CareBot: A Pioneering Full-Process Open-Source Medical Language Model

Lulu Zhao[1*], **Weihao Zeng**[2], **Xiaofeng Shi**[1], **Hua Zhou**[1]

[1]Beijing Academy of Artificial Intelligence (BAAI)
[2]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
llzhao@baai.ac.cn

Abstract

Recently, both closed-source and open-source LLMs have made significant strides, outperforming humans in various general domains. However, their performance in specific professional domains such as medicine, especially within the open-source community, remains suboptimal due to the complexity of medical knowledge. In this paper, we propose CareBot, a bilingual medical LLM, which leverages a comprehensive approach integrating continuous pre-training (CPT), supervised fine-tuning (SFT), and reinforcement learning with human feedback (RLHF). Our novel two-stage CPT method, comprising Stable CPT and Boost CPT, effectively bridges the gap between general and domain-specific data, facilitating a smooth transition from pre-training to fine-tuning and enhancing domain knowledge progressively. We also introduce DataRater, a model designed to assess data quality during CPT, ensuring that the training data is both accurate and relevant. For SFT, we develope a large and diverse bilingual dataset, along with ConFilter, a metric to enhance multi-turn dialogue quality, which is crucial to improving the model's ability to handle more complex dialogues. The combination of high-quality data sources and innovative techniques significantly improves CareBot's performance across a range of medical applications. Our rigorous evaluations on Chinese and English benchmarks confirm CareBot's effectiveness in medical consultation and education. These advancements not only address current limitations in medical LLMs but also set a new standard for developing effective and reliable open-source models in the medical domain.

Code — https://github.com/FlagOpen/CareBot

Introduction

Recently, the advent of generative large language models (LLMs) like ChatGPT (Brown et al. 2020) and LLaMA (Touvron et al. 2023a,b) has revolutionized human-computer interaction. These models excel at basic text understanding and complex problem-solving tasks, demonstrating capabilities akin to human understanding and reasoning. However, in industrial applications, the professionalism and cost-effectiveness of LLMs are more concerned. Although a series of closed-source models such as GPT-4 still perform well in specialized domains, considering the risk of data privacy, it is not convenient to use such APIs to handle domain-specific issues. In the open-source community, a lack of domain-specific knowledge often limits the performance of open-source models in specialized areas, such as medical (Yang et al. 2023a; Xiong et al. 2023; Labrak et al. 2024). The complexity and depth of medical knowledge present significant challenges for developing accurate and secure medical LLMs. Nonetheless, we believe that medical LLMs hold immense potential and can significantly contribute to diagnostic assistance, consultation, drug recommendation, and more. Thus, developing a fully open-source LLM tailored for the medical domain is of paramount importance.

Currently, there are several medical LLMs available in this domain . However, most of these models rely solely SFT (Zhang et al. 2023, 2024a; Han et al. 2023). As is well-known, pre-training is a critical phase for learning domain-specific knowledge, and depending exclusively on SFT results in models that can only produce answers in a fixed format. Another approach attempts to integrate pre-training with SFT by converting pre-training data in specific domains into a unified format similar to SFT data, such as (instruction, output) pairs using GPT-3.5 (Chen et al. 2023). This method of synthesizing large amounts of data can lead to the inclusion of significant amounts of incorrect knowledge that aligns with GPT-4 but diverges from human expertise, as well as high data synthesis costs. Furthermore, Yang et al. (2023b) introduce Zhongjing for Chinese medicine, which employs to implement the pipeline training from pre-training, SFT, to RLHF. However, this approach involves two phases of transformation for the base model, which may lead to issues such as catastrophic forgetting or model degradation (Cheng, Huang, and Wei 2024). Additionally, previous efforts have predominantly focused on data construction during the SFT stage (Li et al. 2023a; Zhang et al. 2024b), while neglecting the importance of data construction during the CPT stage. Yet, a well-designed CPT data strategy is also crucial for inserting medical expertise into the model.

To address these challenges, we propose **CareBot**, a bilingual medical LLM based on LLaMA3-8B, designed to effectively assist doctors with diagnosis, provide personalized treatment plans, and support medical education. Our approach also implements the entire process from CPT,

