SF-Med: Training Large Language Models to Become Competent Doctors with Proficient Chinese Medical Capabilities

Anonymous ACL submission

Abstract

Recent advancements in Large Language Models (LLMs) have shown impressive capabilities, yet their performance in specialized areas like Chinese medicine has been lacking. In this paper, we propose SF-Med, an LLM specifically designed for Chinese medical consultation. SF-Med undergoes an extensive training process that includes continuous pre-training, SFT, and RLHF. During the continuous pre-training phase, we collected a total of 2.7 billion tokens from various sources, including medical books, dialogues, and research papers. In the SFT phase, we compiled around one million real doctor-patient dialogues. Furthermore, we designed a multi-turn dialogue-based instruction augmentation method to improve generalization and robustness. This method generates new instructions by treating existing instructions in the pool as historical dialogues. In the RLHF phase, we employed a voting mechanism with multiple reward models to filter preference data and reduce noise interference. Furthermore, we introduced an online model merging optimization strategy to align the final model with the SFT model. We conducted extensive evaluations on multi-turn dialogues, single-turn dialogues, and general medical knowledge benchmarks. Experimental results demonstrate that SF-Med achieves state-of-the-art performance in Chinese medical consultation among open-source medical LLMs.

1 Introduction

Recent advancements in large language models (LLMs) have led to significant breakthroughs, enabling them to answer a wide variety of questions, sometimes even surpassing human performance in certain areas (Achiam et al., 2023; Wang et al., 2023). Several organizations have introduced LLMs with strong proficiency in Chinese, such as Qwen2.5(Yang et al., 2024a) and GLM4(Team et al., 2024), to address the limitations of LLMs in Chinese language capabilities. Although these models perform well in many tasks, their limited specialized knowledge makes them suboptimal in specific domains like Chinese medicine (Zhao et al., 2023; Huang et al., 2024). LLMs like GPT-4(Achiam et al., 2023) provide very detailed responses to patients but often fail to ask pertinent questions or offer medical conversation in the manner that a doctor would, which frequently results in patients not receiving essential diagnostic information (Zhang et al., 2023b). Additionally, these models fall short in interactive diagnosis and in capturing the nuances of a patient's condition. Despite these challenges, medical LLMs still hold tremendous potential, offering value and convenience in assisting with diagnosis, consultation, and guidance.

To contribute to the development of medical LLMs, we propose SF-Med, a Chinese medical LLM based on Qwen2.5-7B(Yang et al., 2024a), achieving continuous pre-training, supervised fine-tuning (SFT), and reinforcement learning from human feedback (RLHF). Specifically, during the pre-training phase, we collected a vast amount of real medical pre-training corpus, which includes various types of medical data from different sources. These datasets encompass approximately 2.7 billion tokens and can be categorized into four types: books, dialogues, articles, and examination questions.

Subsequently, in the SFT phase, we introduced four types of instruction data for training: single-turn medical dialogue data, multi-turn medical dialogue data, medical natural language processing task data, and general dialogue data. The aim is to enhance the model's generalization and comprehension capabilities and alleviate catastrophic forgetting. To further enhance the generalization ability of the SFT model, we employed the MinHash and LSH algorithms (Bai et al., 2023) for the deduplication of SFT data. Following this, we

utilized a multi-turn dialogue-based instruction expansion method, where we treated instructions in the seed pool as historical data to generate new instruction data.

During the RLHF phase, we used Direct Preference Optimization(Rafailov et al., 2023) (DPO) as the implementation technique. In our experiments, we found that using DPO alone led to a misalignment between the final model and the SFT model. To address this issue, we introduced an online model merging approach (Lu et al., 2024), which merges gradients with the parameter differences between SFT and pre-trained models, effectively steering the gradient towards maximizing rewards in the direction of SFT optimization. Due to incorrect and ambiguous preferences in the dataset that might hinder the model's ability to accurately capture human intent, we employed a voting mechanism of multiple reward models to filter out preference data (Wang et al., 2024a), thereby reducing noise interference.

We conducted extensive evaluation experiments on datasets encompassing single-turn medical dialogues, multi-turn medical dialogues, medical benchmark, and medical terminology explanations. For medical open-source LLMs, we selected HuatuoGPT-II(Chen et al., 2024), Zhongjing(Yang et al., 2024b), WiNGPT2(Winning, 2023), and ChiMed-GPT(Tian et al., 2024) as our baselines, which are basically the same size as our model. For general LLMs, we chose GPT-4(Achiam et al., 2023), ChatGPT(Ouyang et al., 2022), and Qwen2.5(Yang et al., 2024a) as the baselines. Our experiments demonstrate that SF-Med outperforms both open-source medical LLMs and general LLMs across the majority of performance metrics. This indicates that our proposed model, SF-Med, has effectively integrated the advantages of various datasets, delivering exceptional performance in medical tasks. This highlights the potential of SF-Med in advancing the field of medical artificial intelligence, as well as its potential applications in real-world scenarios. Our code can be accessed at the following anonymous address: `https://anonymous.4open.science/r/SFMed-D927`

2 Related Work

The rapid development of large Chinese medical models owes much to the release of large Chinese language models. Initially, researchers trained these models by performing instruction fine-tuning on large language models using medical data. For instance, DoctorGLM(Xiong et al., 2023) collected diverse Chinese and English medical dialogue datasets and fine-tuned the ChatGLM-6B(Team et al., 2024) model using P-tuning(Liu et al., 2022), enabling it to handle Chinese medical consultations. Similarly, DISC-MedLLM(Bao et al., 2023) used the Baichuan-base-13B(Yang et al., 2023) model, performing instruction fine-tuning on over 470,000 medical data points. This data included 420,000 AI-reconstructed real doctor-patient dialogues, 50,000 knowledge Q&A pairs from medical knowledge graphs, and 2,000 manually selected high-quality dialogues to enhance precision.

