompt for the doctor agent is divided into consultation and diagnosis stages, which is shown in Figure 14.

D.2 The prompt for state detection agent

The prompt for the state detection agent is divided into three steps, and different steps are responsible for different effects.

The StageI Prompt The stage I is responsible for the preliminary intent classification, including the inquiry, examination, others, and the end. The prompt is shown in Table 15.

The StageII Prompt The stage II mainly detects whether the question is specific or board, shown in Table 16. The system A prompt is used when the result in stage I is A and the system B prompt is used when the result is B. If the result in stage I is C or D, the stage detection agent will stop working.

The StageIII Prompt The stage III mainly detects whether the question has relevant information in the case report, to better answer the question. The prompt is shown in Table 17, and divided into two different situations. One is for the $A - A-$ result and the other is for the $B - A-$ result.

D.3 The prompt for standard patient agent

The standard patient agent is responsible for answering the questions about the patient himself from the doctor, including the 0, $A-$, and C state. The prompt is shown in Table 18 and there are different prompts for different states to make the response more accurate.

D.4 The prompt for technician agent

The technician agent works when the state is $B-$, and the prompt is shown in Table 19 - Table 21. The agent I is responsible for extracting the exam item from the doctor, the agent II aims at generating the initial response based on the state, and the agent III is used to generate the final response based on the former result.

D.5 The prompt for Actor agent

The actor aims at generating a more characterful response based on the standard patient agent, and the prompt is shown in Table 22. To better display the characters among different roles, we design different prompts based on their characteristics.

Model	Information Gathering				Decision-Making		
	Inquiry	Exam.	MMA.	Avg.	Diagn.	Treat.	Avg.
InternVL-1.5	52.42	13.84	13.94	26.73	26.32	21.44	23.88
Qwen-VL-Chat	57.48	15.06	2.46	25.00	29.47	23.98	26.73
Mini-InternVL-1.5	50.43	10.29	10.48	23.73	23.01	18.97	20.99
HuatuoGPT-Vision-7B	54.87	10.95	5.14	23.65	32.78	25.41	29.10
HuatuoGPT-Vision-34B	59.34	24.52	16.22	33.36	40.54	27.29	33.92
GLM-4V	32.19	10.73	**25.06**	22.66	21.75	19.97	20.86
Qwen-VL-Ma	56.34	19.66	15.20	30.40	28.61	27.08	27.08
Gemini1.5-Flash	63.99	17.95	19.25	33.73	37.81	26.75	32.28
Gemini1.5-Pro	68.19	26.74	11.35	35.43	46.21	**32.95**	39.58
GPT-4V	68.28	25.50	20.59	38.12	44.17	32.19	38.18
GPT-4o	**68.75**	**32.50**	20.76	**40.67**	45.12	32.25	38.69
GPT-4o-mini	68.25	25.20	6.45	33.30	**46.83**	32.17	**39.50**

Table 8: The consultation performance with the **standard *patient-actor* agent** with Rouge-1 metric. We assess the model's performance from two perspectives. 'MMA.' indicates Multi-modal Analysis, 'Diagn.' indicates Diagnosis, and 'Treat' indicates Treatment. The best results of each dimension are **Bold**.

E Case Study

To further illustrate the capability of the consultation among these doctor models, we have shown the different cases of standard patient consultation, actor agent consultation, and one-shot CoT consultation, by using different doctor models. The result is shown in Table 23-Table 29. From these cases, the M^3 framework has shown the advantage in answering the question from the doctor based on the patient cases. Especially, the standard patient agent and the technician agent know different information, which the doctor cannot acquire inspection information through vague questions, and the standard patient knows nothing about the examinations, which is closer to reality.

Model	Information Gathering				Decision-Making		
	Inquiry	Exam.	MMA.	Avg.	Diagn.	Treat.	Avg.
InternVL-1.5	35.82	9.28	12.9	19.33	26.32	17.57	21.95
Qwen-VL-Chat	37.36	4.64	14.09	18.70	29.47	19.57	24.52
Mini-InternVL-1.5	32.43	2.59	11.93	15.65	23.01	15.78	19.40
HuatuoGPT-Vision-7B	31.89	10.58	3.45	15.31	32.78	19.07	25.93
HuatuoGPT-Vision-34B	33.16	15.65	13.17	20.66	40.54	20.10	30.32
GLM-4V	18.32	10.54	11.25	13.37	21.57	15.75	18.66
Qwen-VL-Max	35.00	**22.17**	18.73	25.30	28.61	20.89	24.75
Gemini1.5-Flash	38.76	16.30	24.19	26.42	37.81	22.10	29.96
Gemini1.5-Pro	40.55	17.97	23.78	27.43	**46.12**	**26.77**	**36.45**
GPT-4V	**42.58**	18.95	**27.58**	**29.70**	44.17	26.22	35.20
GPT-4o	42.01	20.83	20.94	27.93	45.12	25.64	35.38
GPT-4o-mini	41.99	13.36	13.34	22.90	47.31	27.18	37.25

Table 9: The consultation performance with the **standard** ***patient-actor*** **agent** with Rouge-L metric. We assess the model's performance from two perspectives. 'MMA.' indicates Multi-modal Analysis, 'Diagn.' indicates Diagnosis, and 'Treat' indicates Treatment. The best results of each dimension are **Bold**.

Model	Information Gathering				Decision-Making		
	Inquiry	Exam.	MMA.	Avg.	Diagn.	Treat.	Avg.
InternVL-1.5	81.31	9.28	38.25	42.95	76.74	77.35	77.05
Qwen-VL-Chat	85.04	4.64	35.58	41.75	**78.96**	**79.72**	**79.34**
Mini-InternVL-1.5	82.40	2.59	38.23	41.07	76.99	77.61	77.30
HuatuoGPT-Vision-7B	81.08	10.58	13.55	35.07	76.63	77.24	76.94
HuatuoGPT-Vision-34B	82.69	15.65	38.13	45.49	77.38	78.03	77.71
GLM-4V	81.00	10.54	51.34	47.63	77.48	78.39	77.94
Qwen-VL-Max	**83.92**	**22.17**	54.44	**53.51**	78.07	78.72	78.40
Gemini1.5-Flash	80.46	16.30	**64.80**	53.85	76.13	76.66	76.40
Gemini1.5-Pro	82.56	17.97	59.69	53.41	77.10	77.75	77.43
GPT-4V	83.03	18.95	62.63	54.87	76.89	77.56	77.23
GPT-4o	82.08	20.83	48.96	50.62	76.46	77.20	76.83
GPT-4o-mini	81.05	13.36	35.39	43.27	76.08	76.74	76.41

Table 10: The consultation performance with the **standard** ***patient-actor*** **agent** with Bert score metric. We assess the model's performance from two perspectives. 'MMA.' indicates Multi-modal Analysis, 'Diagn.' indicates Diagnosis, and 'Treat' indicates Treatment. The best results of each dimension are **Bold**.

