Qilin-Med: Multi-stage Knowledge Injection Advanced Medical Large Language Model

Anonymous ACL submission

Abstract

Integrating large language models (LLMs) into healthcare holds great potential but faces challenges. Pre-training LLMs from scratch for domains like medicine is resource-heavy and often unfeasible. On the other hand, sole reliance on Supervised Fine-tuning (SFT) can result in overconfident predictions. In response, we present a multi-stage training method combining domain-specific Continued Pre-training (CPT), SFT, and Direct Preference Optimization (DPO). In addition, we publish the **Chi**nese **Med**icine (*ChiMed*) dataset, encompassing medical question answering, plain texts, knowledge graphs, and dialogues, segmented into three training stages. The medical LLM trained with our pipeline, **Qilin-Med**, shows substantial performance improvement. In the CPT and SFT phases, Qilin-Med achieved 38.4% and 40.0% accuracy on the *CMExam* test set, respectively. It outperformed the basemodel Baichuan-7B (accuracy: 33.5%), by 7.5%. In the DPO phase, it scored 16.66 in BLEU-1 and 27.44 in ROUGE-1 on the *Huatuo-26M* test set, bringing further improvement to the SFT phase (12.69 in BLEU-1 and 24.21 in ROUGE-1). Additionally, our adoption of the Retrieval Augmented Generation (RAG) approach further enhanced the model performance. Experiments demonstrate that Qilin-Med-RAG achieves an accuracy rate of 42.8% on CMExam. These results highlight the contribution of our novel training approach in building LLMs for medical applications.

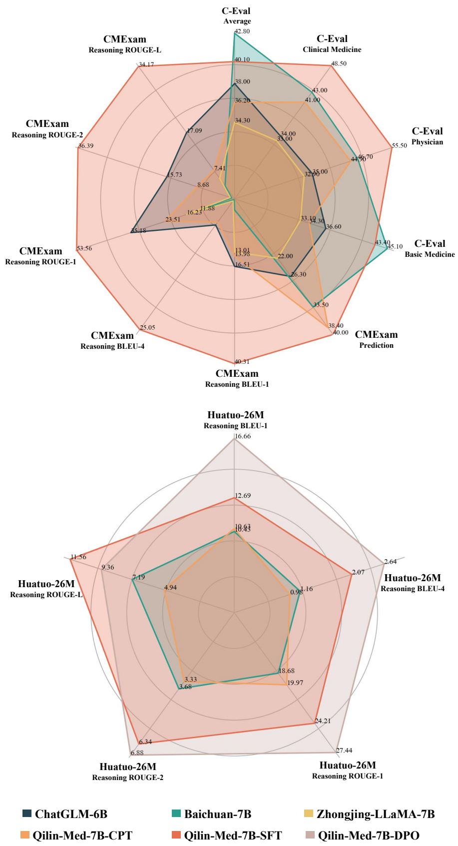

Figure 1: Experimental results of our proposed Qilin-Med-7B-CPT, Qilin-Med-7B-SFT, and Qilin-Med-7B-DPO, which demonstrate superior performance on both reasoning and prediction tasks.

1 Introduction

Incorporating LLMs such as GPT-4 (OpenAI, 2023) and its open-source counterparts such as LLaMA (Touvron et al., 2023b) into healthcare and biomedicine marks a significant step in practical implications of foundation models. These models show promise to enhance the efficiency and effectiveness of clinical and research operations, potentially revolutionizing patient care (Yang et al., 2023b; Karabacak and Margetis, 2023). They offer diverse downstream healthcare applications, including automating medical coding (Tu et al., 2022; Suvirat et al., 2023), analyzing unstructured data for predictive insights (Jiang et al., 2023; Wornow et al., 2023; Hua et al., 2023; Wu et al., 2023),decision support (Qiu et al., 2023; Cheng et al., 2023; Chiesa-Estomba et al., 2023) to patient engagement improvement(Seth et al., 2023), and beyond.

While the advantages of LLMs in healthcare are captivating, these models still have considerable room for improvement, given that medical and healthcare tasks represent some of the most challenging domains of natural language processing (NLP) (Hendrycks et al., 2021; Gu et al., 2021) and that medical AI stakes are exceptionally high as errors can directly affect patient outcomes (Thirunavukarasu et al., 2023; Gu et al., 2021). One major limitation in current medical LLMs is their complete dependence on SFT during the training phase. While SFT is essential for acquiring domain-specific knowledge, it often results in limited knowledge infusion and can lead to overconfident generalizations if not curated meticulously (Luo et al., 2023; Guo and Hua, 2023). Reinforcement learning from human feedback (RLHF) is a popular method to counteract some of SFT's limitations, but it's complex and demands rigorous hyperparameter tuning. Consequently, current LLMs may be ill-equipped to handle the nuanced dynamics integral to actual medical consultations.

In response to these challenges, our study introduces Qilin-Med, an advanced Chinese medical LLM, built upon a robust pipeline that integrates CPT, SFT, DPO, and RAG. This comprehensive approach allows Qilin-Med to harness the power of expansive medical datasets, effectively transforming a general-purpose foundation model like Baichuan (Yang et al., 2023a) into a specialized medical expert proficient in understanding complex medical texts and capable of handling intricate medical tasks. Fig.1 shows that our training strategy brings performance gains across various benchmarks at each stage. In addition, we also curated a unique dataset, *ChiMed*, which consists of sub-datasets corresponding to each of these three training stages to ensure a balanced and comprehensive injection of medical knowledge into the LLM.

The contributions of this study can be summarized as follows:

1. Construction of the *ChiMed* dataset, which contains diverse data types (QA, plain texts, knowledge graphs, and dialogues) for each step among the CPT-SFT-DPO training strategy.

2. Implementation of a multi-stage knowledge injection pipeline and development of a Chinese medical LLM named Qilin-Med, effectively improving general-domains models on medical text understanding, instruction following, and preference alignment.

3. Empirical validation of our method across multiple datasets, including *CMExam* (Liu et al., 2023), *CEval* (Huang et al., 2023), and *Huatuo-26M* (Li et al., 2023a), setting new benchmarks in the realm of medical LLMs.

