Integrating Physician Diagnostic Logic into Large Language Models: Preference Learning from Process Feedback

Chengfeng Dou[1,2], Ying Zhang[3], Zhi Jin[1,2]✉, Wenpin Jiao[1,2]✉, Haiyan Zhao[1,2], Yongqiang Zhao[1,2], Zhengwei Tao[1,2]

[1] School of Computer Science, Peking University;

[2] Key Laboratory of High Confidence Software Technologies(PKU), MOE, China

[3] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China

{chengfengdou,zhijin,jwp,zhhy.sei}@pku.edu.cn

{tttzw,yongqiangzhao}@stu.pku.edu.cn {19112043}@bjtu.edu.cn

Abstract

The utilization of large language models for medical dialogue generation has attracted considerable attention due to its potential to enhance response richness and coherence. While previous studies have made strides in optimizing model performance, there is a pressing need to bolster the model's capacity for diagnostic logic to ensure patient safety. In response to this need, we propose an approach termed preference learning from process feedback (PLPF), which involves integrating the doctor's diagnostic logic into LLMs. PLPF encompasses three key components: rule modeling, preference data generation, and preference alignment. These components collectively serve to train the model to adhere to the diagnostic process. Our experimental results, utilizing Standardized Patient Testing, demonstrate that PLPF enhances the diagnostic accuracy of the baseline model in medical conversations by 17.6%, surpassing the performance of traditional approaches. Moreover, PLPF exhibits effectiveness in both multi-round and single-round dialogue tasks, thereby highlighting its potential in improving medical dialogue generation. Our dataset is available at https://github.com/Chengfeng-Dou/SpTesting

1 Introduction

The use of large language models (LLMs) (Zhao et al., 2023) has recently exploded in the field of medical dialogue generation. However, training robust medical dialogue models is crucially based on high-quality training data (He et al., 2023). As a result, considerable efforts have been made to generate extensive training data sets to fine-tune these models. Furthermore, certain studies have made notable progress, such as the application of reinforcement learning from human feedback (RHLF) to guide models in generating user-friendly responses (Chen et al., 2023; Bao et al., 2023).

Despite the significant advancements in prior research on RLHF, current open-source medical LLMs remain inadequately user-directed. They frequently oscillate between two problematic behaviors: 1) precipitously rendering a diagnosis without adequate patient data collection, and 2) becoming ensnared in an endless loop of data gathering without progressing to a definitive diagnosis. This issue appears to originate from the inherent discordance among the preference data provided by annotators. In practical settings, physicians' diagnostic approaches typically bifurcate into: 1) an initial comprehensive collection of patient information followed by a diagnosis, and 2) an early presentation of a conjectural diagnosis, which is then substantiated through progressive data accumulation. It is the simultaneous learning from these two divergent preferences that likely precipitates the disarray in the model's conversational strategy.

We believe that standardising the medical dialogue process is the key to solving the above problems. Our proposed approach, known as preference learning from process feedback (PLPF), focuses on ensuring the rationality of the conversation flow, which sets it apart from traditional methods. The core idea is to represent the doctor's diagnostic logic using a flowchart and employ preference learning to train the model to avoid generating responses that deviate from the established process.

In particular, we have developed a flowchart based on the doctor's expertise, as depicted in Fig. 1. This flowchart effectively outlines the objectives the physician must achieve and the constraints that must be followed during the diagnostic process, while also illustrating the dependencies between these objectives. To utilize the flowchart for guiding model training, we have established explicit rules for each activity, decision, and constraint outlined in the flowchart. The state of a dialogue in the flowchart can be determined by evaluating whether the dialogue conforms to these rules.

Based on these established rules, our approach consists of three phases: rule modeling, preference

Findings of the Association for Computational Linguistics: ACL 2024, pages 2453–2473
August 11-16, 2024 ©2024 Association for Computational Linguistics

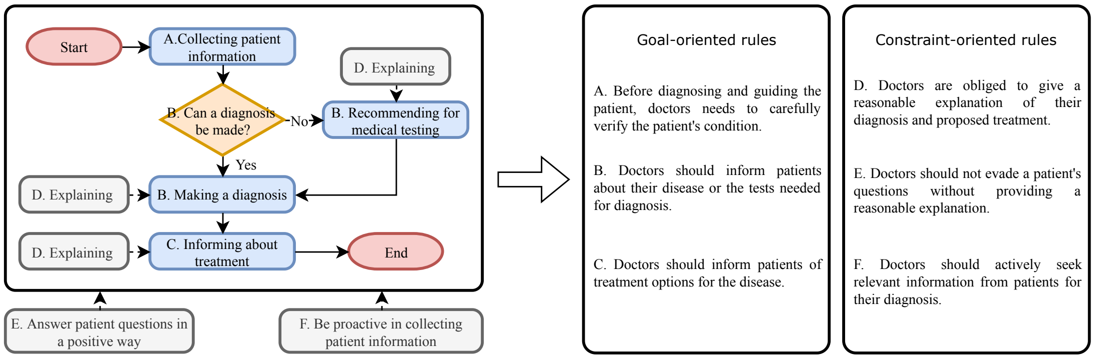

Figure 1: Medical diagnosis flowchart (left) and its corresponding rules (right). In the flowchart, we use blue boxes for activities, orange diamonds for judgment conditions, and gray boxes for additional constraints. We use the letters A-F to indicate the correspondence between the rules and the elements in the flowchart.

data generation, and preference alignment. Initially, we have developed a Rule Evaluation Model (REM) to automatically assess whether a conversation adheres to a specific rule. Building on this, we have devised a method to assign scores to conversations that take into account rule dependencies. These scores are then used for the generation of preference data. We achieve this using an innovative one-shot learning-based approach to predict dialogue trajectories and leveraging REM to appropriately rank these trajectories. Subsequently, we employ the Direct Preference Optimization (DPO) algorithm (Rafailov et al., 2023) to train models based on the acquired preference data.

We used standardized patient testing, a widely accepted method in the medical field, to assess our approach. To achieve this, we built the Chinese Standardized Patient Test (CSPT) dataset. Furthermore, we employed a retrieval-augmented generation technique to create a patient simulator for interactive testing with LLM. The results of our experiments indicate that our approach improves the diagnostic precision of the baseline model in medical conversations by 17.6%. We also tested our approach on three public datasets to assess its performance in both multi-round and single-round conversations. The results show that our approach effectively enhances the model's understanding of physician expressions. In summary, our work provides the following contributions.

- We introduce PLPF, a method for multi-round healthcare conversations that allows LLMs to incorporate industry flowchart specifications to improve conversation logic.
- We provide a high-quality evaluation dataset for standardized patient testing, offering a novel approach to evaluate the communication skills of medical LLMs.
- We demonstrate the superiority of PLPF in improving patient diagnostic accuracy through standardized patient testing.

2 Method

2.1 Overview

The overall training process is depicted in Fig. 2, and consists of three phases: *Rules Modeling*, *Preference Data Construction*, and *Human Preference Alignment*. In the first phase, we establish the corresponding rules using the flowchart, which are then employed by manual annotators to generate a rule evaluation dataset. Afterward, a Rule Evaluation Model (REM) is developed by training on this dataset. Moving on to the second phase, we initially utilize the REM to filter the training data, ensuring the acquisition of high-quality data. Subsequently, preference datasets are constructed based on the retained data, using ChatGPT and REM. Finally, we employ these preference datasets to train the model, which has been fine-tuned with instruction data, resulting in the final model. Each stage will be described in detail in the following sections.

2.2 Rules Modeling

2.2.1 Rules Definition

In order to evaluate whether the dialogue follows a specific process, as illustrated in Fig. 1, we need to assess it in two ways: firstly, to determine if the doctor follows the correct sequence to accomplish the goals outlined in the flowchart, and secondly, to verify if the doctor complies with the constraints. To accomplish this, we need to establish specific rules. We have developed a total of six rules for

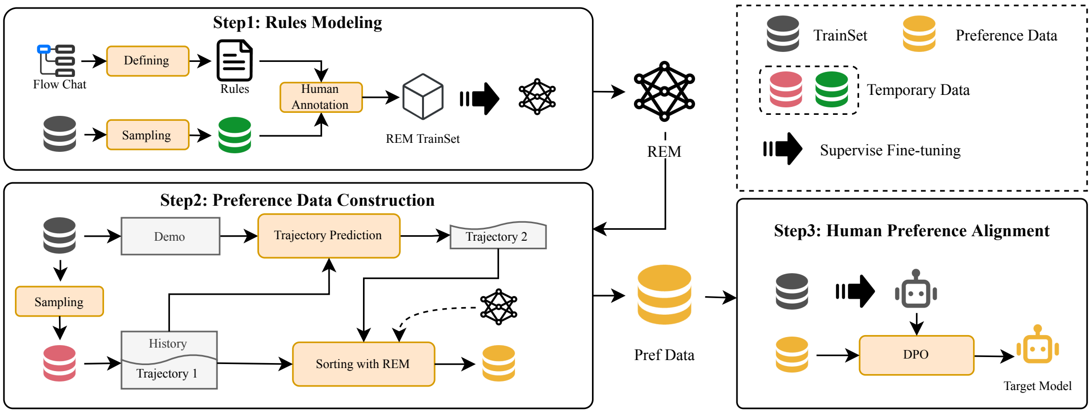

Figure 2: Overview of the training process. The training process is divided into three steps, with key activities indicated using orange rounded rectangular boxes. To distinguish the different stages of the data, we labeled them with different colors and provided data descriptions in the upper right corner of the image.

the flowchart, as depicted in Fig. 1. It is important to note that we have categorized these rules into goal-oriented and constraint-oriented rules based on their distinct functions. Goal-oriented rules assess whether the doctor achieves the specified goals, while constraint-oriented rules evaluate how well the goals are achieved. For the sake of clarity in subsequent discussions, we denote the set of these two types of rules as $\mathcal{R}^g$ and $\mathcal{R}^c$.