Promt for LLM-based Evaluation on Information Gathering

System prompt

You are a strict evaluator. Below, a case will be presented, which has been divided into three modules: case info, dialogue, and case by doctor. Your task is to evaluate the consultation, examination, diagnosis and treatment, and communication skills based on the following evaluation criteria. Provide ratings for each aspect on a scale of 1 to 5 at the beginning in format: 'score: x'. In this case, the case info is the standard case content, which you should consider as a reference answer and use it as a benchmark to evaluate the conversation with the case summarized by the doctor. For cases where the criteria for corresponding scores are not fully met, one can take the decimal between them. You don't need to give any explanation or repeat my input. Use the following scale to evaluate each criterion:

Inquiry

score 1: The doctor's logic is confused, questions are unreasonable, and language is obscure with excessive medical jargon. Medical history collection is incomplete and lacks details.

score 2: The doctor has some logic but still makes unreasonable arrangements and uses obscure language with medical terms. Medical history collection is incomplete with missing key points.

score 3: The doctor has some logic with fewer unreasonable arrangements, speaks more fluently with less jargon, and collects most key medical history details.

score 4: The doctor is logical and clear, focuses on key issues, avoids jargon, and collects a complete medical history.

score 5: The doctor is very clear and logical, focuses on key issues, avoids jargon entirely, and collects a fully complete medical history.

Examination

score 1: The doctor performs no valid tests (i.e., the tests in the 'Examination' item in the standard case) and prescribes highly unnecessary tests.

score 2: The doctor performed the tests in the 'Examination' item in the standard case but some additional tests are unnecessary, with significantly inaccurate interpretations.

score 3: The doctor performed the tests in the 'Examination' item in the standard case but with a few unnecessary ones, and has somewhat inaccurate interpretations.

score 4: The doctor performed the tests in the 'Examination' item in the standard case but very few unnecessary ones, and slightly inaccurate interpretations.

score 5: The doctor performs all necessary tests with reasonable prescriptions and virtually error-free interpretations.

Attention: if doctor doesn't perform the examination in the 'Examination' item in the standard case, the score should be 1.

Multi-modal Analysis:

score 1: Compared to examination results in benchmarks, there is a very large gap between the doctor's analysis of the examination image, e.g., the conclusions given are very different, etc.

score 2: Compared to examination results in benchmarks, there is a large gap between the doctor's analysis of the examination image.

score 3: Compared to examination results in benchmarks, there are some diffrences between the doctor's analysis of the examination image, but the conclusions are similar.

score 4; Compared to examination results in benchmarks, there are small gaps between the doctor's analysis of the examination image.

score 5: Compared to examination results in benchmarks, doctor's analysis of the examination image were almost identical

Table 11: The prompt for LLM-based evaluation on information gathering.

Promt for LLM-based Evaluation on Decision Making

System prompt

You are a strict evaluator. Below, a case will be presented, which has been divided into three modules: case info, dialogue, and case by doctor. Your task is to evaluate the consultation, examination, diagnosis and treatment, and communication skills based on the following evaluation criteria. Provide ratings for each aspect on a scale of 1 to 5 at the beginning in format: 'score: x'. In this case, the case info is the standard case content, which you should consider as a reference answer and use it as a benchmark to evaluate the conversation with the case summarized by the doctor. For cases where the criteria for corresponding scores are not fully met, one can take the decimal between them. You don't need to give any explanation or repeat my input. Use the following scale to evaluate each criterion:

Diagnosis

Consider item 'Diagnosis' in the 'case info' as the standard diagnosis answer

score 1: The doctor is unable to provide a correct primary diagnosis and differential diagnosis, and it's far from the answer. e.g., pancreatitis is diagnosed as a heart attack

score 2: The doctor provides a diagnosis that is completely inconsistent withthe answer, but the deviation is small, as when tonsillitis is diagnosed as bronchitis.

score 3: The doctor provides a diagnosis with minor discrepancies from the answer, but the organ of pathogenesis is correctly diagnosed, e.g., gastric ulcer is diagnosed as gastritis.

score 4: The doctor provides a diagnosis that is close to the answer but there are some minor errors that are difficult to distinguish, such as influenza A being diagnosed as influenza B.

score 5: The doctor provides a diagnosis that is consistent with the answer.

Treatment Consider item 'Treatment' in the 'case info' as the standard diagnosis answer

score 1: The doctor is unable to provide a reasonable treatment plan, or it's far from the answer. e.g., only painkillers are needed but doctor decides to perform surgery.

score 2: The doctor's treatment plan does not resolve the patient's condition, but it also does not cause additional damage to patient, such as the need for doxycycline injections but doctor decides to inject saline

score 3: The doctor's treatment plan is very limited to help the patient's condition, for example, patient needs surgery but the doctor only uses painkillers

score 4: The doctor's treatment is helpful but not optimal for the patient's condition, e.g., patient needs amoxicillin injections but the doctor chooses sulfonamide

score 5: The doctor provides a treatment that is consistent with the answer.

Attention:provide ratings for each aspect on a scale of 1 to 5 at the beginning in format: 'score: x'

Table 12: The prompt for LLM-based evaluation on decision making.

Promt for Patient Evaluation

Standard Patient Agent

$< \mathrm{Requirement} >$: Please evaluate whether the patient can accurately answer the doctor's questions according to the patient's information and conversation history according to the patient information given, and consider the five aspects of description clarity, information completeness, cooperation, understanding and communication ability, and do not require the patient to take the initiative to give treatment suggestions.

The total score is 10 points, every aspect is 2 points, please output the score directly and do not output the analyze.

The output format is: $< \mathrm{descriptionclarity} >$:n , $< \mathrm{informationcompleteness} >$:n , $< \mathrm{cooperation} >$:n , $< \mathrm{understanding} >$:n and $< \mathrm{communicationability} >$:n

Actor Agent

$< \mathrm{Requirement} >$: The existing output is converted into an output with character characteristics according to different occupations, and the quality of the conversion statement is evaluated according to the semantic consistency and character feature richness.

The total score is 10 points, every aspect is 5 points, please output the score directly and do not output the analyze.

The output format is: $< \mathrm{SemanticConsistencyScore} >$: n $< \mathrm{CharacterFeatureScore} >$: n

Table 13: The prompt for patient evaluation.