As open-source Chinese medical large language model research deepens, researchers have discovered that relying solely on instruction fine-tuning is insufficient to make a medical large language model a qualified medical consultation assistant. Consequently, some models adopted a more comprehensive training process, encompassing continued pre-training, instruction fine-tuning, and reinforcement learning. For example, HuatuoGPT(Zhang et al., 2023a) was fine-tuned on Ziya-LLaMA(Yang et al., 2022) using a dataset of 220,000 medical dialogues. These dialogues included instructions and conversations distilled by ChatGPT(Ouyang et al., 2022) and single and multi-turn interactions between real doctors and patients. Additionally, the model introduced a novel phase called hybrid feedback reinforcement learning to integrate the two data types seamlessly. Zhongjing(Yang et al., 2024b) underwent pre-training on various medical datasets, followed by instruction fine-tuning on single and multi-turn dialogues and medical NLP tasks, and further used reinforcement learning to ensure professionalism and safety. HuatuoGPT-II(Chen et al., 2024) proposed a unified domain adaptation protocol, merging the previous two-stage process of continued pre-training and instruction fine-tuning into a single-step procedure.

3 Methods

This section discusses the three training stages involved in constructing SF-Med: continued pre-training, instruction fine-tuning, and reinforcement learning, along with the optimization strategies and training data we introduced in these three stages. The comprehensive method flowchart is shown in Figure 1.

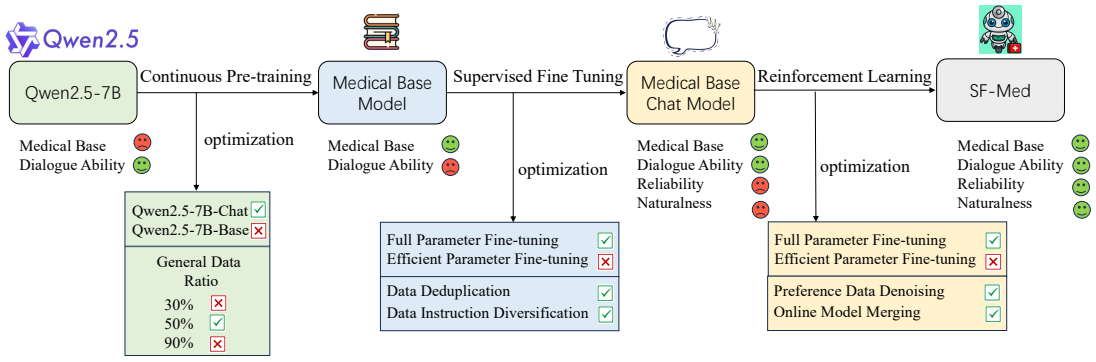

Figure 1: The overall flowchart of constructing SF-Med. Our training process includes continued pre-training, supervised fine-tuning, and reinforcement learning. We have also introduced a series of optimization strategies to enhance the model's capabilities. These strategies include, but are not limited to, mixing in general data to avoid forgetting, data deduplication to improve data quality, diversifying instructions to enhance data generalization, selecting preference data to eliminate noise interference and online model merging to align the DPO model and the SFT model.

Dataset	Type	Size
Medical Books	Books	1.5
CMtMedQA	Medical Dialogues	2.5
ChatMed-Consult-Dataset	Medical Dialogues	7.9
DISC-Med-SFT	Medical Dialogues	14.7
MedDiag	Medical Dialogues	35.1
cMedQA-V2.0	Medical Dialogues	2.2
huatuo-sft-train-data	Medical Dialogues	6.7
webMedQA	Medical Dialogues	5.6
Chinese-medical-dialogue-data	Medical Dialogues	11.5
ShenNong-TCM	Knowledge Graph	2.5
huatuo-knowledge-graph-qa	Knowledge Graph	3.0
CMExam	Examination Questions	1.5
CMB-Exam	Examination Questions	1.8
PromptCBLUE	NLP Task	2.7
Crawler	Articles	20.3
Medical paper	Articles	10.4
baike2018qa	General Corpus	35.8
webtext2019zh	General Corpus	76.3
wiki2019zh	General Corpus	31.4

Table 1: Statistics and sources of continued pre-training data for SF-Med. The unit for Size is ten million tokens. Crawler refers to the medical science articles we obtained from relevant medical websites. Medical paper denotes the abstracts of medical research papers we collected.

3.1 Continuous Pre-training

LLMs accumulate extensive general knowledge through pre-training, enabling them to learn skills across various fields. Current LLMs are pre-trained on a vast amount of unlabelled data, giving them the capability to address issues in multiple domains (Minaee et al., 2024). However, due to the broad scope of pre-training data, these models may struggle to focus on specific specialized areas. Continuous pre-training of LLMs can involve further training with domain-specific data, significantly enhancing the model's focus and problem-solving abilities in that area (Guo et al., 2024).

High-quality pre-training data can markedly improve model performance. In the medical field, we particularly emphasize the diversity, specialization, and comprehensiveness of the data. To ensure the diversity of the pre-training corpus, we have collected a large amount of authentic medical data from various sources, including open-source data and scraped data. These data can be categorized into four types: books, medical dialogues, articles, and examination questions. These data cover various aspects of the medical field, providing the model with rich medical knowledge. We also added approximately 1.4 billion tokens of the general corpus to prevent catastrophic forgetting. The data used for continual pre-training are uniformly presented in Table 1. After continual pre-training based on the Qwen2.5-7B model (Yang et al., 2024a), we obtained a foundational LLM with abundant medical knowledge.