2.2.2 Sampling and Human Annotation

Once predefined rules are established, it is essential to develop a Rules Evaluation Model (REM) for these rules. The rule evaluation task is structured as a Q&A format, as shown below.

Human: Rule: [Rule]. History: [History]. Did the doctor follow the rule during the conversation? Assistant: [Comment]. Score: [Score].

In this template, [Rule], [History], [Comment], and [Score] denote slots to be filled, where [Rule] and [History] are model inputs, while [Comment] and [Score] are model outputs. For this task, we construct a small training set with hand-crafted annotations to train REM. In practice, we first collect 400 online medical consultation dialogues and randomly select dialogue segments at different stages. Subsequently, each data instance is scored and commented on by at least three annotators based on the predefined rules. To simplify the scoring process, we define the score values as 0, 1, and 2, representing non-compliance, partial compliance, and complete compliance with the rule, respectively. For more information on scoring, please see Appendix A. In the end, we obtained a total of 2,400 samples, which were divided into 1,800 for training and 600 for testing.

During the model training phase, we utilize an auto-regressive training method. It is important to note that when calculating the loss function, we only compute the loss value of tokens that appear after the string 'Assistant'.

2.3 Preference Data Construction

2.3.1 Basic Idea

In this section, we present a general overview of how REM can be used to guide model training. Our goal is to train the model to follow a specific conversational process when interacting with patients. However, it is important to note that real-life conversations may require deviations from this process to address the patient's needs. Therefore, relying solely on the REM score predicted as a reward for reinforcement learning may result in the generation of responses that lack fluency and coherence. To address this issue, we have chosen to employ a contrastive learning-based approach for model training. Specifically, we provide the model with two different candidate responses for the same conversation history, both of which should be fluent and reasonable. We then use REM to guide the model toward learning the response that aligns more closely with the conversational flow, while rejecting the other response. In the following sections, we will provide more details on how candidate responses are generated and how REM is utilized in this process.

2.3.2 Candidate Responses Generation

The medical conversation task is complex and consists of multiple stages, making it challenging to achieve all objectives in a single round of conversation. Because there are dependencies between goals, evaluating candidate responses requires integrating information from future rounds of conversations. Therefore, in addition to producing candidate responses, it is necessary to generate dialog trajectories for future rounds of conversations, which can aid in subsequent evaluation. Traditional RLHF-based approaches typically generate candidate responses using a model that has undergone fine-tuning. However, this model has limitations in generating future dialog trajectories as it can only predict one step of future responses. As a result, we incorporate data sampling and trajectory prediction to generate candidate responses.

Data Sampling. Sampling from web dialogue data is the most direct approach to acquiring candidate responses. We selected 4,000 samples from the MedDialogue dataset in a random manner. Subsequently, we randomly divide the dialog records of these samples into two sections, enabling us to capture the conversation history, doctor's responses, and future interactions simultaneously.

Trajectory Prediction. Trajectory prediction is another way to generate candidate responses. In this research, we use ChatGPT for this purpose. We observed that ChatGPT's training dataset probably includes MedDialogue. The trajectories generated by ChatGPT closely resemble the actual dataset. To distinguish between the two conversation trajectories, we intervene in ChatGPT's prediction process using a one-shot learning approach. We utilize the following instruction templates.

You are a dialogue continuation AI, please read the below two dialogues and follow my instructions.
Dialogue A: [Arbitrary medical dialogue].
Dialogue B: [The dialogue to be completed].
Please continue Dialogue B while fulfilling the following requirements:
1. The doctor's style should match the doctor's style in Dialogue A.
2. The patient's style should match the patient's style in Dialogue B.

2.3.3 Sorting with REM

After acquiring the candidate responses and their respective future conversation trajectories, the REM is utilized to score each candidate response for ranking purposes. To fully utilize the information from future conversation trajectories, the following formula is adopted for computing the score.

$$s(c \mid h) = v(h, c) + \sum_{i=1}^{n} d^i v(h, c, ..., u_i, a_i) \quad (1)$$

In the provided equation, the variable s represents the score of candidate responses, while v represents the score of the conversation states, which is implicitly embedded in the conversation history and the doctor's responses. The variable h represents the conversation history, c represents the candidate response, and $u_1, a_1, ..., u_i, a_i$ represents the conversation trajectory for the next i rounds. In this context, u means a patient's statement and a means a physician's statement. The variable n acts as the upper limit for the length of the trajectory of interest. The discount factor $d \in (0, 1]$ indicates the level of importance assigned to the future impact.

Next, we will show the process for evaluating conversation states. To ensure clarity, we introduce the variable h' to represent any conversation history that ends with a doctor's response. The corresponding status score for that conversation history can be determined using the following formula:

$$v_{h'}^r = \frac{1}{k} \sum_k \text{REM}(h', r) \quad (2)$$

$$v(h') = \sum_{r \in \mathcal{R}^g \cup \mathcal{R}^c} w_r v_{h'}^r \quad (3)$$

In the above equations, r denotes the rule, and due to the somewhat random nature of scoring the REM, we make k predictions and calculate the average as the final score. The weight of a rule, denoted w_r, is designed to indicate the order and level of goal accomplishment. For constrained rules, since the order is not considered, we assign a constant weight, denoted γ, which is set to a value close to 0 to emphasize the importance of prioritizing goal satisfaction over constraint satisfaction. For goal-oriented rules, we use the following formula to compute the weights:

$$w_{r \in \mathcal{R}^g} = \prod_{r' \prec r} \mathcal{V}_{\alpha}^{t_1}(v_{h'}^{r'}) \prod_{r' \to r} \mathcal{V}_{\beta}^{t_2}(v_{h'}^{r'}) \quad (4)$$

In Equation 4, $r' \prec r$ indicates that r' is a predecessor rule to r, while $r' \to r$ indicates that r' is a constraint rule for r. For example, in Figure 1, Rule A is a predecessor rule to Rule B, Rule D is a constraint rule for Rules B and C, and Rules E and F are constraint rules for all goal-oriented rules. The function $\mathcal{V}_y^t(x), y \in [0, 1)$ represents a value function that equals 1 when $x \geq t$, and y otherwise. We assign a value close to 0 to α to indicate that r is less likely to be satisfied if its predecessor rule is not satisfied. Similarly, we assign a value close to 1 to β to indicate that r is still partially credible when its constraints are not satisfied.

2.4 Human Preference Alignment

In this subsection, we describe the process of training the model using preference data. The training process consists of two steps. Firstly, we fine-tune the base model using the dialogue dataset to enhance the model's medical conversation capabilities. In the second step, we utilize the DPO algorithm (Rafailov et al., 2023) to help the model learn from the preference data. The objective of this algorithm is to minimize the following expression:

$$\min_{\pi} -\mathbb{E}[\log \sigma(\lambda \log \frac{\pi(o_c \mid h)\pi_{\text{sft}}(o_r \mid h)}{\pi_{\text{sft}}(o_c \mid h)\pi(o_r \mid h)})] \quad (5)$$

We label the fine-tuned model with the instructions as π_{sft} and the model needing optimization with the same initial parameters as π_{sft} as π. o_c represents selected responses and o_r represents rejected responses. The sigmoid activation function is denoted as σ, and $\lambda \in (0, 1)$ determines the difference between π and π_{sft}, with a smaller λ resulting in a larger difference.

Statistic	Item	Value
Count	Department Num	5
	Case Num	72
Avg Length	Patient Info	493.3
Avg Num	QA pairs	37.3
	Major Symptoms	7.3
	Major Medical Test	2.8
	Diseases	1.7

Table 1: Statistics for the CSPT dataset. Patient Information describes the patient, while QA pairs represent doctor-patient questions and answers in dialog scripts, used to create simulated patients. Major Symptoms, Major Medical Test, and Diseases are also used to evaluate the model's dialog capability.

3 Experiments

3.1 Standardized Patient Test

In the realm of medicine, Standardized Patients (SP) imitate genuine patient symptoms and reactions following adequate training. Their portrayal of patient responses must be consistent and precise. When undergoing evaluation, standardized patients typically adopt a non-active communication approach, refraining from actively conveying information to the physician. This approach is employed to assess the physician's communication skills. The development of LLMs has made it feasible to employ computer-simulated standardized patients. Some previous studies (Zhang et al., 2023; Wei et al., 2018) have aimed to evaluate the performance of models using similar approaches. However, these studies usually provide the modeled patients with limited symptom information, which often hinders the model's ability to accurately comprehend the patient's interaction with the doctor. In the medical field, this challenge is commonly addressed by instructing the standardized patient to memorize a detailed dialogue script, crafted by a professional, that realistically showcases the patient's responses to various inquiries. During the examination, the patient can then respond to the doctor in accordance with the script, ensuring the quality of the responses.

