Prefix-Enhanced Large Language Models with Reused Training Data in Multi-Turn Medical Dialogue

Suxue Ma[1]* Zhicheng Yang[2]† Ruei-Sung Lin[2] Youbao Tang[2] Ning Zhang[2]
Zhenjie Cao[3] Yuan Ni[4] Jing Xiao[4] Jieke Hou[5] Peng Chang[2]
[1]Tianjin University, China [2]PAII Inc., USA
[3]Tsinghua SIGS, China [4]Ping An Technology, China
[5]Ping An Healthcare and Technology Company Limited, China
msx@tju.edu.cn; zcyangpingan@gmail.com

Abstract

Large Language Models have made impressive progress in the medical field. In medical dialogue scenarios, unlike traditional single-turn question-answering tasks, multi-turn doctor-patient dialogue tasks require AI doctors to interact with patients in multiple rounds, where the quality of each response impacts the overall model performance. In this paper, we propose PERT to re-explore values of multi-turn dialogue training data after the supervised fine-tuning phase by integrating a prefix learning strategy, further enhancing the response quality. Our preliminary results show that PERT achieves notable improvements on gynecological data, with an increase of up to 0.22 on a 5-point rating scale.

1 Introduction

With the development of large language models (LLMs), there has been increasing attention on their applications in the medical sector. While recent general-purpose models such as GPT series (Hurst et al., 2024), Claude series (Anthropic, 2025), and Qwen series (Yang et al., 2024b) have shown decent capabilities in medical question-answering (QA) tasks (Xie et al., 2024), researchers have leveraged diverse medical datasets to build specialized models tailored to various medical scenarios, such as dedicated SMILE for mental health (Qiu et al., 2023), and comprehensive Med-PaLM series (Singhal et al., 2025), Zhongjing (Yang et al., 2024c), and Baichuan-M1 (Baichuan, 2025). These models offer exciting possibilities for the real-world application of LLMs in the medical domain.

Our scenario is multi-turn doctor-patient dialogues in multiple clinical departments on an online healthcare consultation platform. We aim at deploying LLMs as AI doctors to assist human doctors in collecting adequate prediagnostic information from patients via multi-turn conversations between patients and AI doctors. To train an acceptable LLM for every clinical department, a straightforward idea is to adopt a multi-stage training strategy: pretraining on general medical data (Yang et al., 2024c; Baichuan, 2025), followed by supervised fine-tuning (SFT) using real doctor-patient dialogue history in each clinical department (Yang et al., 2024c). However, the model trained using this simple strategy still falls short of meeting deployment-oriented performance requirements. For instance, we observed that the model occasionally repeats its previous responses. Unfortunately, a repetitive utterance might make patients aware that they are interacting with an AI doctor, destroying their consultation experience.

Since authors in (Zhang et al., 2025) highlighted the effectiveness of appropriate instruction prompts to alleviate this issue, we conduct two pilot experiments: (1) When we apply the instruction prompt "Do not repeat what has already been said" only at the beginning of a multi-turn dialogue, the model tends to forget this constraint after several rounds; (2) When we insert this instruction prompt before every response, the model significantly reduces repetition, but it increases the frequency of irrelevant or off-topic responses, still degrading the overall response quality. We infer that two factors cause this issue: (1) The dataset for each medical department is relatively small, limiting the model's learning capacity; (2) While the prompt-based constraint is effective, the model either forgets it over time or applies it too rigidly.

To mitigate these issues, we propose a novel training strategy **PERT** (Prefix-Enhanced LLMs with Reused Training data) for our multi-turn medical dialogue scenario. Unlike the original single-department SFT paradigm, PERT has two training phases. First, we aggregate data from all departments to train an all-around LLM that benefits from

*This work was done during Suxue Ma's internship at Ping An Technology, Shenzhen, China.
†Corresponding author.

