Interactive Evaluation for Medical LLMs via Task-oriented Dialogue System

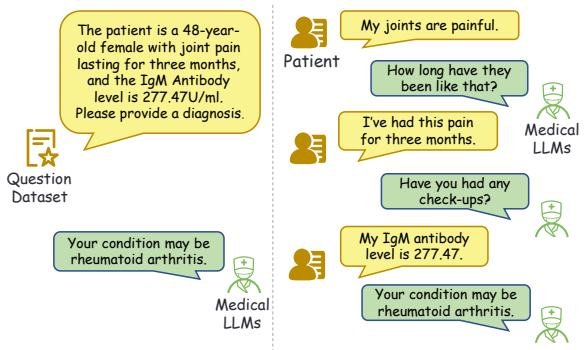
Ruoyu Liu[1], **Kui Xue**[2], **Xiaofan Zhang**[1,2*], **Shaoting Zhang**[2,3]
[1]Shanghai Jiao Tong University
[2]Shanghai AI Laboratory
[3]SenseTime Research
{liury24,xiaofan.zhang}@sjtu.edu.cn
{xuekui,zhangshaoting}@pjlab.org.cn

Abstract

This study focuses on evaluating proactive communication and diagnostic capabilities of medical Large Language Models (LLMs), which directly impact their effectiveness in patient consultations. In typical medical scenarios, doctors often ask a set of questions to gain a comprehensive understanding of patients' conditions. We argue that single-turn question-answering tasks such as MultiMedQA are insufficient for evaluating LLMs' medical consultation abilities. To address this limitation, we developed an evaluation benchmark called Multi-turn Medical Dialogue Evaluation (MMD-Eval), specifically designed to evaluate the proactive communication and diagnostic capabilities of medical LLMs during consultations. Considering the high cost and potential for hallucinations in LLMs, we innovatively trained a task-oriented dialogue system to simulate patients engaging in dialogues with the medical LLMs using our structured medical records dataset. This approach enabled us to generate multi-turn dialogue data. Subsequently, we evaluate the communication skills and medical expertise of the medical LLMs. All resources associated with this study will be made publicly available[1].

1 Introduction

Advanced large language models like OpenAI GPT (OpenAI, 2022), LLaMA (Touvron et al., 2023a), and Gemini (Team et al., 2023) have made significant strides in generating high-quality text. Due to the extensive data utilized during training, these models often possess a vast knowledge base, which has spurred the development of domain-specific models, including medical models (Chen et al., 2023b,a). Medical models based on LLMs have demonstrated impressive performance on various question-answering datasets. However, we believe that single-round Q&A evaluation (Contributors, 2023; Cai et al., 2024) is insufficient to measure the diagnostic capability of medical LLMs, as real-world consultations require multiple rounds of communication between doctors and patients. Figure 1 illustrates this difference, highlighting the doctor's communication competence in clinical consultations.



(a) Single-turn Q&A Evaluation **(b) Multi-turn Dialogue Evaluation**

Figure 1: A comparison of single-turn Q&A task evaluation and multi-turn dialogue evaluation in interactive clinical consultation scenarios.

In a typical medical consultation scenario, patients often lack medical expertise and, therefore, cannot independently and clearly articulate their symptoms. Consequently, the doctor must lead the conversation, asking a series of questions to gather more detailed information. This process can also be regarded as the differential diagnosis phase, where the doctor uses inquiries to rule out other possibilities and ultimately deduce the patient's condition. We categorize this scenario into two stages: the first being **Dialogue Data Collection**, where the doctor poses questions to gather patient information, and the patient is required to provide responses. The second stage is **Diagnostic Analysis**, during which the doctor formulates a diagnosis based on the information collected. Therefore, the evaluation of a medical LLM's interactive consul-

*Corresponding author.
[1]https://github.com/lry00127/MMD-Eval

Proceedings of the 31st International Conference on Computational Linguistics, pages 4871–4896
January 19–24, 2025.

tation capability can also follow these two stages. Firstly, during the dialogue data collection phase, the patient needs to respond to the questions posed by the medical LLM. During this stage, we assess the model's communication skills. Subsequently, in the diagnostic analysis phase, the medical LLM provides a diagnosis. In this phase, we evaluate the model's medical expertise.

The main challenge in evaluating multi-turn dialogues is obtaining multi-turn interaction data. To assess the diagnostic prowess of medical LLMs within authentic interactive settings, it is typically necessary to gather multi-turn dialogue data between practitioners and patients. Since deploying actual humans as patients to engage with these models can be prohibitively expensive, current approaches frequently center on crafting a simulated patient to interact with the medical LLM instead. A typical approach is to use another LLM to act as the patient, interacting with the doctor model in a medical scenario to obtain multi-turn dialogue data. Subsequently, a series of metrics are designed to evaluate the dialogue using an LLM, such as GPT-4 (OpenAI, 2023). For example, Liao et al. (2023) proposed an automatic evaluation framework using a fine-tuned LLaMA-7B (Touvron et al., 2023a) as the patient to interact with the medical LLMs being tested, with GPT-4 evaluating the resulting dialogue. Additionally, Fan et al. (2024) proposed an evaluation framework based on the concept of agents called MVME. However, there are some issues with this approach. Firstly, using LLMs as patients may lead to hallucination problems (Huang et al., 2023), resulting in responses that do not match the patient's information. Secondly, LLMs have strong self-awareness, making it difficult to change their self-perception solely through prompts. That is, LLMs may refuse to play the role of a patient or reveal their identity during the conversation. Lastly, although more capable LLMs perform better, they are often not open-source, accessible only through API calls, and in some cases, only via web interfaces, leading to restricted usage and high costs. Furthermore, such an approach may lead to privacy breaches, which is an even more critical concern.

In this work, we propose **Multi-Turn Medical Dialogue Evaluation (MMD-Eval)**, an interactive evaluation framework based on traditional task-oriented dialogue systems and a structured medical records dataset. We use a task-oriented dialogue system based on T5-small (Raffel et al., 2020; Zhang et al., 2022) as the patient and build a structured medical records dataset to interact with the medical LLM being tested, thus obtaining doctor-patient consultation data in specific medical contexts. When constructing the structured medical records dataset, we not only build structured information for patients but also collect treatment suggestions and diagnostic results from professional doctors for each record, which can be used to evaluate the capabilities of medical LLMs. Task-oriented dialogue systems are designed with advanced intent recognition and high accuracy. Therefore, they are traditionally used as expert systems to answer users' questions by retrieving information from structured data sources. However, we are innovating by using these systems in a unique way: now, they assume the role of users rather than experts, providing answers to the doctors' questions. Furthermore, given that traditional task-oriented dialogue systems are not smooth enough to generate dialogues, we also attempted to use LLMs to rewrite patient responses generated by task-oriented dialogue systems, making the patient's responses more aligned with spoken language characteristics. To our knowledge, we are the first to apply traditional task-oriented dialogue systems to the interactive evaluation of LLMs. In summary, our main contributions include:

- Proposing an evaluation system for assessing the interactive consultation capabilities of LLMs in medical scenarios, which ensures accuracy with minimal resource consumption and allows for convenient local deployment.
- Constructing a structured medical records dataset where patient information is divided into multiple items for querying, and each record is accompanied by corresponding diagnostic results, diagnostic bases, and treatment recommendations.
- Providing a methodology for evaluating LLMs in interactive scenarios using low-resource task-oriented dialogue systems and structured datasets.

2 Related Works

Large Language Models. Large language models, such as the proprietary GPT series (OpenAI, 2022, 2023) and the open-source LLAMA series (Touvron et al., 2023a,b), have demonstrated reasoning and interaction capabilities that significantly surpass those of previous models, enabling them to

tackle complex inferential tasks (Wei et al., 2022; Zhao et al., 2023). Due to their extensive training on diverse corpora, these models possess knowledge across numerous domains, paving the way for the development of genuinely functional medical assistants. However, existing evaluation metrics often focus on single-turn knowledge questions within discrete domains (He et al., 2023; Singhal et al., 2022; Jin et al., 2019; Pal et al., 2022), leading researchers to concentrate predominantly on the models' performance in these areas. Consequently, during the training phase, the multi-turn conversation capabilities of these models might receive less attention than other skills.

Interactive Evaluation. The emergence of large language models has introduced a plethora of new capabilities not seen in prior models, spurring the development of various assessment methodologies, including tool invocation (Qin et al., 2023; Patil et al., 2023) and environmental interactions (Xi et al., 2023). While research on interactive evaluation tailored to medical LLMs is relatively scarce, early work involved fine-tuning LLMs to act as patients engaging with medical LLMs (Liao et al., 2023), with GPT-4 subsequently assessing diagnostic outcomes. Recent studies have shifted focus to the application of multi-agent frameworks (Fan et al., 2024), where such frameworks facilitate collaborative analysis of the interactive diagnostic abilities of medical LLMs, frequently employing GPT-4 to simulate patient models. Nonetheless, whether utilizing GPT-4 or fine-tuning LLMs, the resource consumption is substantial, which we posit might contribute to the hesitance in implementing multi-turn dialogue assessments despite researchers' acknowledgment of their importance (Chen et al., 2024, 2023b; Liu et al., 2024).

Task-Oriented Dialogue Systems. Traditional task-oriented dialogue systems (Hosseini-Asl et al., 2020; Budzianowski and Vulić, 2019) aim to resolve domain-specific issues through modules such as intent recognition, dialogue state tracking, policy learning, and natural language generation. These systems are distinct from chatbots due to their strong functionality and purposefulness. Commonly deployed in scenarios like travel assistance or food services, they serve in an assisting capacity. Innovatively, we repurpose a task-oriented dialogue system as a user, training it to assume the role of a patient for engaging with medical LLMs, thereby responding to their queries.

3 MMD-Eval

We propose an interactive medical consultation evaluation benchmark for medical LLMs. By designing a task-oriented dialogue system and a structured medical records dataset, we are able to simulate doctor-patient dialogues and collect multi-turn dialogue information, illustrated in Figure 2. We then proceed to evaluate medical LLMs from two aspects: communication competence and clinical diagnostic competence. To address the rigid and less fluid outputs of our simulated patients, we employ an LLM to refine the outputs, enhancing the patient's character details and more authentically simulating the consultation scenario without compromising accuracy.