*Corresponding author

SFT to RLHF. Most importantly, we develope a novel two-stage CPT method, consisting of stable CPT and boost CPT. The stable CPT addresses the distribution discrepancy between general and domain-specific data, while boost CPT narrows the gap between pre-training and fine-tuning data. This method facilitates a smooth transition for the model from general data to domain-specific data, and finally to fine-tuning data, thereby enhancing its domain knowledge progressively. Recognizing that data quality is critical to model performance, we also design a data quality assessment model for CPT called **DataRater**. This model employs a comprehensive quality assessment standard, evaluating aspects such as grammatical accuracy, information density, semantic consistency, and domain-specific attributes. DataRater effectively mitigates data bias, ensuring CareBot's performance and generalization capabilities in the medical domain. For SFT stage, we further construct a highly diverse medical SFT dataset, comprising single-turn and multi-turn medical dialogues, as well as medical subject knowledge multiple-choice questions, covering over 15+ departments and 100+ disease specialties. Noted that this corpus is the largest open-source bilingual medical SFT dataset available. It supports a variety of medical applications, clinical tasks, and online consultations, significantly enhancing CareBot's performance across multiple dimensions. Given the importance of data quality during the SFT stage(Zhou et al. 2023), we employ various selection methods. One key innovation is **ConFilter**, a metric designed to measure the correlation between multiple turns, which helps in filtering multi-turn dialogues. The inclusion of high-quality multi-turn dialogues not only improves the model's ability to understand user intent and generate relevant responses but also ensures a natural, smooth dialogue experience that enhances user comfort and satisfaction. Our SFT data is sourced partly from real-world medical diagnosis dialogues and partly from GPT-3.5 generated content. This combination ensures that the model delivers informative, clear, and logically consistent answers, while also providing professional and personalized consultations akin to those of medical experts. Finally, in the RLHF stage, we leverage GPT-4 to create positive and negative medical data pairs based on the SFT results. We then apply the Direct Preference Optimization (DPO) (Rafailov et al. 2023) algorithm to align the model's output with human expression styles, providing personalized answers and recommendations, and enhancing overall user experience and satisfaction.

After extensive training and optimization, we successfully develope the CareBot. We rigorously evaluate the performance of our model using widely used Chinese and English benchmarks in medical domain. The experimental results demonstrate that CareBot excels in both medical consultation and teaching, validating that our constructed datasets significantly enhance the model's performance across multiple dimensions. The main contributions of this paper are as follows: (1) We design a novel two-stage CPT strategy that progressively and stably integrates domain knowledge into the LLM, and effectively addresses data bias and language imbalances in the original pre-training data. (2) We propose DataRater, a model for assessing data quality during CPT, ensuring that the data used for CPT is of high quality. (3) We construct a comprehensive open-source medical SFT dataset with high data diversity and data quality. Additionally, we develope ConFilter, a metric for measuring the correlation between multiple turns, to ensure the quality of multi-turn dialogues. (4) We conduct experiments across multiple Chinese and English benchmarks to validate the effectiveness and reliability of our training strategy and datasets.

Methodology

Continue Pre-training

Data Collection and Decontamination To optimize the use of existing general data resources and minimize the cost of acquiring new medical-related data, we aim to extract medical-specific pre-training data from 15T widely-covered general corpus. These datasets include web content, encyclopedias, books, and academic papers, such as C4 (Raffel et al. 2023), Pile, Wudao, and PubMed, etc. To ensure the high quality of the domain-specific data, we implement a rigorous collection process that includes domain classification and quality assessment[1].

Domain Classification Since the general pre-training corpus is sourced from diverse datasets and lacks clear domain labels, we first conduct domain classification to extract high-quality medical data. Specifically, we sample 40k data from general corpus and use GPT-4 to perform two rounds of domain labeling to enhance accuracy. Data with inconsistent labels across two rounds is removed, leaving us with 36k high-quality seed data. We observe that certain categories, such as artificial intelligence and computers, have long tails. To address this imbalance, we utilize GPT-4 to generate additional synthetic data for these long-tail categories. Finally, we design a domain classifier based on the bge-m3[2]. We try a variety of multilingual pre-training models to train domain classifiers, details can be found in the Appendix A.

Rule-based Data Quality Filtering Quality filtering is a crucial step in processing pre-training corpus. To eliminate noisy data, we use a rule-based filtering solution, including rules for removing data with insufficient tokens, excessive special characters, toxic content, and private information.

LLM-based Data Quality Filtering By sampling and evaluating the data after rule filtering, we identify several issues: (1) The data includes advertising and marketing content, which could significantly skew the model's output preferences; (2) The data contains grammatical errors, semantic inconsistencies, and spliced, unrelated content, as well as image and video clips. Such data is detrimental to model training as it provides minimal valuable information for autoregressive learning. To address these issues, we design a data quality assessment model, DataRater, to assess data quality in terms of grammatical accuracy, information density, semantic consistency, and domain relevance, and further filter out low-quality content. Specifically, we extract 20k data from the rule-based filtered data and score them twice using the GPT-4, ranging from 0 to 5. Data with a

[1]It is worth noting that our data processing methodology is also applicable to any domain.