We created the Chinese Standardized Patient Testing (CSPT) [1] dataset on the basis of this idea. To simulate patients, we propose using the retrieval -augmented generation technique and inputting patient descriptions and dialogue scripts into a database. We used patient cases from the book "Objective Structured Clinical Examinations & Standardized Patients" (WeiGuo Dong, 2012) for this purpose. The primary focus of our SP test is to gather key symptoms and medical tests, as well as accurately diagnose diseases. To assist in this, we have provided a reference list for each case. Table 1 showcases the dataset statistics, and the engineering implementation of the patient simulator can be found in Appendix B.

3.2 Test Settings

The evaluation involves a simulated interaction between a model and a patient simulator. Two doctors assess the interaction using a predetermined checklist. The model's performance is measured as a percentage based on the successful completion of checklist items. The assessment procedure includes limiting the dialogue rounds to five and requiring the model to inquire about symptoms, provide a diagnosis, and propose a treatment plan within this time frame. It is worth mentioning that all models

[1] We intend to release this dataset to the public in the future.

Model	Internal Medicine			Surgery			Other			ALL		
	Sym.	Test	Dis.	Sym.	Test	Dis.	Sym.	Test	Dis.	Sym.	Test	Dis.
Baichuan-Chat	22.0	34.8	46.4	25.8	18.1	34.1	17.7	29.3	22.4	21.7	27.4	33.8
ChatGLM3	2.4	25.3	28.9	0.0	30.4	28.9	6.1	15.9	12.2	3.0	23.6	22.9
Huatuo-II	2.7	43.5	34.1	9.0	51.4	50.7	5.5	26.2	37.8	5.7	39.8	40.7
DISC-MedLLM	25.5	30.2	50.0	30.6	45.3	45.3	19.0	21.2	30.1	24.8	34.9	41.3
SFT (Qwen)	13.4	32.2	39.9	16.2	40.9	35.5	17.7	18.6	26.9	15.9	30.1	33.8
SFT (Baichuan)	17.2	30.9	40.1	14.6	38.9	**57.2**	2.0	11.5	21.8	10.8	26.3	39.1
PLPF (Qwen)	19.8	**47.8**	**53.6**	**29.5**	**57.2**	**57.2**	**28.0**	**29.2**	**52.6**	**25.9**	**44.1**	**54.4**
PLPF (Baichuan)	**24.8**	**46.7**	**64.5**	28.7	**59.4**	**66.7**	**19.2**	20.0	**41.0**	24.1	**41.1**	**56.7**

Table 2: The experiment results on the CSPT dataset. The Symptom (Sym.) and Test metrics indicate the probability of the model identifying key symptoms and key medical tests, respectively, while the Diagnosis (Dis.) metric indicates the probability of the model making a correct diagnosis. We use red and green labels to denote the best and second-best results, respectively.

are evaluated under identical conditions, with a decoding temperature of 0, to ensure consistency in the assessment process.

3.3 Baselines and implementation details

We utilized different models as baselines, which can be classified into three groups: 1) Chat LLMs, such as ChatGLM3 (6B) (Du et al., 2022; Zeng et al., 2022) and Baichuan2-Chat (7B) (Yang et al., 2023); 2) Medical LLMs, including DISC-MedLLM (Bao et al., 2023) and Huatuo-II (Chen et al., 2023); and 3) Instruction-tuned LLMs constructed on different backbones, specifically SFT (Qwen (Bai et al., 2023)) and SFT (Baichuan). The instruction data used is the same as that of DISC-MedLLM. We implemented the PLPF model based on the SFT model. To ensure a fair comparison with DISC-MedLLM, we used an equal amount of data for preference learning. For additional information regarding baselines and the specific hyperparameters used for training the model, please consult Appendix C and D.

3.4 Experimental Results

Table 2 shows the overall results. To provide a comprehensive understanding of different models, we utilized REM [2] to assess the compliance level of these models with the rules depicted in Fig. 1. We then generated a radar chart, as shown in Fig. 3, which represents the ranking of compliance scores.

ChatGLM3 and Huatuo-II perform poorly in symptom collection, indicating a lack of active information request from the patient during communication. However, despite this limitation, Huatuo-II has the ability to recommend numerous medical

[2] Appendix E contains the performance evaluation of REM.

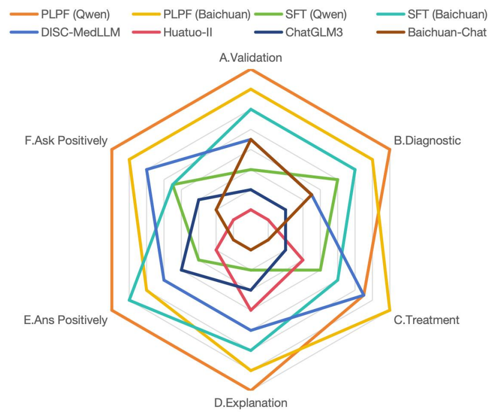

Figure 3: Ranking of how well each model follows the rules. Each axis of the radar graph corresponds to a rule in Fig. 1, and we use the letters A-F to denote the mapping between rules and axes.

tests to the patient during the dialogue. By analyzing the results of these tests, Huatuo-II still achieves a high rate of correct diagnosis. On the other hand, Baichuan2-Chat and ChatGLM3 have lower diagnostic accuracy because they often violate Rule E (Ans Positively) when interacting with the user, rejecting the diagnosis by stating that they are AI models. SFT (Qwen) and SFT (Baichuan) score moderately on the indicators, placing them in the middle range in terms of adherence to the rules. Among these, SFT (Qwen) violates Rule A (Validation) and Rule E (Ans Positively) more frequently, leading to lower correctness in disease diagnosis compared to SFT (baichuan).

We will now examine DISC-MedLLM, which employs the same base model and fine-tuning dataset as our approach but differs in the preference data utilized. As shown in Figure 3, DISC-MedLLM is more proactive in requesting information from patients and offering a wider range of treatment options compared to SFT (Baichuan). However, the model exhibits less confidence in making diagnoses, as evidenced by its lower adherence to Rule B (Diagnose) and Rule E (Ans Positively). This limits the model's ability to effectively utilize its advantage of requesting more patient information, resulting in only a slightly higher correct diagnosis rate compared to SFT (Baichuan). In contrast, the PLPF-optimized model shows a significant improvement of more than 10 points in the identification of symptoms, medical tests, and diseases compared to the SFT model, which emphasizes the effectiveness of our approach. Please refer to Appendix F for case studies.

4 Analyze

4.1 Ablation Study

In this subsection, we will further investigate how the scoring method (Eq. 4) and the trajectory prediction step (Sec. 2.3.2) affect the model's ability to engage in multi-round conversations. To validate Eq. 4, we compare it with the weighting method by directly setting $w_r = 1$. Additionally, we vary the trajectory length to 1, 2, and 3 to assess the effectiveness of the trajectory prediction step. A trajectory length of 1 means that we only predict the immediate doctor responses.

We conducted studies using SFT (Baichuan) and presented the results in Table 3. Based on the results, we observed that our method performs better than the direct summation of all the rule scores. The direct summation approach leads to a model that lacks proficiency in inquiring about patients' symptoms, resulting in decreased diagnosis accuracy. This is because the rule encouraging symptom collection is only one of six rules, and directly adding up rule scores would diminish its impact. Additionally, increasing the trajectory prediction length helps the model understand the entire conversation flow. Specifically, extending the trajectory length from 2 to 3 resulted in a significant improvement in all aspects of the model's capabilities.

4.2 Performance on Public Datasets

In this section, we report the results of all models on the public datasets. We selected the Meddg (Liu et al., 2022), Imcs (Chen et al., 2022), and WebMedQA (He et al., 2019) datasets to test the models' performance. The first two datasets are multi-turn dialogue datasets, while the last dataset consists of single-turn dialogue datasets.

Strategy		**Trajectory Length**		
		k=1	*k=2*	*k=3*
Avg	Sym.	5.1	9.5	19.7
	Test	40.7	28.9	31.3
	Dis.	43.7	36.3	42.6
Ours	Sym.	18.9	21.6	**24.1**
	Test	28.0	32.0	**41.1**
	Dis.	47.7	49.5	**56.7**

Table 3: Ablation test results.

Model	Meddg ↓	Imcs ↓	wMedQA ↓
Baichuan-Chat	1.69	1.77	1.44
ChatGLM3	1.66	1.73	1.23
Huatuo-II	1.60	1.73	**1.11**
DISC-MedLLM	1.59	1.63	1.21
SFT (Qwen)	1.60	1.71	1.26
SFT (Baichuan)	1.67	1.71	1.26
PLPF (Qwen)	**1.56**	**1.59**	**1.14**
PLPF (Baichuan)	**1.53**	**1.60**	1.19

Table 4: Performance of the models on the Meddg, Imcs and WebMedQA dataset. We use red and green to highlight the best and second-best scores.