3.2 Supervised Fine-Tuning

SFT is the key phase that equips the model with conversational abilities (Zhao et al., 2024). By leveraging high-quality dialogue data, the model can effectively utilize the knowledge accumulated during the pre-training phase to understand and respond to patients' questions. Current Chinese medical models, such as Huatuo-GPT (Zhang et al., 2023b), use large amounts of distilled data to make conversations more fluent. However, the lack of real doctor-patient interactions means the model's responses are less accurate than those in actual

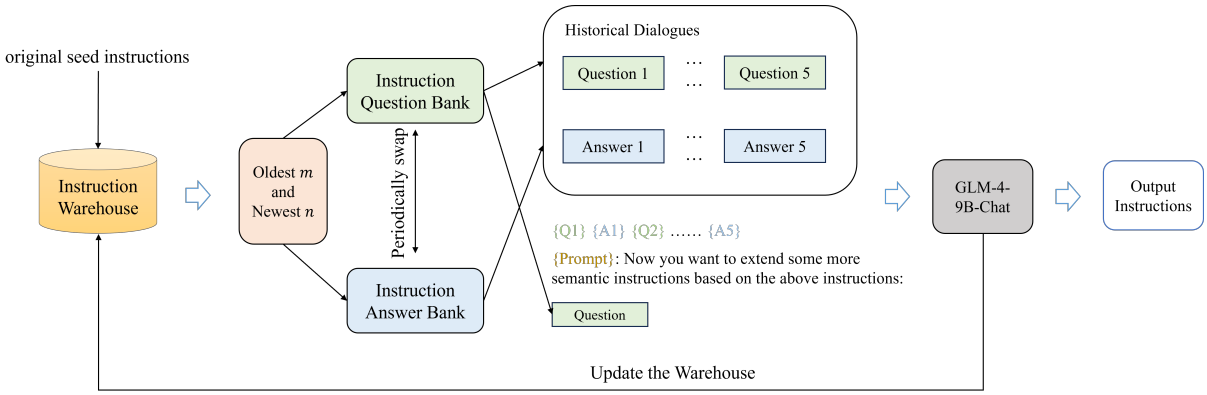

Figure 2: The overall flowchart of instruction diversification. We construct historical dialogues from a question library and an answer library to make the generated current instructions more diverse and rich. The template for the historical dialogue is {Q1}{Prompt}{A1}...{A5}. A specific example is as follows: {Based on your background as a doctor, please combine the patient's description and provide your insights on answering medical questions.} {Now you want to extend some more semantic instructions based on the above instructions:} {As a doctor, accurately evaluate and provide medical advice based on the patient's symptoms and medical history. If the diagnosis is unclear, deeply explore the patient's condition to avoid missing any potentially relevant information, and strictly adhere to the principle of not asking repetitive questions.}

Dataset	Type
ChatMed-Consult-Dataset	Medical Dialogues
DISC-Med-SFT	Medical Dialogues
MedDiag	Medical Dialogues
huatuo-sft-train-data	Medical Dialogues
CMExam	Examination Questions
PromptCBLUE	NLP Task
alpaca-zh	Daily Dialogues
BelleGroup/multiturn_chat_0.8M	Daily Dialogues
BelleGroup/train_0.5M_CN	Daily Dialogues

Table 2: Statistics and sources of SFT data for SF-Med.

scenarios. To address this, we use four types of data in the SFT phase: single-turn and multi-turn real medical dialogues, medical NLP task data, and general daily conversation data. Data statistics can be found in Table 2.

Due to the occurrence of duplicate data from different sources, or even the same data, the diversity of the data is compromised. Therefore, we first utilized the method of Minhash-LSH (Bai et al., 2023) to deduplicate each data individually, followed by further deduplication after merging all the data. After de-duplication, we obtained approximately one million training data.

How to diversify instructions Most existing medical data are in question-and-answer format, lacking directives, which can result in inadequate learning when trained. Single-turn and multi-turn dialogues require different instructions: in multi-turn dialogues, the model can continue to ask questions after the first turn, while in single-turn dialogues, it should respond directly. Therefore, we provide specific instructions for different dialogue scenarios to guide the model's learning. We propose a simple and effective method for instruction augmentation based on multi-turn dialogues. The detailed algorithm flow can be seen in Figure 2. The process is as follows:

- Divide the original seed instructions in the Instruction Warehouse equally into Instruction Question Bank and Instruction Answer Bank. Our original seed instructions are sourced from the PMC_Llama_Instructions dataset. [1]
- Extract 5 instructions from Instruction Question Bank as the questions for instruction generation, and 5 from Instruction Answer Bank as the answers. Pair them to form historical dialogues for current instruction generation.
- Extract one instruction from Instruction Question Bank for expansion and use the GLM-4-9B-Chat (Team et al., 2024) model to generate new instructions.
- Split the generated instructions, and perform a simple evaluation of their completeness and professionalism, ensuring they contain medical-related terms. Half of the satisfactory instructions are returned to the Instruction Question Bank, and the other half to the Instruction Answer Bank.

[1] hf.co/datasets/axiong/pmc_llama_instructions

- Retain only the oldest m and the newest n instructions in both banks. This allows new instructions to reference some newer ones, ensuring diversity, while also referencing some of the more original ones, ensuring standardization. Periodically swap the two instruction banks during iterative instruction generation to maintain randomness in constructing historical dialogues for questions and answers.

