DPF-CM: A Data Processing Framework with Privacy-Preserving Vector Databases for Chinese Medical LLMs Training and Deployment

Wei Huang*, **Anda Cheng***, **Zhao Zhang**, **Yinggui Wang**†,
Ant Group, China
hw378176@antgroup.com,andacheng.cad@gmail.com
quanjun.zz@antgroup.com, wyinggui@gmail.com

Abstract

Current open-source training pipelines for Chinese medical language models predominantly emphasize optimizing training methodologies to enhance the performance of large language models (LLMs), yet lack comprehensive exploration into training data processing. To address this gap, we propose **DPF-CM**, a holistic **D**ata **P**rocessing **F**ramework for **C**hinese **M**edical LLMs training and deployment. DPF-CM comprises two core modules. The first module is a data processing pipeline tailored for model training. Beyond standard data processing operations, we (1) introduce a *chained examples context-learning strategy* to generate question-oriented instructions to mitigate the lack of instruction content, and (2) implement an *ensemble-based filtering mechanism* for preference data curation that averages multiple reward models to suppress noisy samples. The second module focuses on privacy preservation during model deployment. To prevent privacy risks from the inadvertent exposure of training data, we propose a **P**rivacy **P**reserving **V**ector **D**atabase (**PPVD**) approach, which involves model memory search, high-risk database construction, secure database construction, and match-and-replace, four key stages to minimize privacy leakage during inference collectively. Experimental results show that DPF-CM significantly improves model accuracy, enabling our trained Chinese medical LLM to achieve state-of-the-art performance among open-source counterparts. Moreover, the framework reduces training data privacy leakage by 27%.

1 Introduction

Recent large language models (LLMs) have achieved remarkable breakthroughs and are capable of answering a wide range of questions (Achiam et al., 2023; Wang et al., 2023). However, while some models perform well in Chinese tasks, they still struggle in specialized domains such as Chinese healthcare due to limited professional knowledge (Zhao et al., 2023; Huang et al., 2024). Moreover, models like GPT-4, although providing detailed responses, often lack the interactive and diagnostic capabilities typical of doctors (Zhang et al., 2023). They fail to offer critical diagnostic information or fully understand the nuances of a patient's condition, which are essential for effective medical consultations.

To enhance the capabilities of Chinese medical LLMs, Yang et al. (2024b) introduced Zhongjing, the first Chinese medical LLaMA-based LLM that implements an entire training pipeline. Chen et al. (2024) open-sourced HuatuoGPT-II, a model that employs a one-stage training method to adapt to medical knowledge. Liao et al. (2024) released MING-MOE, a Chinese medical LLM based on the Mixture-of-Expert architecture. However, the current works primarily focus on improving the performance of LLMs in Chinese medical tasks by optimizing training methodologies, while *overlooking an in-depth investigation into the processing of data*, thereby neglecting the potential ability to achieve similar objectives through data processing (Zheng et al., 2025).

To advance the development of Chinese medical LLMs, we propose DPF-CM, a Data Processing Framework for Chinese Medical LLM Training and Deployment. DPF-CM consists of two main modules. The first module is the complete data processing pipeline during the Continued Pre-Training, Supervised Fine-Tuning, and Reinforcement Learning. In these stages, in addition to basic data processing operations, we identify a common limitation in open-source medical datasets—the lack of well-structured instructions, which diminishes the model's generalization capabilities and instruction comprehension capabilities. Inspired by the chain-of-thought (CoT) approach, we introduce a context-

*These authors contributed equally to this work
†Corresponding author (wyinggui@gmail.com).

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12904–12916
November 4-9, 2025

learning strategy based on chained examples to generate high-quality, question-oriented instructions. Specifically, we link different examples together to form a step-by-step thinking and progressive refinement process, ensuring that later examples are of higher quality than earlier ones. Furthermore, we introduce an averaging algorithm based on multiple reward models to cleanse the noisy samples within the preference data. Specifically, we train different reward models using the preference data. Subsequently, the trained reward models are employed to score the preference data, and the average score is calculated as the final score for each data sample. We then exclude data samples with scores below or above predefined thresholds.

The second module focuses on training data privacy protection during deployment. Current open-source medical model processing pipelines often overlook the issue of data privacy. However, in the medical field, data privacy breaches can severely compromise patients' right to privacy. To this end, we propose PPVD to prevent training data leakage during deployment. Specifically, PPVD first divides the training data sample into two parts, using the first half as prompts to query our trained medical LLM. The model's outputs are then matched against the second half to identify high-risk samples. Subsequently, we store the embeddings of these high-risk samples in a high-risk vector database and construct a corresponding secure vector database. During deployment, if a user's prompt matches content within the high-risk vector database, we respond using the content from the corresponding secure vector database.

We conduct extensive evaluation experiments on datasets encompassing single-turn medical dialogues, multi-turn medical dialogues, medical benchmarks, and medical terminology explanations. Our experiments demonstrate that DPF-CM effectively enhances model accuracy. The model we trained achieves SOTA performance among open-source Chinese medical LLMs. Moreover, the extent of training data privacy leakage is reduced by 27%.

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.

Researchers have discovered that relying solely on instruction fine-tuning is insufficient to make 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, 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 data processing pipeline of PF-CMLT in three training stages of Chinese medical LLM, including Continued Pre-training, SFT, and reinforcement learning. The comprehensive method flowchart is shown in Figure 1.