3.1 Benchmark Construction

Our evaluation system consists of three parts: a dialogue system designed to act as patients, a structured dataset of medical records, and a set of evaluation criteria. The first two parts are used to recreate doctor-patient conversations in a clinical scenario, while the third part is used for evaluating the performance of the medical LLMs.

Dialogue System. We developed a task-oriented dialogue system designed to simulate a patient, engaging in multi-turn interactions with medical LLMs using a structured records dataset. Our system is capable of recognizing 26 common doctor intentions and capturing 11 different types of slots (detailed in Appendix D.1). Considering that medical consultation scenarios are relatively closed, especially for patients, the difficulty of answering questions is not too high. Therefore, our dialogue system can accurately understand and respond to the doctor's questions, as illustrated in Figure 3. Thanks to our structured medical records dataset, the responses of our dialogue system can be guaranteed to be accurate, which is crucial. Because if the simulated patient's response does not match the record, the diagnosis made by the real doctor for this record will no longer be applicable, which could complicate the evaluation process.

We need annotated doctor-patient dialogue data to train our task-oriented dialogue system. Previously, medical dialogue systems were typically trained to act as doctors, focusing mainly on patients' utterances. However, our goal is to train a task-oriented dialogue system that responds as a patient to questions from medical LLMs. Consequently, the annotated training data available to us

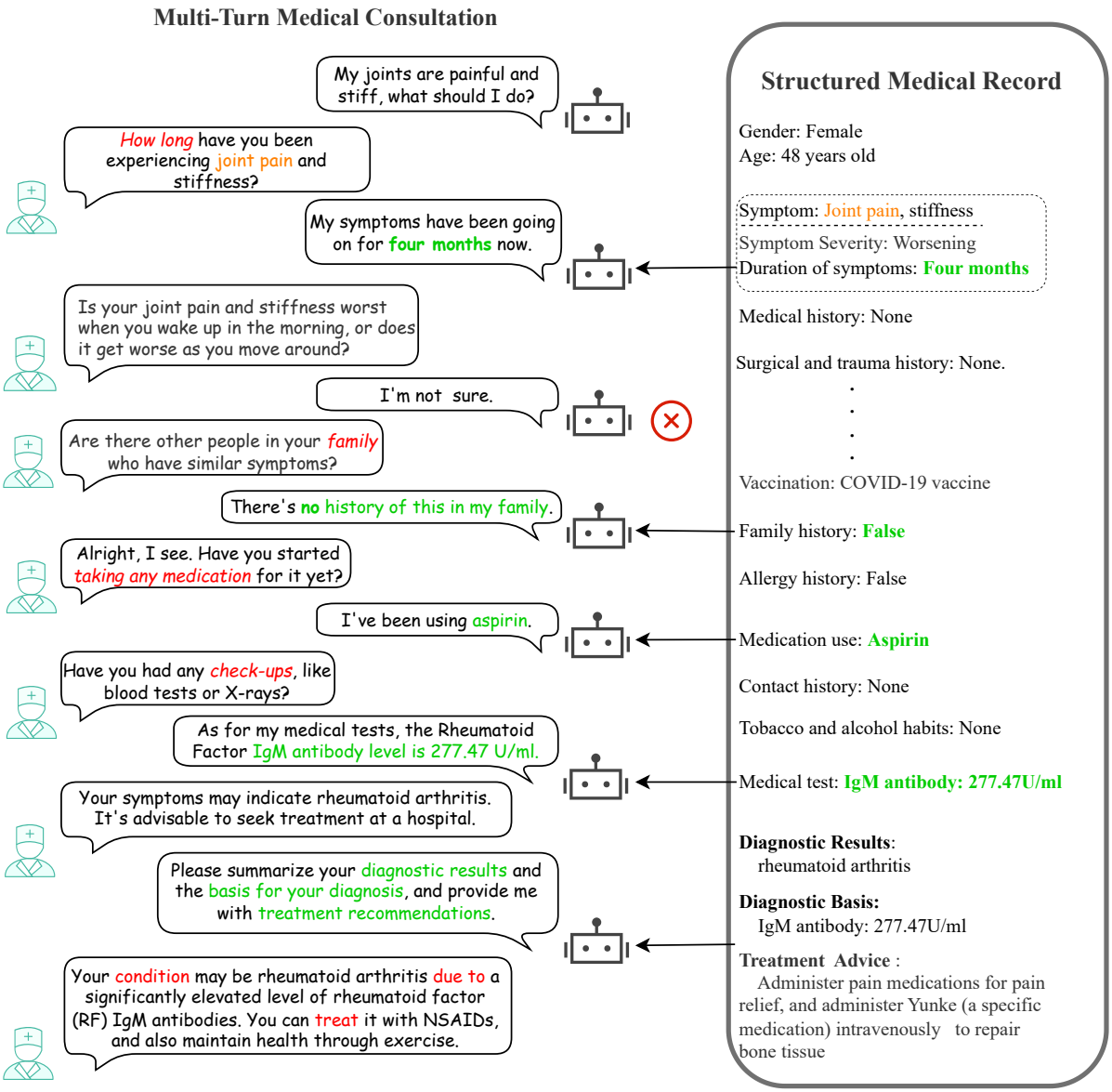


Figure 2: The demonstration of the Interactive Medical Consultation Evaluation Process: Our system acts as the role of the patient, engages with the medical LLM by referencing structured medical records dataset, and ultimately requests a diagnostic summary from the medical LLM for evaluation purposes.

is very limited, which means we need to annotate the data ourselves. We observed significant stylistic differences between the LLM-generated responses and real-world doctor-patient dialogues. In real life, doctors tend to engage in multiple, high-frequency exchanges with short responses each time, whereas LLMs prefer fewer rounds with longer responses each time. Considering our patient model interacts with medical LLMs rather than real doctors, to bridge the generalization gap, we chose to use two GPT-4 models to simulate doctor-patient dialogues. We collected medical consultation information from an online forum[2], processed it, and used it as input for GPT-4. To enrich the dialogue content, we used prompts to give GPT-4 different personalities. We filtered the collected dialogue data and then began annotating it, focusing on the doctor's intent and dialogue state. We treated intent recognition as a multi-label classification problem with 26 labels, while dialogue state tracking was regarded as a generation task with 11 slots. Professional doctors annotated 1,000 instances for testing, and the rest of the data was annotated by our researchers with the assistance of LLMs, initially labeled by the LLMs and then manually corrected. The data annotated by professional doctors served as the test set, and the rest was used for training

[2]https://dxy.com/

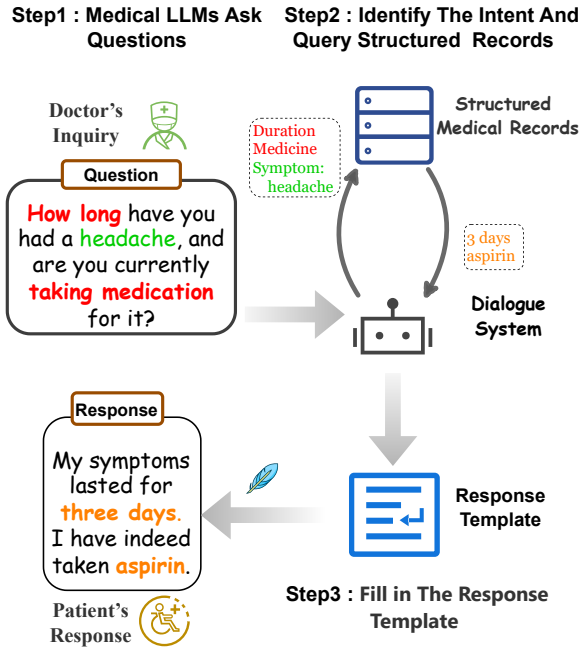


Figure 3: An illustration of our task-oriented dialogue system acting as a patient, handling and responding to a doctor's questions. The system recognizes the doctor's intent and captures the dialogue state, then queries the structured medical records, and finally formulates a response by filling in a reply template.

Task Type	Training	Validation
Intent Recognition	9176	2294
Slot Filling	7576	1894

Table 1: The scale of the dataset used for training our task-oriented dialogue system.

and validation (detailed in Table 1). We found that due to limited requirements for medical knowledge, the data annotated by our researchers with the help of LLMs was sufficient for training purposes. Detailed annotation principles are provided in Appendix A.2.

We trained the task-oriented dialogue system using the aforementioned training data. For intention recognition, we approached it as a multi-label classification problem, fine-tuning DeBERTa-v2-97M (He et al., 2020; Zhang et al., 2022) as the pre-trained model. For dialogue state tracking, we fine-tuned T5-small-77M (Raffel et al., 2020; Zhang et al., 2022), with results shown in Table 2. Hyper-parameter settings and input formats for T5 models are detailed in Appendix B.

Structured Records Dataset. Our structured medical records dataset serves dual purposes as a query resource for the dialogue system and as a source of reference answers for evaluations. We have designed a total of 2,636 structured medical records from various departments including internal medicine, pediatrics, surgery, obstetrics, and others (see in Table 3). Each record includes information such as the patient's symptoms, age, weight, past medications, family history, and more, illustrated in Figure 2. We adapted the initial patient statement from the original complaints, changing the perspective to match that of the patient, enhancing its suitability for oral delivery. Additionally, our structured medical records dataset included diagnostic results, basis, and treatment suggestions derived from professional medical conclusions. These items served as reference answers to evaluate the diagnostic capabilities of medical LLMs. Original records were sourced from professional medical forums[3] with privacy information removed. Further details about the structured medical records dataset are provided in Appendix A.1.

Metric	Result	Metric	Result(%)
F1(%)	82.5	Precision	86.2
ROUGE-1	68.2	Recall	84.9
ROUGE-2	42.9	F1	85.5
ROUGE-L	67.8	Accuracy	98.9

Table 2: Performance of our dialogue system on the validation set: dialogue state tracking results are shown on the left, and intent recognition results are shown on the right.