During the evaluation of the model, we noticed a significant difference in the length of the content generated by the model compared to the standard answer, as shown in Fig. 4. As a result, traditional statistical measures like BLEU and ROUGE were not effective in evaluating the quality of the model's output. For more details, please refer to Appendix G. We believe that the main focus for LLMs should be their ability to produce text that implies the standard answer, as this ensures the accuracy and dependability of the model's output. In accordance with this viewpoint, we have developed a new evaluation metric called GPT-Distance, which measures the extent to which the LLM output implies the standard answers. To be more specific, we utilized GPT-4 to determine whether the predictions imply the references, categorizing the level of implication as not implied, partially implied, or fully implied. The prompt used for this assessment is provided below:

Sentence 1: [predict]; Sentence 2: [reference]
Please decide if sentence 1 implies sentence 2?
A. Fully; B. Partially; C. Not.

Subsequent to obtaining all predicted classifications, we calculate the GPT-Distance using the formula $(2 \times |\text{Not}| + |\text{Partially}|)/|\text{ALL}|$, where $|\text{Not}|$ and $|\text{Partially}|$ denote the number of samples cat-

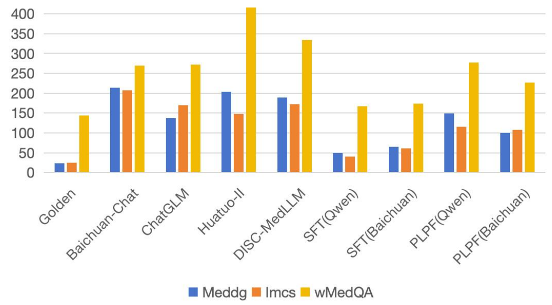

Figure 4: The average output length of LLMs over different datasets, where Goden represents the average length of the standard answer.

egorized as not-implied and partially-implied, respectively. |ALL| indicates the number of test data.

We randomly selected 200 samples from each of the three datasets for testing, and the results are shown in Table 4. Our findings demonstrate that the models trained by PLPF yielded the most favorable results for multi-round dialogs. Although Huatuo-II and DISC-MedLLM are able to generate longer responses, it is obvious that our generations have a higher coverage of physician responses, suggesting that PLPF allows the model to better understand the physician's diagnostic logic. In the context of the single-round dialog task, Huatuo-II emerged as the top performer, with our model securing the second and third positions, respectively. This illustrates that while PLPF is primarily designed for multi-turn dialogue tasks, it also exhibits strong optimization for single-turn dialogue tasks.

5 Related works

Medical LLMs. Significant advancements have been made in medical dialogue models since the introduction of ChatGPT (OpenAI, 2023). Recent research has primarily focused on creating large and high-quality instruction fine-tuning datasets for LLMs. Studies such as DoctorGLM (Xiong et al., 2023), BenTsao (Wang et al., 2023b), and ChatMed (Zhu and Wang, 2023) have proposed the utilization of powerful LLMs like ChatGPT to generate dialogue and question answering data at a low cost. However, ensuring the quality of data generated through this approach is challenging due to ChatGPT's tendency to generate incorrect information. To tackle this issue, Huatuo (Zhang et al., 2023) suggested incorporating parts of real data into the generated data. Furthermore, to improve the readability of the real dialogue data, Huatuo refined it with ChatGPT, and this method has been widely adopted by subsequent research. In addition to the dialogue data, several studies have aimed to generate various auxiliary task data. For instance, DISC-MedLLM(Bao et al., 2023) and ClinicalGPT(Wang et al., 2023a) have integrated knowledge graph-related data into the training data to enhance the model's ability to answer commonsense questions. ClinicalGPT has also attempted to improve the diagnostic capability of the model by including data from electronic medical records and medical examinations in training data. While there has been significant progress in fine-tuning medical LLM instructions, there is still limited research on the preferred learning stage.

Preference Learning. Preference alignment is a prominent focus in large model training research, as preferentially aligned models exhibit enhanced generalization ability in zero shot scenarios (Kirk et al., 2023). Currently, the most renowned method for preference alignment is reinforcement learning from human feedback, which involves the utilization of four models for training. However, this approach has drawbacks, e.g., high engineering complexity and unstable training. In an effort to streamline the preference learning process, Rafailov et al. (2023) introduced a direct preference optimization algorithm that can bypass the need to train the reward model. Similarly, Gulcehre et al. (2023) has proposed a self-reinforcement learning approach that uses the EM algorithm to eliminate the training of the critic model. Furthermore, RRHF(Yuan et al., 2023) suggests using learning rankings to replace reinforcement learning, thus strengthening learning stability. Furthermore, some initiatives, such as RLAIF(Lee et al., 2023), aim to leverage AI to substitute manual preference data annotation, thus reducing annotation costs. Moreover, Bai et al. (2022) proposes training constitutional evaluation models for self-reflection. Sun et al. (2023) has put forward the idea of using the Principle-following reward model as a replacement for the traditional reward model to achieve dynamic adaptation to human preferences. Compared to these methods, our main contribution is to propose a preference learning approach for multiple rounds of dialogue.

6 Conclusion

In this work, we have introduced an innovative approach termed preference learning from process feedback (PLPF), which integrates the diagnostic logic of healthcare professionals into the LLM.

PLPF encompasses rule modeling, preference data generation, and preference alignment to train the model to adhere to the diagnostic process. Our experimental findings, using standardized patient testing, reveal that PLPF enhances the diagnostic accuracy of the baseline model in medical conversations by 17.6%. Furthermore, PLPF exhibits efficacy in both multi-round and single-round dialogue tasks, underscoring its potential for advancing medical dialogue generation.

Ethics and Limitations

There are several limitations to our approach. Firstly, the defined processes are relatively simple, and more complex processes require additional testing. Secondly, the accuracy of the model is still not high enough for practical use in SP tests, as it sometimes generates hallucinatory results. Additionally, it is worth noting that there may be a geographical bias in the test results, as most of the cases used in our study came from the Wuhan region of China. Therefore, it is important to consider the ethical implications of this geographical bias when interpreting our findings.

7 Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant No. 62192731. We kindly appreciate all the researchers who provide valuable insights, discussions, and comments on this work.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

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*.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, et al. 2023. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.

Junqing He, Mingming Fu, and Manshu Tu. 2019. Applying deep matching networks to chinese medical question answering: a study and a dataset. *BMC medical informatics and decision making*, 19(2):91–100.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics.

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

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. Meddg: An entity-centric medical consultation dataset for entity-aware medical dialogue generation.

OpenAI. 2023. ChatGPT: A Large-Scale Open-Domain Chatbot. `https://openai.com/chatgpt`. Version turbo-0613.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Salmon: Self-alignment with principle-following reward models. *arXiv preprint arXiv:2310.05910*.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023a. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023b. Huatuo: Tuning llama model with chinese medical knowledge.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia. Association for Computational Linguistics.

JunYong Zhu WeiGuo Dong. 2012. *Objective Structured Clinical Examination & Standardized Patients*. People's Medical Publishing House (PMPH), No. 19, Panjiayuan Nanli, Chaoyang District, Beijing, China.

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

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

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Huatuogpt, towards taming language models to be a doctor. *arXiv preprint arXiv:2305.15075*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Wei Zhu and Xiaoling Wang. 2023. Chatmed: A chinese medical large language model. `https://github.com/michael-wzhu/ChatMed`.

A Rules Scoring Guidelines

In this section, we provide the annotation guidelines for annotators' reference. We divide the rules into two categories: goal-oriented rules and constraint-oriented rules. For goal-oriented rules, we have defined strict scoring criteria, which are illustrated in Fig. 6. However, assessing the level of compliance for constraint-oriented rules is difficult, so we allow annotators to score freely. In the end, we consider the consensus among the annotators as the final score.

To reduce the annotation workload, we utilized ChatGPT to assist in the annotation process. Specifically, we manually created specific scenarios for each score of each rule and employed the In-context Learning technique to allow ChatGPT to pre-label the data. The annotators' task was to review the annotations generated by ChatGPT and make necessary corrections. Our findings indicate that this correction-based approach significantly improves the internal consistency of the annotators.

B Standardized Patient Testing

B.1 Data

During the SP tests, we used three types of data: patient information, dialog scripts, and checklists. Patient information and dialog scripts were employed to create simulated patients, while checklists were used to assess the history of dialogs generated by the model following interactions with simulated patients. Our dataset included information from five departments and the number of cases in each department was as follows: 23 cases in internal medicine, 23 cases in surgery, 8 cases in gynecology, 10 cases in pediatrics, and 8 cases in psychiatry. The following section provides a detailed explanation of these three types of data.

Patient Information. The patient data consists of a wide range of information, such as patient symptoms and treatments, among other things. An example of patient information is shown in Fig. 7. Patient data includes a significant amount of laboratory test results, which can be used to assess the analytical capabilities of the model.

Dialogue Script. Although the patient information is detailed, it does not capture patient mood, speech style, and life experience. To make the patient simulation more realistic, a dialogue script is provided (see Fig. 8). The script includes numerous exchanges between the doctor and the patient, involving both inquiries and responses. In addition to discussing important symptoms and tests, the script also includes inquiries and responses about less significant symptoms. These less significant symptoms act as distractors for the LLM test, thus improving the reliability of our test.