Single-turn dialogue seed instruction example: Please respond based on the patient's question, speaking like a doctor. Multi-turn dialogue seed instruction example: Please respond based on the patient's question, speaking like a doctor. If the patient's illness cannot be diagnosed, you may ask for more information from the patient. However, please remember not to repeat the questions from previous rounds.

3.3 Reinforcement Learning from Human Feedback

The model may still produce incorrect or unnatural responses despite improvements in medical knowledge and conversational ability through continued pre-training and SFT. We use RLHF to mitigate these issues. First, we randomly select 40,000 samples from SFT data and 20,000 from additional data as annotation data. Then, we generate multiple responses for each annotation using the SFT model. Finally, we employ three open-source models (Qwen2.5-72B-Chat(Yang et al., 2024a), Yi-1.5-34B-Chat (Young et al., 2024), and GLM-4-9B-Chat (Team et al., 2024)) to choose preferred responses, creating a preference dataset.

For reinforcement learning, we use the DPO algorithm(Rafailov et al., 2023). This algorithm bypasses the reward model, directly fine-tuning with preference data. Thus, the training process is simple and efficient, with major improvements in the loss function. The DPO loss function is shown in the following equation.

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[\log \sigma(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{ref}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{ref}(y_l \mid x)})] \quad (1)$$

where (x, y_w, y_l) are preference pairs consisting of the prompt, the winning response, and the losing response from the preference dataset.

During the DPO experiments, we identified two factors affecting model performance: *1) Noisy data in the preference dataset can degrade performance. 2) Misalignment with the SFT model leads to forgetting basic medical capabilities.* To address these issues, we introduced two optimization strategies: preference data selection (Wang et al., 2024a) and online model merging (Lu et al., 2024).

Preference Data Selection: We train ten reward models using the same preference data, with the training order randomized. We define preference strength as the difference between the scores of Chosen and Rejected samples by the reward models. Then, we take the average preference strength from multiple reward models as a metric. For the lowest 10% of average preference strength, we assume these data labels are incorrect and negatively impact model performance. Therefore, we swap the labels of Chosen and Rejected to assist subsequent model learning. For the highest 10% preference strength, we consider the quality gap between these data to be large and overly focusing on them in the alignment process may lead to skewed model distributions. Thus, we remove these data directly.

Online Model Merging: In the DPO process, alignment tax may occur, meaning the model could forget the abilities acquired during the SFT phase. To address this, we introduce an online model merging strategy. Specifically, we merge gradients with the parameter differences between SFT and pretrained models, effectively steering the gradient toward maximizing rewards in the direction of SFT optimization. The formula is as follows:

$$\theta^{(t+1)} = \theta^{(t)} + \mathcal{F}(\Delta\theta^{(t)}) \oplus \mathcal{F}(\tau_r). \quad (2)$$

Where, $\tau_r = \theta_r - \theta_b$, with θ_b representing the pretrained parameters and θ_r representing the SFT model parameters. $F_{R(\cdot)}$, shown in Eq. 3, is a random sparsification operator based on the Bernoulli distribution with a fixed retaining probability p:

$$\mathcal{F}_R(x)_i = \begin{cases} x_i, & \text{Bernoulli}(p) = 1 \\ 0, & \text{Bernoulli}(p) = 0 \end{cases} \quad (3)$$

4 Experiments

4.1 Training Details

Our model is based on Qwen2.5-7B-Chat(Yang et al., 2024a), a versatile large language model with 7 billion parameters. The training process utilized 24 A800-80G GPUs in parallel. We employed full-parameter fine-tuning, and to balance

training costs, we used bfp16 precision alongside ZeRO-3 (Rajbhandari et al., 2020) and gradient accumulation strategies. The length of a single response, including its history, is capped at 2048 tokens. We incorporated the AdamW (Loshchilov, 2017) optimizer, a 0.1 dropout rate, and a cosine learning rate scheduler. The best-performing checkpoint is retained as the final model. We employed LLaMA-Factory (Zheng et al., 2024) as the training platform and vLLM (Kwon et al., 2023) for inference.

4.2 Datasets

To assess the model's capability in medical dialogue, we incorporated specialized medical Q&A data to simulate real doctor-patient interactions. This data includes both single-turn and multi-turn medical conversations. Concurrently, to evaluate the model's understanding and application of fundamental medical knowledge, we introduced a widely-used medical benchmark task and devised a medical terminology explanation task. Through the aforementioned evaluations, the capabilities of our medical model can be clearly reflected.

- Single-turn dialogue. Huatuo-26M(Chen et al., 2024) is currently a large Chinese medical question-and-answer dataset. This dataset contains over 26 million high-quality medical Q&A pairs, covering various aspects such as diseases, symptoms, treatment methods, and drug information. webMedQA(He et al., 2019) is a real-world Chinese medical question-answering dataset collected from online health consultancy websites. We use the test data to evaluate the medical models.

- Multi-turn dialogue. The CMtMedQA[2] test set includes 1000 items for evaluating the model's multi-turn dialogue ability.

- Medical Benchmark. We extracted questions about the medical field from C-Eval(Huang et al., 2023), CMMLU(Li et al., 2024), CMExam(Liu et al., 2023), CMB(Wang et al., 2024b), and part of the 2023 Chinese National Pharmacist Licensure Examination(Chen et al., 2024).

- Medical terminology explanation. We crawled medical terms and specialized explanations on the internet ourselves. For example, from medtiku[3].

4.3 Baselines

We selected four recently released, highly recognized open-source models in the Chinese medical field and conducted a comprehensive comparison with our Chinese medical model using the test dataset.