Promt for Doctor Agent

Consultation prompt

You are an experienced doctor, and a patient will soon consult you about their condition. This consultation will consist of up to 10 rounds of dialogue. To better diagnose the patient, you will ask questions based on prompts each round. Once you feel you have sufficient information about the patient, you will provide a diagnosis.
1. Consultations generally follow this order:
a) Chief complaint, such as main symptoms and duration.
b) Basic information, such as age and gender.
c) Possible causes of the symptoms, such as travel history and past medical history.
d) Medical examinations, such as X-ray and blood pressure.
2. The doctor's questions should be concise and clear, while the tone should be patient and caring for the patient.
3. The patient has already undergone all the necessary examinations for diagnosis, so the doctor can directly inquire about the results of the tests without requiring the patient to do further examinations.
4. There are only a maximum of 10 rounds of consultation dialogue, so the questions asked by the doctor in each round should help to determine the patient's most likely diagnosis or to clarify the next medical examination that should be done as much as possible.

Consultation prompt with zero-shot CoT

Consultation prompt + 5. let's think step by step

Consultation prompt with one-shot CoT

Consultation prompt with zero-shot CoT + < Case >:
[Doctor]: Let's think step by step. Hello, please tell me your main discomfort and its duration.
[Patient]: Chest and back pain for six months, numbness and weakness in both lower limbs for one week.
[Doctor]: Let's think step by step. The patient's symptoms suggest possible involvement of the nervous system or spine. First, we need to understand the patient's basic information to better assess the condition. May I know your age and gender?
[Patient]: Female, 51 years old.
[Doctor]: Let's think step by step. The patient is a 51-year-old female, and women of this age are prone to some degenerative diseases. Next, we need to understand the specific details of the chest and back pain to further evaluate. Please describe the nature and location of the chest and back pain in detail.
[Patient]: Six months ago, I developed chest and back pain without obvious cause, feeling like pricking accompanied by a band-like sensation.
[Doctor]: Let's think step by step. Pricking and band-like sensations suggest possible nerve compression or spinal issues. To further confirm the cause, we need to understand the specific manifestations of the numbness and weakness in both lower limbs. What exactly does the numbness and weakness in both lower limbs feel like?
[Patient]: It appeared one week before admission, with numbness and weakness in both lower limbs. I could lift my upper limbs, and could walk with support, symmetrical on both sides. Occasionally, there are crawling sensations and burning in the lower limbs, especially severe in the soles.
[Doctor]: Let's think step by step. Numbness and weakness in both lower limbs, crawling sensations, and burning suggest a high likelihood of spinal cord lesions. To rule out other nervous system diseases, we need to confirm whether there are any other symptoms. Besides these symptoms, do you have any other symptoms like seizures, speech disorders, headaches, or dizziness?
[Patient]: No other symptoms.
[Doctor]: Let's think step by step. No other nervous system symptoms further support the likelihood of spinal cord or spinal lesions. To comprehensively understand the patient's physical condition, we need to inquire about basic vital signs. What are your temperature, heart rate, respiratory rate, and blood pressure?
[Patient]: T: 36.2°C, P: 76 beats/min, R: 17 breaths/min, BP: 132/74 mmHg.
[Doctor]: Let's think step by step. Basic vital signs are normal. To further confirm the diagnosis, an imaging examination is needed. Please undergo a cranial CT scan to rule out the possibility of brain lesions.
[Patient]: The cranial CT scan showed no significant abnormalities.
[Doctor]: Let's think step by step. The cranial CT scan showed no abnormalities, ruling out the possibility of brain lesions. Next, we need to focus on the spinal condition. Please undergo a spinal MRI examination.
[Patient]: < image00011.jpg >.
[Doctor]: Let's think step by step. Based on the imaging results, the patient has a lesion located at the right side of the spinal cord at the thoracic vertebra level, considered a schwannoma. This highly corresponds with the patient's symptoms. Based on this information, we preliminarily diagnose you with spinal schwannoma. We recommend resection of the spinal lesion for treatment.
[Patient]: Okay, thank you, doctor. < CaseEnd >

Diagnosis prompt

Please generate a medical case report based on the following dialogue, including patient information, present illness, past medical history, examination items and results, diagnosis, and treatment plan.
dialogue:{conversation}
medical case report:

Table 14: The prompt for doctor agent, which is divided into consultation and diagnosis. The blue text denotes the input information.

Promt for State Detection Agent

System prompt

During the consultation process, a doctor's questions can be categorized into five types:

(A) Inquiry: Doctors ask patients for information related to medical conditions, generally with words like 'please', 'please tell', '?', '?' or '?', and those not belonging to type (C) belong to this category.

(B) Examination: Doctors arrange patients for relevant examinations. Any suggestion by the doctor for the patient to undergo a certain medical examination belongs to this category.

(C) Other Topics: Questions from the doctor that do not pertain to the medical consultation scenario, and are unrelated to medical diseases, such as hobbies, movies, cuisine, etc.

(D) End: The doctor has completed the consultation and treatment recommendations have been given.

Based on the descriptions of each question type above, identify the most appropriate category for the following doctor's question:

Doctor's Question: {question}

Output:

Question Type:

Table 15: The prompt for the state detection agent, which is used in the stage I. The blue text denotes the input information.

Promt for State Detection Agent
System A prompt <Definition>: [Specific]: <Question> has a certain specific direction. When asking about symptoms, it should at least inquire about specific body parts, symptoms, sensations, or situations. When asking about examination results, it should mention specific body parts, specific examination items, or abnormal situations. Note that if it's about specific medical conditions, like medical history, family history, chronic illnesses, surgical history, etc., they are always considered [Specific]. Specifically, if the <Question> contain about demonstrative like "these" or "this", then it is related to the above and should belongs to the [Specific]. [Broad]: <Question> such as "Where do you feel uncomfortable?" or "Where does it feel strange?" without any specific information direction are considered [Broad]. <Question>: {question} Based on the <Definition>, determine whether the doctor's <Question> asks for [Specific] medical information from the patient or gives [Specific] advice. If so, directly output [Specific]. If not, output [Broad]. **System B prompt** <Definition>: [Specific]: <Advice> contains specific types of examinations or test (including but not limited to X-rays, MRI, biopsy, etc.), specific treatment plans (including but not limited to specific surgical treatments, exercises, diets, etc.), specific types of medication, etc. [Broad]: <Advice> broadly given without any specific examination/test, treatment plans, doctor's orders, exercises, diets and medication types is considered [Broad]. As long as any of the above information appears, <Advice> does not fall into this category. <Advice>: {question} Based on the <Definition>, determine whether the doctoŕs <Advice> asks for [Specific] medical information from the patient or gives [Specific] advice. If so, directly output [Specific]. If not, output [Broad].

Table 16: The prompt for the state detection agent, which is used in the stage II. The blue text denotes the input information.

Promt for State Detection Agent

System A prompt

<Definition>:

[Relevant Information]: <Patient Information> contains information asked in <Question>, including descriptions of having or not having the symptom, as long as there's relevant content.

[No Relevant Information]: <Patient Information> does not contain information asked in <Question>, and thereś no relevant content in the information.