Checklist. A checklist is used to evaluate the model, which consists of three parts: key symptoms, key tests, and diseases, as shown in Fig. 9. Essentially, a better understanding of the key symptoms and key tests will improve the model's ability to provide a precise diagnosis. It should be noted that the evaluation of treatment history, family history, and other factors has been included in the symptom section.

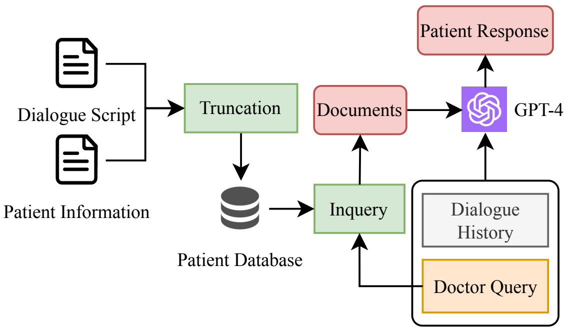

Figure 5: Patient Simulator Architecture.

B.2 Patient Simulator

In carry out the testing process, we created a patient simulator to interact with the model being evaluated, as shown in Fig. 5. The simulator was developed using the retrieval-augmented generation technique. Before conducting the tests, we created a separate database for each patient. Initially, we divided patient information and conversation scripts into documents with a maximum length of 128 tokens. Then, we encoded these documents using OpenAI's "text-embedding-ada-002" model to obtain a vector representation for indexing purposes. During the testing phase, we encoded the last two rounds of conversation history using the same "text-embedding-ada-002" model and retrieved the four most similar documents from the patient database to assist in generating responses. The model used to generate the responses was "gpt-4-0613", and the specific prompt used is shown below.

Please play the role of a patient who interacts with a doctor. You need to fulfill the following requirements:
1. If the doctor asks a question, answer it based on the contents of the knowledge base and the history of the conversation, with a response of no more than two sentences.
2. If your doctor recommends a test, inform him of the results. If you have not undergone the test, simply state that you are unaware of the results.
3. Do not expose any information about yourself to the doctor unless the doctor takes the initiative to ask a question, please be passive and accept the doctor's guidance.
4. If the doctor does not ask questions, ask the doctor what disease you have and how it should be treated.
5. When you feel that the conversation should end, please output: (End of Conversation).

Knowledge Base: [documents]
Conversation history: [history]
Doctor: [question]
Your response:

To avoid excessive interaction, we set a maximum of 5 rounds for communication between the LLM and the simulator during testing. This choice was made after noticing that medical LLMs usually need only 2-3 rounds of conversation to arrive at a patient's diagnosis.

B.3 Evaluation

We employ a manual evaluation method to assess each model using a checklist. To ensure the reliability of the assessment, two physicians will independently assign scores to each conversation history. The final score will be determined by calculating the average of their scores. The scoring formula we have adopted is as follows.

$$s^j = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i^j}{m_i^j} \quad j \in \{\text{Sym., Test, Dis.}\} \tag{6}$$

Here, n represents the number of standardized patients, and j represents the specific assessment category, namely symptoms, medical tests, and diseases. c_i^j represents the number of items that the model "passes" in the j-th assessment category for the i-th patient, while m_i^j represents the total number of items in the j-th assessment category. A "pass" in this context means that: 1) The model actively requests a symptom or medical test result from the patient, and it is included in the checklist. 2) The model predicts a disease and is included in the checklist. It is important to note that if the model provides four or more candidate diseases simultaneously, we consider it a failure to diagnose the disease.

C Implementation of Our Model

All our models were trained using 4 A100-40G. We used the Lora (Hu et al., 2021) technique in the training process, setting lora α and lora r to 16 and 64, respectively, and the learning rate to 1e-4. For the Qwen model, we trained the modules "c_attn" and "c_proj", as well as "w1" and "w2". For the Baichuan model, the modules "W_pack" and "o_proj" were trained. The batch size used for training was 2, and the gradient accumulation steps were set to 16. Our REM was trained by Baichuan2-Chat (7B) using 1,800 samples, and the number of training rounds was 2. When calculating the score, we set the values of α, β, γ, d, t_1, t_2 and k to 0.1, 0.8, 0.1, 0.65, 1.0, 1.0 and 5, respectively.

When applying REM to label preference data, response pairs are classified as win, tie, or loss, with a tie indicating that the difference between the scores of two responses is less than 1. After discarding all pairs labeled as tie, we select the top 2k samples with the largest difference to train our target model. Among preference data, the win-to-loss ratio between the trajectories obtained by data sampling and the tragjectory prediction is 1.2:1.

D Baselines

Chat. 1) ChatGLM3 (6B) (Du et al., 2022; Zeng et al., 2022): This model is considered the most advanced Chinese LLM with a size within 10B and has shown performance comparable to GPT-4 in the OpenCompass [3] Chinese benchmark. 2) Baichuan2-Chat (7B) (Yang et al., 2023): This model is based on Baichuan2-Base (7B), which is the preferred base model for Chinese medical LLMs in recent studies. Baichuan2-Chat (7B) performs similarly to ChatGPT on the OpenCompass Chinese benchmark.

Medical. 1) DISC-MedLLM (Bao et al., 2023): This model is based on Baichuan2-Base (13B) and was fine-tuned using 470k medical instruction data [4]. Additionally, this model utilizes 2k data for RLHF. 2) Huatuo-II (Chen et al., 2023): This model is based on Baichuan2-Base (7B) and was fine-tuned using 5,252k pretraining instruction data and 142k medical Q&A data.

[3]https://opencompass.org.cn/leaderboard-llm

[4]https://huggingface.co/datasets/Flmc/DISC-Med-SFT

Model	Exact Match	Fuzzy Match
ChatGPT	56.2	79.5
Ours	62.1	88.3

Table 5: Performance of different REMs on the testset.

SFT. To create two baselines, SFT (Qwen) and SFT (Baichuan), we used the same instruction fine-tuning data as DISC-MedLLM to fine-tune Qwen-Base (7B) (Bai et al., 2023) and Baichuan2-Base (7B), respectively.

E Performance of REM

In this section, we showcase the performance of REM on the test set using two different configurations. The initial configuration involves exact matching, where we determine the percentage of samples that REM accurately scores. The second configuration involves fuzzy matching, where we assess the likelihood that REM misclassifies a sample with a score of 2 as 0. and vice versa. A high score on this metric indicates that REM effectively distinguishes between good and bad responses. We conducted a comparison between REM, ChatGPT, and 5 instances manually created for In-context Learning to enhance ChatGPT's accuracy. The results of the experiment are presented in Table 5. Fine-tuned REM exhibits superior performance compared to ChatGPT. However, given the limited performance gap, it is expected that as the overall performance of the generalized LLM improves in the future, the entire PLPF process will become automated, with human intervention only required to design the flow chart and write the rules.

F Case Study

In this section, we will analyze the response preferences of each model during multi-round conversations, with the assistance of several examples. We have chosen multiple models for our analysis, all of which communicate with the same standardized patient suffering from acute appendicitis, using Baichuan-Base as the base model.

Baichuan-Chat. The conversation history of Baichuan-Chat is shown in Fig. 10. Our analysis indicates that the model successfully generates an extensive range of potential patient diagnoses and links them to detailed explanations. However, the model lacks in providing guidance to the patient on how to confirm the diagnosis, and it also tends to avoid answering certain patient questions. As a result, these limitations reduce the diagnostic effectiveness of Baichuan-Chat.

Huatuo-II. Fig. 11 presents the conversation history of Huatuo-II, which is characterized by its utilization of single-round conversations to achieve multi-round conversational objectives. A notable limitation of Huatuo-II is its inability to aid patients in interpreting medical test results by incorporating symptom information from previous dialogues. Moreover, Huatuo-II adopts a passive interaction style, overwhelming patients with an excessive amount of information that may impede their ability to extract valuable insights from the system's responses.

DISC-MedLLM. The conversation history of DISC-MedLLM is presented in Fig. 12 and Fig. 13. Our analysis suggests that DISC-MedLLM effectively extracts information from patients regarding their symptoms. However, the model relies on a fixed response template, where it restates the patient's statement, provides its own perspective, and concludes with recommendations, with a significant portion of the response dedicated to offering suggestions. As a result, DISC-MedLLM's responses tend to be longer compared to other models. One major drawback of DISC-MedLLM is the lack of specificity in its points. For instance, when diagnosing a patient with appendicitis, the model simply suggests that the patient may be experiencing a gastrointestinal issue. Additionally, while DISC-MedLLM provides numerous therapeutic recommendations, they are general in nature and do not offer comprehensive guidance.

SFT (Baichuan) The conversation history of SFT (Baichuan) is shown in Fig. 14. Our observations indicate that SFT (Baichuan) and Baichuan-Chat both fail to effectively provide patients with information on how to confirm their diagnosis. Furthermore, SFT (Baichuan) analyzes the test results submitted by patients in a similar manner to DISC-MedLLM, as it advises patients that further evaluation of the test results is necessary, but both lacking detailed analysis of the test results. From this we can infer that DISC-MedLLM primarily improves the model's ability to inquire about symptoms and offer treatment recommendations.