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 26–33
May 4, 2025 ©2025 Association for Computational Linguistics

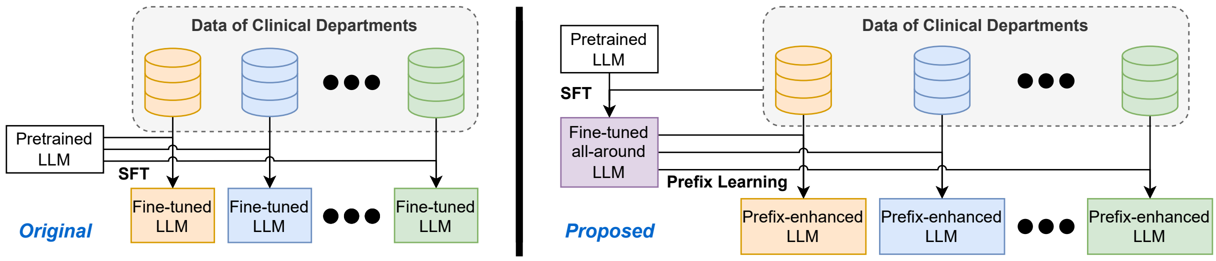

Figure 1: **Framework of the proposed PERT.** Compared with the original LLM training strategy, PERT first trains an all-around LLM using the data of all clinical departments. The prefix learning process is then conducted to leverage the departmental data individually and train their own prefix-enhanced LLMs, respectively.

the data scaling law (Kaplan et al., 2020). Second, since such a generalized LLM needs to retain specialization for individual departments, we design a prefix learning phase by *reusing* the data from each department. Unlike the previous pilot experiments where the prompt was mechanically inserted either at the beginning of the entire dialogue or before each response, prefix learning can provide "soft guidance" for each round of AI doctor responses, improving the overall LLM performance without introducing excessive constraints on response generation. PERT further exploits the values of training data that was used once only in the conventional single-department SFT (original vs. proposed in Fig. 1, described in Sec. 3.1).

Our key contributions are listed as follows: (1) proposing the PERT training strategy combining all-around LLM training with prefix learning by reusing training data from single-department for multi-turn medical dialogues, (2) introducing a strategy for reusing training data from single-department to enhance model performance, and (3) conducting preliminary experiments to validate the effectiveness of our approach in real-world doctor-patient consultations.

2 Related Work

Medical LLMs. Medical LLMs have emerged as a transformative technology in healthcare, with significant advancements in a wide range of applications, including medical summarization (Tang et al., 2023; Van Veen et al., 2024), clinical decision support (Hager et al., 2024), and medical dialogue systems (Li et al., 2023). In dialogue systems, single-turn models provide rapid responses to medical queries, while multi-turn models are always diagnostic-oriented through context-aware interaction. These models can be broadly categorized into fine-tuned general LLMs (Li et al., 2023; Singhal et al., 2025; Yang et al., 2024c) and dedicated LLMs (Luo et al., 2022; Gu et al., 2021). Most of those models are validated on public datasets or in lab-stage settings, but have not been fully studied in deployment-oriented scenarios.

Prefix Learning. The representative prefix-tuning method is a parameter-efficient fine-tuning (PEFT) approach that optimizes a small set of task-specific parameters, called prefixes, while keeping the pretrained model frozen. These prefixes effectively guide the model's behavior during inference without requiring updates to the full model (Li and Liang, 2021). Recent studies have demonstrated the effectiveness of prefix-tuning in medical applications (Van Sonsbeek et al., 2023; Chen et al., 2024; Zhou et al., 2024). For the multi-turn interactive dialogue scenario, the authors in (Li et al., 2024a) introduce an external planner to learn prefix token embeddings. Nevertheless, the efficacy of this method has not been studied in the medical field.

3 Methods

3.1 Framework Overview

Fig. 1 illustrates the framework of our proposed PERT. Compared with the original SFT strategy, we first leverage the data from all clinical departments to achieve an all-around LLM, which plays an intermediate role. We then conduct prefix learning by reusing data from every individual department on the trained all-around LLM. Consequently, each department has its own prefix-enhanced LLM.