<Patient Information>: {patient_info}

<Question>: {question}

Based on the <Definition>, determine whether <Patient Information> contains relevant information asked in <Question>. If [Relevant Information] is present, directly output the relevant text statement, ensuring not to include irrelevant content. If [No Relevant Information], then directly output [No Relevant Information].

System B prompt

<Definition>:

[Relevant Information]: <Patient Information> contains results of the examinations or treatment plans suggested in <Advice>, including any results related to the suggested examination items and treatment plans.

[No Relevant Information]: <Patient Information> does not contain results of the examinations or treatment plans suggested in <Advice>, including no mention of relevant examination items and treatment plans or no corresponding results.

<Patient Information>: {patient_info}

<Advice>: {question}

Based on the <Definition>, determine whether <Patient Information> contains relevant information about the measures suggested in <Advice>. If [Relevant Information] is present, directly output the relevant text statement, ensuring not to include irrelevant content. If [No Relevant Information], then directly output [No Relevant Information].

Table 17: The prompt for the state detection agent, which is used in the stage III. The blue text denotes the input information.

Promt for Standard Patient Agent

State 0 prompt

<Patient's Physical Condition >: {patient_info}

<Current Response Requirement>: Please respond to the doctor's questions using the information provided in <Patient's Physical Condition>. Only include the <Chief Complaint>, and avoid adding extra information. Make sure to use the original text from <Chief Complaint> to respond and keep it as short as possible. Answer in English.

Below is a dialogue between a doctor and a patient. The patient will respond directly to the latest round of questions from the doctor in the first person, without using a [patient] prompt. Do not include any text from <Current Response Requirement> in your response!

State A-A-A prompt

<Patient's Physical Condition >: {patient_info}

<Current Response Requirement>: Please respond to the doctor's questions using all the original text from <Patient's Physical Condition>. Make sure to maintain the accuracy of the patient's information by using the original text from <Patient's Physical Condition> to respond. Deny any information that is not related. Answer in English.

Below is a dialogue between a doctor and a patient. The patient will respond directly to the latest round of questions from the doctor in the first person. Do not include any text from <Current Response Requirement> in your response!

State A-A-B prompt

<Current Response Requirement>: The patient does not have the symptoms the doctor is asking about. Please deny the doctor's current question. Answer in English. {patient_info}

Below is a dialogue between a doctor and a patient. The patient will respond directly to the latest round of questions from the doctor in the first person. Do not include any text from <Current Response Requirement> in your response!

State A-B prompt

<Current Response Requirement>: The doctor's current question is too broad. The patient will request the doctor to ask more specific questions regarding the latest round of questions. Do not fabricate any non-existent information, or ask questions to the doctor. Answer in English. {patient_info}

Below is a dialogue between a doctor and a patient. The patient will respond directly to the latest round of questions from the doctor in the first person. Do not include any text from <Current Response Requirement> in your response!

State C prompt

<Current Response Requirement>: Remind the doctor that they have deviated from the topic of consultation and request them to return to the consultation scenario. Answer in English. {patient_info}

Below is a dialogue between a doctor and a patient. The patient will respond directly to the latest round of questions from the doctor in the first person. Do not include any text from <Current Response Requirement> in your response!

Table 18: The prompt for the standard patient agent. The blue text denotes the input information.

Promt for Technician Agent I

System prompt

Please extract the names of examination items from the questions asked by the DOCTOR, only output the names of the examination items, such as blood routine, electrocardiogram examination.

Table 19: The prompt for technician agent, which is used for exam item detection.

Promt for Technician Agent II

State B-A-A prompt

<Patient's Test Report>: {patient_info}

<Current Response Requirement>: The patient has completed the tests arranged by the doctor. Please respond to the doctor's inquiries using all the original text from <Patient's Test Report>, including the names of the tests and their results, to maintain the accuracy of the test report. Also, pay attention to different expressions for similar tests and include only one for similar test types. Answer in English.

Below is a dialogue between a doctor and a patient. The patient will respond directly to the latest round of questions from the doctor in the first person. Do not include any text from <Current Response Requirement> in your response!

State B-A-B prompt

<Current Response Requirement>: The test mentioned by the doctor is not in the report, indicating that it cannot be performed temporarily due to equipment issues. Answer in English. {patient_info}

Below is a dialogue between a doctor and a patient. The patient will respond directly to the latest round of questions from the doctor in the first person. Do not include any text from <Current Response Requirement> in your response!

State B-B prompt

<Current Response Requirement>: The doctor's request for tests is too broad. The patient will request the doctor to ask more specific questions regarding the latest round of tests. Do not fabricate any non-existent information, or ask questions to the doctor. Answer in English. {patient_info}

Below is a dialogue between a doctor and a patient. The patient will respond directly to the latest round of questions from the doctor in the first person. Do not include any text from <Current Response Requirement> in your response!

Table 20: The prompt for technician agent, which is used for generating an initial response based on the detected state.

Promt for Technician Agent III

System prompt for final result

You are a technician in charge of medical examinations at a hospital. Below is the <examination report> for the patient: {patient_info}

You need to generate responses for the examinations listed under <category of examination> based on the <examination report>:

(1) Only respond about the items listed under <category of examination>, do not mention items that are not included in <category of examination>.

(2) The responses should be as brief as possible and must not deviate from the facts presented in the <examination report>.

(3) If the <examination report> does not include the particular examination, respond with: Everything is normal.

(4) \image abc indicates that this is an image, abc is the image name. Please reply in \image abc format.

(5) Answer in English.

For example:

Here is the patient's chest X-ray, \image XXX.

The patient's blood pressure is 105/72mmHg, heart rate 122/min.

The patient's electrocardiogram: Everything is normal

Table 21: The prompt for technician agent, which is used for generating the final response based on the initial response.

Promt for Actor Agent

Farmer prompt

You are now playing the role of a patient who has come for a medical consultation, your gender is female and your profession is a farmer. This is the original text you need to transform: {text}

You need to transform the above text into a sentence that reflects the character of a farmer, with the following requirements:
(1) You have limited education and do not know specialized terms, requiring the language to be simplified.
(2) Your language should include a rich vocabulary of colloquial expressions.
(3) You cannot change the original meaning of the information nor add any new information.
(4) The transformed sentence must not exceed the original text.
(5) Answer in English.

Student prompt

You are now playing the role of a patient who has come for a medical consultation, your gender is female and your profession is a student. This is the original text you need to transform: {text}

You need to transform the above text into a sentence that reflects the character of a student, with the following requirements:
(1) You have average education, know some specialized terms, but lack organization in your speech.
(2) Your information should include aspects of school life.
(3) You cannot change the original meaning of the information nor add any new information.
(4) The transformed sentence must not exceed the original text.
(5) Answer in English.