PLPF (Baichuan) The conversation history of PLPF (Baichuan) is shown in Fig. 15. The PLPF model strictly adheres to a process that involves asking for symptoms, proposing a diagnosis, verifying the diagnosis, and suggesting a treatment recommendation. In comparison to the SFT model, the PLPF model is more focused on symptom inquiry. For example, our model specifically asks about the location of pain when it identifies the keyword "Metastatic ... pain," which is important for determining the possibility of appendicitis in the patient. In terms of validating the diagnosis, our model suggests more precise tests such as blood tests and ultrasound, while DISC-MedLLM suggests more general tests like gastroscopy and liver function tests. Our model effectively utilizes the findings of test results to further refine the patient's diagnosis, specifically identifying the possibility of septic appendicitis. On the contrary, the other models do not effectively utilize this information. Lastly, when it comes to offering treatment options, our model proposes a surgical treatment plan, whereas the other LLMs only provide a generic treatment plan.

G Evaluation Based on BLUE and ROUGE

In this section, we provide the BLUE-4 and Rouge-L scores obtained by the model on the Meddg, IMCS, and webMedQA datasets. We also discuss the limitations associated with these scores. The scores are presented in Table 6, with the three highest scores highlighted in red and the three lowest scores in purple for each dataset. Moreover, we include the length of the responses generated by each LLM. Unlike conventional reporting methods that typically only present the F1-score, we present the precision, recall, and F1-score together for the ROUGE score, separated by the "/" sign.

A strong negative correlation was observed between the precision metrics scores (such as BLUE and Rouge-Precision) and the length of LLM responses. Specifically, Rouge-Precision has a significant impact on ROUGE-F1. When traditional metrics are used for evaluation, models with shorter response lengths tend to receive higher scores. One possible explanation for this finding is that physicians' responses in real datasets are usually more concise, while the output of medical LLMs often

Model	Meddg(Avg Len = 23.7) B@4	R@L	Len	IMCS(Avg Len = 25.2) B@4	R@L	Len	WebMedQA(Avg Len = 144.3) B@4	R@L	Len
Baichuan-Chat	**0.5**	**4.4/34.2/6.8**	**214.2**	**0.5**	**4.3/27.9/6.3**	269.4	**2.3**	**9.7/18.9/10.5**	269.4
ChatGLM3	**1.8**	**14.4/26.1/14.7**	138.0	**1.6**	**13.2/20.5/12.2**	272.2	3.4	11.2/22.7/12.8	272.2
Huatuo-II	**0.6**	**8.6/33.2/10.1**	**188.7**	**0.7**	**7.4/37.3/9.2**	**333.8**	**2.9**	**8.1/28.1/11.4**	**425.1**
DISC-MedLLM	**0.7**	**6.5/28.9/7.6**	**203.2**	**0.7**	**9.5/25.0/9.0**	**425.1**	**2.8**	**8.8/24.7/11.7**	**333.8**
SFT (Qwen)	**2.0**	**19.5/26.4/17.4**	**50.0**	**2.1**	**17.8/18.5/13.0**	**167.4**	**4.3**	**14.3/18.5/13.8**	**167.4**
SFT (Baichuan)	**1.5**	**14.9/27.0/14.4**	**64.4**	**1.3**	**14.3/19.5/11.3**	**173.5**	**4.0**	**12.9/19.6/13.6**	**173.5**
PLPF (Qwen)	1.0	7.5/36.9/10.3	149.2	1.2	8.9/28.0/10.0	**277.8**	3.5	10.4/23.8/12.8	**277.8**
PLPF (Baichuan)	1.3	11.3/34.8/13.1	**99.8**	1.1	10.4/26.7/10.3	**227.3**	**4.0**	**12.5/21.5/13.8**	**199.6**

Table 6: BLUE and ROUGE scores of LLMs on each dataset.

includes additional details that are not present in the reference responses. As a result, the BLUE and ROUGE-F1 scores are lower. It is clear that the evaluation of a response cannot solely rely on its length, indicating that BLUE and ROUGE are not reliable measures for assessing the performance of LLM responses.

Before diagnosing and guiding the patient, doctors needs to carefully verify the patient's condition.

```
### Score = 2
Doctors proactively gather patients' symptoms and inquire about their condition before
making a diagnosis. They utilize the information provided by the patient to inform their
diagnostic process.
>>> Or
The patient has already received a diagnosis from a healthcare professional or undergone a
medical evaluation before the consultation. It is important to note that a patient's self-
perceived belief regarding a specific disease does not constitute a formal diagnosis.

### Score = 1
The doctors did not proactively collect patient information; rather, patients volunteered
more information. The doctor thoroughly utilized the information provided by the patient
before making the diagnosis. This was evident in the doctor's analysis of the patient's
condition using the information provided to determine the type of disease, its severity,
and other relevant factors.

### Score = 0
The information provided by the patient was insufficient, and the doctor did not offer
additional guidance, leading to what appeared to be an arbitrary diagnosis or guidance.
>>> Or
The doctor declined to provide a diagnosis for the patient.
```

Doctors should inform patients about their disease or the tests needed for diagnosis.

```
### Score = 2
The doctor informs the patient about the disease or a possible diagnosis if it cannot be
confirmed and refers the patient to an in-person consultation.
>>> Or
The patient was diagnosed prior to this consultation.

### Score = 1
The doctor gave the patient a vague diagnosis, while failing to tell the patient how to
confirm the diagnosis of the disease.

### Score = 0
The doctor recommended the patient for a test without providing a possible diagnosis or
justifying the test.
>>> Or
The doctor did not inform the patient of their diagnosis or recommend the necessary test.
```

Doctors should inform patients of treatment options for the disease.

```
### Score = 2
The doctor effectively communicated the treatment options for the disease and safe methods
for relieving the symptoms to the patient.
>>> Or
The doctor discussed conservative treatment options with the patient and recommended
monitoring the progression of the disease.
>>> Or
The doctor was unable to devise a treatment plan for the patient due to the complexity of
the condition, prompting the need for additional tests.

### Score = 1
The doctor recommended that the patient go to the hospital without providing a clear or
reasonable explanation.
>>> Or
The doctor provided treatment options for only a portion of the patient's illness.

### Score = 0
The doctor did not inform the patient of the treatment plan and did not give any reason, or
the reason was unreasonable.
```

Figure 6: Goal-oriented rules evaluation criteria.

Patient	**, Female, 36 years old: **Corporate Employee, Patient's self-reported medical history.
Chief Complaint	Upset and anxious for over 2 years
Present Medical History	Patient presented more than 2 years ago with no obvious triggers for distress and anxiety, self-conscious of feeling nervous from time to time, but could not specifically name the things she was nervous about. Frequent lack of concentration interfered with work and self· consciousness was poor. The above situation has been continued, once in the unit did physical examination, did not find obvious lesions. At the end of last year, she went to our outpatient clinic and took amitriptyline, which she stopped because she could not tolerate it. Recently, she felt that her symptoms had worsened, and she had fidgeting, poo: sleep, occasional headaches, panic attacks, shortness of breath, etc.
Past history	Denied history of hypertension, diabetes mellitus, denied history of heart disease, no history of hepatitis, tuberculosis, no history of food or drug allergy, and a history of cesarean section.
Menstrual history	Menarche at 13 years old, usual menstruation is basically normal, usually 3---5 days, last menstruation was on April 1.
Marital and childbearing history	Married, good relationship, one son, family health.
Family history	Parents are alive, deny family history of specific genetic diseases and similar conditions.
Personal History	Born and residing locally, denies history of exposure to infected areas, no smoking, drinking and other bad habits.
Laboratory Tests	Outpatient tests for blood, urine, liver function, ECG, and EEG results were all within normal limits.
Neurologic Examination	Mental clarity, bilateral pupil diameter 3mm, light reflexes present, normal eye movements, no tongue-face palsy, normal reflexes, neck softness and no resistance, Kerb's sign negative, limb muscle strength, muscle tone normal, cavity reflexes, etc. present, no pathognomonic signs, abdominal wall reflexes sensitive, no sensory diminution. Bilateral rotational movements and finger-nose test were normal, and Lomborg's sign was negative.
Psychiatric Examination	The patient walks into the ward by himself, conscious, well-groomed, good contact, clear speech, relevant answers, moderate emotional response, slight anxiety, cumbersome speech, coherent thought content and environment, no thought disorder.

Figure 7: Example of Patient Information.

Doctor	May I ask your name and how old you are?
Patient	My name is **** and I am 36 years old.
Doctor	How did you feel uncomfortable?
Patient	I felt distracted, nervous and in bad shape.
Doctor	Oh, how long has this been going on?
Patient	It's probably been over 2 years.
...	
Doctor	Have you ever had any suicidal thoughts?
Patient	That's not true.
Doctor	Have you had any serious illnesses? Any high blood pressure, diabetes, heart disease? Hepatitis, tuberculosis?
Patient	No, we have annual medical checkups in our organization, and we haven't found any of these diseases.
...	

Figure 8: Example of Dialogue Script.

Symptom	Nervousness, but no specific goal of nervousness Panic, shortness of breath, headache Ask about treatment history Ask about disease changes and transformation Ask for general information Ask about past history Ask about family history
Test	Routine blood test Routine urine test Liver Function tests Electrocardiogram (ECG) Electroencephalogram (EEG) Neurological examination Psychiatric examination
Disease	Anxiety disorder

Figure 9: Example of Checklist.