[2]https://huggingface.co/BAAI/bge-m3

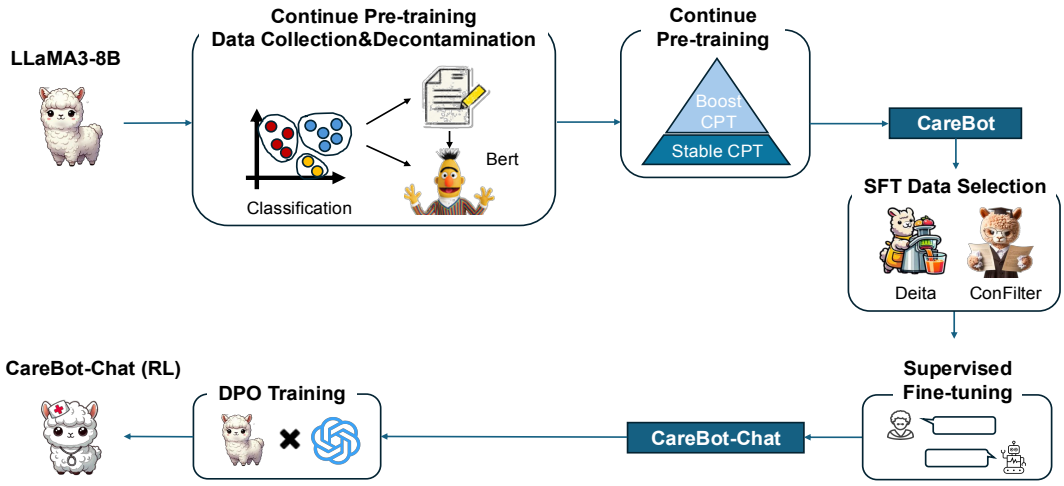


Figure 1: The overall pipline of CareBot-Chat (RL), which includes the two-stage continue pre-training, supervised fine-tuning, and the DPO process.

score discrepancy of about 2 points between the two assessments is removed, resulting in a final set of 15k high-quality training data. Finally, we train the DataRater based on the bge-m3. We also try a variety of multilingual pre-training models, details can be found in the Appendix B.

Training Strategy To gradually align the data distribution between pre-training and fine-tuning and minimize the loss of knowledge acquired during pre-training, we design a novel two-stage CPT strategy. This approach ensures a stable integration of medical knowledge into the LLM.

Stable CPT To balance medical domain knowledge with general knowledge, we first implement a Stable CPT stage, which ensures the model maintains and enhances its general language understanding while concentrating on medical information. In this stage, we combine a high-quality medical pre-training corpus with general data via the ratio as 19:1, with a token-level distribution of 1:9 for Chinese:English. We conduct adequate experiments to search both ratios, detailed results are available in the Appendix C.

Boost CPT To integrate medical knowledge during the model pre-training phase and facilitate a smooth transition to domain-specific tasks, we then design a Boost CPT phase. In this phase, we combine a very high-quality medical pre-training corpus with open-source medical SFT data at a 1:1 ratio, with a token-level distribution of 4:6 for Chinese:English. Notably, throughout these two phases, we progressively increase the proportion of Chinese data.

Supervised Fine-Tuning

To enhance model's ability to follow medical instructions and better adapt to specific medical scenarios, we conduct the SFT. This process involves using conversational-style data (comprising both queries and responses) to finetune the pretrained LLM. In the following sections, we will explore the details of data construction and training methods.

Data Construction Our SFT dataset comprises a diverse array of question types, including multiple-choice questions from medical exams, single-turn disease diagnoses, and multi-turn health consultations. It integrates data from seven publicly available sources: Chinese Medical Dialogue Data[3], Huatuo26M (Li et al. 2023a), MedDialog (Zeng et al. 2020), ChatMed Consult Dataset (Tian et al. 2023), ChatDoctor (Li et al. 2023b), CMB[4], and MedQA (Jin et al. 2021). We preserve portions of authentic doctor-patient conversations and augment the dataset by rewriting the remaining content. For these rewrites, we use real-world medical scenarios as prompts and generate responses via GPT-4. We believe this ensures the diversity of the SFT dataset, which can help the CareBot better adapt to different types of medical problems and patient situations, thereby improving its performance in a variety of scenarios.

As stated in Zhou et al. (2023), a relatively small, high-quality dataset can be sufficient for fine-tuning LLMs, our focus is on efficiently filtering "good data" from massive data to achieve competitive performance with a minimal amount of data. Like standard data cleaning processes, our approach begins by removing duplicates and eliminating data associated with security concerns such as violence, bias, and pornography. In the following sections, we specifically introduce the data selection methods.

Single-turn Medical Dialogue Data Following Liu et al. (2024); Zeng et al. (2024), we believe that "good data" should have a complex instruction and a high-quality response. Therefore, We adopt the approach from Deita (Liu et al. 2024), which scores each instance along two dimensions: instruction complexity score c_i and response quality score q_i. By multiplying c_i with q_i, we combine the complexity score and quality score to obtain a comprehensive score. Finally, we set a score threshold to select the most effective data instances in the massive data pool.