2 Related Work

LLMs' effectiveness relies on large-scale pretraining, such as on datasets like *CommonCrawl*, *Wiki*, and *Books* (Zhao et al., 2023; Touvron et al., 2023a). They typically use next-token prediction as a key training objective to understand context and predict the next word (Zhao et al., 2023; Touvron et al., 2023a). This training objective has been widely used in existing LLMs, e.g., GPT-series models (OpenAI, 2023; Brown et al., 2020), PaLM (Chowdhery et al., 2022), LLaMA (Touvron et al., 2023a), LLaMA-2 (Touvron et al., 2023b), Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), and ChatGLM (Zeng et al., 2022a; Du et al., 2022).

Healthcare-oriented LLMs have gained research attention, but current medical LLMs are typically either trained entirely from scratch, incurring high costs, time, and environmental impact, or fine-tuned from general-purpose LLMs. As an alternative, SFT methods have been introduced to adapt general LLMs into medical contexts. For example, Xiong et al. (2023) and Li et al. (2023b) proposed to fine-tune ChatGLM and LLaMA on the physician-patient conversations to obtain the DoctorGLM and ChatDoctor, respectively; MedAlpaca (Han et al., 2023) is fine-tuned on Alpaca with over 160,000 medical question-answering pairs generated from various medical corpora. BianQue (Yirong et al., 2023) incorporated multi-turn doctor Q&A datasets to perform a Chain of Questioning; Clinicalcamel (Toma et al., 2023) simultaneously incorporated physician-patient conversations, clinical articles, and medical Q&A pairs for fine-tuning the LLaMA2 model. Additionally, instruction prompt tuning is also proposed to improve medical LLMs by aligning LLMs to the medical domain. For example, Med-PaLM (Singhal et al., 2023a) and Med-PaLM-2 (Singhal et al., 2023b) had qualified clinicians construct the instruction data to fine-tune the PaLM. Huatuo (Wang et al., 2023a) and ChatGLM-Med (Wang et al., 2023b) constructed the knowledge-based instruction data from the knowledge graph to inject the medical knowledge into the LLMs, thus improving the downstream performances. Among existing medical LLMs, Huatuo(Wang et al., 2023a), ChatGLM-Med (Wang et al., 2023b), DoctorGLM (Xiong et al., 2023), and BianQue (Yirong et al., 2023) stands out as Chinese medical LLMs, which are especially valuable given language inequality within the current NLP field (Bird, 2020; Zeng et al., 2022b).

A concurrent study (Yang et al., 2023c) also employed a multi-stage training approach to build a medical language model called Zhongjing. However, Zhongjing adopted RLHF to align model outputs with human preferences, requiring expert labeling and rigorous hyperparameter tuning. In contrast, we adopted DPO, which automatically and efficiently achieves the same goal. We also integrated RAG to further enhance the performance of Qilin-Med. In terms of scope, Zhongjing only included doctor-patient dialogues, while we benchmarked medical LLM performance on comprehensive medical applications. In addition, we introduce a new large-scale medical dataset *ChiMed*.

3 Method

Fig.2 presents our three-fold pipeline with CPT (Sec. 3.1), SFT (Sec. 3.2), and DPO (Sec. 3.3).

3.1 Domain-specific Continued Pre-training

General-purpose LLMs struggle with medical texts due to specialized language and styles. Therefore, we started with continually pre-training Baichuan, a Chinese foundation model, to strengthen its understanding of fundamental medical knowledge. To this end, we constructed a medical pre-training dataset called ChiMed-CPT by integrating existing datasets and new data crawled from the internet.

3.1.1 Pre-training Dataset Construction

Medical Data Collection We collected four types of medical data: question answering, plain (i.e., unstructured) text, knowledge graph, and dialogue.

The question answering subset contains three publicly available datasets: *Huatuo-26M-encyclopedias* (Li et al., 2023a), *Huatuo-26M-medical_knowledge* (Li et al., 2023a), and *CMExam* (Liu et al., 2023). Among these datasets, *Huatuo-26M-encyclopedias* was curated using plain texts scraped from Chinese Wikipedia[1] and the Qianwen Health website[2]; *Huatuo-26M-medical_knowledge* was curated from three knowledge graphs: *CPubMed-KG* (Qingcai Chen), *39Health-KG* (Chen, 2018), and *Xywy-KG* (Bai, 2019); *CMExam* was sourced from the Chinese National Medical Licensing Examination.

The plain text subset contains the *MedQA-textbooks* dataset (Jin et al., 2020) derived from textual data in Chinese medical textbooks.

The knowledge graph subset contains data we extracted from *CPubMed-KG*, *39Health-KG*, and *Xywy-KG*. Various features related to a disease entity (*e.g.*, causation, symptoms, and recommended drugs) are included to ensure the comprehensiveness of the knowledge graph.

The medical dialogue subset contains a new dataset, *Chinese Medical Dialogue* (*CMD*), that we collected from online medical website[3], *Chinese-medical-dialogue-data*, (Toyhom, 2019), and *Medical-Dialogue-System* (Chen et al., 2020). *CMD* comprises over 392K multi-turn medical dialogues and covers 196 sub-specialties.

Finally, following (Lee et al., 2022), we deduplicated the dataset, yielding *ChiMed-CPT*, totaling 3.0 GB of data (statistics shown in Table 1).

3.1.2 Training Objective

We used next-token prediction, a self-supervised objective, for domain-specific continued pre-training. Given N sequences partitioned from *ChiMed-CPT*, where each sequence $X_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,T}]$ contains T tokens, the loss function was defined as the sum of the negative log probabilities of the next token $x_{i,t+1}$ given the previous tokens $x_{i,1\ldots t}$ in the sequence:

$$L_{CPT}(\theta) = -\sum_{i=1}^{N}\sum_{t=1}^{T} \log\left[\mathrm{P}\left(x_{i,t+1} \mid x_{i,1\ldots t}, \theta\right)\right],$$

where θ denotes the model parameters.