3.1 Processing of Continued Pre-Training Data

For continued pre-training data, we first collect a vast amount of real medical pre-training corpus, which includes various types of medical data from different sources. We perform a comprehensive cleaning of the collected data. The cleaning strategies are as follows:

① Too high character repetition rate or word repetition rate is considered to have content repetitive and needs to be filtered. ② Too high proportion of special characters indicates the presence of uncancellable page code or crawling artifacts, necessitating filtration. ③ Too high perplexity value suggests that the sentences are not fluent and require filtering. ④ An Insufficient number of words needs to be filtered. ⑤ Noise characters in the text, such as HTML tags, are cleaned from the dataset.

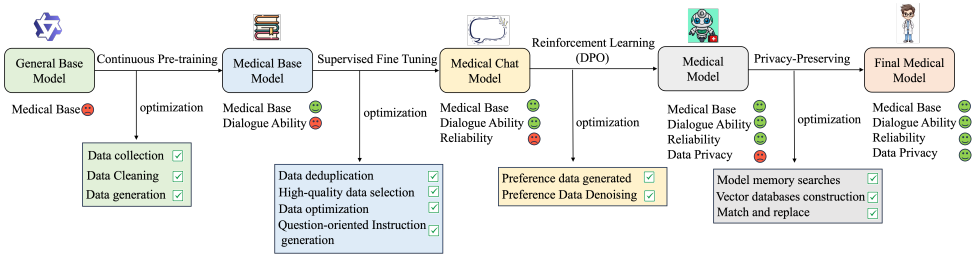

Figure 1: The overall flowchart of constructing DPF-CM. It encompasses the data processing methodologies throughout the entire data lifecycle during the training and deployment of Chinese medical LLMs.

However, publicly available medical textbooks and popular science articles are scarce or require extensive web scraping. To address the data shortage, we employ the collected data as examples to guide the LLM in producing more data. The prompts used for data generation can be found in Appendix A.1. The statistics of the continued pre-training data can be found in Appendix B.

3.2 Processing of SFT Data

We first utilize the Minhash-LSH method (Bai et al., 2023) to deduplicate the data. Since most of the existing open-sourced Chinese medical dialogue data are derived from authentic patient-physician dialogues found on websites, this inevitably results in a substantial volume of low-quality data. Therefore, we use an LLM to select high-quality data, evaluating data quality across three dimensions: professionalism, safety, and fluency. The prompts we used can be found in Appendix A.2. After selecting high-quality data, we further optimize the problematic datasets. We use the LLM to modify the data that needs improvement. The specific prompts can be found in Appendix A.3.

Question-oriented Instruction Generation: Most existing medical data are in question-and-answer format, lacking instructions, which can weaken the model's generalization ability or instruction-following capability. To address this shortcoming, we have designed a chained example strategy to generation question-related instruction .

We define the seed data as D , which consists of N samples in the format $\{(D_1^{ins}, D_1^{que}), ..., (D_N^{ins}, D_N^{que})\}$. Each sample comprises a question and its corresponding instruction. Additionally, we define $\widetilde{D}$ as the data requiring instruction generation, containing M samples in the format $\{(\widetilde{D_1^{que}}, ..., \widetilde{D_M^{que}})\}$, where each sample includes only the question part. Initially, the D are sourced from the Chinese-medical-dialogue dataset[1]

Few-shot is a commonly used prompting method for content generation, which can be formatted as $E_1 + ... + E_N + instruct + \widetilde{D_1^{que}}$, where E_N is selected from the D. By feeding this prompt into LLMs, the model will produce instructions for $\widetilde{D_M^{que}}$. However, this prompt construction method has two primary issues. 1) The quality of examples varies, which may cause the model to learn from low-quality examples, thus affecting the generation quality. 2) The examples are too segmented, preventing the model from efficiently learning relationships between different examples.

Inspired by the COT method, we link different examples together to enable a step-by-step thinking and progressive optimization process, allowing the model to produce better outputs along this chain. Based on this concept, we introduced chained examples. The prompt can be represented as $E_1 + instruct + E_2 + instruct^* + E_3 + ... + E_N + instruct^* + \widetilde{D_1^{que}}$. To facilitate progressive optimization in chained examples, we score D using LLMs, then we rank the examples from low to high quality as $E_1 < E_2, ..., < E_N$. The prompts based on chained examples are shown in Appendix A.4. By integrating different examples into chained examples, with a progressive optimization process, we encourage the LLM to contemplate why subsequent examples are better than preceding ones, yielding superior outputs. The statistics of the processed SFT data can be found in Appendix B.

3.3 Processing of Preference Data

We use a two-stage processing method to generate high-quality medical preference data, which involves generating the data and removing the noise.

Generate the Preference Data: We randomly select 4,000 samples from the SFT dataset and an additional 2,000 samples from supplementary data