[3]https://github.com/Toyhom/Chinese-medical-dialogue-data
[4]https://github.com/FreedomIntelligence/CMB

Multi-turn Medical Dialogue Data For multi-turn dialogues, we initially use Deita to compute the score s_i for each individual turn and then average these scores to derive the final score for the entire dialogue. However, we identify two specific challenges in multi-turn dialogues compared to single-turn dialogues: (1) The low correlation between different turns can negatively affect the relevance of earlier information for subsequent turns; (2) Excessive correlation between turns can lead to significant context duplication and redundant information. To address these issues, we propose the ConFilter method, which uses a score CF based on cross-entropy loss, to assess the influence of historical information on each turn. The details of this approach are outlined as follows:

In the instruction-tuning process, the loss of a sample pair (H, T) is calculated by continuously predicting the next tokens in the current turn T given their previous tokens and the history information H:

$$L_\theta(t_i|H) = -\frac{1}{N}\sum_{j=1}^{N} logP(w_i^j|H, w_i^1, w_i^2, ..., w_i^{j-1}; \theta) \tag{1}$$

where $H = \{t_1, t_2, ...t_{i-1}\}$, t_i is the current turn, w_i^j is the j-th token in the i-th turn, and N is the number of tokens of the current turn. We define $L_\theta(t_i|H)$ as the Conditioned Information Score, which measures the ability to generate the current turn under the guidance of corresponding historical information.

To measure the ability of LLM to generate this turn alone, we also define a Direct Information Score:

$$L_\theta(t_i) = -\frac{1}{N}\sum_{j=1}^{N} logP(w_i^j|w_i^1, w_i^2, ..., w_i^{j-1}; \theta) \tag{2}$$

We believe that the higher Direct Information Score may indicate that the turn is more challenging or complex. Finally, we try to estimate CF by calculating the ratio between $L_\theta(t_i)$ and $L_\theta(t_i|H)$.

$$CF_\theta(H, T) = \frac{L_\theta(t_i|H)}{L_\theta(t_i)} \tag{3}$$

Here, if $CF > 1$, it means historical information has a negative impact on current turn, that is, the correlation between contexts is very low. If $CF < 1$, it means historical information has a positive impact on current turn, that is, the correlation between contexts is high. However, too small CF means that the context is highly repeated and the information is highly redundant. We also set a threshold to filter the data.

RLHF

We enhance the model's capabilities using Direct Preference Optimization (DPO) (Rafailov et al. 2023) after the SFT stage. To align the model's output with human preferences while preserving the foundational abilities gained during the CPT and SFT stages (Lu et al. 2024), we construct subjective preference data and objective preference data using samples that have the same distribution as the SFT dataset:

Subjective Preference Data We aim to construct dpo pairs where the chosen response aligns closely with human preferences. For each prompt, we first ask GPT-4 to respond as a professional and helpful doctor. Then, using GPT-4, we evaluate the superiority or inferiority of the our CareBot-Chat's response and GPT-4's response. The evaluation considers four aspects: fluency, relevance, completeness, and proficiency in medical. We select the superior response as the chosen response for the dpo pair and the inferior response as the rejection response.

Objective Preference Data While RLHF can guide LLMs to align with human expectations, numerous studies show that this method can cause LLMs to forget abilities acquired during pre-training and SFT stages (Bai et al. 2022; Dong et al. 2023a), leading to an "alignment tax" (Dong et al. 2023b; Sun et al. 2024). To mitigate this issue, we construct objective preference data. Specifically, for objective prompts with known ground truth answers, we consider the ground truth as the chosen response and randomly select incorrect answers from the remaining options as rejection responses. For instance, in multiple-choice questions, if the ground truth is option A, we randomly select from options B, C, and D to construct the rejection response.

Experimental Setup

Baselines

We conduct a comparative analysis of our model against most representative open-source medical LLMs including HuatuoGPT-7B (Zhang et al. 2023), Zhongjing-13B (Yang et al. 2023b), MedAlpaca-7B (Han et al. 2023), BioMistral-7B (Labrak et al. 2024), and HuatuoGPT ll-7B (Chen et al. 2023). These models are specifically designed for medical applications, showcasing robust open-domain chat capabilities and applicability to various medical scenarios. Additionally, we also compare results from the closed-source model GPT-3.5-turbo. More details can be found in Appendix E.

Medical Benchmark

We comprehensively evaluate CareBot's medical capabilities from two aspects, one is medical concept knowledge, and the other is medical consultation ability. For medical concept knowledge, CareBot is evaluated using three popular Chinese medical benchmarks (CMB (Wang et al. 2024), CMMLU-Med (Li et al. 2024), C-Eval-Med (Huang et al. 2023)) and four English medical benchmarks (MedQA (Jin et al. 2021), MMLU-Med (Hendrycks et al. 2021), MedMCQA (Pal, Umapathi, and Sankarasubbu 2022), and PubMedQA (Jin et al. 2019) test set). Accuracy is served as the primary evaluation metric for this aspect. For single-turn medical consultation questions, we use the Huatuo26M-test (Li et al. 2023a), evaluating responses via HuatuoEval (Chen et al. 2023) for pairwise comparisons.Additionally, multi-turn medical consultation questions are assessed using CMtMedQA (Yang et al. 2023b) and CMB-Clin (Wang et al. 2024). Consistent with Wang et al. (2024), the model's responses are rated based on the fluency, relevance, completeness and medical proficiency of the reference answers. More details can be found in Appendix E.