3.2 All-Around LLM Training Phase

We aggregate data from all departments and train the all-around LLM using the same SFT strategy as the original one. We find that this all-around LLM overall outperforms the single-department LLM (shown in Table 2).

3.3 Prefix Learning Phase

Inspired by prefix learning designed for the multi-turn dialogue scenario (Li et al., 2024a), which adopted an extra planner to update the prefix token features, we design two stages in our prefix learning phase. The first stage involves cloning the behavior of the pretrained all-around LLM to ensure that the LLM steered by the prefixes behaves similarly to the LLM itself. The prefixes are generated by a planner. In the second stage, we fine-tune the planner by using responses from real doctors, collected through our online consultation platform. This allows the LLM's behavior to become more aligned with the communication style and expertise of real medical professionals.

3.3.1 Self-Cloning Stage

Behavior cloning (Bratko et al., 1995) is a technique in imitation learning where an agent learns to replicate the actions of an expert. Inspired by this approach, we aim to make an LLM with prefix tokens behave consistently with the all-around LLM. To achieve this, we train the planner from scratch using the responses generated by the all-around LLM as training data. This stage ensures the prefix-equipped LLM retains the capacities of the all-around LLM, offering a robust starting point.

To prepare the corpus for self-cloning, we provide the fine-tuned all-around LLM with real doctor-patient dialogue history which ends with the patient's utterance, and ask the LLM to generate the response as a doctor. Formally, the corpus is denoted as $\{p_1^i, q_1^i, p_2^i, q_2^i, \cdots, p_{N_i}^i, q_{N_i}^i\}_{i=1}^M$, where M is the number of collected dialogues, N_i is the number of rounds of the i-th dialogue, and p_j and q_j ($j = \{1, 2, \ldots, N_i\}$) are the patient's and the doctor's utterance at the j-th round, respectively. Note that a dialogue with n rounds can be split into n individual datapoints with $\{p_1, q_1, \cdots, p_j\}$ being the dialogue history and q_j being the ground truth for $j = \{1, 2, \cdots, n\}$.

Now we describe the process of prefix generation. Initially, the embedding of the dialogue history at the j-th round of the i-th dialogue is obtained by the LLM, which produces an embedding:

$$e_j^i = \mathrm{Emb}(\{p_1^i, q_1^i, p_2^i, q_2^i, \cdots, p_j^i\}). \quad (1)$$

Next, the planner extracts the last-token embedding from the output of the LLM's last layer, and then transforms this token embedding into the prefix space by an MLP. Formally, the planner is defined as:

$$\phi(e) = \mathrm{MLP}(g_\theta(e)), \quad (2)$$

where θ is learnable parameters of the transformer and g_θ denotes the extraction operation. We train the planner by minimizing conditional language modeling objective as follows:

$$\mathcal{L}_{sc} = -\sum_{i=1}^{M}\sum_{j=1}^{N_i} \log f_\theta(\tilde{q}_j^i \mid \phi(e_j^i) \| e_j^i), \quad (3)$$

where $\|$ denotes concatenation of the dialogue action tokens with token embeddings e, and f_θ denotes the autoregressive distribution of generated strings. Here, the ground truth $\tilde{q}_j$ is generated by the all-around LLM.

3.3.2 Supervised Fine-Tuning Stage

In the supervised fine-tuning stage, we refine the prefix embeddings to better align the LLM's behavior with real doctors' communication styles and expertise. Unlike the self-cloning stage, which uses responses generated by the all-around LLM, this stage reuses the real doctors' responses from clinical department data as ground truth to fine-tune the planner in a supervised manner. Note that the dialogue history remains the same as that in the self-cloning stage, but the ground truth for fine-tuning is now the real doctors' responses rather than those generated by the LLM. That is, the ground truth for the real doctor's response at the j-th round is q_j instead of $\tilde{q}_j$. The loss function in this stage is

$$\mathcal{L}_{sft} == -\sum_{i=1}^{M}\sum_{j=1}^{N_i} \log f_\theta(q_j^i \mid \phi(e_j^i) \| e_j^i). \quad (4)$$