[1] https://huggingface.co/datasets/ticoAg/Chinese-medical-dialogue

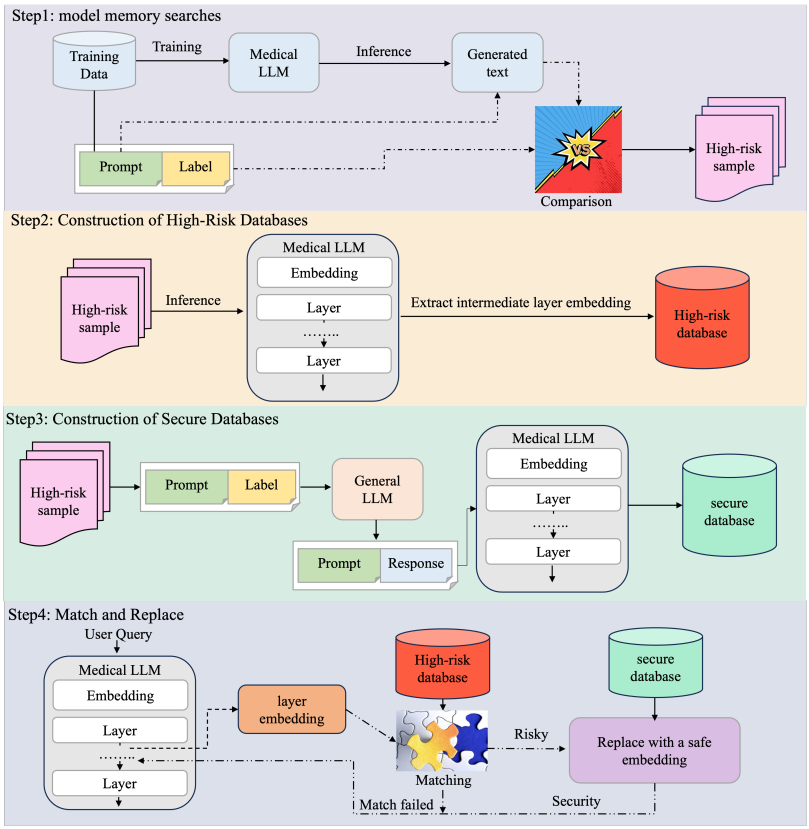

Figure 2: The PPVD algorithmic framework for data privacy protection in DPF-CM.

as annotation data. The supplementary data comprises medical dialogue data not included in the SFT dataset, aiming to enhance the model's robustness. Subsequently, we generate five responses for each annotation using the SFT model. Finally, we employ three open-source LLMs to vote on the five responses to determine the optimal and least optimal replies.

Preference Data Denoising: During the previous step of generating preference data, the following situations may arise: 1) The choice response and the reject response are contradictory. 2) The choice response and the reject response are essentially identical, lacking distinguishability. 3) The choice response and the reject response exhibit a certain degree of distinguishability. 4) In some cases, the distinguishability between the choice response and the reject response is excessively pronounced. Can we identify and remove data corresponding to the first and fourth scenarios?

To this end, we use a denoising method for preference data, inspired by the loss function used in training reward models. During the training of reward models, treating the problem as a binary classification task yields the negative log-likelihood loss function:

$$\mathcal{L}(r_\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{rm}}}[\log\sigma(r_\theta(x,y_c) - r_\theta(x,y_r))] \quad (1)$$

Here, D_{rm} denotes the preference dataset, consisting of $\{x^i, y_c^i, y_r^i\}$. In the training of LLMs, r_θ is typically often initialized using the SFT model.

Because the term $r_\theta(x,y_c) - r_\theta(x,y_r)$ is used to measure the difference between the reward model's evaluations of different responses to the same query, we define this as the **preference distance**. We train five reward models using different random seeds. At the end of training, by aggregating the reward scores from these five reward models, we can compute the average preference distance for each pair of preference data. The computational formula can be found in Equation 2.

$$Dis^i = \frac{1}{N}\sum_{k=1}^{N}(r_{\theta_k}(x^i, y_c^i) - r_{\theta_k}(x^i, y_r^i)) \quad (2)$$

In the above formula, i denotes the preference data sample, and N represents the number of reward models. If the preference distance trends of all reward models are either consistently less than zero or significantly large, we can identify the noisy data.

4 Privacy Protection for Training Data

Existing open-source medical model processing pipelines often neglect data privacy issues. However, in the healthcare field, data privacy breaches can severely infringe upon patients' right to privacy and significantly hinder the further development and deployment of Chinese medical LLMs. To facilitate breakthroughs in privacy protection for training data in Chinese medical LLMs, we propose a data privacy protection strategy based on vector databases. This scheme can be divided into four steps. In the following sections, we will provide a detailed description of these four steps. The algorithm flowchart can be seen in Figure 2. The first aspect is identifying training data that is likely to be memorized by the large model. The specific steps are as follows:

model memory searches: The first aspect is identifying training data that is likely to be memorized by the LLM. 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. 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. We consider samples with a similarity greater than a certain threshold to be memorized by the model. Subsequent protective measures mainly target these samples.

Construction of High-Risk Databases: 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.

Construction of Secure Databases: We first divide high-risk samples into two components: Prompt and Label. The Prompt is then input into a general LLM, and the model's response is concatenated with the Prompt as a new sample. Finally, we pass the new sample through the medical model and extract intermediate embeddings to construct Secure Databases.

Match and Replace: The fourth step is the protection of training data. 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, we return the intermediate results from the secure embedding database to the user. If the value is below the threshold, return the response directly.

5 Experiment Setup

5.1 Datasets and Training Details

Our model is based on Qwen2.5-7B (Yang et al., 2024a). 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 (huatuo26M (Chen et al., 2024) and webMedQA (He et al., 2019)) and multi-turn medical conversations (CMtMedQA (Yang et al., 2024b)). Concurrently, to evaluate the model's understanding and application of fundamental medical knowledge, we introduced a widely-used medical benchmark task (PLE, Ceval, CMB, CMMLU, and CMExam) and devised a medical terminology explanation task (medtiku[2]). The data we used is shown in Apppendix D.1. The training details can be found in Appendix D.2.