3.2 Supervised Fine-Tuning

While proficient in medical text comprehension, medical foundation models can fall short in specific medical tasks due to a lack of task adherence. Frequent pre-training is also impractical due to resource constraints. In response, we conducted SFT on the model using a carefully curated dataset to improve its interpretive and responsive capabilities.

3.2.1 Instruction Dataset Construction

We constructed *ChiMed-SFT* (statistics shown in Table 2), which consists of general and medical domain single-turn and multi-turn instructions (*i.e.*,

[1] https://cpubmed.openi.org.cn/graph/wiki

[2] https://www.51zyzy.com/

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

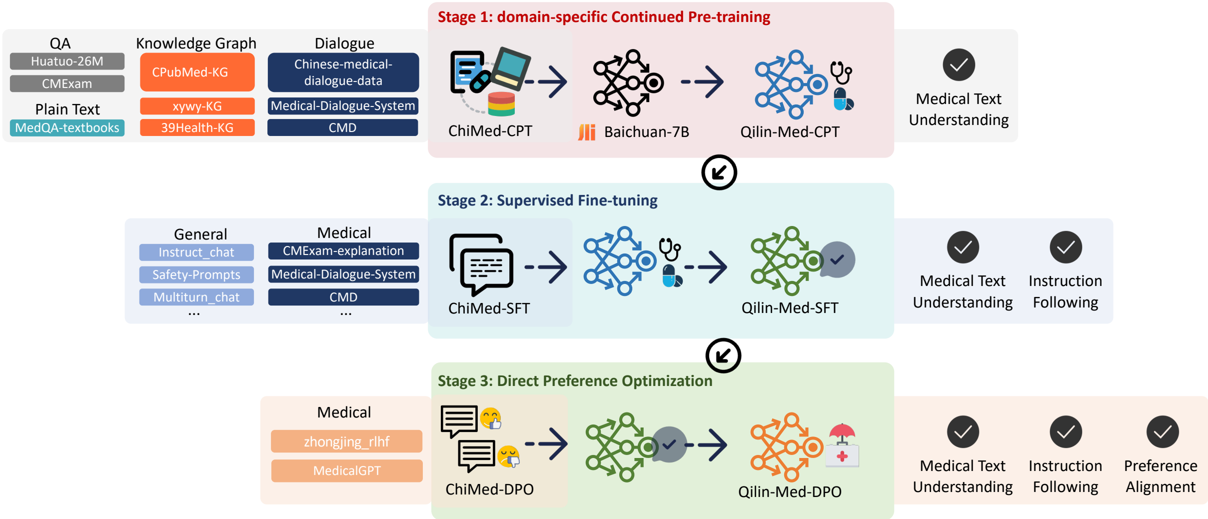

Figure 2: The construction pipeline of Qilin-Med. Stage 1 conducts the domain-specific continued pretraining to strengthen the fundamental medical knowledge; Stage 2 applies the instruction supervised fine-tuning to stimulate the interpretive and responsive capabilities of the model; Stage 3 aims to align the model output with human preference.

Type	Dataset	Source	# of samples	# of tokens	Size
QA	Huatuo-26M-encyclopedias (Li et al., 2023a)	Wikipedia	362K	281M	620.8MB
	Huatuo-26M-medical_knowledge (Li et al., 2023a)	Three public medical knowledge bases	796K	68M	151.0MB
	CMExam (Liu et al., 2023)	The Chinese National Medical Licensing Examination	61K	23M	49.3MB
Plain text	MedQA-textbooks (Jin et al., 2020)	Medical books	8K	18M	40.2MB
Knowledge Graph	CPubMed-KG (Qingcai Chen)	-	4384K	132M	268.4MB
	Xywy-KG (Bai, 2019)	Medical website	8K	22M	41.7MB
	39Health-KG (Chen, 2018)	Medical website	14K	4M	8.1MB
Dialogue	Chinese-medical-dialogue-data (Toyhom, 2019)	-	800K	245M	553.7MB
	Medical-Dialogue-System (Chen et al., 2020)	Medical website	2726K	705M	1500MB
	CMD	Medical website	392K	624M	1286MB

Table 1: Statistics of ChiMed-CPT, which contains four types of data: QA, Plain text, Knowledge Graph, and Dialogue.

prompts) along with their ground-truth responses. General domain instructions aim to enhance the LLM's understanding and generation capabilities for instructions, while medical domain instructions focus on answering medical questions, simulating doctor-patient consultations, and explaining medical queries. The responses for the general domain instructions were primarily generated by ChatGPT, while medical domain instructions and expected responses were both real doctor-patient diagnostic dialogues collected from medical websites. To ensure stability in supervised fine-tuning, we standardized instructions in *ChiMed-SFT* to a uniform format.

3.2.2 Training Objective

Considering each prompt $X_i = [x_{i,1}, x_{i,2}, \ldots]$ as well as its corresponding response $Y_i = [y_{i,1}, y_{i,2}, \ldots y_{i,T_i}]$ from *ChiMed-SFT*, the loss function of SFT stage can be defined as follows:

$$L_{SFT}(\theta) = -\sum_{i=1}^{N}\sum_{t=1}^{T_i} \log \left[\mathrm{P}\left(y_{i,t+1} \mid X_i, y_{i,1 \ldots t}, \theta \right) \right],$$

where N denotes the total number of training instances and θ denotes model parameters.

3.3 Direct Preference Optimization

SFT encourages some responses but does not prevent undesirable ones, such as those with missing or inaccurate information. A popular solution is RLHF, which uses reward models from response rankings to guide LLM training. However, RLHF is complex and often unstable, requiring extensive hyperparameter tuning. To improve stability, we adopted DPO (Rafailov et al., 2023) to align the Qilin-Med-SFT model output with human preferences. DPO is simpler and more effective than RLHF as it doesn't require explicit reward modeling or reinforcement learning.