Evaluation Criteria. After collecting dialogue data from medical LLMs, we primarily assess them from two aspects: communication competence and clinical diagnostic competence. The former represents the ability of medical LLMs to communicate with patients, while the latter represents their professional medical diagnostic capabilities. Since our evaluation simulates real medical consultation scenarios, if the medical LLM's communication skills are inadequate, it will not be able to obtain sufficient information during interactions with patients, which will be detrimental to the subsequent diagnostic analysis. See Section 4 for details.

3.2 Evaluation Process

The interaction process is illustrated in Figure 2. In structured medical records datasets, each entry consists of key-value pairs, showing patient information. The patient's initial statement in the

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

Department	Num
Internal Medicine	944
Surgery	1019
Obstetrics and Gynecology	404
Otorhinolaryngology	38
Pediatrics	87
Others	144
Total	2636

Table 3: Departmental distribution of the structured medical record dataset. For a more detailed breakdown of departments, see Figure 4.

structured record is used to initiate the dialogue process. This statement is passed as input to the medical LLMs, which generates a response. This response is received by the patient model in the system, which identifies the doctor's intent and dialogue state, then searches the structured record database. Based on the intent and dialogue state, the patient model selects an appropriate template, fills it with the found record data as a response, and passes it back to the medical LLMs, thus starting the next round of interaction. The patient model briefly records the dialogue state to prevent repetitive answers and interrupts the questioning when the doctor model no longer poses new questions, instead asking the medical LLMs to summarize the diagnostic findings. We have designed questions in three aspects: diagnostic results, treatment recommendations, and diagnostic basis. Given that we already have reference answers for these three questions, we can evaluate the medical LLM's capabilities in diagnosis, treatment, and analysis with greater accuracy.

4 Experiment

The multi-turn interactive diagnostic capability of medical LLMs is reflected in two dimensions: **Communication Competence** and **Clinical Diagnostic Competence**. In this section, we will evaluate medical LLMs around these two aspects. We have tested various general and medical-specific LLMs, including ChatGLM and HuaTuoGPT (detailed in Table 10). Next, we designed experiments to verify the effectiveness of our evaluation system. First, we compared the performance of our system with that of LLMs on intent recognition and dialogue state tracking test datasets. Then, we examined whether our template-based responses affected the judgment of medical LLMs. The results show that we achieved high accuracy while maintaining low resource consumption.

4.1 Evaluation of Communication Competence

The communication competence of medical LLMs encompasses both the depth of the conversation and the tone in which it is delivered.

Depth of Conversation. The depth of conversation refers to the ability of medical LLMs to obtain more information through interaction with patients. This involves whether the discussion about the patient's condition is sufficiently detailed, which is reflected in the number of dialogue turns, the quantity and type of questions posed by the medical LLM, and the length of the medical LLM's responses. For these metrics, we directly counted the dialogue information.

Expressive Nuances. Higher-level communication skills do not solely rely on the content itself but on the manner of expression, tone, language style, and communication techniques. Doctors should exhibit professionalism, empathy, patience, and respect in their tone when communicating with patients. These nuances in expression are difficult to measure with rule-based methods; thus, we attempted to assess the expression style of medical LLMs using LLM-based approaches.

Settings. To evaluate the communication competence of medical LLMs, we constructed multi-turn dialogue data between the medical LLMs and simulated patients on the structured medical records dataset. We measured the depth of communication by statistics on the number of questions, the number of dialogue rounds, and the length of responses from the medical LLMs, and we assessed the expression nuances of the medical LLMs using GPT-4 (detailed in Appendix C.1). The results are presented in Table 4. We also statistically analyzed the types of questions asked by the medical LLMs during the dialogue. Detailed statistical information can be found in Appendix C.1.

Results. Through our experiments, we found that even in medical scenarios, general LLMs exhibit superior proactive communication abilities with users compared to specialized medical LLMs. We speculate that this might be due to the fact that the base models utilized by medical-specific models are not of sufficient quality, and they have not been

Model	Average sentence length	Average Number of Dialogue Turns	Average Number of Questions Asked	Win Rate
BianQue2	142.14	1.12	0.22	0.02
HuatuoGPT2-7b	266.14	1.13	0.22	0.34
Ming	180.22	2.25	2.12	0.06
Pulse	232.22	2.31	3.13	0.04
HuatuoGPT2-13b	240.09	1.13	0.21	0.34
Baichuan2	186.56	2.00	2.67	0.38
ChatGLM3	215.28	**2.44**	3.59	0.24
Qwen1.5	**376.70**	1.93	**4.07**	**0.68**
ChatGPT-3.5	217.72	2.2	3.42	0.50

Table 4: The results of testing LLMs on a structured medical records dataset, including conversational statistics and the communication skills of the medical LLMs evaluated using GPT-4. *Win Rate* refers to the probability that, when compared to ChatGPT 3.5, the model prevails in terms of communication ability

specifically optimized for communication capabilities. Notably, some medical LLMs tend to ask multiple questions at once, potentially reducing the number of dialogue rounds.

4.2 Evaluation of Clinical Diagnostic Competence

Settings. We believe that the medical capability of LLMs lies mainly in their ability to reasonably infer the patient's condition and propose suggestions based on this inference. Therefore, we plan to start by evaluating these aspects. Our task-oriented dialogue system, acting as the patient, has a simple dialogue management module capable of determining whether the doctor-patient interaction has largely concluded. Specifically, if the medical LLMs no longer pose new questions, we consider the information collection phase complete and proceed to the next stage. In this stage, the medical LLMs are required to summarize the diagnosis results, diagnostic bases, and treatment recommendations based on the dialogue information. When constructing our structured medical records dataset, we considered these three aspects, deriving answers from professional medical analyses. These answers serve as reference points for evaluating the medical analytical skills of LLMs. We employed various evaluation methods, including rule-based approaches represented by ROUGE, neural network-based methods represented by BERTScore, and LLMs-based methods exemplified by GPT-4. The results are presented in Table 5. For more detailed information, please refer to Appendix C.2.

Results. Our experiment found that in medical consultation scenarios, even after multiple rounds of dialogue information collection, traditional evaluation metrics such as ROUGE and BERTScore struggle to effectively assess medical LLMs. This is primarily due to the inherent complexity of medical diagnoses, for example, there may be multiple treatment options for a single patient. Consequently, even with reference answers, they cannot serve as definitive benchmarks for evaluation. This poses a significant challenge in evaluation tasks. Future evaluation efforts may necessitate the creation of more detailed datasets. This could result in a narrower evaluation scope, concentrating on specific departments, due to the complexity and high labor costs associated with producing such datasets.

4.3 Performance of Our Task-Oriented Dialogue System

Settings. To validate that our task-oriented dialogue system acting as patients can fully understand questions from doctors, we compared its performance against LLMs in intent recognition and dialogue state tracking on the test set annotated by professional physicians. In our design, the primary role of the patient model is to accurately answer the doctor's questions. However, since we use a template-based response method, the simulated patient's statements can sometimes be rigid and mechanical. Therefore, we attempted to use LLMs to enhance the patient's responses, making them more akin to those of real patients. Given the LLMs' excellent performance in generation tasks, it is relatively easy for them to stylize the pa-

Model	ROUGE1	ROUGE2	ROUGEL	BERTScore	Win Rate
BianQue2	8.83	0.95	6.75	54.04	0.12
HuatuoGPT2-7b	9.81	1.47	6.93	55.53	0.34
Ming-7b	10.81	1.8	7.5	56.06	0.28
Pulse-7b	9.67	1.88	7.59	55.29	0.08
HuatuoGPT2-13b	9.44	1.32	6.74	55.29	0.30
Baichuan2-7B-Chat	9.39	1.44	7.19	54.61	0.14
ChatGLM3-6b	10.91	2.04	7.74	55.78	0.36
Qwen1.5-7B-Chat	12.0	**2.64**	7.55	**57.22**	**0.76**
ChatGPT-3.5	**12.05**	2.5	**8.22**	56.93	0.50

Table 5: Evaluation results of the medical capabilities of LLMs, assessed on three issues: diagnosis results, treatment recommendations, and diagnostic reasons, and then averaged the results. The win rate refers to the probability of winning compared to ChatGPT-3.5

Model	Precision (Intent,%)	Recall (Intent,%)	F1 (Intent,%)	ROUGE-1 (Slot Value)	ROUGE-L (Slot Value)	F1 (Slot Key,%)
Ours	**51.4**	**76.9**	**61.6**	**73.0**	**72.7**	**80.1**
ChatGPT-3.5	25.4	64.4	36.4	42.7	42.5	18.5
Qwen1.5-7B	47.6	67.0	55.7	58.5	58.1	50.8

Table 6: Comparing the two typical LLMs, **Qwen1.5-7B** and **ChatGPT-3.5**, on a dataset annotated by professional doctors, where *Intent* refers to the intent recognition classification task, and *Slot* represents the dialogue state tracking task. Given that dialogue state tracking involves slots and values, we separately calculate the accuracy of the slots and the co-occurrence rate of the values.

tient model's responses into colloquial expressions. We used Qwen1.5-4B for this optimization and repeatedly evaluated the communication abilities of LLMs. Details can be found in Appendix C.3.

Results. The results in Table 6 show that our system outperforms ChatGPT-3.5 and Qwen1.5-7B in both intent recognition and dialogue state tracking. This indicates that our system is capable of accurately identifying the doctor's intent and understanding the context. Comparing the responses of our simulated patients before and after being polished by an LLM, we found that the communication ability of the medical LLMs slightly improved after being polished, with an average increase of 1.23%, while the medical capability slightly declined, averaging 2.54%. We believe this may be because input with a conversational style is more likely to stimulate the model's proactive responses, whereas structured input can convey more accurate information. Furthermore, we found that different medical LLMs react differently to polished inputs. When interacting with simulated patients whose dialogues have been polished, some models become more proactive, while others become more passive; however, most remain largely unchanged in their responses. We hypothesize that this could be due to varying preferences among different models for input styles. Some models may prefer conversational or colloquial input, whereas others might favor structured information. This might be because different models focus on different corpora during training. The detailed results are shown in Appendix C.3.