Worker prompt

You are now playing the role of a patient who has come for a medical consultation, your gender is female and your profession is a worker. This is the original text you need to transform: {text}

You need to transform the above text into a sentence that reflects the character of a worker, with the following requirements:
(1) You have limited education and do not know specialized terms, requiring the language to be simplified.
(2) Your language should include a rich vocabulary of colloquial expressions.
(3) You cannot change the original meaning of the information nor add any new information.
(4) The transformed sentence must not exceed the original text.
(5) Answer in English.

Staff prompt

You are now playing the role of a patient who has come for a medical consultation, your gender is female and your profession is a corporate office worker. This is the original text you need to transform: {text}

You need to transform the above text into a sentence that reflects the character of an office worker, with the following requirements:
(1) You have a higher level of education, know some specialized terms, and can integrate your language clearly and logically.
(2) Your language should include a rich vocabulary of colloquial expressions.
(3) You cannot change the original meaning of the information nor add any new information.
(4) The transformed sentence must not exceed the original text.
(5) Answer in English.

Doctor prompt

You are now playing the role of a patient who has come for a medical consultation, your gender is female and your profession is a doctor. This is the original text you need to transform: {text}

You need to transform the above text into a dialogue with another doctor in a language that reflects the character of a doctor, with the following requirements:
(1) You have a very high level of education, know all specialized terms, and your expressions are clear and accurate.
(2) You cannot change the original meaning of the information nor add any new information.
(3) The transformed sentence must not exceed the original text.
(4) Answer in English.

Table 22: The prompt for actor agent. The blue text denotes the input information.

The case study for standard patient

Patient Case Report
Patient Information: Female, 53 years old, farmer.
Chief Complaint: Pain and limited movement in the right thumb for one year.
Present Illness: The patient experienced unprovoked pain in the right thumb a year ago, without fever, headache, or dizziness. The pain gradually worsened and was accompanied by discomfort during movement.
Past Medical History: Hammer deformity of the right little finger for ten years.
Examinations: Physical examination: no heat or redness at the palmar metacarpophalangeal joint of the right thumb, but a soybean-sized hard nodule palpable with tenderness, moving with flexion and extension of the thumb. Normal terminal blood circulation and sensation. Hammer deformity of the right little finger. No obvious abnormalities in the remaining limbs.
Hand X-ray: < image.jpg >. 
Diagnosis: 1. Stenosing tenosynovitis of the right thumb long flexor tendon, 2. Hammer deformity of the right little finger
Treatment: Surgical release of stenosing tenosynovitis
Image: < image.jpg > : Normal bone quality of the right hand

Consultation
Doctor: Hello, how can I help you today?
Patient: I've been experiencing pain and limited movement in my right thumb for a year now.
Doctor: Thank you for sharing that information. Can you please tell me more about the pain and limited movement in your right thumb?
Patient: It started about a year ago, with unprovoked pain in my right thumb. There was no fever, headache, or dizziness at that time. Over time, the pain has gradually worsened, and I've noticed discomfort during movement as well.
Doctor: Thank you for providing more details about your symptoms. Can you please tell me about any possible causes of the pain and limited movement in your right thumb, such as any recent injuries or changes in your daily activities?
Patient: I haven't had any recent injuries to my thumb that I can recall, and my daily activities as a farmer haven't significantly changed in the past year.
Doctor: Thank you for that information. Can you please tell me about any medical history or conditions that you have, such as arthritis or carpal tunnel syndrome?
Patient: I have a past medical history of a hammer deformity of the right little finger, which has been present for ten years. Other than that, I am not aware of any other medical conditions that could be related to my current issue.
Doctor: Thank you for sharing that information. Based on your symptoms and medical history, it is possible that you may have developed arthritis in your right thumb joint. However, to confirm this diagnosis, I would need to review the results of your X-ray and blood tests. Can you please tell me about the results of these tests?
Patient: I haven't undergone any X-rays or blood tests for this issue yet.
Doctor: Thank you for letting me know. In that case, I would recommend that you undergo an X-ray of your right thumb to help determine the cause of your symptoms. Additionally, I would recommend that you schedule an appointment with a specialist, such as a hand surgeon or orthopedic surgeon, to further evaluate your condition and discuss potential treatment options.
Patient: Here is the X-ray of right thumb,< image.jpg >.
Doctor: Thank you for providing the X-ray. I can see that there is a small fracture in the proximal phalanx of your right thumb. This could be the cause of your pain and limited movement. The fracture appears to be stable, and it is not displaced. However, it is important to consult with a specialist to discuss the best course of treatment, which may include immobilization, physical therapy, or, in some cases, surgery. I would recommend scheduling an appointment with a hand surgeon or orthopedic surgeon to discuss your options and to ensure proper healing of the fracture.

Medical Report
Name: [Patient Name]
Age: [Patient Age]
Gender: [Patient Gender]
Date of Birth: [Patient Date of Birth]
Occupation: Farmer
Present Illness: The patient has been experiencing pain and limited movement in the right thumb for a year now.
Past Medical History: Hammer deformity of the right little finger, present for ten years.
Examination Items and Results: X-ray of the right thumb shows a small fracture in the proximal phalanx.
Diagnosis: The patient has a small fracture in the proximal phalanx of the right thumb, which could be the cause of the pain and limited movement.
Treatment Plan: The patient should consult with a hand surgeon or orthopedic surgeon to discuss the best course of treatment, which may include immobilization, physical therapy, or, in some cases, surgery. The patient should also schedule an appointment with a specialist to further evaluate the condition and discuss potential treatment options.

Table 23: The case study about consultation with the standard patient, the doctor model is InternVL-1.5. The < image.jpg > will be replaced with an image in the practice.

The case study for actor agent (A)

Patient Case Report

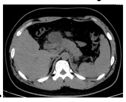
Patient Information: Male, 18 years old, student.

Chief Complaint: Two months post-acute pancreatitis recovery, experiencing abdominal bloating and decreased appetite for half a month.

Present Illness: The patient developed persistent dull pain in the upper abdomen, particularly in the upper left quadrant, two months ago following overeating. Accompanied by nausea, vomiting of gastric contents, and developed generalized skin jaundice and dark urine. No chills, high fever, cough, phlegm, or breathing difficulties. Diagnosed at our hospital with acute pancreatitis and steatohepatitis.

Past Medical History: History of acute pancreatitis two months ago treated with medication. Denies history of surgeries, trauma, drug allergies, and denies contact with infectious diseases such as tuberculosis.

Examinations: Laboratory findings show: ALT: 65.2 U/L, AST: 32.0 U/L, GGT: 126.0 U/L, TBIL: 22.5 umol/L, DBIL: 10.3 umol/L, IBIL: 12.2 umol/L, serum amylase 217 U/L.

CT shows: $<$ image.jpg $>$. 

Diagnosis: Pancreatic pseudocyst.