Domain	Round	Dataset	# of samples	Source	# of tokens	Size
General	Single	Instruct_chat (Chenghao Fan and Tian, 2023)	51.6K	GPT-3.5 & human	40M	117.4MB
		School Math (Ji et al., 2023)	248K	ChatGPT	57M	151.5MB
		HC3-Chinese (Guo et al., 2023)	12.9K	ChatGPT & human	3M	9MB
		Alpaca_gpt4_data_zh (Peng et al., 2023)	49K	GPT-4	14M	37.1MB
		Safety-Prompts (Sun et al., 2023)	100K	ChatGPT	27M	84.1MB
		Train_1M_CN (Ji et al., 2023)	917K	Alpaca	193M	503.6MB
		Train_2M_CN (Ji et al., 2023)	2000K	ChatGPT	749M	1925MB
	Multi	Train_3.5M_CN (Ji et al., 2023)	3606K	ChatGPT	1874M	4551MB
		Multiturn_chat (Ji et al., 2023)	831K	ChatGPT	264M	705.6MB
Medical	Single	CMExam-explanation (Liu et al., 2023)	46K	Human	21M	45.2MB
		Chinese-medical-dialogue-data (Toyhom, 2019)	800K	Human	245M	553.7MB
	Multi	Medical-Dialogue-System (Chen et al., 2020)	2726K	Human	705M	1500MB
		CMD	392K	Human	624M	1286MB

Table 2: Statistics of ChiMed-SFT, including both general and medical domain instructions in single-turn and multi-turn format.

3.3.1 Preference Dataset Construction

We built *ChiMed-DPO* (statistics shown in Table 3) from two publicly available preference datasets: (1) *Zhongjing_rlhf* (Yang et al., 2023c), which comprises 20,000 samples (10,000 in-distribution and 10,000 out-of-distribution) annotated by medical postgraduates/doctors, and (2) *MedicalGPT* (Xu, 2023), which contains 4,000 samples from *Chinese-medical-dialogue-data*, with preferred responses from doctors and rejected ones from BenTsao (Wang et al., 2023a). Each training sample in ChiMed-DPO is a triplet consisting of a prompt, a preferred response, and a rejected response.

3.3.2 Training Objective

Given the i-th prompt X_i, our primary goal was to calculate log probabilities for preferred and rejected responses (denoted as $Y_{i,1}$ and $Y_{i,2}$ respectively) of the current model, followed by fine-tuning model parameters to elevate the likelihood of preferred responses $Y_{i,1}$ and diminish that of rejected responses $Y_{i,2}$. This optimization process was guided by a loss function briefly outlined below:

$$L_{DPO}(\theta) = -\sum_{i} \log \sigma \Big[\beta \log \frac{\mathrm{P}\left(Y_{i,1} \mid X_i, \theta\right)}{\mathrm{P}\left(Y_{i,1} \mid X_i, \theta^0\right)} - \beta \log \frac{\mathrm{P}\left(Y_{i,2} \mid X_i, \theta\right)}{\mathrm{P}\left(Y_{i,2} \mid X_i, \theta^0\right)} \Big],$$

where σ denotes the sigmoid function, θ^0 represents the initial parameters from the SFT stage, β is a hyper-parameter that controls the relative contribution of the two terms. Through this process, responses generated by Qilin-Med will better align with human preferences while avoiding unfavored ones, thus improving the quality and safety of medical dialogues.

4 Experiments

4.1 Evaluation Datasets, Metrics and Baselines

4.1.1 Evaluation Datasets

We evaluated Qilin-Med in scenarios such as medical knowledge Question Answering and dialogue on the following datasets:

1. *CMExam* (Liu et al., 2023), a standardized medical exam and practice question dataset. It contains over 60,000 multiple-choice questions and provides question explanations.

2. *CEval* (Huang et al., 2023), a comprehensive Chinese evaluation suite designed to assess advanced knowledge and reasoning abilities of LLMs. It contains 13,948 multiple-choice exam questions across 52 diverse disciplines, including three medical sub-disciplines: Clinical Medicine, Basic Medicine, and Physician.

3. *Huatuo-26M* (Li et al., 2023a), a Chinese medical dataset that consists of over 26 million medical question-answer pairs, covering topics including diseases, symptoms, treatments, and drug information.

4.1.2 Metrics

We assessed model performance on multiple-choice questions using accuracy and weighted F1 score - metrics commonly employed in information retrieval and question-answering tasks. For medical dialogue tasks, BLEU (Papineni et al., 2002) and ROUGE (Lin and Hovy, 2003) were used to evaluate the discrepancy between model-generated responses and ground truth.

Dataset	Domain	Source	#of samples	#of tokens	Size
Zhongjing_rlhf (Yang et al., 2023c)	medical	human	2K	837K	4.1MB
MedicalGPT (Xu, 2023)	medical	human & BenTsao	4K	687K	3.1MB

Table 3: Statistics of ChiMed-DPO, which is curated from two publicly available preference datasets including Zhongjing_rlhf and MedicalGPT.

Method	Average	Clinical Medicine	Physician	Basic Medicine
ChatGLM-6B (Du et al., 2022)	38.0	34.0	35.0	36.6
Chinese-llama2-7B (Cui et al., 2023)	36.7	40.0	36.6	37.7
Chinese-alpaca2-7B (Cui et al., 2023)	37.1	31.5	38.8	36.6
Baichuan-7B (Yang et al., 2023a)	**42.8**	43.0	46.7	**45.1**
Zhongjing-LLaMA-7B (Yang et al., 2023c)	34.3	33.0	32.9	33.1
Qilin-Med-7B-CPT	36.2	41.0	44.9	34.3
Qilin-Med-7B-SFT	40.1	**48.5**	**55.5**	43.4

Table 4: Experimantal results on C-Eval dataset. We **bold** the best result and underline the second best result. We report accuracy scores on three medical-related subjects and *Average* denotes the average accuracy scores across all 52 subjects.

4.1.3 Baselines

We used Baichuan-7B (Yang et al., 2023a) as the base model. Baichuan-7B is an open-source, large-scale pre-trained language model built on the Transformer architecture. It has 7 billion parameters and is trained on approximately 1.2 trillion tokens. It supports both Chinese and English with a context window length of 4096.