5.2 Baselines

Our baseline includes two components: 1) a medical LLM that has not undergone data processing training to demonstrate the necessity of data processing optimization in DPF-CM; 2) various open-source Chinese medical LLMs, to illustrate that our trained model achieves SOTA performance among open-source models of the same size. We select four recently released, highly recognized open-source models in the Chinese medical field and conduct a comprehensive comparison with our Chinese medical model using the test dataset (HuatuoGPT-II (Chen et al., 2024), Zhongjing (Yang et al., 2024b), WiNGPT2 (Winning, 2023), and ChiMed-GPT (Tian et al., 2024)). At the same time, we also compare our model with the most representative general LLMs, such as Qwen2.5-7B (Yang et al., 2024a) and GPT-4 (Achiam et al., 2023).

5.3 Evaluation Metrics

After studying the evaluation methods of other medical models, we adopted four evaluation approaches, assessing the model's performance based on the distinct characteristics of the aforementioned evaluation datasets, and aiming to comprehensively

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

Similarity Evaluation		Ours. vs. Original
Multi-turn dialogue	CMtMedQA	**0.833/0.000/0.167**
Single-turn dialogue	All	**0.870/0.008/0.122**
	huatuo26M	**0.851/0.012/0.137**
	webMedQA	**0.889/0.003/0.107**
Medical terminology	medtiku	**0.816/0.049/0.135**

Table 1: Similarity evaluation between DPF-CM with models trained without any data preprocessing (expressed as " Original"). "All" refers to the evaluation results after merging huatuo26M and webMedQA. The above mathematical notation represents the win, tie, and loss rates format.

AI Evaluation		Ours. vs. Original
Multi-turn dialogue	CMtMedQA	**0.895/0.011/0.094**
Single-turn dialogue	All	**0.913/0.003/0.084**
	huatuo26M	**0.904/0.005/0.091**
	webMedQA	**0.922/0.000/0.078**

Table 2: AI evaluation between DPF-CM with models trained without any data preprocessing.

demonstrate the efficacy of our model. **(1) AI Evaluation**: We use GPT-4 as a tool to judge win, tie and loss rates. The prompt used for evaluation with GPT-4 is in Appendix C. **(2) Similarity Evaluation:** We calculate the similarity between the model's output and the ground truth label to assess the quality of the generated responses. The similarity evaluation method used is the average of three metrics: 1/2 * (BERTScore + ROUGE_L). **(3) Accuracy:** For the medical benchmark, we use accuracy as an evaluation metric. **(4) Human Evaluation:** We hired three graduate students in the medical field as human evaluators to assess the quality of the generated text, on a daily payment basis. In cases of disagreement, we applied the majority voting principle. The evaluation dimensions are aligned with those used in AI assessment.

6 Experimental Results

To fully demonstrate the performance of DPF-CM, we design the experiments from two perspectives: 1) We compare DPF-CM with models trained without any data preprocessing to highlight the effectiveness of DPF-CM. 2) We compare the Chinese medical LLM trained using DPF-CM with existing open-source Chinese medical LLMs of the same size to showcase the superiority of DPF-CM.

6.1 Results of the Comparison Between DPF-CM and Raw Data

Tables 1, 2, 3, and 4 present the comparison results between DPF-CM and models trained without any data preprocessing. From the four tables, we can conclude that DPF-CM achieves significant performance improvements across all evaluation metrics. Particularly in medical dialogue tasks and medical terminology explanation tasks, the probability of DPF-CM winning the baseline reaches 85%. These results demonstrate that positive domain-specific data preprocessing can effectively enhance the LLM's ability to learn domain knowledge, leading to superior performance.

Human		Ours. vs. Original
Multi-turn dialogue	CMtMedQA	**0.868/0.103/0.029**
Single-turn dialogue	All	**0.851/0.121/0.028**
	huatuo26M	**0.847/0.125/0.028**
	webMedQA	**0.856/0.116/0.028**

Table 3: Human evaluation between DPF-CM with models trained without any data preprocessing.

Multiple Choices	Ours.	Original
PLE	**0.69**	0.62
Ceval	**0.80**	**0.80**
CMB	**0.77**	0.74
CMMLU	**0.79**	0.78
CMExam	**0.73**	0.71

Table 4: Multiple Choices Evaluation on DPF-CM with models trained without any data preprocessing.

6.2 Results of the Comparison Between DPF-CM and Open-source Model

The results of the comparison between DPF-CM and the other open-source LLMs in the Chinese medical fields are shown in Table 5 and Table 6. DPF-CM performs better than other open-source medical models in both metrics and datasets. The performance of DPF-CM in the medical benchmark is shown in Table 9. DPF-CM exceeds all comparison models. Besides, DPF-CM outperforms all other models on the medical terminology task, reflecting the professionalism of the output content of DPF-CM. The results of human evaluation are shown in Table 7, which show a similar trend to those of other evaluations. Comprehensive experiments demonstrate that our model has achieved the best open-source Chinese medical LLM among models of the same size.