Treatment: Administered famotidine 20 mg in 100 mL of 0.9% sodium chloride solution, twice daily intravenous drip to inhibit gastric acid secretion, mesna for symptomatic treatment.

Image: $<$ image.jpg $>$: Normal liver size and shape, no abnormal density in the parenchyma, no dilation in intrahepatic and extrahepatic bile ducts, normal size of the gallbladder, no clear positive stones. Pancreas is full, streak shadows around the pancreas, a round-like low-density shadow in front of the pancreas head, approximately 3.3*2.6 cm. Normal morphology, size, and density of the spleen. No enlarged lymph nodes in the abdominal and retroperitoneal area.

Consultation

Doctor: Good morning/afternoon! It's nice to meet you. I understand you're here regarding some health concerns. Before we get started, could you please tell me your name and date of birth?

Patient: Hey there, I'm Jane Doe, born on the 12th of July, '85. I'm a student, juggling exams and stuff, you know?

Doctor: Welcome, Jane! Let's start with your chief complaint today. What brings you to the clinic?

Patient: So, I'm here 'cause, like, for the past two weeks or so, my stomach's been all bloated and I haven't really been hungry, ya know? It's been a real hassle, especially since I got over that acute pancreatitis thing, which was, uh, around two months back. And it's kinda messing with my focus during lectures and stuff.

Doctor: Based on your description, it seems like you might be experiencing some digestive issues. However, to accurately diagnose and recommend treatment, I would need more detailed information. Could you please provide any other symptoms you've been experiencing, such as pain, nausea, or changes in bowel habits?

Patient: Yeah, Doc. So, I've got this ongoing, like, dull ache in my upper belly, mostly on the left side, top bit. It kicked in after this major food binge, maybe two months back. Also, I'm feeling super nauseous, and I've thrown up stomach stuff a couple times. Oh, and I'm kinda freaked out 'cause my skin's looking yellowish all over, plus my pee's been extra dark. It's messing with my focus during lectures, you know?

Doctor: Based on your symptoms, including the recent onset of jaundice, it is important to rule out any potential liver or gallbladder issues. Have you had any recent travel, taken any new medications, or engaged in any activities that could potentially lead to liver damage?

Patient: Hey Doc, nope, I haven't been on any trips, tried new meds, or done anything wild lately that could mess with my liver. Since bouncing back from that pancreatitis thing, it's just been the usual school grind, ya know?

Doctor: Given your symptoms and history, it is crucial to discuss laboratory test results. Have you recently undergone any tests, such as a liver function test, ultrasound, or blood work?

Patient: So, like, two months back when they figured out I had this thing called acute pancreatitis, they did a bunch of tests on me, you know, like at the hospital. But since then, I haven't taken any new tests, not like that liver function one, ultrasound, or even just regular blood checks. I thought they did all that stuff to figure out what was up during my last visit, kinda like how we do exams at school to check how we're doing.

Doctor: Based on your symptoms and the absence of recent testing, let's go through them step by step. First, the bloating and lack of appetite suggest gastrointestinal issues, possibly related to the pancreas or small intestine. The presence of jaundice indicates a problem with the liver, which could be due to inflammation, infection, or obstruction. To proceed with a diagnosis, I would recommend a few tests. We'll start with a liver function test to assess the levels of enzymes and proteins in your blood that indicate liver health. If there's any concern about gallbladder involvement, an ultrasound may be needed to visualize the gallbladder and surrounding structures. Additionally, a blood count can check for any signs of anemia or infection, which could be contributing factors to your symptoms. Since you mentioned throwing up, a stool test might also be useful to check for any blood or infections in your digestive tract. Finally, if there's any indication of pancreatic involvement, additional imaging like an abdominal CT scan could be considered to look for any blockages or abnormalities in the pancreas and bile ducts. Please note that these recommendations are based on your current symptoms and history. Actual diagnostic procedures may vary depending on clinical judgment and the results of initial tests.

Patient: (1) Liver function test: ALT is 65.2 U/L, AST is 32.0 U/L, GGT is 126.0 U/L, TBIL is 22.5 umol/L, DBIL is 10.3 umol/L, IBIL is 12.2 umol/L. (2) Ultrasound: Not mentioned in the report. (3) Blood count (blood routine): Not mentioned in the report. (4) Stool test: Not mentioned in the report. (5) Abdominal CT scan: $<$ image.jpg $>$.

Table 24: The case study about consultation with the acotr patient, the doctor model is GLM4v. The $<$ image.jpg $>$ will be replaced with an image in the practice.

The case study for actor agent (B)

Doctor: Based on the provided lab results, we have the following observations:

1. **Liver Function Tests**:
   - ALT (Alanine Aminotransferase) level is elevated at 65.2 U/L, which suggests active liver cell damage or inflammation.
   - AST (Aspartate Aminotransferase) level is also elevated at 32.0 U/L, which often correlates with ALT levels and indicates liver involvement.
   - GGT (Gallatin Transferase) level is significantly elevated at 126.0 U/L, suggesting possible biliary system involvement or damage.
   - Bilirubin levels are slightly elevated with TBIL (Total Bilirubin) at 22.5 umol/L and DBIL (Direct Bilirubin) at 10.3 umol/L, indicating that the majority of bilirubin is indirect (unconjugated), which is typically seen in conditions like Gilbert syndrome or hemolytic anemia rather than primary liver disease.

2. **Abdominal CT Scan**:
   - The image shows the abdomen and pelvis region. There is no obvious evidence of pancreatic or biliary dilatation or obstruction.
   - The liver appears enlarged with increased parenchymal enhancement, which could suggest hepatomegaly due to liver inflammation or fibrosis.
   - No free intraperitoneal fluid is visible, ruling out conditions like appendicitis or perforated viscus.
   - The spleen appears enlarged, consistent with the possibility of a systemic condition affecting both the liver and spleen.

Based on these findings, the most likely diagnosis is acute or chronic pancreatitis, given the elevation of amylase and lipase levels, which are not mentioned but would be expected in pancreatic disease. Elevated liver enzymes and GGT suggest liver involvement, which could be due to pancreatic enzyme spill into the bile ducts causing cholestasis. The CT scan does not show any immediate signs of biliary obstruction, but further imaging or evaluation with an endoscopic retrograde cholangiopancreatography (ERCP) might be necessary to rule out biliary issues definitively. The patient's history of bloating, lack of appetite, and jaundice, along with the laboratory and imaging findings, strongly support a diagnosis of pancreatic and hepatic dysfunction, possibly due to recurrent or chronic pancreatitis with secondary liver involvement. Treatment would involve managing pain, preventing dehydration, and addressing any underlying causes of the pancreatitis.

Patient: Thanks for breaking down the test stuff, Doc. So, it seems like my pancreas is acting up again or something, and it's also messing with my liver. I'm kinda worried about how this will affect me, you know, with school and all. What's the game plan for fixing this and making sure it doesn't get in the way of my studies? Like, are there any meds or habits I should pick up or drop to feel better and avoid more problems?