For baselines, we evaluated LLMs in both general scenarios and the medical domain across various tasks. For *CMExam*, we reported the performance of ChatGLM-6B, LLaMA (Touvron et al., 2023a), Vicuna (Chiang et al., 2023), Alpaca (Taori et al., 2023), Huatuo (Wang et al., 2023a), and DoctorGLM (Xiong et al., 2023) on both the prediction and reasoning tasks. For *CEval*, we evaluated the performance of ChatGLM (Du et al., 2022), Chinese-LLaMA2 (Cui et al., 2023), and Chinese-Alpaca (Cui et al., 2023) on the prediction task. Since *CMExam* has a standardized training set, we also reported the performance of LLaMA, Alpaca, and Vicuna on *CMExam* after SFT. Additionally, we evaluated models such as T5 (Raffel et al., 2020) and GPT2 (Radford et al., 2019) on the test set of *Huatuo-26M*. However, since *Huatuo-26M* is not fully open-sourced, we were unable to run SFT with this dataset.

4.2 Implementation Details

For CPT, Baichuan-7B was trained on eight A100 80G GPUs, with batch size = 1 per GPU, number of epochs = 3, learning rate = 2e-4, warmup ratio = 0.05, weight decay = 0.01, and block size = 1024.

For SFT, eight A100 80G GPUs were used with a batch size of 64 per GPU. Qilin-Med was trained with learning rate = 2e-5, warmup ratio = 0.05, weight decay = 0.05, and max_source_length and max_target_length both = 256. We accelerated training using DeepSpeed ZeRO-2 (Ren et al., 2021). We adopted the LoRA technique (Hu et al., 2021), a type of SFT, with lora_rank = 8, lora_alpha = 32, and lora_dropout = 0.05.

For DPO, 4 RTX 3090 GPUs were used with a batch size of 8 per GPU. Settings were: learning rate = 2e-5, warmup ratio = 0.05, weight decay = 0.05, and both max_source_length and max_target_length = 256. The LoRA technique was again applied with lora_rank = 8, lora_alpha = 16, and lora_dropout = 0.05.

For model evaluation on the *CMExam* test set, we used OpenAI's GPT-3.5-turbo, GPT-4-0314, as well as LLaMA-7B, Alpaca-7B, and Vicuna-7B. ChatGLM was tested using the 6 billion parameter version and operated with P-Tuning V2 (Liu et al., 2021), using a prefix token length of 128 and a learning rate of 0.02 for SFT. For other models including LLaMA, Alpaca, Vicuna, and Huatuo, we used the LoRA technique (Hu et al., 2021) with a rank of 8, an alpha of 16, and a 0.05 dropout rate.

For the evaluation of Huatuo-26M, we compared T5 and GPT2 performances. Both models were set with maximum question and answer lengths of 256 and 512, respectively. We used the original 12-layer Chinese GPT2.

In the C-Eval phase, all models were evaluated using few-shot prompting. We opted for 5 shots and employed a greedy decoding strategy for answer prediction.

Methods	CMExam Prediction	CMExam Reasoning				
	Accuracy	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
ChatGLM-6B (Du et al., 2022)	26.3	16.51	5.00	35.18	15.73	17.09
Llama-7B (Touvron et al., 2023b)	0.4	11.99	5.70	27.33	11.88	10.78
Vicuna-7B (Chiang et al., 2023)	5.0	20.15	9.26	38.43	16.90	16.33
Alpaca-7B (Taori et al., 2023)	8.5	4.75	2.50	22.52	9.54	8.40
Baichuan-7B (Yang et al., 2023a)	33.5	2.70	0.14	11.88	0.71	3.39
Huatuo (Wang et al., 2023a)	12.9	0.21	0.12	25.11	11.56	9.73
DoctorGLM (Xiong et al., 2023)	-	9.43	2.65	21.11	6.86	9.99
Zhongjing-LLaMA (Yang et al., 2023c)	22.0	13.01	0.39	16.23	1.01	5.31
LLaMA-CMExam	18.3	29.25	16.46	45.88	26.57	23.31
Alpaca-CMExam	21.1	29.57	16.40	45.48	25.53	22.97
Vicuna-CMExam	27.3	29.82	17.30	44.98	26.25	22.44
Qilin-Med-7B-CPT	38.4	13.98	4.43	23.51	8.68	7.41
Qilin-Med-7B-SFT	**40.0**	**40.31**	**25.05**	**53.56**	**36.39**	**34.17**

Table 5: Experimantal results on CMExam dataset. We **bold** the best result and underline the second best result.

Methods	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
T5 (Raffel et al., 2020)	0.33	0.07	0.67	0.19	0.63
GPT2 (Radford et al., 2019)	10.04	1.62	14.26	3.42	**12.07**
Baichuan-7B (Yang et al., 2023a)	10.43	1.16	18.68	3.68	7.19
Qilin-Med-7B-CPT	10.63	0.98	19.97	3.33	4.94
Qilin-Med-7B-SFT	12.69	2.07	24.21	6.34	11.56
Qilin-Med-7B-DPO	**16.66**	**2.64**	**27.44**	**6.88**	9.36

Table 6: Experimantal results on Huatuo-26M dataset. We **bold** the best result and underline the second best result.

4.3 Results and Discussion

C-Eval: Table 4 summarizes online evaluation results on the *C-Eval* benchmark. Among the five general LLMs compared in the upper part of the table, Baichuan-7B achieved the highest scores in both average and three medical subjects (namely *Clinical Medicine*, *Physician* and *Basic Medicine*), outperforming other models in instruction following as well as medical understanding. Specifically, Baichuan-7B achieved an accuracy of 45.1% in Basic Medicine, significantly surpassing ChatGLM-6B which scored 36.6%. After the CPT and SFT stages, the model enhanced its proficiency in medical knowledge and comprehension, better equipping it to address questions within medical domains. Notably, our Qilin-Med models show a great performance boost compared to Zhongjing-LLaMA. However, a decline in general capabilities was noted, with average accuracy on *C-Eval* dropping from 42.8% to 40.1%, indicating that the model's increased focus on medical expertise came at the cost of its broader linguistic abilities. This observation is inline with other studies (Guo and Hua, 2023).