5 Conclusion

We construct a task-oriented dialogue system acting as a patient to engage in multi-turn interactions with Medical LLMs on the structured medical records dataset, thereby accumulating multi-turn dialogue data. Subsequently, we evaluate the Medical LLMs under the scenarios of interactive diagnosis using two metrics: communication competence and clinical diagnostic competence. This outlines our evaluation system's procedure. Our experimental results substantiate the effectiveness of our evaluation system. Contrasted with LLM-based multi-turn evaluations, our system offers greater accuracy, can be deployed locally, and is more resource-friendly.

Limitations

Our study comes with a few limitations. Firstly, in the development of our task-oriented dialogue system, we utilized LLMs for assistance in annotation. Despite conducting a comprehensive manual review, the initial reliance on LLMs for labeling, combined with potential annotator complacency, may have undermined the reliability of the annotations. Consequently, this could have adversely impacted the performance of our dialogue system, indicating a need for further refinement.

Secondly, Constrained by the number of model parameters, traditional neural network models struggle with handling out-of-vocabulary (OOV) words. This limitation potentially restricts the generalizability of our model across various medical specialties.

Lastly, our patient model is conceptualized as an ideal entity, designed to cooperate with the physician's queries and actively participate in the dialogue. This design was primarily motivated by the need for fairness and simplification of the task. While incorporating diverse patient personality traits could add realism, it would also introduce significant complexity due to the wide range of human personalities. This could potentially be addressed through LLMs enhanced with prompt engineering. However, designing the system around an ideal patient helps to avoid biases. Although LLMs enable the generation of more varied patient responses, it is imperative that these responses are strictly based on the provided clinical information. If the patient model fabricates details not contained in the medical records to achieve diversity in responses, the validity of the diagnostic outcomes and other record-specific information could be compromised.

Potential Risks

The raw medical records we collect may contain patient privacy information. To address this, we employ various methods to eliminate such information, such as rule-based approaches to remove names and locations, as well as large model-based methods for screening and transformation. Additionally, our datasets undergo manual verification during annotation, which reduces the risk of privacy breaches.

Ethical Considerations

The data we collected from medical websites was uploaded by doctors for learning and communication, and it is not proprietary data of the websites. According to the copyright requirements of the websites, non-commercial reproduction is allowed as long as the source is cited. Our work does not involve commercial activities and complies with the reproduction requirements of the websites, and we have also cited the sources. After collecting this data, we filtered it to ensure that it does not contain any patient privacy information.

References

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. Medbench: A large-scale chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16, pages 17709–17717.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, et al. 2023a. 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*. Btac817.

Xiaolan Chen, Jiayang Xiang, Shanfu Lu, Yexin Liu, Mingguang He, and Danli Shi. 2024. Evaluating large language models in medical applications: a survey. *arXiv preprint arXiv:2405.07468*.

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

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou.

2024. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv preprint arXiv:2402.09742*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Zexue He, Yu Wang, An Yan, Yao Liu, Eric Y Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2023. Medeval: A multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. *arXiv preprint arXiv:2310.14088*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

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*.

Lei Liu, Xiaoyan Yang, Fangzhou Li, Chenfei Chi, Yue Shen, Shiwei Lyu Ming Zhang, Xiaowei Ma, Xiangguo Lyu, Liya Ma, Zhiqiang Zhang, et al. 2024. Towards automatic evaluation for llms' clinical capabilities: Metric, data, and algorithm. *arXiv preprint arXiv:2403.16446*.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. 2023. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Jeffrey Zhao, Raghav Gupta, Yuan Cao, Dian Yu, Mingqiu Wang, Harrison Lee, Abhinav Rastogi, Izhak Shafran, and Yonghui Wu. 2022. Description-driven task-oriented dialog modeling. *arXiv preprint arXiv:2201.08904*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Appendix

A Data Construction and Preparation

A.1 Construction of the Structured Medical Case Dataset

In this study, we primarily employ LLMs and prompt engineering to transform raw medical records data into structured medical records data. We source the records information shared by doctors on professional medical forums, ensuring a high level of expertise as these contributions come from medical professionals. The original medical records data encompasses numerous items, which are detailed in Table 16. Our transformation process predominantly utilizes prompt engineering techniques, supplemented by regular expression matching to handle fixed-format responses, such as age and gender.

We discovered that a large model's tagging ability significantly decreases when asked to handle multiple tagging tasks at once. For structured medical records, we use the large model for grouped tagging with similarly phrased prompts. Finally, the structured medical records we obtain are illustrated in Table 17.

During clinical consultations, although the patient's lifestyle and other background information are also taken into consideration, paramount importance is given to the patient's symptoms and their specific details. To provide targeted responses focusing on symptoms, we have devised a set of descriptive metrics for each symptom extracted, such as duration and location of onset, among others. When a medical model inquires about the detailed aspects of a symptom, our dialogue system first matches the symptom name, identifies the query concerning the detailed aspects of that symptom from the medical model, and finally generates a response. For instance, when the medical model seeks specifics about the timing of a patient's headache, the dialogue state tracking model within our dialogue system extracts the symptom as headache, while the intent recognition model identifies the intent as querying the duration of the symptom. Upon acquiring this information, a response can be generated by consulting a structured medical case database.

To annotate these medical records data, we employed Large Language Models (LLMs) and the methodology of prompt engineering for annotation purposes. Specifically, we utilized the Qwen1.5-72B model for this task. From the aforementioned table, it becomes evident that raw medical records inherently possess certain structural information. Drawing inspiration from this observation, when employing LLMs for annotation, we refrained from feeding the entirety of the medical records content into the LLMs as input. Instead, we selectively extracted portions relevant to the annotation requirements. For instance, in annotating patient allergy information, we drew upon historical and current medical histories to gather pertinent details, which were then presented to the LLMs along with specialized prompts to facilitate annotation.

We endeavored to minimize the complexity of the annotation tasks, thereby enabling the LLMs to generate more accurate annotation outcomes.

A.2 Construction of Train Dataset

This subsection introduces how we constructed the training set, validation set, and test set for our task-oriented dialogue system, including the specific processes of original data collection and cleaning, annotation specifications, and so on.

Collection and Cleansing of Training Datasets. In our research, we observed that the response style of LLMs significantly differs from human linguistic styles. By analyzing the IMCS-V2 (Chen et al., 2022) dataset and the dialogue data constructed using GPT-4, we found stark contrasts in terms of dialogue length and number of turns. Specifically, the average sentence length in authentic doctor-patient conversations was 13 words, with the number of turns primarily ranging from 10 to 30. Conversely, when LLMs engage in medical contexts, their average dialogue length extends to 99 words, and most conversations conclude within five turns. This discrepancy underscores the notable divergence between the conversational styles of LLMs and human interactions. Given that our objective is to evaluate large medical models rather than real doctors, we opted not to annotate genuine doctor-patient dialogues but instead focused on conversations generated by medical LLMs. We amassed a substantial amount of doctor-patient interaction data from medical forums such as Dingxiang Doctor[4]. These interactions were refined and used as seeds for prompts inputted to GPT-4, which acted as the patient. Accompanying these prompts were various patient personality traits, which facilitated GPT-4 in generating responses with diverse tones and styles. As a result, we generated ap-

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

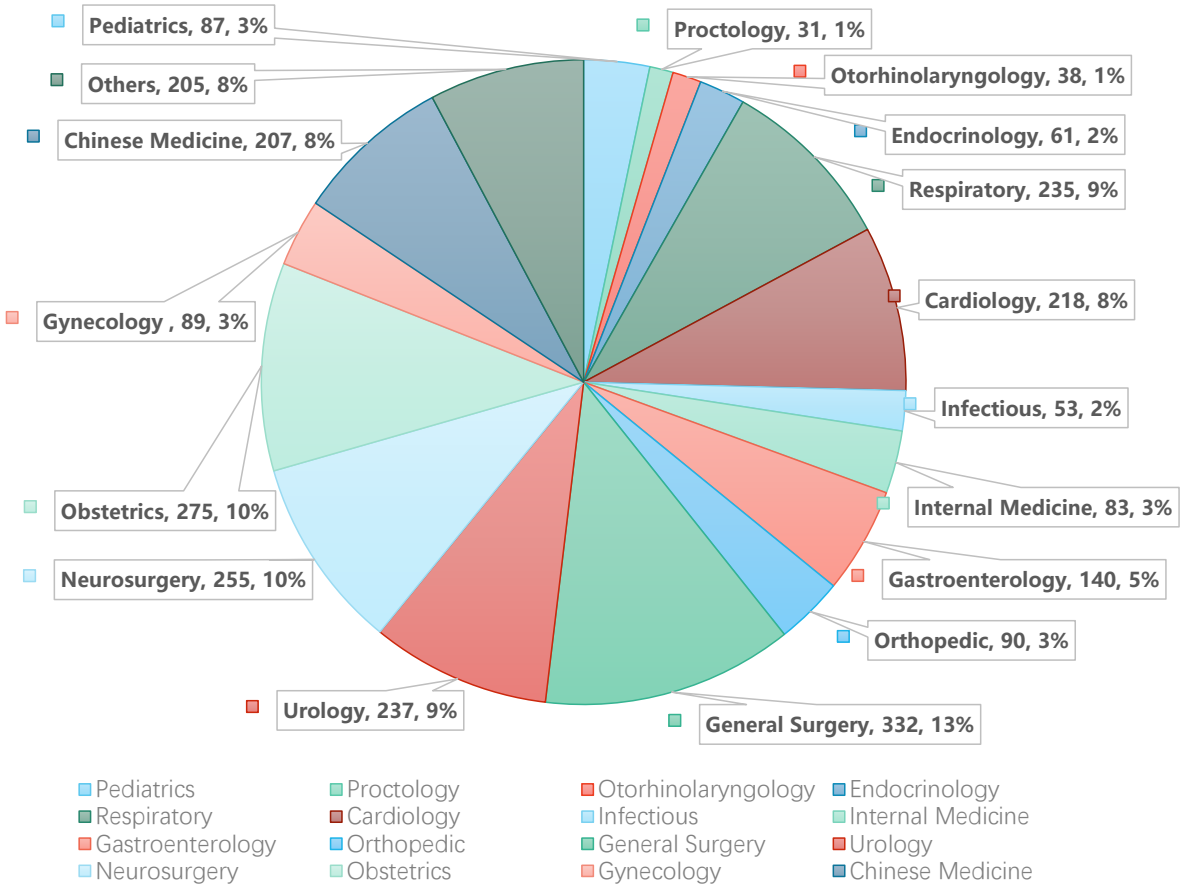


Figure 4: The detailed distribution of departments in the structured medical record dataset.