- HuatuoGPT-II(Chen et al., 2024) employs an innovative domain adaptation method to boost its medical knowledge and dialogue proficiency significantly. It showcases state-of-the-art performance in several medical benchmarks, especially surpassing GPT-4 in expert evaluations and fresh medical licensing exams.

- Zhongjing(Yang et al., 2024b) is the first Chinese medical model to implement the complete training process of pre-training, supervised fine-tuning, and RLHF, demonstrating impressive generalization capabilities. In some dialogue scenarios, it even approaches the professional level of medical doctors.

- WiNGPT2(Winning, 2023) is an LLM in the medical field, aimed at integrating professional medical knowledge, medical information, and data, providing intelligent medical Q&A, diagnostic support, and medical knowledge information services for the medical industry.

- ChiMed-GPT(Tian et al., 2024) is a Chinese medical LLM built by continually training Ziya-v2 on Chinese medical data, where pre-training, supervised fine-tuning (SFT), and RLHF are comprehensively performed on it.

At the same time, we also compare our model with the most representative general LLMs, such as Qwen2.5-7B(Yang et al., 2024a), ChatGPT(Ouyang et al., 2022) and GPT-4(Achiam et al., 2023).

4.4 Evaluation Metrics

After studying the evaluation methods of other medical models, we adopted three evaluation approaches, assessing the model's performance based on the distinct characteristics of the aforementioned

[2] https://huggingface.co/datasets/zhengr/CMtMedQA

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

QA-Rouge		Ours vs. HuatuoGPT-II	Ours vs. Zhongjing	Ours vs. ChiMed-GPT	Ours vs. WiNGPT2
Multi-turn dialogue	CMtMedQA	**0.754/0.000/0.246**	**0.592/0.002/0.406**	**0.861/0.002/0.137**	0.368/0.000/0.633
Single-turn dialogue	All	**0.558/0.008/0.434**	**0.539/0.014/0.447**	**0.500/0.013/0.487**	**0.545/0.010/0.445**
	huatuo26M	**0.584/0.008/0.408**	**0.506/0.016/0.478**	**0.506/0.008/0.486**	**0.532/0.008/0.460**
	webMedQA	**0.532/0.008/0.460**	**0.572/0.012/0.416**	**0.494/0.018/0.488**	**0.558/0.012/0.430**
Medical terminology	medtiku	**0.760/0.003/0.237**	**0.638/0.002/0.360**	**0.687/0.005/0.308**	**0.654/0.001/0.345**

Table 3: Rouge-L between SF-Med and other Chinese medical LLMs. "All" refers to the evaluation results after merging huatuo26M and webMedQA. The above mathematical notation represents the win, tie, and loss rates format, where the number before the slash indicates the number of times our model won, the number in the middle indicates the number of ties between our model and the compared model, and the number after the slash indicates the number of times our model lost.

QA-GPT		Ours vs. HuatuoGPT-II	Ours vs. Zhongjing	Ours vs. ChiMed-GPT	Ours vs. WiNGPT2
Multi-turn dialogue	CMtMedQA	**0.391/0.487/0.122**	**0.621/0.337/0.043**	**0.988/0.010/0.002**	**0.621/0.313/0.066**
Single-turn dialogue	All	**0.242/0.567/0.186**	**0.702/0.248/0.045**	**0.975/0.013/0.005**	**0.861/0.114/0.019**
	huatuo26M	**0.282/0.540/0.178**	**0.712/0.244/0.044**	**0.972/0.016/0.008**	**0.876/0.106/0.014**
	webMedQA	**0.202/0.594/0.194**	**0.692/0.252/0.046**	**0.978/0.010/0.002**	**0.846/0.122/0.024**

Table 4: AI evaluation between SF-Med and other Chinese medical LLMs.

QA-Rouge		Ours vs. Qwen2.5-7B-Instruct	Ours vs. ChatGPT	Ours vs. GPT-4
Multi-turn dialogue	CMtMedQA	**0.689/0.000/0.311**	0.342/0.002/0.656	**0.654/0.000/0.346**
Single-turn dialogue	All	**0.613/0.007/0.380**	0.451/0.010/0.539	**0.540/0.007/0.453**
	huatuo26M	**0.662/0.010/0.328**	0.416/0.008/0.576	**0.550/0.006/0.444**
	webMedQA	**0.564/0.004/0.432**	0.486/0.012/0.502	**0.530/0.008/0.462**
Medical terminology	medtiku	**0.863/0.003/0.134**	**0.605/0.006/0.389**	**0.842/0.000/0.158**

Table 5: Rouge-L between SF-Med and general LLMs.

QA-GPT		Ours vs. Qwen2.5-7B-Instruct	Ours vs. ChatGPT	Ours vs. GPT-4
Multi-turn dialogue	CMtMedQA	0.182/0.516/0.302	**0.265/0.586/0.149**	0.164/0.518/0.317
Single-turn dialogue	All	0.119/0.523/0.354	**0.325/0.554/0.117**	0.037/0.505/0.452
	huatuo26M	0.124/0.518/0.358	**0.364/0.506/0.130**	0.056/0.516/0.426
	webMedQA	0.114/0.528/0.350	**0.286/0.602/0.104**	0.018/0.494/0.478

Table 6: AI evaluation between SF-Med and general LLMs.