4 Experiments

4.1 Dataset

Our dataset is sourced from a real-world online doctor-patient consultation platform in China, including more than 10 clinical departments, such as pediatrics, ophthalmology, etc. This data source consists of authentic doctor-patient multi-turn dialogues, covering a range of medical inquiries and responses. In this paper, we present preliminary results using the data from the gynecology department because of its large number of consultations (300k+ dialogues), while the available data across all departments (800k+ dialogues) are for training the all-around LLM. Table 1 lists the statistics of

Dataset	#Dialog.	#Rounds
Original		
gynecology	310k	1.77m
For prefix learning		
self-cloning	10,000	58,105
supervised fine-tuning	10,000	54,133
test set	1,000	5,463

Table 1: Statistics of dialogues from the gynecology department and those used for prefix learning during self-cloning, supervised fine-tuning, and inference, respectively.

Method	Avg. s	s>2(%)	s>3(%)	s>4(%)
Original				
gynecology LLM	3.5824	97.21	57.93	7.15
Proposed				
all-around LLM	3.6353	97.74	58.65	7.48
random prefix w/o learning	3.6437	98.32	58.44	7.91
prefix w/ self-cloning only	<u>3.7584</u>	**98.86**	<u>68.38</u>	<u>8.34</u>
PERT (prefix w/ self-cloning & SFT)	**3.8013**	<u>98.41</u>	**71.66**	**10.36**

Table 2: Performance comparison among different methods by the average score s and the percentage of dialogues with scores exceeding 2, 3, and 4. Bold and underlined text represent the best and the second best, respectively.

our used dialogue data, including the number of dialogues (#Dialog.) and the total number of rounds (#Rounds). Specifically, we use 10,000 dialogues for both self-cloning and supervised fine-tuning, with average rounds per dialogue of 5.8 and 5.4, respectively. For evaluation, we use 1,000 dialogues as the test set.

4.2 Implementation Details

The fine-tuned all-around LLM in PERT is obtained by fine-tuning Qwen2-14B-Instruct (Yang et al., 2024a) with aggregated data from all clinical departments. For training, We used a learning rate of 0.001 and Adam optimizer to minimize the loss. We used a prefix token length of 2, with prefix embedding size of 128. The dimension of the hidden state of the LLM is 5120. The planner for generating prefix tokens was trained for 10 epochs for self-cloning and 5 epochs for supervised fine-tuning, while the all-around LLM was frozen. All experiments were conducted on servers with 8 NVIDIA V100 GPUs, each with 16 GB VRAM.

4.3 Preliminary Results

We compared several methods for doctor-patient dialogue generation to validate the effectiveness of our method in Table 2. The methods tested for comparison include (i) the original gynecology LLM; (ii) the all-around LLM that generates responses without any prefix learning stages; (iii) a random prefix without learning, where the planner is randomly initialized to generate prefix tokens; and (iv) update the prefix embeddings using self-cloning only, referring to no fine-tuning with real doctor responses. Finally, our proposed PERT, which combines the self-cloning stage of the planner to generate prefix embeddings with the supervised fine-tuning stage using real doctor responses, was also evaluated. We utilized a general-purpose LLM (Qwen2-7B-Instruct) to assess dialogue responses. Each response was rated on a scale from 1 to 5, with higher scores indicating better quality. The evaluation considered factors including safety, professionalism, and friendliness. The complete prompt template is provided in Appendix A. For each dialogue, the highest turn score was taken as the dialogue's overall score. We then calculated the average score and the proportions of dialogues with scores exceeding 2, 3, and 4 in Table 2.