Models	English		Chinese			
	MedQA	MMLU-Med	CMB	CMMLU-Med	C-Eval-Med	Avg.
ChatGPT	52.24	69.96	43.26	50.37	48.80	52.93
HuatuoGPT-7B	16.63	25.62	19.68	23.11	25.66	22.14
Zhongjing-13B	11.28	16.90	20.39	23.85	30.09	20.50
MedAlpaca-7B	49.74	62.72	23.29	25.38	27.43	37.71
Biomistral-7B	50.60†	59.08†	23.83	26.55	25.67	37.15
HuatuoGPT ll-7B	41.13	51.44	**60.39**	**59.08**	62.40	54.89
CareBot-Chat	**63.71**	**71.53**	52.50	56.42	**63.72**	**61.58**
CareBot-Chat (RL)	63.63	71.44	52.45	56.58	62.83	61.39

Table 1: The results of five medical concept knowledge benchmarks. † means the result of 3-shot (consistent with the original paper) and others are 0-shot. All scores are averaged over three random runs. ($p<0.05$ under t-test)

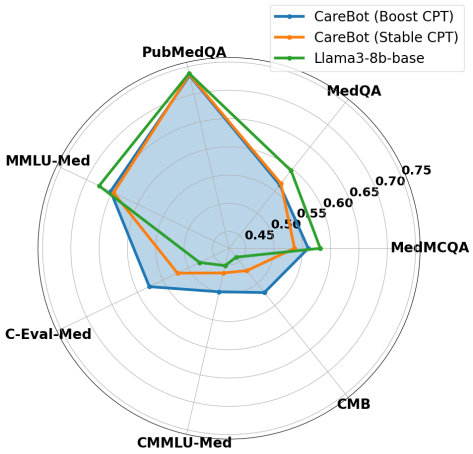


Figure 2: The performance of seven benchmarks for our CPT model, CareBot.

Experimental Results for CPT

In Figure 2, we evaluate our CPT model, CareBot, on seven common medical benchmarks. Considering that our goal is to train a medical model that performs well in both Chinese and English, we strive to improve Chinese medical ability while ensuring that English medical ability of the model is slightly reduced. We observe that for English benchmarks (MMLU-Med, PubMedQA, MedQA, MedMCQA), the performance of CareBot (Stable CPT) and CareBot (Stable CPT & Boost CPT) shows a slight decrease. This is expected, given that the LLaMA-8B-base model already has strong English capabilities. However, for Chinese benchmarks (C-Eval-Med, CMMLU-Med, CMB), our models demonstrate significant improvements, with particularly notable gains in models trained using the two-stage approach. This confirms that our two-stage CPT strategy effectively integrates medical domain knowledge into the model, resulting in robust enhancements to its Chinese medical capabilities.

Experimental Results for Alignment

Results for Medical Concept Knowledge

We present the results from five widely used benchmarks in Table 1. For English benchmark MedQA, our model CareBot-Chat outperforms ChatGPT by 11.47 points and BioMistral, the strongest open-source medical LLM, by 13.11%. For MMLU-Med, CareBot-Chat achieves a 1.57% improvement over ChatGPT and surpasses MedAlpaca by 8.81 points. For English benchmarks CMB and CMMLU-Med, HuatuoGPT II emerges as the top-performing model and our model does not have an advantage over it. For C-Eval-Med, our CareBot only achieves competitive results with HuatuoGPT II. This is because CareBot is built on LLaMA3-8B, an LLM with inherent strengths in English. Therefore, its performance in Chinese aligns with expectations. Nevertheless, in terms of average scores, CareBot-Chat exceeds HuatuoGPT II, the leading open-source medical model, by 6.69%, and ChatGPT by 8.65%. These outcomes underscore CareBot's exceptional performance in medical applications, establishing it as a significant contributor to the field of medical AI.

Results for Medical Consultation Ability

Multi-turn Dialogue In Table 2 and 3, we present the results for the multi-turn dialogue benchmarks CMtMedQA and CMB-clin, respectively. Overall, the performance of two baselines is basically the same, with notable performance from HuatuoGPT and HuatuoGPT ll. Across four dimensions, fluency and proficiency scores are particularly high, indicating coherent responses and a solid grasp of medical terminology by the medical LLMs. However, relevance and completeness scores are lower, suggesting room for improvement in providing highly relevant and comprehensive answers tailored to specific questions. Our model, CareBot-Chat (RL), achieves strong performance across all dimensions, averaging scores of 4.58 and 4.31 respectively, with notable strengths in relevance and completeness. This underscores the effectiveness of our high-quality SFT dataset and our proposed multi-turn dialogue selection method, significantly enhancing the model's contextual understanding and ensuring dialogue coherence and consistency. We further analyze the advantages of our models in multi-turn dialogues in Section Advantages of Multi-turn Dialogue.