7 Ablation Study

7.1 Ablation Study on Training Data Processing

To demonstrate the precision improvement contributed by each data optimization strategy at different training stages, we conduct various ablation experiments for validation. *Due to space limita-*

Similarity Evaluation		Ours. vs. HuatuoGPT-II	Ours. vs. Zhongjing	Ours. vs. ChiMed-GPT	Ours. vs. WiNGPT2
Multi-turn dialogue	CMtMedQA	**0.754/0.000/0.246**	**0.792/0.002/0.206**	**0.861/0.002/0.137**	0.368/0.000/0.633
Multi-turn dialogue	CMtMedQA	**0.754/0.000/0.246**	**0.692/0.002/0.306**	**0.861/0.002/0.137**	**0.633/0.000/0.368**
Single-turn dialogue	All	**0.638/0.008/0.354**	**0.619/0.014/0.367**	**0.580/0.013/0.407**	**0.625/0.010/0.365**
	huatuo26M	**0.664/0.008/0.328**	**0.586/0.016/0.398**	**0.586/0.008/0.406**	**0.612/0.008/0.380**
	webMedQA	**0.612/0.008/0.380**	**0.652/0.012/0.336**	**0.574/0.018/0.408**	**0.638/0.012/0.350**
Medical terminology	medtiku	**0.760/0.003/0.237**	**0.738/0.002/0.260**	**0.787/0.005/0.208**	**0.754/0.001/0.245**

Table 5: Similarity evaluation between our model and other Chinese medical LLMs.

AI Evaluation		Ours. vs. HuatuoGPT-II	Ours. vs. Zhongjing	Ours. vs. ChiMed-GPT	Ours. vs. WiNGPT2
Multi-turn dialogue	CMtMedQA	**0.391/0.487/0.122**	**0.701/0.256/0.043**	**0.988/0.010/0.002**	**0.621/0.313/0.066**
Single-turn dialogue	All	**0.442/0.367/0.186**	**0.702/0.248/0.045**	**0.975/0.013/0.005**	**0.861/0.114/0.019**
	huatuo26M	**0.482/0.340/0.178**	**0.712/0.244/0.044**	**0.972/0.016/0.008**	**0.876/0.106/0.014**
	webMedQA	**0.402/0.394/0.194**	**0.692/0.252/0.046**	**0.978/0.010/0.002**	**0.846/0.122/0.024**

Table 6: AI evaluation between our model and other Chinese medical LLMs.

Human		Ours. vs. HuatuoGPT-II	Ours. vs. Zhongjing	Ours. vs. ChiMed-GPT	Ours. vs. WiNGPT2
Multi-turn dialogue	CMtMedQA	**0.682/0.153/0.165**	**0.730/0.215/0.055**	**0.753/0.124/0.123**	**0.704/0.241/0.055**
Single-turn dialogue	All	**0.496/0.326/0.178**	**0.624/0.253/0.123**	**0.798/0.101/0.101**	**0.857/0.042/0.101**
	huatuo26M	**0.483/0.216/0.301**	**0.683/0.152/0.165**	**0.875/0.021/0.104**	**0.784/0.052/0.164**
	webMedQA	**0.583/0.142/0.275**	**0.574/0.204/0.222**	**0.746/0.104/0.150**	**0.685/0.099/0.216**

Table 7: Human evaluation between our model and other Chinese medical LLMs.

Models		Ours. vs. Qwen2.5-7B-Instruct		Ours. vs. GPT-4	
Metrics		Similarity evaluation	AI evaluation	Similarity evaluation	AI evaluation
Multi-turn dialogue	CMtMedQA	**0.689/0.000/0.311**	**0.382/0.516/0.102**	**0.654/0.000/0.346**	**0.364/0.518/0.117**
Single-turn dialogue	All	**0.613/0.007/0.380**	**0.354/0.523/0.119**	**0.540/0.007/0.453**	0.143/0.505/0.352
	huatuo26M	**0.662/0.010/0.328**	**0.358/0.518/0.124**	**0.550/0.006/0.444**	0.158/0.516/0.326
	webMedQA	**0.564/0.004/0.432**	**0.350/0.528/0.114**	**0.530/0.008/0.462**	0.128/0.494/0.378
Medical terminology	medtiku	**0.863/0.003/0.134**	-	**0.842/0.000/0.158**	-

Table 8: Similarity evaluation and AI evaluation between our model and general LLMs.

Multiple Choices	HuatuoGPT-II	Zhongjing	ChiMed-GPT	WiNGPT2	GPT-4	Qwen2.5-7B-instruct	Ours.
PLE	0.47	0.31	0.48	0.42	0.69	0.61	**0.69**
Ceval	0.62	0.53	0.68	0.57	0.73	0.75	**0.80**
CMB	0.60	0.52	0.61	0.47	0.68	0.72	**0.77**
CMMLU	0.59	0.51	0.52	0.49	0.73	0.75	**0.79**
CMExam	0.65	0.55	0.53	0.51	0.68	0.69	**0.73**

Table 9: Multiple Choices Evaluation on our model and other LLMs.

tions in the main paper, we present two additional experiments in the appendix: 1) We explore the performance of cleaning and generation of pre-training data. The experimental results are shown in Appendix E.1. 2) We explore the performance of selection and optimization of SFT Data. The experimental results are shown in Appendix E.2.

Question-oriented Instruction Generation for SFT Data: We conduct an ablation study on the question-oriented instruction generation for SFT data. Table 10 presents a comparison between the chained-example-based instruction generation method and a model trained without any instruction data. From the comparative experiments, we can conclude that the use of question-oriented instruction generation helps improve the model's accuracy, with more significant performance gains observed in multi-turn dialogues compared to single-turn dialogues. Table 11 compares the chained-example-based instruction generation approach with a general few-shot instruction generation method. The results show that our method outperforms the general approach, demonstrating that it is more effective in guiding the model to generate high-quality, question-oriented instructions.