As we can see, PERT achieves the highest average score of 3.8013, significantly surpassing the baselines (gynecology LLM and all-around LLM), which have an average score of 3.5824 and 3.6353, respectively. The random prefix method also shows a comparable result (3.6437), but it remains lower than the prefix learning approaches. Meanwhile, our method generally accomplishes the best results in the percentage of responses with scores above various thresholds ($s > 2 \sim 4$), except the comparable percentage with the self-cloning stage only for $s > 2$. These results indicate that the inclusion of prefix learning by reusing real doctors' replies from the training data is significant for generating more coherent and contextually appropriate responses.

5 Conclusions and Discussion

In this paper, we propose PERT, which leverages a prefix learning strategy to re-explore multi-turn dialogue training data after the SFT training phase, leading to further LLM performance improvement. Our preliminary results show that PERT achieves noticeable improvements on gynecological data.

Since our model is designed for deployment, the performance of the medical LLM needs to be continuously improved through iterative updates. Once the existing data has been effectively utilized, a key question is whether we can further explore its potential for specific medical scenarios. This paper presents a novel model-based approach to achieving this objective. In fact, prefix learning is often compared side by side with low-rank adaptation (LoRA) SFT (Van Sonsbeek et al., 2023) in terms of model performance. However, we cascade these two stages and adapt them to our multi-turn interactive dialogue scenario to achieve further improvements.

In medical scenarios, the tolerance for hallucinations is much stricter than in general contexts. During interactions with patients, responses from a medical LLM must not contain blatantly commonsense-violating errors. For example, if a male patient is asked about menstruation, such an error represents a critical *red line* that cannot be crossed. A response like that could lead the patient to entirely abandon the use of the online medical consultation platform. However, such issues are difficult to directly measure through standard performance evaluation metrics (e.g. the rating scale used in this paper). Since these issues are crucial considerations in determining whether a medical LLM is suitable for real-world deployment, we plan to leverage reinforcement learning to address these red-line issues.

There has recently been considerable research on retrieval-augmented generation (RAG) to mitigate hallucination issues, such as GraphRAG (Edge et al., 2024). However, building a precise and professional knowledge graph in the medical field requires a significant investment of time and effort from medical professionals. This research direction is currently also underway in our project.

Limitations

A limitation of our work is that we did not involve medical specialists in rating the responses at this point, since the scope of this preliminary study is within our internal research team. We will continue to test PERT in other departments. Once its effectiveness is demonstrated thoroughly, medical professionals from the online consultation platform will perform further evaluation.

Many medical LLMs used ChatGPT/GPT-4 series for scoring or included them for performance comparison (Moor et al., 2023; Yang et al., 2024c; Chen et al., 2023; Singhal et al., 2023). Unfortunately, in compliance with our platform's safeguarding medical data privacy policies, we are restricted from accessing external API interfaces, including ChatGPT/GPT-4 series.

In this work, we focus only on the pure textual content rather than multi-modal dialogue data, even though the appearance of images sent by patients to better illustrate their symptoms is common in practice (Li et al., 2024b). Meanwhile, incorporating the paradigm of the conventional medical imaging diagnosis or screening tasks such as our previous studies (Yang et al., 2021; Cao et al., 2024, 2025; Tang et al., 2021; Yi et al., 2022) into the LLM/VLM-powered multi-turn interactive dialogue setting still remains a challenging and ongoing area of research.

Ethical Considerations

All personal data were anonymized to ensure participant privacy. This study was reviewed and approved by the Institutional Review Board (IRB) of Qingdao Ping An Kangjian Internet Hospital, China (IRB number: LLSC2024A01). The authors declare no competing interests.

References

Anthropic. 2025. Meet claude. https://www.anthropic.com/claude. Accessed: 2025-01-31.

Baichuan. 2025. Baichuan-m1-14b. https://github.com/baichuan-inc/Baichuan-M1-14B. Accessed: 2025-01-31.