proximately 200,000 instances of doctor-patient exchanges using GPT-4, encompassing both single-turn and multi-turn dialogues. From this pool, we selected 20,000 sentences uttered by doctors for annotation, ensuring data quality through rigorous screening from multiple perspectives. Filtering criteria included:

- **Dialogue Turns**: We prioritized multi-turn dialogues over single-turn ones and discarded dialogues with excessive turns showing instability or repetition.
- **Model's Self-awareness**: Although GPT-4 has strong directive-following abilities, we found that in certain cases GPT-4 would forget its role as a doctor or patient and instead reveal its identity as a language model. This may interfere with the diagnosis process, so we discarded this part of the data.
- **Oral Dialogue Style**: Some of the seeds we collected to initiate doctor-patient conversations did not conform to the norms of spoken language, such as containing URLs, overly structured dates and times, and nonsense characters. Additionally, some seeds provided as prompts contained insufficient information, making it difficult for GPT-4, acting as the patient, to elaborate on its condition based on these cues. Therefore, these parts were also discarded.
- **Repeated Dialogue**: In some cases, when GPT-4 plays both the patient and the doctor, it can fall into a pattern of repetitive dialogues, where the exchanges are exactly the same. These redundant conversations lack depth and fail to add any significant value to the discourse.
- **Excessive Sentence Length**: We have found that in medical scenarios, if GPT-4 produces outputs longer than 500 words under our settings, most of the content is devoted to listing more detailed circumstances. Although these responses are lengthy, they do not provide much useful information.
- **Incomplete Dialogue**: GPT-4 incorrectly judged the conditions for stopping the dialogue, halting the conversation while it was still ongoing, resulting in incomplete dialogues, often with the doctor not responding to the patient's

questions.

- **Improper Length of Dialogue Turns**: Most LLMs do not have conversational abilities like humans. In our setup, interactions between GPT-4 instances often conclude within five turns. This may be due to LLMs tending to provide detailed responses, which can quickly bring dialogues to an end. We have observed that when the number of dialogue turns is too high, the conversation becomes unstable, for example, deviating from the medical consultation theme or becoming repetitive.
- **Inappropriate Prompts**: We collected medical consultation information online and used summaries of patient information as part of the input for large language models. However, the quality of this information might not be sufficient. We filtered out patient information that did not conform to colloquial habits and descriptions that were too vague, thereby improving the quality of the dialogue.

We employed methods based on LLMs and rule-based methods for filtering, reducing the original dataset of approximately 200,000 dialogue entries to about 20,000 single sentences. This alleviated the burden of annotation and enhanced the quality of the training data.

Annotation of the Training Dataset. To reduce the cost of annotation, we employ multiple LLMs and rule-based annotation schemes for labeling the training and validation sets. Subsequently, we use a large model to arbitrate the annotation results obtained from different methods, followed by manual verification for final validation. For the test set, we hire professional doctors for annotation. Since we aim to train two models, a classification model for intent recognition and a generative model for dialogue state tracking, we need to annotate each sentence in two aspects.

The prompts used for our annotations are shown in Appendix A.2. To fully leverage the annotation capabilities of LLMs, we design a prompt for each label. However, due to space limitations, here we introduce one prompt for slot annotation and one for intent annotation.

Instruction (System Prompt for Annotating the 'Symptom' Slot):

```
In medicine, symptoms refer to the
abnormal states or experiences
subjectively felt by patients during the
 course of a disease. These are direct
experiences and descriptions of
discomfort reported by the patient. For
instances, headache, fatigue, nausea,
etc.
The following sentence is a question
posed by a doctor to a patient. Please
assist me in completing the sequence
labeling task to determine whether the
names of the patient's symptoms are
mentioned (note that symptoms differ
from diseases; symptoms are the patient'
s subjective feelings and descriptions
of discomfort, while diseases are
diagnosed through a comprehensive
analysis of a series of symptoms and
signs in medicine, characterized by
objectivity and universality. Doctors
diagnose diseases through a
comprehensive analysis of symptoms and
signs, combined with laboratory and
imaging tests).
If the names of the patient's symptoms
are mentioned, please reply with the
specific symptom names, using the
content from the original text as much
as possible, and reply with the name of
the symptom, not a descriptive
explanation. Please analyze objectively
without adding your personal
interpretation. If there is no inquiry
related to the name of the symptom,
please reply 'None'. Please reply in
Chinese, only providing the answer
without explaining the reasons or
providing explanations.
Here is an example:
  'Do you have dizziness?'
The answer would be:
  'Dizziness'.
The sentence you need to evaluate is:
  [Start of Sentence]
  {question posed by a doctor}
  [End of Sentence]
Please provide your answer below.
```

Instruction (System Prompt for Annotation of Medical Examination Inquiry Intent):

```
The following sentence is a question
posed by a doctor to a patient. Please
assist me in completing the intent
recognition task to determine whether
the doctor is requesting the patient's
medical examination results, referring
to medical auxiliary examinations and a
few physical examinations that require
instruments. For example,
  1. Bacteriological examination,
  2. Blood pressure,
  3. Heart rate,
  4. X-ray examination,
  5. CT scan,
  6. MRI examination,
  7. Ultrasound and electrocardiogram
  examination,
  8. Blood tests (complete blood count,
```

```
  biochemical tests, coagulation
  function tests, etc.),
  9. Urine tests,
  10. Sputum tests,
  11. Cerebrospinal fluid examination,
  12. Endoscopic examination,
  13. Endocrine function tests. If asked
  , please reply with {'label': 1}.
Try to use the content from the original
 text in your response, and objectively
analyze without adding your personal
interpretation. If there is no request
or inquiry about medical examination
results, please reply with {'label': 0}.
Only the answer is needed; no
explanation or justification is required
.
Here is an example:
  'Have you had your blood pressure
  checked?'
The answer is
  {'label': 1}.
The sentence you need to judge is:
  [Start of Sentence]
  {question posed by a doctor}
  [End of Sentence]
Please provide your answer below.
```

To ensure the quality of the annotations, we tried to employ multiple LLMs for labeling along with manual annotations. These results are comprehensively considered and arbitrated.

To reduce labor costs, we primarily use LLMs for arbitration, supplemented by rule-based labeling and manual verification. For each piece of annotation content, we have designed prompts to assist the LLMs in arbitration. Due to space limitations, we present one example as an illustration.

Instruction (System Prompt for Arbitration of Symptom):

```
In medicine, symptoms refer to the
abnormal states or experiences
subjectively perceived by patients
during the course of a disease,
representing the patients' direct
experience and description of discomfort
. Examples include headache, fatigue,
nausea, etc.
Please assist me in completing a
sequence labeling task, which involves
extracting symptom names from a sentence
. There are already answers from two
individuals, which may not necessarily
be correct and should only be used as
references. You need to consider the
given sentence and the answers provided
by the two individuals to formulate your
 own response. Remember, you are
completing a sequence labeling task, so
provide a precise answer without any
additional explanation.
The sentence to be evaluated is:
  {Question}
The first person's answer is:
  {Answer of the first person},
The second person's answer is:
  {Answer of the second person}.
Now, please provide your answer, your
answer is:
```

Annotation of the Test Dataset. We employed professional physicians to annotate the test set. However, since the annotation of the test set was completed earlier than that of the training set, some items mentioned in the training set are absent in the test set. Additionally, due to improvements made in the experimental design later on, the details of the annotations between the two sets are not entirely consistent, and thus should only be considered as a reference. The results on the validation set are not comparable to those on the test set.

Annotation Guidelines for Symptom Names Provided to Physicians:

```
Symptom names are categorized into four
scenarios. The first scenario is when
the symptom name appears and the
question pertains specifically to that
symptom name; in this case, the symptom
name should be annotated and appended
with "1". The second scenario is when
the symptom name appears in the question
, but the question is not about the
symptom name; here, the symptom name
should be annotated and appended with
"0". The third scenario involves
questions about the presence of a
symptom without specifying which symptom
; in this case, simply mark it with "1".
 The fourth scenario is when there is no
 relevant question, and the item should
be left unmarked.
A symptom refers to an abnormal physical
 or psychological experience or
sensation caused by a disease or
condition, such as headache, cough, or
fever, representing specific
physiological phenomena. It is important
 to distinguish between a symptom and an
 illness; the latter refers to the signs
 or specific manifestations of a disease
, such as pneumonia, diabetes, or cancer
, and does not include symptoms like
weight.
Example 1: "How long have you had a
fever?"
Annotation result: "Symptom Name": Fever
 0
Explanation: The symptom of fever is
mentioned, hence it is annotated as
fever. However, the question is about
the duration, not the presence of fever,
 thus a "0" is appended.
Example 2: "Do you have symptoms of
fever?"
```

```
Annotation result: "Symptom Name": Fever
 1
Explanation: The symptom of fever is
mentioned, hence it is annotated as
fever, and the question indeed pertains
to whether the symptom is present,
therefore a "1" is appended.
Example 3: "May I ask your age?"
Annotation result: This item should be
left blank, as there is no relevant
inquiry.
Example 4: "Do you have any other
symptoms?"
Annotation result: "Symptom Name": "1"
Explanation: The question asks about the
 presence of symptoms in general without
 specifying which, thus only "1" is
marked.
```

B Training Details

We trained two models for intent recognition and dialogue state tracking. We fine-tuned DeBERTa-v2-97M-Chinese as our intent recognition model, which handles a multi-label classification problem with 26 categories. For dialogue state tracking, we employed T5-small-77M, utilizing a generative task format to complete sequence labeling.

B.1 Demonstration of Training Samples

It is noteworthy that T5 is highly sensitive to prefixes when addressing different types of tasks, which may be seen as a rudimentary form of prompt engineering. However, due to the significantly smaller parameter size of T5 compared to LLMs, the prompts used with T5 differ from those used with LLMs. For T5, the prefix should not detail the task extensively; instead, the simpler, the better. Following the principles outlined in (Zhao et al., 2022) we designed the following prefixes.