Single-turn Dialogue Figure 3 (a) displays the comparison results between our CareBot-Chat and various baselines on the single-turn dialogue benchmark Huatuo26M-test. It is evident that CareBot-Chat achieves comparable performance to ChatGPT and HuatuoGPT, indicating that

Models	Fluency	Relevance	Completeness	Proficiency	Avg.
ChatGPT	4.94	4.21	3.98	4.04	4.29
HuatuoGPT-7B	4.94	3.28	3.22	4.19	3.91
Zhongjing-13B	2.26	1.68	1.62	2.63	2.05
MedAlpaca-7B	4.35	2.18	1.97	3.29	2.95
Biomistral-7B	4.48	2.45	2.17	3.55	3.16
HuatuoGPT ll-7B	4.96	3.41	3.47	4.27	4.03
CareBot-Chat	**4.99**	4.67	4.20	4.26	4.53
CareBot-Chat (RL)	4.96	**4.70**	**4.31**	**4.34**	**4.58**

Table 2: The scores of different models on CMtMedQA.

Models	Fluency	Relevance	Completeness	Proficiency	Avg.
ChatGPT	4.78	3.75	3.77	4.01	4.08
HuatuoGPT-7B	4.69	3.25	3.08	4.00	3.76
Zhongjing-13B	3.14	2.14	1.89	3.11	2.57
MedAlpaca-7B	3.31	1.69	1.49	2.58	2.27
Biomistral-7B	3.79	2.10	1.82	3.09	2.70
HuatuoGPT ll-7B	4.75	3.43	3.43	4.22	3.96
CareBot-Chat	**4.82**	4.20	3.68	4.16	4.22
CareBot-Chat (RL)	4.74	**4.28**	**4.01**	**4.21**	**4.31**

Table 3: The scores of different models on CMB-Clin.

the baseline HuatuoGPT performs similarly to ChatGPT under this evaluation framework. Moreover, our model significantly outperforms Zhongjing, MedAlpaca, and BioMistral. The lower performance of MedAlpaca and BioMistral can be attributed to their limited Chinese language capabilities and inadequate medical SFT data. It is noteworthy that Zhongjing, incorporating pre-training, SFT, and RLHF stages, performs poorly. This is likely due to the two training stages of the base model, leading to catastrophic forgetting or model degradation. Additionally, its instruction fine-tuning data is both limited and of inferior quality, which explains why a model with larger parameters than our CareBot performs substantially worse. However, we observe that our model only competes with HuatuoGPT ll-7B in the HuatuoEval framework. We will further discuss this phenomenon in Section Case Study. In Figure 3 (b), we also compare our CareBot-Chat (RL) model with other baselines, and basically achieve the same performance as CareBot-Chat.

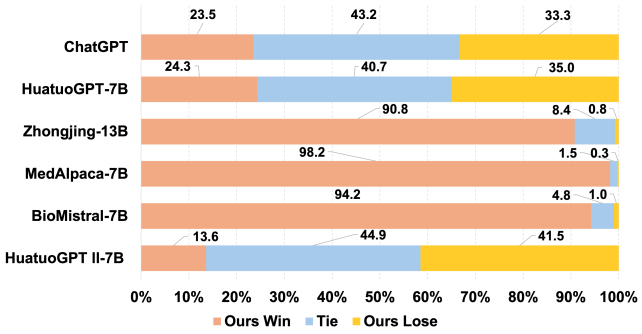
Analysis and Discussion

One Stage CPT vs Two Stage CPT

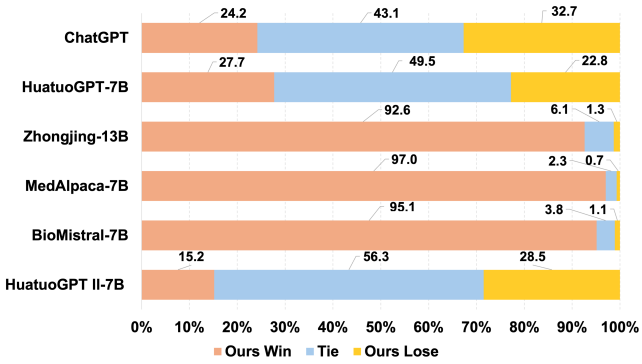
Figure 4 illustrates the changes in Acc for the standard CPT (represented by the orange line) and our two-stage CPT method (shown with both the orange and blue lines). Initially, as the number of training tokens increases, the Acc rises but fluctuates significantly. After reaching 45B tokens, the Acc stabilizes around the LLaMA-8B-base level, indicating that the medical CPT has reached a relatively stable state. Introducing the Boost CPT stage results in a marked and consistent improvement in Acc. In contrast, continuing training with 20B additional tokens using the Stable CPT approach shows minimal changes in performance. These findings strongly demonstrate that our two-stage CPT method effectively facilitates a smooth transition from general knowledge to domain-specific knowledge, from English to Chinese, and from PT to SFT.