Ivan Bratko, Tanja Urbančič, and Claude Sammut. 1995. Behavioural cloning: phenomena, results and problems. *IFAC Proceedings Volumes*, 28(21):143–149.

Zhenjie Cao, Zhuo Deng, Zhicheng Yang, Jie Ma, and Lan Ma. 2025. Supervised contrastive pre-training models for mammography screening. *Journal of Big Data*, 12(1):24.

Zhenjie Cao, Zhuo Deng, Zhicheng Yang, Jialin Yuan, Jie Ma, and Lan Ma. 2024. Asydisnet: Scalable mammographic asymmetry and architectural distortion detection with angle-based quadruplet loss. *IEEE Transactions on Medical Imaging*.

Jiawei Chen, Yue Jiang, Dingkang Yang, Mingcheng Li, Jinjie Wei, Ziyun Qian, and Lihua Zhang. 2024. Can llms' tuning methods work in medical multimodal domain? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 112–122. Springer.

Junying Chen, Xidong Wang, et al. 2023. Huatuogpt-ii, one-stage training for medical adaption of llms. *preprint arXiv:2311.09774*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kenneth Li, Yiming Wang, Fernanda Viégas, and Martin Wattenberg. 2024a. Dialogue action tokens: Steering language models in goal-directed dialogue with a multi-turn planner. *arXiv preprint arXiv:2406.11978*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Zhangpu Li, Changhong Zou, Suxue Ma, Zhicheng Yang, Chen Du, Youbao Tang, Zhenjie Cao, Ning Zhang, Jui-Hsin Lai, Ruei-Sung Lin, et al. 2024b. Zalm3: Zero-shot enhancement of vision-language alignment via in-context information in multi-turn multimodal medical dialogue. *arXiv preprint arXiv:2409.17610*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Michael Moor, Qian Huang, et al. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*. PMLR.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Karan Singhal, Tao Tu, et al. 2023. Towards expert-level medical question answering with large language models. *preprint arXiv:2305.09617*.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158.

Yuxing Tang, Zhenjie Cao, Yanbo Zhang, Zhicheng Yang, Zongcheng Ji, Yiwei Wang, Mei Han, Jie Ma, Jing Xiao, and Peng Chang. 2021. Leveraging large-scale weakly labeled data for semi-supervised mass detection in mammograms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3855–3864.

Tom Van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees GM Snoek, and Marcel Worring. 2023. Open-ended medical visual question answering through prefix tuning of language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 726–736. Springer.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Wenya Xie, Qingying Xiao, Yu Zheng, Xidong Wang, Junying Chen, Ke Ji, Anningzhe Gao, Xiang Wan, Feng Jiang, and Benyou Wang. 2024. Llms for doctors: Leveraging medical llms to assist doctors, not replace them. *arXiv preprint arXiv:2406.18034*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.1067*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024b. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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

Zhicheng Yang, Zhenjie Cao, Yanbo Zhang, Yuxing Tang, Xiaohui Lin, Rushan Ouyang, Mingxiang Wu, Mei Han, Jing Xiao, Lingyun Huang, et al. 2021. Momminet-v2: Mammographic multi-view mass identification networks. *Medical Image Analysis*, 73:102204.

Chunyan Yi, Yuxing Tang, Rushan Ouyang, Yanbo Zhang, Zhenjie Cao, Zhicheng Yang, Shibin Wu, Mei Han, Jing Xiao, Peng Chang, et al. 2022. The added value of an artificial intelligence system in assisting radiologists on indeterminate bi-rads 0 mammograms. *European Radiology*, pages 1–10.

Chen Zhang, Xinyi Dai, Yaxiong Wu, Qu Yang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025. A survey on multi-turn interaction capabilities of large language models. *arXiv preprint arXiv:2501.09959*.

Lu Zhou, Yiheng Chen, Xinmin Li, Yanan Li, Ning Li, Xiting Wang, and Rui Zhang. 2024. A new adapter tuning of large language model for chinese medical named entity recognition. *Applied Artificial Intelligence*, 38(1):2385268.