Prefix for T5 for Dialogue State Tracking:

```
1. Symptom Name
2. Past Diagnoses
3. Past Medications
4. Physical Changes
5. Trauma Surgery
6. Preventive Vaccinations
7. Allergy History
8. Exposure History
9. Family History
10. Lifestyle Habits
11. Medical Examinations\n
```

We annotated each sentence from two aspects, meaning that one sentence can be used to train two models. An example of such a training sample is as follows.

Example of Training Sample for Intent Recognition Model:

```
Question:
  I would like to know, did this itching
   sensation and the red spots appear
  suddenly or did they gradually worsen
  ?"
Label:
  ["ask_symptom", "
  ask_symptom_description_degree", "
  ask_symptom_description_time"]
```

Example of Training Sample for Dialogue State Tracking Models:

```
Question:
  1.Symptom Name
  2.Past Diagnoses
  3.Past Medications
  4.Physical Changes
  5.Trauma/Surgery
  6.Immunizations
  7.Allergy History
  8.Exposure History
  9.Family History
  10.Lifestyle Habits
  11.Medical Examinations

  Firstly, do you regularly take
  antihypertensive medication and follow
   the directions correctly?"
Label:
  "3:Antihypertensive Medication"
```

B.2 Hyper-Parameter Details

The key hyperparameters used during training are shown in Table 9. By adjusting strategies such as learning rate and early stopping, we fine-tuned DeBERTa-v2-97M-Chinese as our intent recognition model and fine-tuned T5-cn as our dialogue state tracking model. We employed these two models to compose our task-oriented dialogue system, which acts as a virtual patient interacting with a doctor.

C Evaluation Details

C.1 Evaluation of Communication Ability in Medical LLMs

Considering that assessing communication skills does not require extensive medical knowledge, we hire laypeople to conduct manual evaluations. We selected 200 data cases and used LLMs for evaluation alongside human evaluation. The resulting Cohen's Kappa coefficient was 0.697, indicating a high degree of correlation between the assessments made by human evaluators and LLMs. In 88.0% of the cases, the conclusions drawn from

Model	Average sentence length	Average Number of Dialogue Turns	Average Number of Questions Asked
BianQue2	0.31	0.91	**39.24**
HuatuoGPT2-7b	1.92	1.80	5.88
Ming	**-4.25**	2.22	0.28
Pulse	**3.23**	-0.86	-2.64
HuatuoGPT2-13b	1.07	-3.48	**-10.32**
Baichuan2	-2.84	1.50	5.03
ChatGLM3	-2.91	**-4.84**	-5.45
Qwen1.5	-0.45	**2.62**	1.18
ChatGPT-3.5	-0.02	1.36	2.70

Table 7: The growth rates of the average response length, average number of conversation turns, and average number of questions asked when medical LLMs communicate with patients before and after refinement.

Model	Rouge1	Rouge2	RougeL	BERTScore
BianQue2	**1.44**	**2.08**	-1.60	-0.09
HuatuoGPT2-7b	0.61	-4.93	-0.14	0.00
Ming-7b	-1.56	-9.39	-1.58	0.09
Pulse-7b	-1.24	-6.82	-1.72	-0.22
HuatuoGPT2-13b	0.00	1.54	-1.32	-0.07
Baichuan2-7B-Chat	-0.42	-9.49	**-6.21**	**1.35**
ChatGLM3-6b	**-4.96**	**-18.09**	-4.01	**-0.70**
Qwen1.5-7B-Chat	-3.09	-8.98	-2.64	-0.42
ChatGPT-3.5	-0.75	-3.75	**-0.12**	-0.02

Table 8: Changes in metrics of diagnostic outcomes made by medical LLMs for patients before and after refinement, where the average values of diagnostic results, diagnostic reasons, and treatment recommendations are considered.

Model	Epochs	Batch Size	LR	LR Scheduler	Optimizer	Max Length	Class Number
DeBERTa-v2-97M	7	64	1e-5	Fixed	AdamW	-	26
T5-small	3	64	1.25e-3	Linear	AdamW	192	-

Table 9: Training Hyper-Parameters applied during the training of the intent recognition model and dialogue state tracking model.LR denotes the learning rate

the LLMs' evaluations were consistent with those of human evaluators. We measure the depth of interaction between medical models and patients by quantifying the number and types of questions asked, the number of rounds of communication, and the length of evaluated responses during their exchanges. Our dialogue system can recognize various questions posed by medical LLMs. Here, we statistically present the frequency of each type of question. These questions reflect the aspects of a patient's condition that medical large language models tend to focus on during consultations. The detailed statistics are shown in Table 15.

As mentioned earlier, our system is more resource-efficient compared to directly simulating patient using LLMs. Here, we have compiled the costs associated with using GPT-4 as the patient during the doctor-patient interaction phase. The results are shown in Table 11. The currency used is USD.

C.2 Evaluation of Medical abilities

We focus our evaluation of the medical capabilities of medical LLMs on the analysis of their diagnostic results, diagnostic reasoning, and treatment recommendations. Due to space constraints, we did not provide the specific performance of these medical LLMs in these three areas in the main text, but rather presented an average result. Here, we provide more detailed results and analysis in Ta-

Model	Parameter Size	Model Type
BaiChuan2	7B	General
BianQue2	7B	Medical
ChatGLM3	6B	General
HuatuoGPT2-7B	7B	Medical
HuatuoGPT2-13B	13B	Medical
Pulse	7B	Medical
Qwen1.5	7B	General
Ming	7B	Medical
ChatGPT-3.5	-	General

Table 10: The Models to be Tested.*General* refers to the general LLMs, while *Medical* refers to the medical LLMs. In our experiment, these models all assumed the identity of a doctor.

Model	Price(USD)
Baichuan2	435.34
BianQue2	235.76
ChatGLM3	547.72
HuatuoGPT2-7B	440.45
HuatuoGPT2-13B	407.54
Pulse	538.24
Qwen1.5	803.27
Ming	408.11
ChatGPT-3.5	511.82

Table 11: The costs incurred from using GPT-4 to simulate patients and interacting with medical LLMs on our structured medical record dataset.

ble 12. Although the current metrics cannot serve as definitive benchmarks, it can still reflect the medical capabilities of LLMs to some extent. In the early stages of the development of medical LLMs, it at least provides some guidance for the training iterations of these models. Due to the specialized nature of medicine, definitive evaluations can only be provided after discussions among professional doctors.

C.3 Performance of Our Task-Oriented Dialogue System

The responses of our dialogue system are generated by filling in templates. The main advantage of this method is its simplicity, allowing us to avoid designing a dialogue generation model, which aligns with our initial goal of low resource consumption. However, generating responses in this manner can result in replies that are not very smooth and may come across as somewhat rigid. Given the strong generative capabilities of LLMs, we are experimenting with using a large model to polish the outputs of our dialogue system, making them more fluent and more in line with conversational habits. Specifically, we are using Qwen-4B for this polishing, with the prompts used detailed as follows.

Instruction (System Prompt for Polishing):

```
Please help me polish the following
sentence to make it more fluent and in
line with spoken language habits,
without changing the original meaning.
Please directly output the polished
sentence without adding any other
content. The sentence you need to polish
 is:
[Start of Sentence]
{response generated by the dialogue
system}
[End of Sentence]
```

The Effectiveness of Polishing We analyzed and statistically examined the changes in communication and medical capabilities of medical LLMs before and after refinement, presenting the growth rates in Table 7 and Table 8. We found that different medical LLMs have varying degrees of acceptance for responses before and after refinement. Some models seem more willing to interact with users who exhibit personality traits, while others prefer more structured input. This may be related to the type of corpus used during the training of the LLMs. Overall, after refinement, the interaction between our dialogue system and the medical LLMs slightly improved.

Regarding medical capabilities, the medical LLMs' judgments when receiving refined patient responses slightly declined. This could be because colloquial information is not as rigorously structured as more formal input, meaning that the information conveyed in a conversational manner may not be as accurate, potentially affecting the medical LLMs' judgments.

In summary, the changes in both communication and medical capabilities of the medical LLMs before and after refinement are not particularly significant. Since our designed system focuses more on completing interactions with the medical LLMs, the step of refining the output can be omitted. In the future, if we aim to enrich the personality traits of simulated patients to make them closer to real scenarios, we might consider adding the refinement step.

Task	Model	Rouge1	Rouge2	RougeL	BERTScore
Diagnostic Results	Baichuan2-7B-Chat	6.98	2.11	5.70	55.55
	BianQue2	4.75	0.77	3.93	55.57
	ChatGLM3-6B	7.42	2.16	5.77	58.42
	HuatuoGPT2-7b	6.61	1.56	5.05	57.71
	Ming-7B	6.70	1.80	5.22	58.07
	Pulse-7B	7.54	2.61	**6.48**	56.59
	Qwen1.5-7B-Chat	**8.59**	**2.68**	6.17	**60.10**
	ChatGPT-3.5	8.43	**2.68**	6.41	59.43
	HuatuoGPT2-13b	5.54	1.12	4.40	58.13
Rreatment Advice	Baichuan2-7B-Chat	9.63	0.88	7.11	55.55
	BianQue2	9.74	0.83	7.43	55.57
	ChatGLM3-6B	9.69	1.12	7.08	58.42
	HuatuoGPT2-7b	9.06	0.76	6.62	57.71
	Ming-7B	**10.95**	1.37	**7.80**	58.07
	Pulse-7B	9.38	1.20	7.42	56.59
	Qwen1.5-7B-Chat	9.71	**1.46**	6.34	**60.10**
	ChatGPT-3.5	10.66	1.38	7.45	59.43
	HuatuoGPT2-13b	8.83	0.72	6.51	58.13
Diagnostic Bases	Baichuan2-7B-Chat	11.56	1.34	8.76	55.55
	BianQue2	12.02	1.25	8.90	55.57
	ChatGLM3-6B	15.61	2.84	10.36	58.42
	HuatuoGPT2-7b	13.76	2.09	9.13	57.71
	Ming-7B	14.81	2.22	9.48	58.07
	Pulse-7B	12.11	1.82	8.88	56.59
	Qwen1.5-7B-Chat	**17.69**	**3.78**	10.12	**60.10**
	ChatGPT-3.5	17.06	3.42	**10.79**	59.43
	HuatuoGPT2-13b	13.95	2.12	9.31	58.13

Table 12: The performance of medical LLMs in the three tasks of diagnostic results, diagnostic reasons, and treatment recommendations.