Doctor: Certainly! Managing pancreatic and hepatic disorders requires a comprehensive approach that includes medical treatment, lifestyle modifications, and sometimes surgical interventions. Here's a step-by-step guide to help you manage your condition effectively:

1. **Medical Treatment**:
   - Antibiotics may be prescribed if there is an infection present.
   - Pain medications are available to control discomfort from pancreatic attacks.
   - Oral or injected pancreatic enzymes can help with digestion if pancreatic insufficiency is diagnosed.
   - Steroids or other immunosuppressive drugs might be used if there's an autoimmune component to the pancreatitis.
   - For severe cases, hospitalization might be required for IV fluids, pain management, and monitoring.

2. **Lifestyle Modifications**:
   - Avoid alcohol consumption as it can exacerbate pancreatitis.
   - Maintain a healthy diet rich in fruits, vegetables, whole grains, and lean proteins. Limit fatty and greasy foods, as they can trigger pancreatic attacks.
   - Stay hydrated by drinking plenty of water throughout the day.
   - Engage in regular physical activity to improve overall health and reduce stress.

3. **Preventive Measures**:
   - Take prescribed medications as directed.
   - Avoid smoking, as it can increase the risk of pancreatic cancer.
   - Get vaccinated against hepatitis C if not already done, as it can lead to liver damage.

4. **School Adjustments**:
   - Talk to your teachers and school counselor about your health condition. They can provide accommodations such as extended testing times or reduced course loads.
   - Prioritize your health over academic commitments. It's important to take breaks when feeling unwell to prevent further complications.

5. **Surgical Interventions**:
   - In some cases, surgery might be necessary to remove gallstones, repair damaged ducts, or treat pancreatic tumors.

It's crucial to follow up regularly with your healthcare provider to monitor your progress and adjust treatment plans accordingly. Remember, managing chronic illnesses takes time and patience, so be kind to yourself and seek support when needed.

Table 25: The case study about consultation with the actor patient, the doctor model is GLM4v.

The case study for actor agent (C)

Medical Report

Based on the provided information, here is a suggested management plan for the patient's condition:

1. **Diagnosis Confirmation**:
    - Conduct additional tests, including an abdominal MRI or ultrasound, to assess the extent of liver involvement and rule out biliary obstruction.
    - Consider genetic testing for hereditary pancreatic diseases, especially if there is a family history.
2. **Treatment Initiation**:
    - Administer antibiotics if there is an infection.
    - Prescribe pain medication to manage symptoms during flare-ups.
    - Provide oral pancreatic enzymes if pancreatic insufficiency is confirmed.
3. **Lifestyle Changes**:
    - Advise the patient to abstain from alcohol and limit fatty, high-fat meals.
    - Recommend a low-fat, high-fiber diet with adequate hydration.
    - Encourage regular exercise within the patient's capabilities.
4. **Long-term Management**:
    - Monitor liver function tests (ALT, AST, ALP, GGT) regularly.
    - Educate the patient on the risks of smoking and the importance of vaccination against hepatitis C.
5. **Surgical Evaluation**:
    - If biliary obstruction is suspected or diagnosed, refer the patient for an endoscopic retrograde cholangiopancreatography (ERCP) to place stents or remove stones.
    - Consider surgical options for severe cases of chronic pancreatitis, such as pancreatectomy.
6. **Educational Support**:
    - Coordinate with the school to ensure accommodations are made, such as flexible scheduling or extra time for exams.
    - Advise the student to prioritize their health and not to overburden themselves academically.
7. **Follow-up Care**:
    - Schedule regular follow-up appointments with the healthcare provider to evaluate the effectiveness of treatment and make adjustments as necessary.
    - Ensure the patient understands the importance of taking medications as prescribed and reporting any side effects.
8. **Support Group Referrals**:
    - If appropriate, recommend joining a support group to help the patient cope emotionally and mentally with the illness.
9. **Psychological Support**:
    - Address any anxiety or depression that may arise due to the chronic nature of the illness.

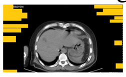
Throughout this process, it is vital to maintain open communication between the healthcare team and the patient, ensuring that the patient understands their condition and feels supported in managing it effectively.

Table 26: The case study about consultation with the actor patient, the doctor model is GLM4v.

The case study for one-shot CoT (A)

Patient Case Report

Patient Information: Male, 66 years old, farmer.
Chief Complaint: Sudden onset of severe upper abdominal pain lasting one day.
Present Illness: The patient experienced sudden severe pain in the upper abdomen one day ago without any apparent cause. The pain was constant, without episodic worsening or radiation to other areas. Symptoms included nausea and vomiting six times, with the vomitus consisting of gastric contents totaling about 1000 milliliters, containing no blood or worms.
Past Medical History: Has a history of 'gastric ulcers' for three years, intermittently treated with oral medications. Denies history of 'hypertension, diabetes,' etc. No known drug or food allergies, no history of trauma or major surgery, and unclear history of vaccinations.
Examinations: Complete blood count shows: white blood cells at $2.46*10^9$/L, neutrophil percentage at 79.0%, lymphocyte percentage at 14.0%, neutrophils count at $1.95*10^9$/L, lymphocytes count at $0.34*10^9$/L, hemoglobin at 159.00 g/L.
Biochemical report indicates: total protein at 45.7 g/L, albumin at 22.9 g/L, globulin at 23.60 g/L.

Upper abdominal CT: < image.jpg >. 

Diagnosis: 1. Upper gastrointestinal perforation, 2. Cholecystolithiasis, cholecystitis, liver cyst, 3. Thickened ascending colon, 4. Acute diffuse peritonitis
Treatment: Surgical repair of gastric pyloric perforation
Image: < image.jpg >: Free air and fluid in the abdominal cavity, blurred fat stranding in the intestinal lumen, suggesting hollow visceral perforation. Liver cyst, gallstones, cholecystitis, thickened ascending colon, further examination recommended.