A Appendix

A.1 Prompt Template

In this section, the complete prompt template for the 5-point rating scale is provided. Since our data source is in Chinese, the original language of this prompt template is Chinese. We have translated it into English.

You will act as an evaluator and rate the doctor's next response based on the dialogue history between the patient and the doctor. Please provide a score from 1 to 5 according to the following scoring criteria.
你将作为评估员，根据患者和医生之间的对话历史，对医生的下一轮回复进行打分。请根据以下评分标准给出1到5分的评分。

Scoring Criteria

评分标准

- 1 Point - Very Dissatisfied:
  1分 - 非常不满意：
  - The response is completely irrelevant to the patient's question or contains obvious errors;
    回复与患者的问题完全无关或明显错误；
  - Lacks basic medical knowledge and common sense, potentially misleading the patient;
    缺乏基本的医疗知识和常识，可能误导患者；
  - The response could negatively impact the patient's health.
    回复可能对患者的健康造成负面影响。
- 2 Points - Dissatisfied:
  2分 - 不满意：
  - The response is partially correct but contains significant errors or omits key information;
    回复部分正确，但包含明显的错误或遗漏关键信息；
  - Fails to adequately address the patient's concerns and lacks depth;
    未能充分解决患者的问题，缺乏深度；
  - Lacks professionalism and does not provide effective diagnosis or recommendations.
    回复缺乏专业性，未能提供有效的诊断或建议。
- 3 Points - Average:
  3分 - 一般：
  - The response is generally correct but lacks detailed explanations or supporting information;
    回复基本正确，但缺少详细的解释或支持；
  - The question is addressed, but the expression is not entirely clear and could be improved;
    解决了问题，但表达不够清晰，有改进的空间；
  - The response is neutral, without major errors, but also does not exceed expectations.
    回复态度中立，没有明显错误，也没有超出期望的表现。
- 4 Points - Satisfied:
  4分 - 满意：
  - The response is accurate and provides sufficient information and explanations;
    回复准确，提供了足够的信息和解释；
  - Considers the patient's condition and offers personalized advice;
    考虑了患者的情况，提供了个性化的建议；
  - Demonstrates professionalism and provides effective diagnosis or recommendations.
    回复展现了专业性，能够针对患者的问题提供有效的诊断或建议。
- 5 Points - Very Satisfied:
  5分 - 非常满意：
  - The response is not only accurate but also exceeds patient expectations, offering in-depth analysis and recommendations;
    回复不仅准确，而且超出了患者的期望，提供了深入的分析和建议；
  - Demonstrates a high level of professional knowledge and a deep understanding of the patient's condition;
    展现了高水平的专业知识和对患者情况的深刻理解；
  - The response is encouraging and instills confidence and reassurance in the patient.
    回复态度积极，能够给予患者信心和安慰。

Steps

步骤

- Read the dialogue history between the patient and the doctor;
  阅读患者和医生之间的对话历史；
- Read the doctor's next response;
  阅读医生的下一轮回复；
- Evaluate the response based on the scoring criteria;
  根据评分标准，对回复进行评估；
- Assign a score.
  给出一个评分。