Analytical Capabilities of Our Task-Oriented Dialogue System. To validate that our task-oriented dialogue system can fully understand doctors' questions, we compared the performance of our system with that of LLMs on two tasks: intent recognition and dialogue state tracking. The intent recognition task measures whether the model can understand the doctors' questions, while dialogue state tracking reflects the model's understanding of the dialogue content. The entire experiment was conducted on our test set, which was annotated by professional doctors and is highly specialized. Our experimental results show that our designed task-oriented dialogue system can fully understand doctors' questions. In these two experiments, we primarily compare Qwen-7B and ChatGPT-3.5, as these two models are highly representative. The prompts we used are as follows.

Instruction (System Prompt for Slot Filling):

```
The following paragraph is given to you,
 which is what the doctor said to the
patient. Please help me complete the
labeling task of sequence labeling and
check whether there are the following
slots and corresponding values in the
doctor's words:
1. Symptom Name: The name of the patient
's symptom
2. Previous diagnosis: What diseases the
 patient has been diagnosed with before
3. Past medications: What medications
the patient has taken before
4. Physical changes: Changes in the
patient's weight, diet and sleep
5. Trauma surgery: What trauma surgery
has the patient had before
6. Vaccination: What vaccines have the
patient received before
7. Allergy history: What allergy history
 does the patient have
8. Contact history: What contact history
 does the patient have
```

```
9. Family history: What is the history
of disease in the patient's family
10. Living habits: the patient's living
habits, including sleeping habits,
eating habits and hygiene habits
11. Medical tests: What medical tests
the patient has had
If so, output the slot and its value.
Output a JSON output, where each slot
corresponds to a key and the value
corresponds to the value.If not, output
an empty dictionary. Please output the
answer directly without explanation.
Note that the slots are represented by
numbers and the corresponding values are
 the corresponding values of the slots.
To give you an example:
[Start of Example]
Have you taken any cold medicine
recently? Are you allergic to penicillin
?
[End of Example]
The correct answer is:
{
"1": "cold medicine ",
"7": "penicillin"
}
The statements you need to judge are:
[Start of Stentence]
{content}
[End of Stentence]
```

Instruction (System Prompt for Intent Recognition):

```
The following paragraph is given to you,
 which is what the doctor said to the
patient. Please help me complete the
annotation task of intention recognition
, and check whether the doctor's words
contain the following intention:
"ask_symptom_duration": This asks for
the duration of the symptom,
"ask_symptom_description_color": Ask for
 the color of the symptom description,
"ask_symptom_description_smell": Ask the
 symptom about the smell,
"ask_symptom_description_degree": Asks
for the degree of the symptom
description,
"ask_symptom_description_position": Ask
for the location of the symptom
description,
"ask_symptom_description_shape": Ask for
 the shape of the symptom description,
"ask_symptom_description_time": The time
 when the symptom description was asked,
"ask_coitus": asking about sex,
"ask_tobacco": Asks about smoking,
"ask_alcohol": Asks about alcohol
consumption,
"ask_drug": Asks about the drug,
"ask_receiving_treatment": Ask what
treatment is being received,
"ask_age": Asks for the age,
"ask_weight": Asks for the weight,
"ask_gender": Asks for the gender,
"ask_symptom": Asks for the name of the
symptom,
"ask_medical_examination": asks about a
medical examination,
"ask_living_habit": Asks about lifestyle
 habits,
"ask_past_medication": Ask for past
medication,
"ask_past_diagonsis": Ask for past
diagnoses,
"ask_physical_change": Asks if the body
has changed,
"ask_trauma_surgery": Ask about trauma
surgery,
"ask_preventive": Asks for preventive
status,
"ask_allergic_history": Ask about
allergy history,
"ask_contact_history": Ask for the
contact history,
"ask_family_history": This asks for the
family history.
If yes, please output the intention,
please output a list of the intention in
 the doctor's words, if not, please
output an empty list. Please output the
answer directly without explanation.
The statements you need to judge are:
[Start of Stentence]
{content}
[End of Stentence]
```

D Details of Task-oriented Dialogue System

To simulate patient interactions, we constructed a task-oriented dialogue system designed to engage in conversations with doctors, replicating authentic consultation scenarios. In this section, we provide a detailed exposition of the system's operational principles and construction specifics. This includes the configuration of intents and slots, as well as the methodology for selecting appropriate templates.

D.1 Settings of Intents and Slots

On one hand, we have statistically analyzed the common questions in doctor-patient interactions within real medical consultation scenarios; on the other hand, we have also referred to the opinions of professional doctors, ultimately selecting 26 doctor intents and 11 types of slots. All intents and slots are listed in Table 13 and Table 14, where F1 represents the recognition effect of the item. The specific numbers do not carry special significance. Intent recognition is a traditional multi-label classification task, while slot filling is a sequence labeling task.

D.2 Selection of Templates

We designed a total of 62 templates across 26 categories to adequately cover the basic questions that

Intent Name	F1(%)
symptom duration	85.4
symptom color	90.9
description smell	85.2
symptom degree	66.7
symptom position	81.6
symptom shape	66.7
symptom time	84.9
ask coitus	77.5
ask tobacco	94.7
ask alcohol	94.7
ask drug	40.0
receiving treatment	75.7
ask age	99.2
ask weight	63.2
ask gender	95.8
ask symptom	91.1
medical examination	80.3
living habit	83.9
past medication	85.7
past diagonsis	70.6
physical change	87.9
trauma surgery	57.8
ask preventive	92.3
allergic history	83.1
contact history	72.2
family history	97.1

Table 13: Recognition effectiveness of each slot.

Slot Name	Explanation
Symptom Name	The name of the patient's symptom
Past Diagnoses	The previous diagnosis of the patient
Past Medications	Medications previously taken by the patient
Physical Changes	Changes in weight, diet, and sleep
Trauma or Surgery	Whether there has been a history of trauma or surgery
Vaccination History	Which vaccines have been administered
Allergy History	Allergies to what substances
Exposure History	Exposure to which pathogenic factors
Family History	The medical history within the family

Table 14: The names and meanings of the slots used in the dialog system.

doctors might ask during consultations. Initially, we compiled a list of common questions that doctors typically ask patients. After consulting with doctors, we developed the intent recognition and slot-filling modules for the task-oriented dialogue system. Subsequently, we designed response templates based on these intents and slots. The selection and filling of response templates were guided by the following elements:

- The doctor's intent.
- Extracted key information.
- Relevant content from the structured medical records.

First, we select the template category based on the doctor's intent. For example, if the doctor inquires about family history, we identify the template category as related to family history. Within the family history category, there are multiple templates available for selection. Next, we choose the most appropriate template and fill in the response based on the key information extracted by the dialogue state tracking model and the relevant content from the structured medical records. Specifically, if the question is "Do you have a family history of heart disease?" and the family history section of the structured medical record indicates heart disease, we can select an affirmative response template by calculating text relevance. If the medical record shows a family history of another disease, such as diabetes, we would respond with a family history of diabetes. If there are no relevant records in the medical history, we would respond with "I'm not clear about my family medical history."

D.3 Avoiding Hullucinations

GPT4 is used to construct the dialog dataset, which is sent to the labelers for labeling the doctor's intent and slots. Then, a task-oriented dialogue system is trained based on these annotations. Even if GPT4 hallucinates, it does not affect the labeling and training of intent recognition and slot-filling tasks. During the evaluation stage, our patient simulator, i.e., the task-oriented dialogue system, recognizes the intent and key slots and generates the response **strictly according to the structured medical records**, which can avoid hallucination issues.

E Response Style Analysis

We use a task-oriented dialogue system to interact with medical LLMs as a patient. To build such a system, we need to annotate doctor-patient dialogue data as a training dataset. We found that the style of real doctor-patient dialogues differs significantly from the dialogues generated by LLMs. As shown in Figure 5, real-world doctor-patient dialogues tend to have high-frequency interactions with less content conveyed per turn. Additionally, in some cases, one party in the doctor-patient interaction may speak continuously without waiting for the other's response. Moreover, real-world doctor-patient interactions are often more colloquial, meaning that doctors' descriptions of issues may not be very precise and may omit many details. In stark contrast, LLM-generated responses have shorter interaction frequencies but convey a lot of information in each turn. Furthermore, LLM responses are more accurate, formal, and well-organized compared to colloquial responses, sometimes even appearing verbose. Since our task-oriented dialogue system is designed to communicate with medical LLMs, to more accurately recognize the intentions of medical LLMs, we chose not to use real doctor-patient dialogue data for annotation. Instead, we directly used GPT-4 to simulate doctor-patient dialogues and annotated these dialogues. This helps reduce generalization error and improves our dialogue model's ability to communicate with medical LLMs.

F Annotation Time and Cost

We hired professional doctors to annotate a test dataset, which is used to evaluate the model's performance on the tasks of intent recognition and dialogue state tracking. A total of 1,000 data entries were annotated. Each question required approximately 3 minutes to annotate and review. The cost was $30 per hour for the four junior physicians and $50 per hour for the senior physicians.