Advantages of Multi-turn Dialogue

To investigate the reasons behind our model's strong performance in multi-turn dialogues, we analyze the quality of each turn of responses, as shown in Table 4 (Appendix G). Our analysis reveals that, except for the first turn, CareBot-Chat's average score was 0.25 lower than that of HuatuoGPT II. However, in subsequent turns, CareBot-Chat consistently outperforms HuatuoGPT II. Furthermore, our advantage becomes increasingly pronounced as the number of dialogue turns increases. Specifically, in terms of fluency and proficiency, our model performs comparably to HuatuoGPT II. In contrast, HuatuoGPT II's performance in relevance and completeness significantly deteriorates with more turns, while CareBot-Chat maintains stable performance. This highlights CareBot-Chat's superior capability in multi-turn dialogues, demonstrating its better understanding of dialogue context. We attribute this advantage to our carefully curated multi-turn dialogue SFT data, which retains dialogues with contextual relevance without excessive repetition. Additionally, Figure 10 and 11 in Appendix G provide comparisons of responses from CareBot-Chat and HuatuoGPT II in the same multi-turn dialogue, further confirming our model's effectiveness. Notably, in the first turn, HuatuoGPT II's response is longer and more detailed than ours. However, from the second turn, while HuatuoGPT II's answers remain detailed, they become increasingly irrelevant and lack coherence with the ongoing dialogue. Generally, our results strongly indicate that CareBot-Chat has a significant advantage in multi-turn dialogues, demonstrating superior contextual understanding and coherence.



(a) CareBot-Chat



(b) CareBot-Chat (RL)

Figure 3: Comparison of CareBot-Chat and CareBot-Chat (RL)'s predicted answers and other baselines' predicted answers on single-turn dialogues from Huatuo26M-test.

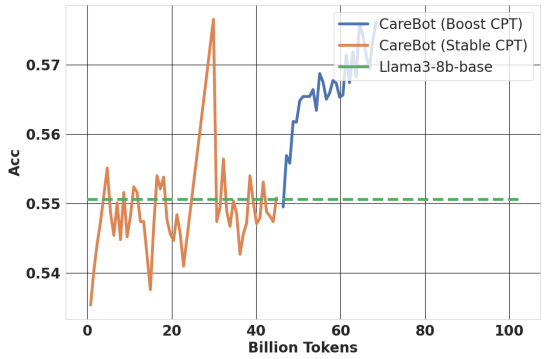


Figure 4: Comparison of the loss between our proposed two-stage CPT and the plain CPT.

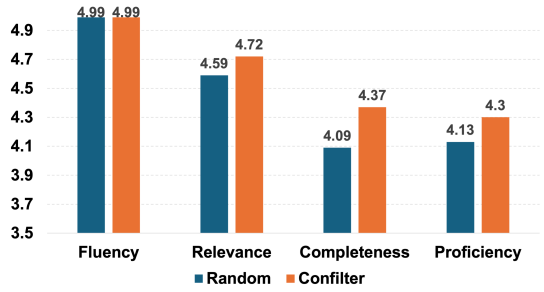


Figure 5: The comparison of the high-quality multi-turn dialogues selected by our ConFilter and the randomly selected multi-turn dialogues. These are all from CMtMedQA.

Effect on ConFilter

To evaluate the effectiveness of our multi-turn SFT data selection method, ConFilter, we conduct a comparative experiment. We finetune the pre-trained CareBot using two datasets: 110k high-quality multi-turn dialogues selected by ConFilter and 110k randomly selected samples from the original CMtMedQA dataset. As shown in Figure 5, both datasets achieve comparable fluency scores. However, in the other three dimensions, fine-tuning with the high-quality data significantly outperforms the randomly selected data. These results underscore the effectiveness of ConFilter in enhancing dialogue coherence and relevance by focusing on contextual correlations, thereby helping the model better understand and maintain conversation history.

Case Study

Figure 3 (a) illustrates our Carebot does not hold a significant advantage over HuatuoGPT II. Here, we present a case study. As depicted in Figure 12 (Appendix H), we provide an example where the output of CareBot-Chat is deemed inferior to that of HuatuoGPT II under the HuatuoEval evaluation framework. However, upon closer examination, we found there are some issues in HuatuoGPT II's response, particularly in the statement "such as 100% skim milk or low-fat milk. These milks usually contain more protein and calcium, and have lower sugar and fat content." Firstly, this statement itself is an hallucination, that is, skim milk or low-fat milk does not contain more protein and calcium, but only has lower fat content. Furthermore, medical professionals confirm that whole milk is more suitable for infants due to its nutritional benefits, including fats crucial for development. This aligns with responses generated by GPT-4. Additionally, HuatuoGPT II often includes repetitive content such as "It is better to choose milk designed specifically for children," which, despite mimicking a doctor's tone and offering longer responses, sometimes lacks relevance and completeness. This approach occasionally introduces ambiguities.