Examples

示例

- 1-Point Example:
  1分示例
  - *Patient*: "My menstrual blood has been dark brown for the past few months, and my period lasts longer than usual."
    患者："我最近几个月的月经颜色都是深褐色的，而且经期时间也延长了。"
  - *Online Doctor's Response*: "It might be due to fatigue. Just get some rest."
    在线医生回复："这可能是疲劳引起的，多休息就好。"
  - *Score*: 1
    评分：1
  - *Reasoning*: The response is overly simplistic and does not consider possible gynecological conditions such as endometriosis or fibroids. It also fails to ask about other symptoms or medical history.
    理由：回复过于简单，没有考虑到可能的妇科疾病，如子宫内膜异位症或子宫肌瘤，也没有进一步询问患者的其他症状或病史。
- 2-Point Example:
  2分示例
  - *Patient*: "I've been feeling dizzy lately, especially when I stand up."
    患者："我最近总是感到头晕，尤其是在站起来的时候。"
  - *Online Doctor's Response*: "It might be low blood pressure. Drink more water and eat more salt."
    在线医生回复："这可能是低血压，多喝水，多吃盐。"
  - *Score*: 2
    评分：2
  - *Reasoning*: The response does not inquire about additional symptoms, such as fainting or blurred vision, and lacks a recommendation for further medical evaluation. It also does not provide personalized advice.
    理由：回复没有询问患者的其他症状，如是否有晕厥或视力模糊，也没有建议患者进一步检查，缺乏个性化建议。
- 3-Point Example:
  3分示例
  - *Patient*: "I've been experiencing chest tightness, especially at night."
    患者："我最近经常感到胸闷，尤其是在晚上。"
  - *Online Doctor's Response*: "Chest tightness could be a heart issue or caused by anxiety. You should go to the hospital for further evaluation."
    在线医生回复："胸闷可能是心脏问题，也可能是焦虑引起的。建议你去医院做进一步检查。"
  - *Score*: 3
    评分：3
  - *Reasoning*: The response is correct but lacks detailed explanations and does not ask about relevant symptoms or medical history. However, it at least suggests further medical evaluation.
    理由：回复正确但缺乏详细解释，没有询问患者是否有心脏病史或其他相关症状，但至少建议了进一步检查。
- 4-Point Example:
  4分示例
  - *Patient*: "I've been feeling constantly fatigued, even after getting enough sleep."
    患者："我最近总是感到疲劳，即使睡眠充足也感觉累。"
  - *Online Doctor's Response*: "Persistent fatigue could be a sign of anemia or hypothyroidism. Have you noticed pale skin or changes in heart rate? I recommend a complete blood count and thyroid function tests."
    在线医生回复："持续的疲劳可能是贫血或甲状腺功能减退的迹象。你有没有注意到皮肤苍白或心率变化？建议你做血常规和甲状腺功能测试。"
  - *Score*: 4
    评分：4
  - *Reasoning*: The response is accurate, offers advice, and asks relevant follow-up questions. The tone is proactive, though it does not provide specific over-the-counter treatment options or further diagnostic guidance.
    理由：回复准确，提供了一些建议，并询问了相关症状，态度积极，但没有提供具体的非处方药名称或进一步的诊断建议。
- 5-Point Example:
  5分示例
  - *Patient*: "I recently noticed blood in my stool, and I'm really worried."
    患者："我最近发现我的大便中带血，我很担心。"
  - *Online Doctor's Response*: "Blood in stool could be due to hemorrhoids, but it might also indicate more serious conditions such as colon polyps or cancer. Do you have abdominal pain, weight loss, or changes in bowel habits? I strongly recommend scheduling a colonoscopy as soon as possible. In the meantime, I can provide guidance on how to prepare for the procedure."
    在线医生回复："大便带血可能是痔疮，也可能是更严重的问题，如结肠息肉或癌症。你是否有腹痛、体重下降或大便习惯改变？我建议你尽快预约结肠镜检查，并在等待期间注意观察大便的任何变化。同时，我会为你提供一些关于如何准备结肠镜检查的信息。"
  - *Score*: 5
    评分：5
  - *Reasoning*: The response is thorough, professional, and considers multiple potential causes. It asks relevant follow-up questions, provides a clear action plan, and reassures the patient with expert guidance.
    理由：回复非常详细，提供了专业的建议，询问了相关症状，展现了高水平的专业知识和对患者情况的深刻理解，态度积极，能够给予患者信心和安慰。

Output Format

输出格式

Please print the evaluation score following the format below, where x = {1, 2, 3, 4, 5}:
请按照以上要求打印评分，你的答案格式如下，其中 x={1, 2, 3, 4, 5}：

Score: x
评分：x