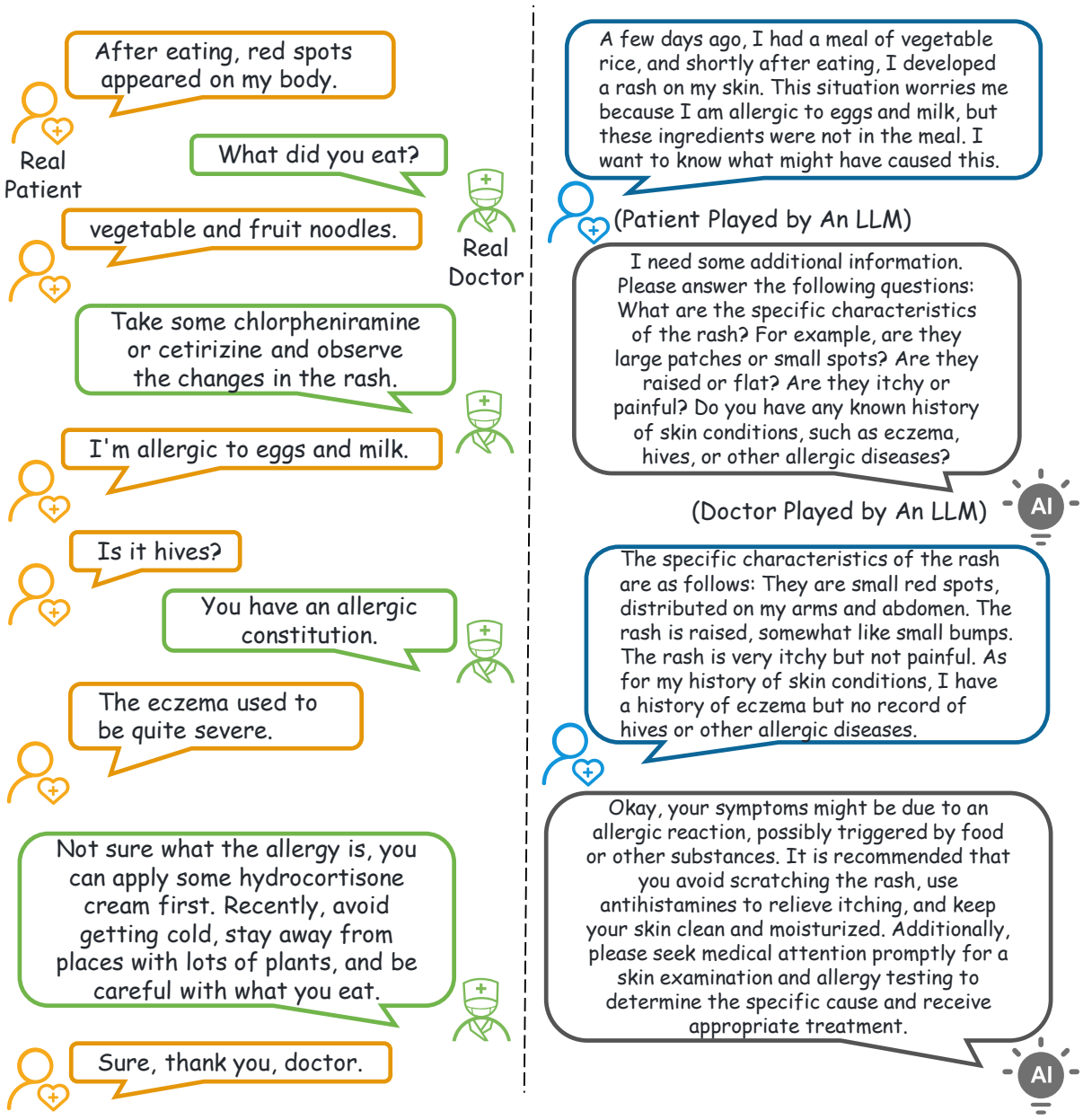


Figure 5: Comparison of the style between real doctor-patient conversations and AI-generated doctor-patient conversations

Model	Baichuan	BianQue	ChatGLM	Huatuo	Ming	Pulse	Qwen	GPT3.5
Symptom Name	0.64	0.09	0.81	0.06	0.74	0.87	0.69	**0.97**
Symptom Duration	0.16	0.01	0.15	0.02	0.2	0.11	**0.22**	0.08
Symptom Description	0.34	0.03	0.32	0.04	0.29	0.26	**0.48**	0.24
Past Diagnosis	0.12	0.0	**0.30**	0.0	0.06	0.19	0.16	0.29
Past Medication	0.08	0.00	0.23	0.01	0.04	0.18	**0.26**	0.16
Physical Change	0.24	0.01	0.22	0.01	0.16	0.22	**0.47**	0.36
Trauma Surgery	0.09	0.01	0.23	0.0	0.07	0.16	0.17	**0.29**
Allergic History	0.07	0.00	**0.22**	0.00	0.02	0.06	0.04	0.06
Contact History	0.06	0.0	0.1	0.0	0.02	0.07	0.08	**0.13**
Habit Tobacco	0.08	0.00	**0.10**	0.00	0.02	0.08	**0.10**	0.06
Habit Wine	0.06	0.00	0.06	0.00	0.02	0.05	**0.09**	0.03
Habit Living	0.23	0.01	0.16	0.01	0.08	0.14	**0.34**	0.31
Coitus History	0.01	0.00	0.02	0.00	0.01	**0.06**	0.03	0.04
Family History	0.13	0.00	0.12	0.00	0.02	0.08	0.08	**0.14**
Personal Age	0.04	0.0	0.05	0.00	0.05	0.13	**0.16**	0.01
Personal Weight	0.01	0.00	0.00	0.00	0.01	0.01	**0.02**	0.0
Personal Sex	0.03	0.0	0.03	0.0	0.03	0.08	**0.14**	0.01
Medical Examination	0.06	0.02	0.12	0.01	0.08	0.17	**0.23**	0.09
Receiving Treatment	0.03	0.01	**0.13**	0.02	0.01	0.04	0.05	0.06
Symptom Color	0.01	0.0	0.0	0.0	**0.02**	0.01	**0.02**	0.01
Symptom Degree	**0.07**	0.0	0.1	0.01	0.06	0.06	0.1	0.03
Symptom Position	0.07	0.01	0.06	0.01	0.05	0.06	**0.08**	0.04
Symptom Time	0.04	0.0	0.05	0.0	**0.06**	0.03	0.05	0.02

Table 15: Statistics on the types of questions asked by the medical LLMs

Case Item	Detail	Example
Treatment Outcomes	Following medical treatment by a physician, the patient's treatment outcome	The patient's clinical status is currently stable, indicating no significant deterioration or improvement in their condition.
Treatment Plan	The diagnostic recommendations and treatment plan provided by a specialist physician.	The administration of treatments aimed at supplementing vitamins, nourishing nerves, protecting cardiac function, and safeguarding liver health.
Basic Information	Patient identification information, including age and gender.	Female, 56 years old, occupation: farmer.
Etiology	The cause of the disease.	Bacterial infection.
Key Points of the Case	The critical points of the medical case.	Intrauterine infection.
Clinical Diagnosis	Disease identified by history, exam; no lab tests.	Severe acute chorioamnionitis (intrauterine infection)
Analysis and Summary	A critical review of the case, summarizing the findings and outcomes.	HIV attacks CD4+ T cells, causing GI symptoms. Patient shows blood cell drop, liver enzyme spike, HIV+. Treatment eases symptoms, specialist referral made.
History of Present Illness	Detailed account of the current health issue, including onset, duration, and progression.	The patient experienced dizziness, blackouts, and weakness in all four limbs for no obvious reason two days before admission.
Past Medical History	Information about the patient's previous health conditions and treatments.	The patient has a history of hypertension for about 10 years and has been taking Nifedipine SR tablets and Indapamide regularly.
Chief Complaint	The primary reason the patient seeks medical attention.	Bronchial asthma, left heart failure with nocturnal paroxysmal dyspnea.
Physical Examination	Findings from the doctor's hands-on assessment of the patient's body.	T: 36.8°C, P: 71 beats/min, R: 10 breaths/min, BP: 160/80 mmHg.
Ancillary Tests	Additional diagnostic procedures or tests conducted to aid in diagnosis.	CT of the head shows ischemic changes in the right frontal lobe.
Personal History	Lifestyle, habits, and other non-medical factors that may impact health.	Born in the place of origin, long-term resident of the local area, no history of visiting epidemic areas or pastoral areas
Diagnostic Criteria	Specific standards used to confirm the presence of a disease.	Numbness in the right upper and lower limbs for one week.Head CT shows an ischemic lesion in the left corona radiata.
Diagnostic Results	Outcomes of the diagnostic tests and examinations.	Lacunar infarction, Hypertension (Stage 3).
Differential Diagnosis	List of possible conditions considered before arriving at a definitive diagnosis.	Hemorrhage: sudden, with hemiplegia, CT dense. Spondylosis unlikely, no history, muscles fine. TIA: acute, symptoms vanish in 24 hrs, CT clear.
Preliminary Diagnosis	Initial suspected condition based on initial assessment.	Lacunar infarction, hypertension.
Department	The specific medical specialty or unit involved in the patient's care.	Neurolog

Table 16: Original Case Data Items.

Case Item	Detail
Symptom-Name	Name of the primary symptoms
Past-Diagnosis	Record of previous diagnoses, including past illnesses
Past-Medication	Medications currently being taken
Physical-Change	Perceptible changes in physical state, including weight, appetite, sleep
Trauma-Surgery	History of significant trauma or surgery
Preventive	Vaccination status
Allergic-History	History of allergies to food or medications
Contact-History	History of contact with infectious diseases or environments in the recent past
Habit-Tobacco	Smoking habits
Habit-Wine	Drinking habits
Habit-Drug	Drug abuse
Habit-Living	Lifestyle habits, including diet, hygiene, sleep
Coitus-History	History of unprotected sexual activity
Family-History	Family history, including life experiences, occupations, health conditions, and genetic diseases of family members across generations
Personal-Age	Patient's age
Personal-Weight	Patient's weight
Personal-Sex	Patient's gender
Medical-Examination	Medical examinations, including physical and auxiliary tests such as CT, MRI, blood tests, etc.
Department	Department of consultation
Receiving-Treatment	Whether the patient is currently receiving treatment
Symptom-Degree	Severity of the symptoms
Symptom-Position	Location of the symptoms
Symptom-Time	Timing of symptom onset
Symptom-Shape	Shape information, seen in symptoms like patches, swelling
Symptom-Texture	More detailed description of symptoms, e.g., thickness of phlegm, whether it is filamentous
Symptom-Smell	Odor information related to the symptoms
Symptom-Color	Color information of the symptoms

Table 17: Structured Case Data Items.