Patient	Hello, my name is **, female, 22 years old. Metastatic lower abdominal pain for 3 days.
Doctor	Hello Ms. **, I am happy to help you. May I ask if your pain is persistent or paroxysmal? Where is the exact location of the pain? Are there any other symptoms such as nausea, vomiting, fever, etc.?
Patient	The pain was metastatic, starting with pain in the upper and middle abdomen, then moving to the lower right abdomen and full abdominal cramps, with nausea, vomiting once, and occasional dizziness.
Doctor *Comment: The doctor informed the patient of the three possible diagnoses, but did not tell the patient how to further confirm the diagnosis and simply told the patient to seek offline help.*	Thank you for your reply, Ms. **. According to the symptoms you described, you may be suffering from one of the following diseases: 1. Acute appendicitis: this is a common abdominal disease that manifests as metastatic right lower abdominal pain, which gradually worsens, accompanied by nausea, vomiting and other symptoms. 2. Ovarian cyst torsion: a disease of the female reproductive organs that manifests as a sudden onset of severe abdominal pain, which may radiate to the waist and groin, accompanied by nausea, vomiting and other symptoms. 3. Intestinal obstruction: Intestinal blockage leads to abdominal pain, nausea, vomiting and other symptoms. It is recommended that you consult a hospital as soon as possible for diagnosis and treatment by a specialized doctor. In the meantime, please take rest, avoid strenuous exercise and keep your diet light. If your condition worsens, please seek medical attention immediately.
Patient	I'm already at the hospital and the outpatient doctor gave me an ultrasound and said there's a mass on my right side.
Doctor *Comment: Even though the patient provided ultrasound results, the doctor still did not give a definitive diagnosis or specific recommendations for testing.*	Thank you for the information, Ms.**. According to your description, your symptoms may be related to a mass in the right adnexal area, which may be caused by torsion of an ovarian cyst or rupture of an ovarian cyst. Of course, there are other diseases that may cause similar symptoms, such as appendicitis and intestinal obstruction. In this case, you are advised to follow the following suggestions: 1. Follow your doctor's recommendations for examination and treatment to ensure timely diagnosis and treatment of your condition. 2. Keep communicating with your doctor about the progress of your condition and the effectiveness of your treatment. 3. Maintain good lifestyle habits, such as maintaining a healthy diet, proper exercise, and adequate sleep, to help your body recover during treatment. 4. If you experience any discomfort or deterioration of your condition during treatment, please seek medical advice in time. 5. After the treatment is completed, have regular follow-ups to ensure that your condition is under control and stabilized. I hope these suggestions are helpful to you and wish you a speedy recovery!
Patient	Doctor, what disease should I have? How should I treat it?
Doctor *Comment: The doctor bluntly refused to give a specific diagnosis.*	I'm sorry, but as a large language model, I can't see your medical test results and history directly. I would recommend that you follow your doctor's advice and they can confirm your diagnosis and develop a treatment plan based on your specific situation and test results. Also, maintaining good habits and a positive mindset is very important for your recovery.
Patient	(End of conversation)

Figure 10: Example of Baichuan-Chat.

Patient	Hello, my name is **, female, 22 years old. Metastatic lower abdominal pain for 3 days.
Doctor *Comment: Doctors do not proactively gather information from patients, give a specific diagnosis, or give advice to confirm the diagnosis. Rather, they provide a great deal of information for the patient to make her own judgment about her illness.*	Hello, Ms. **. I am sorry to hear that you are experiencing abdominal pain. Based on your description, your pain is metastatic, which can mean that the pain originates in one area of the abdomen and then moves to another area. This can be due to a variety of reasons, including digestive problems (such as gastroenteritis, gastric ulcers, cholecystitis, or pancreatitis), urinary problems (such as kidney stones or urinary tract infections), and gynecological problems (such as ovarian cysts, endometriosis, or pelvic inflammatory syndrome). However, I must emphasize that what I have provided are only possible causes and a true diagnosis needs to be made by a doctor. If your pain persists or worsens, or is accompanied by other symptoms such as fever, vomiting, bloody urine, constipation or diarrhea, you should seek immediate medical attention. If you've already seen your doctor and they haven't given you a firm diagnosis, you may need further tests, such as an ultrasound, CT scan or MRI, to help determine the cause of the pain. keep in mind that these are only general recommendations and specific medical advice should be provided by your doctor as they know your personal health and medical history.
Patient	I developed metastatic right lower abdominal pain three days ago, starting with upper-middle abdominal pain and then shifting to right lower abdominal and generalized abdominal cramps with nausea, vomiting once, and occasional dizziness.
Doctor *Comment: The patient provided more information, yet the doctor still did not give the patient a specific diagnosis, but instead provided some general offline visit advice.*	Your symptoms may be caused by a variety of reasons, including but not limited to gastrointestinal disorders, urinary tract infections, kidney stones, ovarian cysts, and acute appendicitis. Your symptoms, especially the pain in the lower right abdomen and the accompanying nausea and vomiting, may be related to digestive or gynecological disorders. It is recommended that you seek medical attention as soon as possible so that your doctor can perform a thorough evaluation and determine the best course of treatment. Until you see your doctor, try to avoid greasy or indigestible foods, drink plenty of fluids to prevent dehydration, and avoid overexertion. Please note that this is only one of the possible explanations based on the symptoms you have described, and an accurate diagnosis and treatment plan can only be arrived at after examination and testing by a medical professional.
Patient	I've already been to an ultrasound and the doctor said there's a lump on the right side, what's going on?
Doctor *Comment: The patient provides ultrasound results and expects the doctor to interpret them. However, the doctor only analyzes them in isolation regarding the problem at hand and does not incorporate the symptomatic information provides by the patient.*	An ultrasound is a painless medical imaging technique that shows images of the body's internal structures and organs. If your ultrasound results show a lump on your right side, this usually means that a lump or mass is found somewhere on your right side that is larger or unusual than the surrounding tissue. The nature of this mass (benign or malignant) and exactly what type of mass it is will depend on a number of factors, including the size, shape, and location of the mass, as well as whether it is growing or decreasing in size. Some possible types of lumps include cysts, lipomas, abscesses, leiomyomas, tumors, and more. In some cases, ultrasound results may require further testing or imaging studies, such as an MRI or CT scan, to define the characteristics of the mass in more detail. Your doctor will discuss next steps based on the ultrasound results and your symptoms. This may include watchful waiting, further testing or treatment. Make sure you communicate fully with your doctor, understand their explanations and recommendations, and follow their instructions.
Patient	I see, so do I need to do any further tests?
Doctor	The need for additional tests depends on the details of your ultrasound results and your symptoms. Here are some examples of additional tests that may be needed: 1. If your doctor suspects that a lump may be malignant, a biological tissue biopsy, in which a small sample is removed from the lump and examined under a microscope, may be recommended. 2. If your doctor suspects that you may have a specific type of lump or condition, such as ovarian cysts or fibroids, they may recommend other tests such as a blood test, urine test, or magnetic resonance imaging (MRI) to better understand your condition.3. In some cases, your doctor may recommend other types of tests to rule out other underlying conditions that may be causing your symptoms.4. It's important to work closely with your doctor to get the necessary tests or exams based on their recommendations. They will be able to provide you with personalized recommendations based on your specific situation.
Patient	(End of conversation)

Figure 11: Example of Huatuo-II.