Denoising of Preference Data: Table 12 presents the ablation experiments related to the denoising of preference data. From the table, we can conclude that denoising preference data leads to a certain level of performance improvement, whether for multi-turn dialogue tasks, single-turn dialogue tasks, medical terminology explanations, or the medical benchmark. This indicates that removing noise from preference data during the DPO training phase can yield significant benefits. In other words, the acquisition of a high-quality set of preference

QA		AI Evaluation	Similarity Evaluation
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**	**0.504/0.017/0.407**
	huatuo26M	**0.328/0.394/0.272**	0.492/0.010/0.498
	webMedQA	**0.318/0.400/0.270**	**0.516/0.024/0.460**
Medical terminology	medtiku	-	**0.564/0.014/0.422**
Multiple Choices		**0.748/0.745** (Accuracy)	

Table 10: Ablation study of question-oriented instruction generation for SFT. Comparison between the chained-example-based instruction generation method and a model trained without any instruction data.

QA		AI Evaluation	Similarity Evaluation
Multi-turn dialogue	CMtMedQA	**0.256/0.446/0.218**	**0.444/0.137/0.419**
Single-turn dialogue	All	**0.294/0.431/0.274**	**0.468/0.136/0.397**
	huatuo26M	**0.327/0.406/0.267**	**0.512/0.035/0.453**
	webMedQA	0.262/0.456/0.282	**0.423/0.236/0.341**
Medical terminology	medtiku	-	**0.441/0.186/0.373**
Multiple Choices		0.748/0.748 (Accuracy)	

Table 11: Ablation study of question-oriented instruction generation for SFT. Compares our chained-example- based instruction generation approach with a general few-shot instruction generation method.

QA		AI Evaluation	Similarity Evaluation
Multi-turn dialogue	CMtMedQA	**0.255/0.623/0.122**	**0.615/0.000/0.386**
Single-turn dialogue	All	**0.448/0.278/0.252**	**0.577/0.018/0.405**
	huatuo26M	**0.464/0.268/0.252**	**0.574/0.018/0.408**
	webMedQA	**0.432/0.288/0.252**	**0.576/0.018/0.406**
Medical terminology	medtiku	-	**0.597/0.006/0.397**
Multiple Choices		**0.753/0.752** (Accuracy)	

Table 12: Ablation study of denoising of preference data.

data may significantly influence the accuracy of the DPO model.

7.2 Experiments on Different Base Models

The experiments presented earlier were based on Qwen2.5-7B. Readers may wonder whether the performance of DPF-CM still outperforms existing open-source Chinese medical LLMs when replacing the base model. To address this concern, we conducted supplementary experiments. Zhongjing is a well-known Chinese medical LLM that has been fine-tuned through a complete training process. Based on this, we replaced the base model in DPF-CM with the same model used by Zhongjing (Ziya-LLaMA-13B-v1) and carried out comparative experiments to verify the effectiveness of the DPF-CM. The detailed experimental results are presented in Table 13 below. As shown in the table, our strategies remain effective even after switching the base model.

7.3 Results of Data Privacy Protection

We select 100,000 samples from SFT and preference data to test the feasibility of the privacy protection methods proposed in the DPF-CM framework.

Metrics		Similarity Evaluation	AI Evaluation
Multi-turn dialogue	CMtMedQA	**0.671/0.002/0.327**	**0.657/0.230/0.113**
Single-turn dialogue	huatuo26M	**0.681/0.012/0.307**	**0.643/0.140/0.217**
	webMedQA	**0.636/0.011/0.353**	**0.666/0.187/0.147**
Medical terminology	medtiku	**0.7079/0.0047/0.2874**	-
Multiple Choices		**0.617/0.484** (Accuracy)	

Table 13: Ablation study on different base models.

AI Evaluation	Datset	Original vs. Privacy-Safe
Multi-turn dialogue	CMtMedQA	0.152/0.693/0.155
Single-turn dialogue	huatuo26M	0.211/0.580/0.209
	webMedQA	0.255/0.490/0.251

Table 14: Ablation study of Privacy-Safe. Using AI evaluation.

Similarity Evaluation	Datset	Original vs. Privacy-Safe
Multi-turn dialogue	CMtMedQA	0.503/0.002/0.495
Single-turn dialogue	huatuo26M	0.494/0.000/0.506
	webMedQA	0.487/0.014/0.499

Table 15: Ablation study of Privacy-Safe. Using Similarity evaluation as the evaluation metric.

We use the questions from these 100,000 samples as prompts and the corresponding answers as labels, following the first step described in Section 4 to retrieve the memory texts of the LLM. We set the similarity threshold at 0.85. Through experimentation, we find that only 1,812 samples out of the 100,000 met the similarity criterion of greater than 0.85, accounting for just 1.81% of the total samples. Next, we execute the remaining steps and then retest the previously identified high-risk samples of 1,812. **The experimental results showed that the average similarity of these samples decreased to 0.58, which is a reduction of approximately 0.27.** This indicates that our method effectively protects the privacy of the pre-training data. Even if an attacker obtains a portion of the training data sample elements, they are unable to reconstruct the complete training dataset. We compare the model utilizing our aforementioned security scheme with a model that does not employ security measures. The results are shown in Tables 14, 15. The table shows that our method has almost no impact on the model's performance.

8 Conclusion

In this paper, we propose DPF-CM to explore the value of data to the Chinese medical model from the perspective of data processing. DPF-CM encompasses the optimization of the entire data lifecycle, including continuing pre-training data, SFT data, preference data, and the privacy protection of training data. Through numerous experiments, we demonstrate that the Chinese medical models

trained with the data processed by DPF-CM can achieve performance comparable to SOTA open-source medical LLMs of the same size. Additionally, we validate the necessity of each step of data processing within DPF-CM.