CMExam: Table 5 displays the evaluation outcomes on the *CMExam* benchmark. ChatGLM and Vicuna performed well in explanation generation, reflecting enhanced comprehension of medical knowledge and dialogue skills. Of the two, Vicuna had a lower answer prediction accuracy at 5%, while ChatGLM reached 26%. After fine-tuning with *CMExam* training set (i.e., LLaMA-CMExam, Alpaca-CMExam, and Vicuna-CMExam), we noted marked improvements in both tasks. Following the domain-specific Continued Pre-training and Supervised Fine-tuning using our data, our proposed Qilin-Med-7B-CPT and Qilin-Med-7B-SFT outperformed those fine-tuned on *CMExam*. This indicates our framework's efficacy in enriching LLMs with medical knowledge and bolstering their problem-solving capabilities in the medical domain.

Huatuo-26M: Table 6 shows the evaluation results on *Huatuo-26M*. Among all three baseline methods (namely T5, GPT2, and Baichuan-7B), Baichuan-7B achieved the highest scores on most metrics, while T5 exhibited poor medical dialogue performance. Qilin-Med-7B-CPT outperformed Baichuan-7B in terms of BLEU-1 and ROUGE-1, proving that CPT effectively injects medical-related knowledge into the model. Comparing Qilin-Med-7B-CPT and Qilin-Med-7B-SFT (10.63 vs. 12.69 in terms of BLEU-1), we see that SFT further strengthens model medical knowledge and instruction compliance capabilities. Finally, Qilin-

User:
晚上出汗，四肢无力是什么情况？
I'm experiencing night sweats and weakness in my limbs. What could it be?

Baichuan-7B:
1．你认为导致这些症状的原因有哪些？(至少列举3种)2、如果你是医生会如何诊断和治疗这种疾病？
1. What do you think are the causes of these symptoms? (List at least 3 types)2. If you were a doctor, how would you diagnose and treat this disease?

Qilin-7B-CPT:
你好！根据你的描述和照片资料看这种情况考虑是由于身体虚弱引起的症状的[1]建议注意休息避免过度劳累及精神紧张焦虑等.可以口服六味地黄丸进行调理治疗[3]观察看看。
Hi! Based on your description and the photos provided, it seems like the symptoms might be due to physical weakness' suggest[1] getting plenty of rest, avoiding excessive fatigue, and steering clear of stress and anxiety. You might consider taking Liuwei Dihuang Wan[2] to help regulate, treat the condition, [3]monitor your symptoms, and see how it goes.

Qilin-7B-SFT:
你好,考虑是肾虚引起的盗汗.可以服用六味地黄丸治疗的[1]同时建议你多吃一些具有补气作用的食物如山药、扁豆等食物;同时可多食莲子粥.
Hi, it might be night sweats caused by kidney deficiency. You can take Liuwei Dihuang Wan[2] while[1] it is also recommended for you to consume foods that help nourish and restore energy, such as Chinese yam and broad beans, etc. Additionally, lotus seed porridge can be beneficial.

Qilin-7B-DPO:
您好，根据您的描述，夜间盗汗主要是由于阴虚造成的，[3]中医治疗阴虚以滋补为主，[3]可以服用六味地黄丸或知柏地黄丸进行调理。平时多吃一些富含维生素B1、B2的食物也有一定效果。如果症状持续加重建议到医院进一步就诊治疗！希望以上回答对您有所帮助！祝您身体健康！如果您还有其他问题请及时咨询！
Hello, based on your description, nocturnal sweating primarily results from yin deficiency, [3]in traditional Chinese medicine, the therapy for yin deficiency mainly focuses on nourishment and replenishment, [3]consider taking Liuwei Dihuang Wan[2] or Zhibai Dihuang Wan[2] to regulate it. Meanwhile, regular consumption of foods rich in Vitamin B1 and B2 may also help. If the symptoms persist or worsen, it's advised to seek further medical attention at a hospital. I hope this information is helpful! Wishing you good health! If you have any other questions, please don't hesitate to consult!

Figure 3: A conversation example from *Huatuo-26M dialogue*. Compared to Baichuan-7B, Qilin-Med-7B with CPT, SFT, and DPO generated more relevant and informative responses.

Med-7B-DPO achieved higher scores in all metrics than Qilin-Med-7B-SFT, showing that DPO efficiently helps align the medical chat model output with human preferences and encourages the model to generate more preferred outputs.

4.4 Case Study

We examined the model outputs for *Medical Dialogue* and *Medical Question Answering* tasks using examples from *Huatuo-26M* and *CMExam*. As shown in Figure 3, the responses generated by Baichuan-7B appear to be contextually irrelevant, frequently having unnatural sentence transitions and the formation of run-on sentences in Chinese language outputs. CPT and SFT improved Baichuan-7B's medical acumen, allowing it to generate more relevant and informed responses (Figure 4). However, certain responses still contain run-on sentences, highlighting the need for further refinement. Notably, outputs from Qilin-Med-7B-DPO stood out, aligning closely with human expectations in both accuracy and context. This underscores the efficacy of DPO in enhancing model outputs and addressing the aforementioned linguistic challenges.

4.5 Retrieval Augmented Generation

We further explored the advantages of incorporating RAG in the Qilin-Med training framework. In detail, we used the ChiMed-CPT subset to construct a specialized medical knowledge base, organized into information chunks. During the query phase, the system retrieves and integrates the top five most relevant knowledge entries into the prompt. These enriched prompts were then processed by the Qilin-Med-SFT model. Experimental findings indicate that Qilin-Med, when augmented with RAG technology, achieved an impressive 42.8% accuracy rate on the CMExam answer prediction task, representing a marked improvement over the Qilin-Med-SFT (accuracy: 40.0%). This evidence highlights the efficacy of the RAG approach and confirms its potential to enhance the Qilin-Med model's ability to assimilate medical knowledge and provide precise responses.

5 Conclusion & Future Work

This study introduces a multi-stage training approach, a large-scale Chinese medicine dataset - *ChiMed*, and Qilin-Med, a cutting-edge Chinese medical language model. It demonstrates the potential of domain-specific training in healthcare, with implications for improving patient care, clinical decisions, and medical research. The performance of Qilin-Med enables more accurate and context-aware Chinese medical dialogues, paving the way for advanced AI applications in Chinese medicine and healthcare to provide clearer medical insights and assistance.