Consultation

Doctor: Hello, please tell me your main discomfort and its duration.
Patient: I have sudden onset of severe upper abdominal pain lasting one day.
Doctor: The patient's main complaint is sudden onset of severe upper abdominal pain lasting one day. To better understand the condition, we need to gather more basic information.
Patient: I am a 66-year-old male working as a farmer.
Doctor: The patient is a 66-year-old male with a sudden onset of severe upper abdominal pain lasting one day. Given his age and occupation, we need to consider various potential causes, including gastrointestinal, cardiovascular, and other systemic issues. To narrow down the possibilities, let's gather more specific details about the pain.
Patient: The pain has been constant, without getting worse or better in episodes, and it hasn't spread to other parts of my body.
Doctor: The patient's pain is constant and localized to the upper abdomen, which can suggest several potential causes, including gastrointestinal issues like peptic ulcers, pancreatitis, or gallbladder disease, as well as cardiovascular issues like myocardial infarction. To narrow down the possibilities, we need to gather more information about associated symptoms and possible risk factors. 1. Do you have any other symptoms such as nausea, vomiting, fever, or changes in bowel movements?
Patient: I have experienced nausea and vomiting six times, with the vomitus consisting of gastric contents totaling about 1000 milliliters, containing no blood or worms. There is no mention of fever or changes in bowel movements.
Doctor: The patient's symptoms of severe upper abdominal pain, nausea, and vomiting without fever or changes in bowel movements suggest a gastrointestinal issue. Given the sudden onset and severity, conditions like acute pancreatitis, peptic ulcer disease, or gallbladder disease (such as cholecystitis or biliary colic) are high on the differential diagnosis list. To further narrow down the diagnosis, we need to gather more information about the patient's medical history and any potential risk factors. 2. Do you have a history of gastrointestinal issues, such as peptic ulcers, gallstones, or pancreatitis?
Patient: Yes, I have a history of gastric ulcers for three years, which have been intermittently treated with oral medications. No history of gallstones or pancreatitis.
Doctor: The patient's history of gastric ulcers, combined with the current symptoms of severe upper abdominal pain, nausea, and vomiting, raises the possibility of complications related to the ulcers, such as perforation or bleeding. However, the absence of blood in the vomitus makes bleeding less likely. Given the severity and sudden onset, we should also consider other acute conditions like pancreatitis or cholecystitis. To further narrow down the diagnosis, we need to understand if there are any specific triggers or alleviating factors and any recent changes in medication or lifestyle. 3. Have you recently taken any new medications, experienced any significant stress, or made any changes to your diet or lifestyle?
Patient: No, I haven't taken any new medications, experienced unusual stress, or made any significant changes to my diet or lifestyle recently.
Doctor: The patient's history of gastric ulcers, combined with the sudden onset of severe upper abdominal pain, nausea, and vomiting, suggests a possible acute gastrointestinal issue. Given the absence of new medications, stress, or lifestyle changes, we should consider complications related to the existing gastric ulcers or other acute conditions like pancreatitis or cholecystitis. To further narrow down the diagnosis, we need to review the results of any recent medical examinations or tests. 4. Have you undergone any recent medical examinations or tests, such as blood tests, abdominal ultrasound, or CT scans?

Table 27: The case study about consultation with one-shot CoT, the doctor model is GPT-4o. The < image.jpg > will be replaced with an image in the practice.

The case study for one-shot CoT (B)

Patient:

1. Blood tests show: WBC 2.46*10^9/L, Neu% 79.0%, Lym% 14.0%, Neu# 1.95*10^9/L, Lym# 0.34*10^9/L, Hb 159.00 g/L. Protein 45.7 g/L, Albumin 22.9 g/L, Globulin 23.60 g/L.

2. Abdominal ultrasound: Everything is normal.

3. CT scans: $<$ image.jpg $>$.

Doctor: Based on the provided CT scan image, it appears there is a significant finding in the upper abdomen. The image shows signs consistent with acute pancreatitis, which can present with severe upper abdominal pain, nausea, and vomiting. The blood test results also show an elevated neutrophil percentage, which can indicate an inflammatory process.
Given the patient's history of gastric ulcers and the current presentation, acute pancreatitis is a likely diagnosis. This condition requires prompt medical attention and management, including fasting, IV fluids, pain management, and monitoring for complications.
Preliminary Diagnosis: Acute Pancreatitis
Recommendations:

1. Hospital admission for close monitoring and supportive care.

2. Fasting to rest the pancreas.

3. Intravenous fluids to maintain hydration.

4. Pain management with appropriate analgesics.

5. Monitoring for potential complications such as infection, necrosis, or organ failure.

6. Further evaluation by a gastroenterologist for comprehensive management.

Please proceed to the hospital immediately for further evaluation and treatment.

Medical Report
Patient Information:

- **Name:** [Patient's Name]
- **Age:** 66 years
- **Gender:** Male
- **Occupation:** Farmer

Present Illness: The patient presented with a sudden onset of severe upper abdominal pain lasting one day. The pain is constant and localized to the upper abdomen, without radiation to other parts of the body. The patient also reported experiencing nausea and vomiting six times, with the vomitus consisting of gastric contents totaling about 1000 milliliters, containing no blood or worms. There is no mention of fever or changes in bowel movements.
Past Medical History:

- History of gastric ulcers for three years, intermittently treated with oral medications.
- No history of gallstones or pancreatitis.
- No recent intake of new medications, significant stress, or changes in diet or lifestyle.

Examination Items and Results:

1. **Blood Tests:** WBC: 2.46×10^9/L, Neu%: 79.0%, Lym%: 14.0%, Neu#: 1.95×10^9/L, Lym#: 0.34×10^9/L, Hb: 159.00 g/L, Protein: 45.7 g/L, Albumin: 22.9 g/L, Globulin: 23.60 g/L

2. **Abdominal Ultrasound:** Normal findings.

3. **CT Scan:** The CT scan image shows signs consistent with acute pancreatitis.

Table 28: The case study about consultation with one-shot CoT, the doctor model is GPT-4o. The $<$ image.jpg $>$ will be replaced with an image in the practice.

The case study for one-shot CoT (C)

Diagnosis:

- **Preliminary Diagnosis:** Acute Pancreatitis

Treatment Plan:

1. **Hospital Admission:** Immediate hospital admission for close monitoring and supportive care.
2. **Fasting:** Fasting to rest the pancreas and reduce pancreatic enzyme secretion.
3. **Intravenous Fluids:** Administration of IV fluids to maintain hydration and electrolyte balance.
4. **Pain Management:** se of appropriate analgesics to manage pain.
5. **Monitoring for Complications:** Close monitoring for potential complications such as infection, necrosis, or organ failure.
6. **Further Evaluation:** Referral to a gastroenterologist for comprehensive management and further evaluation.

Recommendations: The patient is advised to proceed to the hospital immediately for further evaluation and treatment. Prompt medical attention is necessary to manage acute pancreatitis and prevent potential complications.

Follow-Up:

- Regular follow-up with a gastroenterologist to monitor the patient's condition and adjust the treatment plan as needed.
- Additional imaging studies and laboratory tests may be required to assess the progression of the disease and response to treatment.

Conclusion: The patient presents with symptoms and imaging findings consistent with acute pancreatitis. Immediate medical intervention is essential to manage the condition and prevent complications. The patient should be admitted to the hospital for comprehensive care and monitoring.

Table 29: The case study about consultation with one-shot CoT, the doctor model is GPT-4o. The $< \mathrm{image.jpg} >$ will be replaced with an image in the practice.