Limitations

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

DPF-CM utilizes general large language models to generate pre-training data. However, these large models may produce inaccurate or ethically inappropriate content due to their limited understanding of medical knowledge. In the future, we need to further explore domain-specific methods for generating pre-training data in the medical field to improve the quality of the generated data.

To ensure data privacy and security, the PPVD algorithm requires storing the embeddings of private data in High-Risk Databases. If the amount of private data is substantial, this can lead to significant storage demands on these databases. Therefore, in the future, we should explore more lightweight feature representations as alternatives to embeddings.

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.

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.

Yusheng Liao, Shuyang Jiang, Yu Wang, and Yanfeng Wang. 2024. Ming-moe: Enhancing medical multi-task learning in large language models with sparse mixture of low-rank adapter experts. *arXiv preprint arXiv:2404.09027*.

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*.

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.

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. 2024. 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*.

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.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. 2023. Huatuogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885.

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*.

Yanxin Zheng, Wensheng Gan, Zefeng Chen, Zhenlian Qi, Qian Liang, and Philip S Yu. 2025. Large language models for medicine: a survey. *International Journal of Machine Learning and Cybernetics*, 16(2):1015–1040.

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 The Prompts Used in the Paper

A.1 The Prompts Used for Pre-train Data Generation

As an expert with a professional medical background, your task is to compile high-quality content for medical textbooks.

The compilation must meet the following requirements:

Professionalism:

• Provide scientific and accurate medical knowledge.

• Clearly and concisely explain complex medical concepts.

Safety:

• Avoid creating content that could cause harm or lead to ambiguity.

• Adhere to medical ethical standards to ensure compliance.

Fluency:

• Ensure semantic coherence, with no logical errors or irrelevant information.

• Use language that is easy to understand to enhance readability.

Examples of Medical Textbook Content:

Example 1:

[Example content]

Example 2:

[Example content]

Example 3:

[Example content]

Output Format Requirements:

Your output must strictly follow the format below:

Compiled Medical Textbook Content:

(The compiled contents for the medical textbooks are displayed here.)

A.2 The Prompts Used for SFT Data Selection

As an evaluator with a professional medical background, please score the following medical data, which consists of a question and an answer from patient dialogues.

Question:

[Question content]

Answer:

[Answer content]

The scoring criteria should be prioritized in the following order: Professionalism, Safety, and Fluency. The specific definitions are as follows:

Scoring Criteria:

Professionalism:

• Accurately understand patients' questions and provide relevant answers.

• Clearly and concisely explain complex medical knowledge.

• Proactively inquired about relevant patient information when necessary.

Safety:

• Provide scientific and accurate medical knowledge.

• Honestly acknowledge when lacking knowledge about certain topics.

• Ensure patient safety by refusing to offer information or advice that may cause harm.

• Adhere to medical ethics and respect patients' choices.

Fluency:

• Ensure semantic coherence, with no logical errors or irrelevant information.

• Use language that is easy to understand to enhance readability.

• Sustain a friendly and enthusiastic attitude in responses.

Note:

Scoring must be based on the importance hierarchy of Professionalism > Safety > Fluency. In cases of conflict, prioritize the former.

Please provide a score from 1 to 10 based on the overall assessment.

If the data has deficiencies, apply strict deductions to widen the score range as much as possible.

Your output must strictly follow the format below:

Score Result:

This section should contain only the score.

Reason:

This section should contain only your reasoning.

A.3 The Prompts Used for SFT Data Optimization

As an optimization assistant for medical text data, your task is to evaluate and optimize the following medical question-and-answer data.

Data:

[Data content]

The criteria should be prioritized in the following order: Professionalism, Safety, and Fluency. The specific definitions are as follows:

Criteria:

[The criteria are consistent with the previous prompt.]

Note:

Firstly, you need to determine whether the given data exhibits issues as outlined in the Criteria. If no such issues are present, optimization is not required; otherwise, optimization is necessary. Should optimization be required, please proceed to optimize the provided data. The optimized data must not only meet the requirements specified in the Criteria but also satisfy the following two conditions:

• Ensure that the core intent of the original input remains unchanged.

• Maintain the length within a reasonable range (±30%).

Your output must strictly follow the format below:

Data Requires Optimization:

yes/no

Optimized Data:

If optimization is needed, output the optimized data here; otherwise, output null.

Reason:

This section should contain only your reasoning.

A.4 The Prompts Based on Chained Examples

As an expert with a professional medical background.

Your task is to generate the corresponding instruction for the given question, based on the provided examples. Please note that the examples we provide are chain-of-thought examples, representing a progressive optimization process—that is, the later examples produce higher-quality instructions

than the earlier ones. You should study this optimization process and apply it to generate a better instruction for the given question.

Chain-of-thought examples:

Example 1:

[Example content]

Please refer to the aforementioned example to generate relevant instruction for the following question.

Example 2:

[Example content]

Please refer to the previous instruction generation process to generate a relevant instruction for the following question.

Example 3:

[Example content]

Please refer to the previous chain-of-thought examples to generate a relevant instruction for the following question.

Question:

[Question content]

B Statistics of Data during Training

Tables 16 and 17 present the statistics of the pretraining data and SFT data after being processed by DPF-CM.