Multiple Choices	HuatuoGPT-II	Zhongjing	ChiMed-GPT	WiNGPT2	ChatGPT	GPT-4	Qwen2.5-7B-instruct	Our
All	0.58	0.49	0.53	0.48	0.49	0.71	0.72	**0.76**
PLE	0.47	0.31	0.48	0.42	0.41	0.69	0.61	**0.69**
Ceval	0.62	0.53	0.68	0.57	0.56	0.73	**0.75**	0.71
CMB	0.60	0.52	0.61	0.47	0.49	0.68	0.72	**0.77**
CMMLU	0.59	0.51	0.52	0.49	0.50	0.73	0.75	**0.79**
CMExam	0.65	0.55	0.53	0.51	0.50	0.68	0.69	**0.73**

Table 7: Multiple Choices Evaluation on SF-Med and other LLMs. "Avg." represents the average results across the five datasets. "All" refers to the evaluation results after merging five datasets

evaluation datasets, and aiming to comprehensively demonstrate the efficacy of our model.

AI evaluation The most authoritative evaluation method for various indicators of LLMs is manual evaluation, which uses human resources to compare the outputs of different models. In the absence of additional human resources, we currently consider that the GPT series models are closer to human evaluation results, so they can be replaced with the commonly used closed-source model APIs of the GPT series, such as GPT-4(Achiam et al., 2023).

Our evaluation method is as below,

- When referencing (Yang et al., 2024b), using GPT-4 as a tool to judge wins and losses can introduce bias. Specifically, the winning probability of model responses earlier in the prompt is significantly higher than those later in the prompt, which is not a fair comparison. To address this, we used GPT-4 to score each model's output individually and then compared the wins and losses based on these scores.

- Adding standard labels to the prompt content of GPT-4 and referring to them during evaluation can alleviate the bias of GPT-4 to some extent.

Rouge-L In the above evaluation process, we found that GPT models tend to consider the output length, and the reason given is generally "more detailed". After analyzing the characteristics of medical data, we believe there is not necessarily a positive correlation between output length and output quality. In real scenarios, doctors' responses are often accurate and brief, and overly detailed answers may not be friendly to patients with poor linguistic abilities.

In response to the above issue, as both the dialogue data and the medical terminology explanation data contain standard answers, we used Rouge-L (Lin, 2004), a widely used evaluation metric for generative tasks.

Accuracy For the medical benchmark, the model's answer options can be directly extracted, and the accuracy of the answers is calculated because of the standard answers. There are two key aspects here:

- Post-processing script. Although the model may contain an option in the answer, it does not explicitly provide the sequence number of the option, the post-processing script needs to have the ability to accurately extract the intent of the model option. We referred to a large number of other similar scripts and designed a comprehensive option post-processing plan based on the characteristics of the medical model.

- Some models may not have strong command capabilities, and if complex Prompts are inputted, the model may not be able to understand and provide answers. So we designed a relatively simple Prompt for inference. On the one hand, we tried to eliminate the impact of complex prompts on the model output as much as possible.

4.5 Results

The results of comparison between SF-Med and the other four open-source models in the Chinese medical fields are shown in Table 3 and Table 4. SF-Med performs better than other open-source medical models in both metrics and datasets except for a gap compared to WiNGPT2(Winning, 2023) on CMtMedQA datasets.

The performance of SF-Med in the medical benchmark is shown in Table 7. SF-Med exceeds all comparison models including open-source models and even GPT-4, etc. Overall, our model demonstrates a stronger capability in medical foundational knowledge.

The results of comparison between SF-Med and other general LLMs are shown in Table 5 and Table 6. To be honest, the output of ChatGPT(Ouyang et al., 2022) is relatively short and accurate, with higher textual similarity to standard answers. AI Evaluation is more inclined to longer content from Qwen2.5-7B(Yang et al., 2024a) and GPT-4(Achiam et al., 2023), which is always considered more detailed.

Besides, SF-Med outperforms all other models on the medical terminology task, reflecting the professionalism of the output content of SF-Med.

We also explored the performance improvements brought by three optimization strategies: instruction diversification, preference data selection, and online model merging. Detailed ablation studies can be found in Appendix A.

5 Conclusion

In this paper, we introduce SF-Med, an LLM specifically designed for Chinese medical consultations to advance the development of open-source Chinese medical models. SF-Med covers the complete model training process, including continued pretraining, supervised fine-tuning, and reinforcement learning. To enhance the model's generalization and expertise, we implemented optimizations such as data deduplication, instruction construction, preference data selection, and online model merging. We extensively evaluated SF-Med on medical benchmarks, demonstrating its superior performance compared to current open-source medical LLMs.

6 Limitations

Despite SF-Med demonstrating good performance across multiple test datasets and showing strong potential in many tasks, the model still has several limitations that need to be addressed and improved in future work.

Firstly, although our pre-training data includes 2.7 billion tokens of text, it still cannot cover all aspects of the medical field. Consequently, the model's performance might not be as strong when dealing with certain specific medical domains and rare diseases as it is with common disease datasets.

Secondly, the datasets and experimental setups used in this study may not fully represent the diversity of real clinical environments. Our model might overfit to specific patterns present in the training data, resulting in reduced generalization capability when faced with different data distributions from other medical institutions or regions.

Additionally, privacy and data security are also issues that must be taken seriously. Our training process relies on a large amount of patient data, and the privacy and security of this data need special attention in future practical applications. We will further explore how to effectively combine medical model training with the safeguarding of patient privacy to achieve an effective balance between accuracy and security.

Lastly, although we have demonstrated the potential of the model in our experiments, extensive validation and testing are still required for its practical application in clinical settings. This is necessary to evaluate the model's practicality and reliability in real-world scenarios.

References

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

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

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

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Huatuogpt-ii, one-stage training for medical adaption of llms. *Preprint*, arXiv:2311.09774.

Yiduo Guo, Jie Fu, Huishuai Zhang, Dongyan Zhao, and Yikang Shen. 2024. Efficient continual pre-training by mitigating the stability gap. *arXiv preprint arXiv:2406.14833*.