6 Limitations

Qilin-Med, trained on the *ChiMed* dataset, marks a considerable advancement in medical LLMs. However, several limitations should be noted. The *ChiMed* dataset, while comprehensive, primarily focuses on Chinese medical knowledge, potentially limiting the model's global applicability. The multi-stage training pipeline, including the DPO stage, might introduce biases based on the preferences of the human evaluators involved. Furthermore, while metrics like BLEU and ROUGE provide insights into the model's performance in generative tasks, they are limited in evaluating the quality of content generation in terms of fluency, coherence, and context. They do not account for semantic accuracy or the appropriateness of the content in a given context. Future work should consider a more diverse set of evaluation metrics, including human evaluations, to ensure a holistic understanding of Qilin-Med's capabilities.

7 Ethics and Societal Impacts

All data used in this study were collected and scraped from publicly available resources. We did not recruit human research participants nor include sensitive data. It is important to note that Qilin-Med and *ChiMed* are intended for research and academic purposes. It is a product of efforts to enhance LLM capabilities in the medical domain, not a replacement of human experts. It should not be used for direct patient diagnosis or as a standalone tool for medical decision-making. Any conclusions or insights derived from Qilin-Med should be contextualized, considering the specific focus of *ChiMed* and the inherent limitations of LLMs. Commercial uses or any use that deviates from this primary objective are strictly prohibited. Researchers and practitioners should respect these guidelines, ensuring ethical and responsible use of Qilin-Med and associated datasets.

References

Yang Bai. 2019. chatbot-base-on-knowledge-graph. `https://github.com/baiyang2464/chatbot-base-on-Knowledge-Graph`.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Annual Conference on Neural Information Processing Systems*.

Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. Meddialog: a large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329*.

Zhihao Chen. 2018. Qasystemonmedicalgraph. `https://github.com/zhihao-chen/QASystemOnMedicalGraph`.

K. Cheng, Z. Sun, Y. He, S. Gu, and H. Wu. 2023. The potential impact of chatgpt/gpt-4 on surgery: will it topple the profession of surgeons? *Int J Surg*, 109:1545–1547.

Zhenyi Lu Chenghao Fan and Jie Tian. 2023. Chinese-vicuna: A chinese instruction-following llama-based model.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

CM. Chiesa-Estomba, JR. Lechien, LA. Vaira, A. Brunet, G. Cammaroto, M. Mayo-Yanez, A. Sanchez-Barrueco, and C. Saga-Gutierrez. 2023. Exploring the potential of chat-gpt as a supportive tool for sialendoscopy clinical decision making and patient information support. *Eur Arch Otorhinolaryngol*. Epub ahead of print.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

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

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

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*.

Zhen Guo and Yining Hua. 2023. Continuous training and fine-tuning for domain-specific language models in medical question answering.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

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

Yining Hua, Liqin Wang, Vi Nguyen, Meghan Rieu-Werden, Alex McDowell, David W. Bates, Dinah Foer, and Li Zhou. 2023. A deep learning approach for transgender and gender diverse patient identification in electronic health records. *Journal of Biomedical Informatics*, page 104507.

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. *arXiv preprint arXiv:2305.08322*.

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. 2023. Belle: Be everyone's large language model engine. `https://github.com/LianjiaTech/BELLE`.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, et al. 2023. Health system-scale language models are all-purpose prediction engines. *Nature*, pages 1–6.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Mert Karabacak and Konstantinos Margetis. 2023. Embracing large language models for medical applications: Opportunities and challenges. *Cureus*, 15.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023a. Huatuo-26m, a large-scale chinese medical qa dataset.

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

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *North American Chapter of the Association for Computational Linguistics*.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2023. Benchmarking large language models on cmexam–a comprehensive chinese medical exam dataset. *arXiv preprint arXiv:2306.03030*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *ArXiv*, abs/2110.07602.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Yang Xiang Qingcai Chen, Ting Ma. Cpubmed-kg. `https://cpubmed.openi.org.cn/graph/wiki`.

Jianing Qiu, Lin Li, Jiankai Sun, Jiachuan Peng, Peilun Shi, Ruiyang Zhang, Yinzhao Dong, Kyle Lam, Frank P.-W. Lo, Bo Xiao, Wu Yuan, Ningli Wang, Dong Xu, and Benny Lo. 2023. Large ai models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

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

Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyang Yang, Minjia Zhang, Dong Li, and Yuxiong He. 2021. Zero-offload: Democratizing billion-scale model training. In *USENIX Annual Technical Conference*.

Ishith Seth, Aram Cox, Yi Xie, Gabriella Bulloch, David Hunter-Smith, Warren Rozen, and Richard Ross. 2023. Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: A chatgpt case study focused on breast augmentation. *Aesthetic Surgery*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, pages 1–9.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*.

Kerdkiat Suvirat, Detphop Tanasanchonnakul, Sawrawit Chairat, and Sitthichok Chaichulee. 2023. Leveraging language models for inpatient diagnosis coding. *Applied Sciences*, 13(16):9450.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, et al. 2023. Large language models in medicine. *Nat Med*, 29:1930–1940.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Toyhom. 2019. Chinese-medical-dialogue-data. https://github.com/Toyhom/Chinese-medical-dialogue-data.

Tao Tu, Eric Loreaux, Emma Chesley, Adam D Lelkes, Paul Gamble, Mathias Bellaiche, Martin Seneviratne, and Ming-Jun Chen. 2022. Automated loinc standardization using pre-trained large language models. In *Machine Learning for Health*, pages 343–355. PMLR.

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

Haochun Wang, Chi Liu, Sendong Zhao, Bing Qin, and Ting Liu. 2023b. Chatglm-med. https://github.com/SCIR-HI/Med-ChatGLM.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135.

Jiageng Wu, Xian Wu, Yining Hua, Shixu Lin, Yefeng Zheng, and Jie Yang. 2023. Exploring social media for early detection of depression in COVID-19 patients. In *Proceedings of the ACM Web Conference 2023*. ACM.

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. *ArXiv*, abs/2304.01097.