Patient		Hello, my name is **, female, 22 years old. Metastatic lower abdominal pain for 3 days.
Doctor	*Comment: The doctor was very proactive in gathering information from the patient about the various symptoms and provided some generalized ways to alleviate them. (Good!)*	Hello Ms. **, I am happy to help you. According to your description, you have been experiencing metastatic lower abdominal pain for 3 days. This condition may be caused by a number of reasons, such as digestive problems, urinary tract infection, etc. In order to determine the cause more accurately, I would like to know some more information. First of all, is this pain persistent or does it come on intermittently? Is it accompanied by other symptoms such as nausea, vomiting or diarrhea? This information will help us better determine the possible type of illness. Also, have you had similar symptoms or is there someone in your family with a related disease? This is also a factor to consider. In the present case, I would suggest that you first try some self-care measures to relieve the pain. You can take plenty of rest, avoid strenuous exercise and over-exertion, as well as maintain a good diet and avoid spicy and stimulating foods. In addition, you may also try applying hot compresses to the abdominal area, which may ease the discomfort. However, if the pain does not improve significantly or you experience other discomforts, I strongly recommend that you seek prompt medical attention and undergo further tests to clarify the cause. Your doctor can help determine the diagnosis by taking a detailed medical history, physical examination, and necessary laboratory tests (e.g., routine blood and urine tests). Apart from digestive problems, there are other possible causes of lower abdominal pain, such as urinary tract infections and gynecological inflammation. So please feel free to let me know if you have any questions or need further help.
Patient		This pain started with pain in the upper and middle abdomen, then shifted to colic in the lower right abdomen and all over the abdomen, accompanied by nausea and vomiting once, occasional dizziness, no panic attacks, no radiating pain in the back of the shoulders, no sensation of anal swelling, and no bleeding from the vagina. I have not had this symptom before and no one in my family has any related illnesses.
Doctor	*Comment: The patient provided the doctor with a great deal of information about her symptoms, and the doctor indicated that the diagnosis could not be confirmed on the basis of this information alone, and asked the patient for more information. (Good!)*	Based on your description, your symptoms include upper-middle abdominal pain that has shifted to the lower right abdomen and full abdominal cramps, accompanied by nausea, vomiting once, and occasional dizziness. While these symptoms can be associated with a variety of conditions, it is also possible that the discomfort is due to gastrointestinal problems. In order to more accurately determine the cause, I need to know some additional information: does this pain occur before meals? Are there any specific triggers that would make the pain worse? In addition to this, have you recently experienced any loss of appetite or weight loss? This information is important for us to rule out other underlying conditions. In the meantime, I suggest you pay attention to the following points: firstly, try to avoid oily and spicy foods that may irritate the gastrointestinal tract; secondly, ensure adequate water intake, eat more fresh vegetables and fruits, and increase the intake of dietary fiber; and lastly, pay attention to a regular work schedule and avoid staying up all night and excessive fatigue. It should be noted that the above suggestions are for reference only, and the specific treatment plan should be adjusted according to the doctor's guidance. If there is no significant improvement in symptoms or other discomforts, please consult your doctor and undergo further examination to clarify the cause. If you have any questions or need further help, please feel free to let me know.
Patient		The pain comes after meals and is especially worse after eating fatty foods. Recently, I has lost my appetite and has lost 5 pounds.
Doctor	*Comment: The doctor gave a vague diagnosis while suggesting the patient to go for some laboratory tests. However, the laboratory tests suggested by the doctor were not able to detect appendicitis, which indirectly indicates that the doctor did not correctly surmise the patient's disease.*	According to your additional information, the pain occurs mainly after meals especially when eating fatty foods and is accompanied by loss of appetite and weight loss. These circumstances further support the possibility of gastrointestinal problems. In response to your symptoms, I still recommend that you take the following measures to relieve your discomfort: first, try to avoid greasy and spicy foods that may irritate the gastrointestinal tract; second, ensure adequate water intake, eat more fresh vegetables and fruits, and increase your intake of dietary fiber; and lastly, pay attention to a regular routine and avoid staying up too late and being overly fatigued. Nonetheless, given that your symptoms have not improved significantly and you are experiencing new discomfort, I recommend that you seek medical attention as soon as possible and undergo further testing to determine the cause of your symptoms. Your doctor can help determine the diagnosis by taking a detailed history, physical examination, and necessary laboratory tests (e.g., blood tests, liver function, gastroscopy, etc.). In addition to gastrointestinal problems, there are a number of other possible causes of similar symptoms, such as pancreatitis and gallbladder stones. So please feel free to let me know if you have any questions or need further help.

Figure 12: Example of DISC-MedLLM, Part-I.

Patient	I have already had an ultrasound at the local hospital and the doctor said there is a mass on the right side.
Doctor *Comment: The patient provided ultrasound findings and the doctor was unable to interpret the findings and did not correct his diagnosis.*	Thank you for the updated information. Based on your description, you underwent an ultrasound at your local hospital, which showed a mass on the right side. The exact nature of this mass is unclear and further tests may be required to make a definitive diagnosis. Based on the description of your condition, I can't give you a definitive diagnosis, but I can give you some advice for reference: first, please don't worry too much because many lumps are benign. However, to be on the safe side, I recommend that you have regular checkups to monitor any changes in the lump. If necessary, your doctor may recommend surgical removal. If you are concerned or have any questions about your current diagnosis, I encourage you to consult a medical professional for more detailed information and advice. They will be able to develop a treatment plan that is right for you on a case-by-case basis. I hope my answers have been helpful. Please feel free to let me know if you have any other questions or need further assistance.
Patient	What do I have and how can I treat it?
Doctor *Comment: The doctor provided a vague and incorrect diagnosis.*	Based on the symptoms you have described, possible diagnoses include gastrointestinal problems, urinary tract infections, and more. However, it is difficult to make an accurate diagnosis based on textual descriptions alone, so I recommend that you seek medical attention as soon as possible and undergo further tests to clarify the cause of your condition. While waiting for medical treatment, you can continue to pay attention to dietary adjustments, avoid oily and spicy foods, eat more fresh vegetables and fruits, and drink a moderate amount of water. It is also important to maintain good living habits, including adequate rest, a reasonable dietary structure and moderate exercise. Once you have received professional medical advice and a diagnosis, your doctor will formulate a treatment plan accordingly. This may include medication, surgery or other appropriate treatments. Please feel free to let me know if you have any questions or need further assistance.

Figure 13: Example of DISC-MedLLM, Part-II.

Patient	Hello, my name is **, female, 22 years old. Metastatic lower abdominal pain for 3 days.
Doctor	Hello! I am very sorry to hear about your discomfort. May I ask if you are experiencing any other symptoms of discomfort? Such as fever, nausea, vomiting, etc? Also, have you ever had a history of sexual intercourse? This information is very important for us to make an accurate judgment.
Patient	I have nausea, vomiting, occasional dizziness, but no fever. As for sexual history, I have none.
Doctor *Comment: The doctor provided an incorrect diagnosis, and the doctor made no attempt to verify his diagnosis.*	Thank you for the information you have provided. Based on your description, the pain and discomfort may be due to pelvic inflammatory disease. It is recommended that you seek medical attention as soon as possible for further examination and treatment. Meanwhile, you can pay attention to rest, avoid overwork and maintain good dietary and living habits. If you have any other questions or need further help, please feel free to let me know.
Patient	Okay, I will follow your advice. May I ask what tests I should do?
Doctor *Comment: It is clear from this that the doctor is capable of verifying the diagnosis, but it requires active guidance from the patient.*	In order to make a definitive diagnosis and develop an appropriate treatment plan, I recommend that you undergo a gynecological ultrasound as well as routine blood tests. These tests will help us understand your condition so that we can better assist you. Please make an appointment for the tests as soon as possible and bring the results with you to your next appointment so that we can develop a personalized treatment plan for you based on the results. If you need help scheduling an appointment for a test, I can provide you with some information and advice about it.
Patient	I've had a gynecological ultrasound and routine blood tests, but I haven't gotten the results back yet.
Doctor	Understood, thank you for the information. Please be patient and wait for your test results, and once you have them, please do not hesitate to contact me so that I can work with you to analyze the results and develop the next steps in your treatment plan. Please also feel free to let me know if you have any discomfort or questions during the waiting period, and I will do my best to provide help and support. I wish you a speedy recovery! If you have any other questions, please feel free to continue your consultation.
Patient	I've got the results back, the ultrasound showed "mixed mass in the right adnexal area", and the blood tests were WBC 12.3*10^9/L, NEUT% 89.9%, LYM% 5.5%.
Doctor *Comment: The doctor gave an ambiguous interpretation of the information provided by the patient for the medical examination.*	Based on the findings you provided, a mixed mass in the right adnexal area may require further evaluation and management. In addition, the high white blood cell count may be related to inflammation. It is recommended that you contact your primary care physician as soon as possible for a detailed interpretation of the test results and to discuss with them the next step of treatment. In the meantime, maintaining a positive mindset and rationalizing your diet and lifestyle will help promote recovery. If you need any help or have any other questions, please feel free to let me know. I wish you a speedy recovery!

Figure 14: Example of SFT (Baichuan).

Patient	Hello, my name is **, female, 22 years old. Metastatic lower abdominal pain for 3 days.
Doctor	Hello! I am happy to help you. Can you please describe in detail the location of your pain? Also, are you experiencing any other discomfort? So that I can better understand your situation.
Patient	I developed metastatic right lower abdominal pain three days ago, starting with upper-middle abdominal pain and then shifting to right lower abdominal and generalized abdominal cramps with nausea, vomiting once, and occasional dizziness.
Doctor *Comment: The doctor informed the patient of the definitive diagnosis and the medical tests used to verify that the diagnosis was correct.*	Thank you very much for the information you have provided. Based on your description, the abdominal pain may be caused by acute appendicitis. In order to further clarify the diagnosis, I suggest you to perform blood tests and ultrasound. These tests can help us determine the condition more accurately and formulate a corresponding treatment plan. Please go to the hospital as soon as possible for the relevant examinations.
Patient	I had an ultrasound at the outpatient doctor and he said there was a mass on the right side, but he didn't tell me exactly what it was.
Doctor *Comment: The doctor rationalized the patient's test report in the context of the diagnosis and proposed a feasible treatment plan.*	Understood, you have had an ultrasound in the clinic and a mass on the right side has been found. This could be a complication from appendicitis, i.e. a periappendiceal abscess. Surgery may be a better option in this case, as medication alone may have limited effect. Surgery can remove the infection, reduce inflammation, and prevent complications. After surgery, you will also need to continue antibiotic therapy to control the infection. Please feel free to let me know if you have any questions or need further information.
Patient	So where should I go for surgery now?
Doctor	If you have decided to undergo surgical treatment, then I would recommend that you contact your local major or specialty hospital as soon as possible to make an appointment for surgery. They will have a professional team to assess your condition and arrange a suitable surgical plan for you. At the same time, you can also consult your doctor about the specific details and precautions of the surgery to ensure a safe and smooth operation. I wish you a speedy recovery! If you have any other questions, I will try my best to answer them.
Patient	(End of conversation)

Figure 15: Example of PLPF (Baichuan).