Junqing He, Mingming Fu, and Manshu Tu. 2019. Applying deep matching networks to chinese medical question answering: a study and a dataset. *BMC Medical Informatics and Decision Making*, 19(2):52.

Wei Huang, Yinggui Wang, Anda Cheng, Aihui Zhou, Chaofan Yu, and Lei Wang. 2024. A fast, performant, secure distributed training framework for llm. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4800–4804. IEEE.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pages 62991–63010. Curran Associates, Inc.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11260–11285, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, LEI ZHU, and Michael Lingzhi Li. 2023. Benchmarking large language models on cmexam - a comprehensive chinese medical exam dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 52430–52452. Curran Associates, Inc.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th*

Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Keming Lu, Bowen Yu, Fei Huang, Yang Fan, Runji Lin, and Chang Zhou. 2024. Online merging optimizers for boosting rewards and mitigating tax in alignment. *arXiv preprint arXiv:2405.17931*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

GLM Team, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv e-prints*, pages arXiv–2406.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2024. ChiMed-GPT: A Chinese medical large language model with full training regime and better alignment to human preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7156–7173, Bangkok, Thailand. Association for Computational Linguistics.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, et al. 2024a. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*.

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

Yinggui Wang, Wei Huang, and Le Yang. 2023. Privacy-preserving end-to-end spoken language understanding. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5224–5232.

Winning. 2023. Wingpt2.

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

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

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:2407.10671*.

Ping Yang, Junjie Wang, Ruyi Gan, Xinyu Zhu, Lin Zhang, Ziwei Wu, Xinyu Gao, Jiaxing Zhang, and Tetsuya Sakai. 2022. Zero-shot learners for natural language understanding via a unified multiple choice perspective. *Preprint*, arXiv:2210.08590.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19368–19376.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023a. HuatuoGPT, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo

Wu, Zhang Zhiyi, Qingying Xiao, et al. 2023b. Huatuogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*.

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*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Ablation Study

To investigate the contribution of SFT and RLHF to the performance of SF-Med, we conduct a series of ablation experiments on the evaluation benchmarks.

Diversified Instructions We verify the effects of the instruction augmentation in Table 8. As seen, incorporating the instructions can achieve a larger improvement in essentially all metrics across all datasets.

Preference Data Selection Table 9 shows the experimental results of SF-Med trained with and without the noisy preference data. We find that it marginally improves the metrics since our preference data has been selected by open-source models carefully early. The performance gains will be more noticeable if the preference data is collected from other sources.

Online Model Merging We compare the performance before and after online model merging during the RLHF phase in Table 10. The improvements in the accuracy of medical benchmarks are the most significant, further demonstrating that online model merging can mitigate the catastrophic forgetting phenomenon while reserving abilities gained by DPO.

In summary, these ablation experiments reveal the importance of diversified instructions, preference data selection, and online model merging in the training of medical LLMs, providing valuable experience and guidance for future research and applications in this field.

B The Privacy and Security of Medical Training Data

In addition to the training process of large medical models, we have conducted further research on the privacy leakage of medical models. The issue we aim to address is preventing the leakage of training data from large models. Our research primarily focuses on three aspects. The algorithm flowchart can be seen in Figure 3.

The first aspect is identifying training data that is likely to be memorized by the large model. The specific steps are as follows:

1) We split a sample from the training set into two parts. The first part serves as a prompt for the attacker to input into the model, and the second part acts as the label to check whether the model's output matches this training sample. For example, if the original text is A+B, then the prompt is A and the label is B, where A and B are sequences of tokens.

2) We use ROUGE-L to measure whether the model's output is sufficiently similar to the label. A higher similarity indicates that the model has a stronger "memory" of this training sample, thereby increasing the risk of privacy leakage.

3) We consider samples with a similarity greater than a certain threshold to be memorized by the model. Subsequent protective measures mainly target these samples. We selected 61,225 samples from the training data as test samples and found that only 512 samples, about 0.84% of the total, had a ROUGE-L score greater than 0.8.

The second aspect is generating a high-risk embedding database by passing high-risk training data through the medical model and extracting intermediate embeddings. We use intermediate embeddings mainly to address potential leakage issues of the high-risk embedding database.

The third aspect is the protection of training data in large medical models. The specific steps are as follows:

1) Based on the high-risk embedding database, We manually create a carefully designed secure embedding database in advance.

2) When the medical model is used for each user call, we compare the model's intermediate embeddings with the entries in the high-risk embedding database using cosine similarity. If the similarity exceeds a certain threshold (>0.8), we return the intermediate results from the secure embedding database to the user.

QA		QA-GPT	QA-Rouge	Accuracy
Multi-turn dialogue	CMtMedQA	0.298/0.491/0.199	0.594/0.004/0.402	-
Single-turn dialogue	All	0.323/0.397/0.271	0.479/0.017/0.504	-
	huatuo26M	0.328/0.394/0.272	0.498/0.010/0.492	-
	webMedQA	0.318/0.400/0.270	0.460/0.024/0.516	-
Medical terminology	medtiku	-	0.564/0.014/0.422	-
Multiple Choices		-	-	0.753/0.747

Table 8: Ablation study of diversified instructions.

QA		QA-GPT	QA-Rouge	Accuracy
Multi-turn dialogue	CMtMedQA	0.122/0.723/0.155	0.486/0.000/0.515	-
Single-turn dialogue	All	0.252/0.378/0.348	0.497/0.018/0.485	-
	huatuo26M	0.252/0.368/0.364	0.488/0.018/0.494	-
	webMedQA	0.252/0.388/0.332	0.506/0.018/0.476	-
Medical terminology	medtiku	-	0.397/0.006/0.597	-
Multiple Choices		-	-	0.737/0.721

Table 9: Ablation study of preference data selection.