Ming Xu. 2023. Medicalgpt: Training medical gpt model. https://github.com/shibing624/MedicalGPT.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang

Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023a. Baichuan 2: Open large-scale language models.

Rui Yang, Ting Tan, Wei Lu, Arun Thirunavukarasu, Daniel Ting, and Nan Liu. 2023b. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2.

Songhua Yang, Hanjia Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2023c. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *arXiv preprint arXiv:2308.03549*.

C Yirong, W Zhenyu, X Xiaofen, X Zhipei, F Kai, L Sihang, W Junhong, and X Xiangmin. 2023. Bianque-1.0: Improving the" question" ability of medical chat model through finetuning with hybrid instructions and multi-turn doctor qa datasets," 2023.

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

Qingcheng Zeng, Lucas Garay, Peilin Zhou, Dading Chong, Yining Hua, Jiageng Wu, Yikang Pan, Han Zhou, Rob Voigtand, and Jie Yang. 2022b. Greenplm: Cross-lingual transfer of monolingual pre-trained language models at almost no cost. *the 32nd International Joint Conference on Artificial Intelligence*.

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*.

A Appendix

Instruction:
返回限制：只返回两行。假设你是一位医疗行业专家，请回答下列问题，注意是单选题，只需要返回一个最合适的选项。
甘味的作用特点是：
A 能补、能缓
B 能泄、能坚
C 能软、能下
D 能收、能涩
E 能散、能行
注意，结果只有两行。第一行只需要返回答案的英文选项(注意只需要返回一个最合适的答案)，第二行进行简要的解释。输出格式限制为："答案："，"解释："。
注意，题目是单选题。若有多个合适的答案，只返回最准确的即可。
Return limitation: Only return two lines in your answer. Assuming you are a medical expert, please answer the following single-choice question. Just give the most suitable choice.
What are the features of sweetness?
A Nourishing and soothing
B Draining and solidifying
C Softening and descending
D Contracting and astringent
E Dispersing and moving
Note that only two lines are permitted in the response. The first line should state the answer's English letter only, and the second line should provide a brief explanation. The format should be: "Answer:", "Explanation:". Only return the most accurate answer if multiple seem appropriate.

LLaMA:
A 甘味的作用特点是能补、能缓。
B 甘味的作用特点是能泄、能坚。
C 甘味的作用特点是能软、能下。
D 甘味的作用特点是能收、能涩。
E 甘味的作用特点是能散、能行。 解释：甘味是一种味噌，它是一种味噌
A The features of sweetness are nourishing and soothing.
B The features of sweetness are draining and solidifying.
C The features of sweetness are softening and descending.
D The features of sweetness are contracting and astringent.
E The features of sweetness are dispersing and moving.
Explanation: Sweetness is a kind of miso, it is a kind of miso

ChatGLM:
答案：B 能泄、能坚。解释：甘味的作用特点是能泄、能坚，这主要是因为甘味的食物能够促进体内液体的排泄，从而增强身体的耐力和耐久性。同时，甘味还具有收敛的作用，能够增强身体的韧性和强度。
Answer: B Draining and solidifying. Explanation: The features of sweetness are draining and solidifying. This is primarily because sweetness can promote the excretion of bodily fluids, thus enhancing the body's endurance and durability. Meanwhile, sweetness also has a constricting effect, enhancing the body's resilience and strength.

GPT-4:
答案：A 解释：甘味的作用特点是能补、能缓，能滋养身体，缓和药物的效果。
Answer: A Explanation: The features of sweetness are nourishing and soothing. It can enrich the body and moderate the effects of medicine.

Qilin-Med-CPT:
答案：A。解释：甘味的作用特点是能补、能缓。甘味能补虚、能缓急，还能益气、润燥。甘味能补虚，能补益五脏之气，补益脾胃之气，补益心气。
Answer: A. Explanation: The features of the sweetness are nourishing and soothing. Sweetness can replenish deficiencies and alleviate urgencies, as well as benefit the Qi and moisturize dryness. It can supplement the Qi of the five organs, nourish the Qi of the spleen and stomach, and benefit the Qi of heart.

Qilin-Med-SFT:
答案：A。\n 解释：本题考查的是五味对人体的效果。甘味的作用特点是能补、能缓（A对）。五味对人体的效果：(1)辛：能散、能行(E错)，有发散、行气、活血作用。(2)甘：能补、能缓(A对)、能和，有补虚、和中、缓急、调和药性等作用。(3)酸：能收、能涩，有收敛固涩作用。(4)苦：能泄、能燥、能坚(B错)。(5)咸：能软、能下(C错)，有软坚散结、泻下通便作用。(6)涩：能收、能敛，同酸味一样有收敛固涩作用。(7)淡：能渗、能利，有渗湿利水作用。此外，还有芳香味，其能散、能行、能开，有化湿、辟秽、开窍、醒脾等作用
Answer: A.
Explanation: This question examines the effects of the five flavors on the human body. The features of sweetness are nourishing and soothing (Option A is correct). The effects of the five flavors on the human body are:
1) Spicy: Can disperse and act (Option E is incorrect), having the effects of dispersing, moving Qi, and invigorating blood.
2) Sweet: Can nourish and soothe (Option A is correct), harmonize, with effects like replenishing deficiencies, harmonizing the middle, soothing urgencies, and moderating the properties of medicines.
3) Sour: Can constrict and be astringent, having effects of constriction and astringency.
4) Bitter: Can drain, dry, and solidify (Option B is incorrect).
5) Salty: Can soften and descend (Option C is incorrect), with effects like softening hard masses, dispersing accumulations, and inducing bowel movements.
6) Astringent: Can constrict and constrict, similar to the sour flavor in having astringent effects.
7) Bland: Can seep and benefit, having the effects of draining dampness and promoting diuresis. Additionally, there's an aromatic flavor, which can disperse, act, and open, with effects like transforming dampness, repelling foulness, opening orifices, and invigorating the spleen.

Figure 4: A conversational case on CMExam dataset. Compared to LLaMA, ChatGLM, and GPT-4. Qilin-Med-7B-CPT and Qilin-Med-7B-SFT generated more relevant and informative responses.