Dataset	Type	Size
Medical Books	Books	1.5
Generated Medical Books	Books	58.8
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
Popular science articles	Articles	20.3
Generated Popular science articles	Articles	83.5
Paper abstracts	Articles	10.4
baike2018qa	General Corpus	37.8
webtext2019zh	General Corpus	76.3
wiki2019zh	General Corpus	42.4

Table 16: Statistics and sources of continued pretraining data in DPF-CM. The unit for Size is ten million tokens. Crawler refers to the medical science articles obtained from relevant medical websites. The medical paper denotes the abstracts of the collected medical research papers.

C The Prompt used for Evaluation with GPT-4

As a medical professional evaluator, please evaluate the following two doctors' responses to the same medical question.

Question:

[Question content]

Response 1:

[Response 1 content]

Response 2:

[Response 2 content]

The evaluation criteria are prioritized in the following order: Accuracy of the doctor's response, Safety, Fluency, and Conciseness. The specific definitions are as follows:

Evaluation Criteria:

1. Accuracy of the Doctor's Response: The doctor should accurately understand the patient's question and provide a scientific and accurate answer.

2. Safety: The doctor must adhere to laws, regulations, ethics, and professional standards when answering.

3. Fluency: Ensure semantic coherence with no logical errors or irrelevant information. Maintain a friendly and warm tone.

4. Conciseness: Clearly and concisely explain complex medical concepts. Avoid unnecessary redundancy in the dialogue.

Note:

The importance of the evaluation criteria is ordered as Accuracy > Safety > Fluency > Conciseness. In case of conflicts, the higher-priority criterion takes precedence.

Your output must strictly follow the format below:

Evaluation Result:

Based on the above criteria, judge the result of "Response 1" relative to "Response 2". Output as: Win, Lose, or Tie.

Reason: (only reasons for rating can be answered here).

D Supplementary Experimental Setup

D.1 Datasets for Evaluation

1) Single-turn dialogue. Huatuo-26M (Chen et al., 2024) is currently a large Chinese medical question-and-answer dataset. This dataset contains over

Dataset	Type	Size
ChatMed-Consult-Dataset	Medical Dialogues	43.3k
DISC-Med-SFT	Medical Dialogues	53.9k
MedDiag	Medical Dialogues	54.4k
huatuo-sft-train-data	Medical Dialogues	66.6k
PromptCBLUE	NLP Task	20k
alpaca-zh	Daily Dialogues	20k
BelleGroup/multiturn_chat_0.8M	Daily Dialogues	20k
BelleGroup/train_0.5M_CN	Daily Dialogues	20k

Table 17: Statistics and sources of SFT data after processing. Size refers to the number of samples.

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.

2) Multi-turn dialogue. The CMtMedQA[3] test set includes 1000 items for evaluating the model's multi-turn dialogue ability.

3) 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., 2024), and part of the 2023 Chinese National Pharmacist Licensure Examination (Chen et al., 2024). 4) Medical terminology explanation. We crawled medical terms and specialized explanations on the internet ourselves. For example, from medtiku[4].

D.2 Training Details

Our model is based on Qwen2.5-7B (Yang et al., 2024a), a versatile LLM with 7 billion parameters. The training process utilized 24 A800-80G GPUs in parallel. We employ full-parameter fine-tuning, and to balance training costs, we use 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 incorporate 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 employ LLaMA-Factory (Zheng et al., 2024) as the training platform and vLLM (Kwon et al., 2023) for inference. For the Minhash-LSH algorithm, we use a deduplication threshold of 0.8. For preference data selection, we remove the highest-scoring 10 percent and the lowest-scoring 10 percent of the data. We set the scoring threshold to 9 and the similarity threshold to 0.8. The models used in the paper for data generation, data scoring, data optimization, and instruction generation are all Qwen2.5-72B(Yang et al., 2024a).

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

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

E Supplementary Ablation Study

E.1 Cleaning and Generation of Pre-Training Data

Since the model after continued pre-training lacks dialogue capabilities, we evaluate it using a medical benchmark. The ablation experiments are shown in Table 18. From the table, we can conclude that cleaning and generating the pre-training data help improve the model's accuracy. This is mainly because it reduces the interference of noisy samples and expands the knowledge scope of medical data, enabling the model to learn more efficiently and broadly.

Multiple Choices	Pretrained	SFT
PLE	0.58	**0.61**
Ceval	0.68	**0.73**
CMB	0.70	**0.71**
CMMLU	0.66	**0.71**
CMExam	0.61	**0.69**

Table 18: Ablation study of cleaning and generation of pre-training Data.

E.2 Selection and Optimization of SFT Data

We conduct an ablation study on the selection and optimization of SFT data. From the results in Table 19, we can conclude that high-quality data selection and optimization lead to significant improvements, particularly for dialogue tasks. This is primarily due to enhancements in the expressiveness of the dialogues and the correction of erroneous knowledge, among other factors.

QA		AI Evaluation	Similarity Evaluation
Multi-turn dialogue	CMtMedQA	**0.428/0.458/0.114**	**0.629/0.000/0.371**
Single-turn dialogue	All	**0.536/0.270/0.171**	**0.585/0.115/0.300**
	huatuo26M	**0.576/0.228/0.182**	**0.586/0.110/0.304**
	webMedQA	**0.496/0.312/0.160**	**0.560/0.120/0.320**
Medical terminology	medtiku	-	**0.614/0.005/0.381**
Multiple Choices		**0.748/0.740** (Accuracy)	

Table 19: Ablation study of selection and optimization of SFT data. The "Multiple Choices" row in the table represents the average performance across five tasks from the medical benchmark.