QA		QA-GPT	QA-Rouge	Accuracy
Multi-turn dialogue	CMtMedQA	0.122/0.727/0.151	0.663/0.004/0.333	-
Single-turn dialogue	All	0.252/0.411/0.319	0.568/0.014/0.418	-
	huatuo26M	0.266/0.408/0.312	0.576/0.008/0.416	-
	webMedQA	0.238/0.414/0.326	0.560/0.020/0.420	-
Medical terminology	medtiku	-	0.682/0.001/0.317	-
Multiple Choices		-	-	0.721/0.656

Table 10: Ablation study of online model merging.

We compared the model utilizing our aforementioned security scheme with a model that does not employ security measures. The results are shown in Table 11. The table shows that our method has almost no impact on the model's performance.

Example of high-similarity text:

Label: Yes, you have understood accurately. Only individuals who have been bitten by animals or otherwise infected with the rabies virus can potentially transmit the virus through contact, saliva, etc. Since you have not been bitten by an animal and have not had contact with a patient, there is no need to worry about contracting the rabies virus.

Model-generated response: That's right, only people who have been bitten by animals or otherwise infected with the rabies virus can possibly transmit it to others through contact. You have not been bitten by an animal and have not had contact with a patient, so there is no need to worry about contracting the rabies virus.

Example of low similarity text:

Label: Okay, before discovering the child's eye disease, did the child ever have a high fever or a serious illness? Additionally, what you referred to as "pale optic nerve" might be "pale optic disc or papilla," which is often associated with optic nerve atrophy. It's important to note that many optic nerve disorders can lead to optic nerve atrophy, so it's crucial to find the cause once optic nerve atrophy is detected.

Model-generated response: Thank you for the information. If we can rule out the above possibilities, I can consider a treatment plan. Currently, oral murine nerve growth factor, injections of murine nerve growth factor, and citicoline are all possible treatment options. However, it should be noted that these medications cannot guarantee effectiveness.

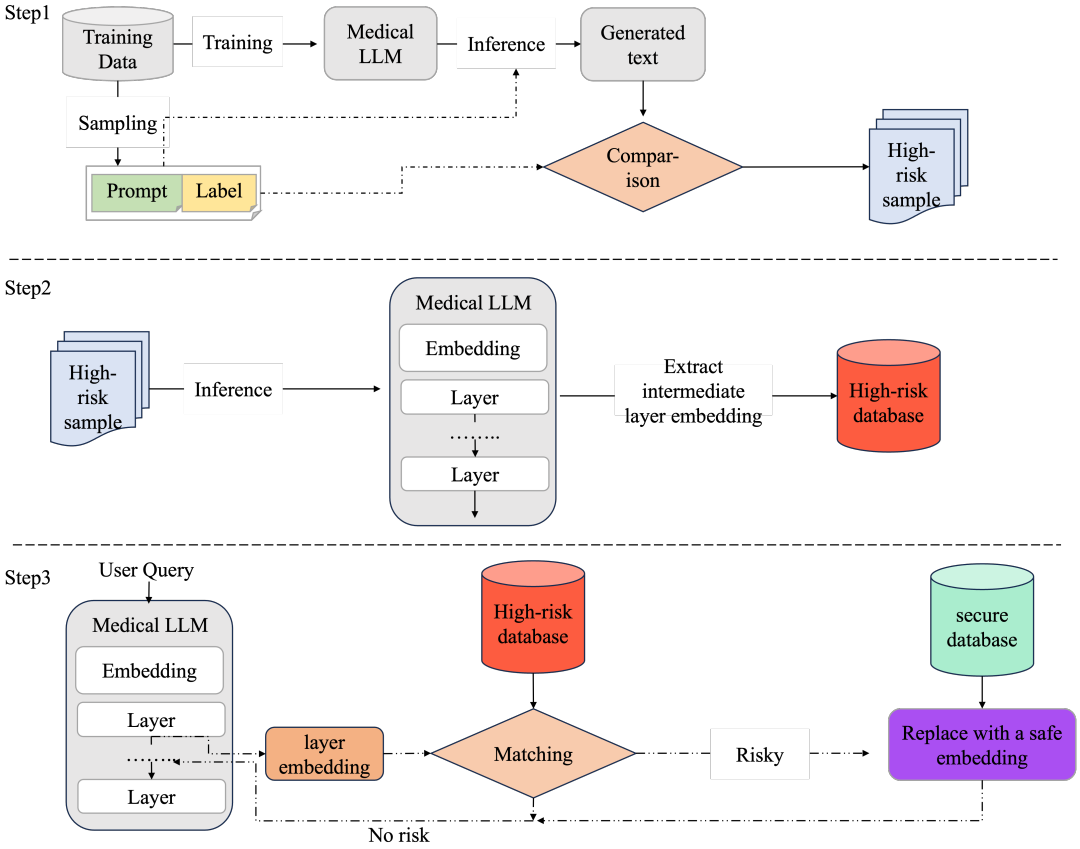

Figure 3: The algorithmic framework for privacy leakage and protection in SF-Med.

QA-GPT	Datset	Original Model vs. Privacy-Safe Model
Multi-turn dialogue	CMtMedQA	0.152/0.698/0.150
Single-turn dialogue	huatuo26M	0.211/0.580/0.209
	webMedQA	0.255/0.49/0.255

Table 11: Ablation study of diversified instructions.