Conclusion

In this paper, we propose CareBot, a bilingual medical LLM, which is designed to enhance medical diagnostics, treatment planning, and medical education. To bridge gaps between data with different distributions, we design a novel two-stage continuous pre-training (CPT) approach, i.e., stable CPT and boost CPT. A model for evaluating data quality during CPT, DataRater, is also proposed. Besides, we further present ConFilter for selecting high-quality multi-turn dialogues, which is crucial to improving the model's ability to handle more complex dialogues. CareBot's performance, validated through extensive testing on Chinese and English medical benchmarks, demonstrates significant improvements in medical consultation and teaching, showcasing the effectiveness of its training strategies and datasets.

Acknowledgments

We thank all anonymous reviewers. This work was supported by National Science and Technology Major Project No.2022ZD0116314.

References

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chen, J.; Wang, X.; Gao, A.; Jiang, F.; Chen, S.; Zhang, H.; Song, D.; Xie, W.; Kong, C.; Li, J.; Wan, X.; Li, H.; and Wang, B. 2023. HuatuoGPT-II, One-stage Training for Medical Adaption of LLMs. arXiv:2311.09774.

Cheng, D.; Huang, S.; and Wei, F. 2024. Adapting Large Language Models via Reading Comprehension. In *The Twelfth International Conference on Learning Representations*.

Dong, G.; Yuan, H.; Lu, K.; Li, C.; Xue, M.; Liu, D.; Wang, W.; Yuan, Z.; Zhou, C.; and Zhou, J. 2023a. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Dong, H.; Xiong, W.; Goyal, D.; Zhang, Y.; Chow, W.; Pan, R.; Diao, S.; Zhang, J.; Shum, K.; and Zhang, T. 2023b. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Han, T.; Adams, L. C.; Papaioannou, J.-M.; Grundmann, P.; Oberhauser, T.; Löser, A.; Truhn, D.; and Bressem, K. K. 2023. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. arXiv:2304.08247.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; jiayi lei; Fu, Y.; Sun, M.; and He, J. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14).

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W. W.; and Lu, X. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-A.; Rouvier, M.; and Dufour, R. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv:2402.10373.

Li, H.; Zhang, Y.; Koto, F.; Yang, Y.; Zhao, H.; Gong, Y.; Duan, N.; and Baldwin, T. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. arXiv:2306.09212.

Li, J.; Wang, X.; Wu, X.; Zhang, Z.; Xu, X.; Fu, J.; Tiwari, P.; Wan, X.; and Wang, B. 2023a. Huatuo-26M, a Large-scale Chinese Medical QA Dataset. arXiv:2305.01526.

Li, Y.; Li, Z.; Zhang, K.; Dan, R.; Jiang, S.; and Zhang, Y. 2023b. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. arXiv:2303.14070.

Liu, W.; Zeng, W.; He, K.; Jiang, Y.; and He, J. 2024. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *The Twelfth International Conference on Learning Representations*.

Lu, K.; Yu, B.; Huang, F.; Fan, Y.; Lin, R.; and Zhou, C. 2024. Online Merging Optimizers for Boosting Rewards and Mitigating Tax in Alignment. *arXiv preprint arXiv:2405.17931*.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260. PMLR.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683.

Sun, Z.; Shen, Y.; Zhou, Q.; Zhang, H.; Chen, Z.; Cox, D.; Yang, Y.; and Gan, C. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.

Tian, Y.; Gan, R.; Song, Y.; Zhang, J.; and Zhang, Y. 2023. ChiMed-GPT: A Chinese Medical Large Language Model with Full Training Regime and Better Alignment to Human Preferences. arXiv:2311.06025.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023a. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale,

S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wang, X.; Chen, G.; Dingjie, S.; Zhiyi, Z.; Chen, Z.; Xiao, Q.; Chen, J.; Jiang, F.; Li, J.; Wan, X.; Wang, B.; and Li, H. 2024. CMB: A Comprehensive Medical Benchmark in Chinese. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6184–6205. Mexico City, Mexico: Association for Computational Linguistics.

Xiong, H.; Wang, S.; Zhu, Y.; Zhao, Z.; Liu, Y.; Huang, L.; Wang, Q.; and Shen, D. 2023. DoctorGLM: Fine-tuning your Chinese Doctor is not a Herculean Task. arXiv:2304.01097.

Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D.; Yang, F.; Deng, F.; Wang, F.; Liu, F.; Ai, G.; Dong, G.; Zhao, H.; Xu, H.; Sun, H.; Zhang, H.; Liu, H.; Ji, J.; Xie, J.; Dai, J.; Fang, K.; Su, L.; Song, L.; Liu, L.; Ru, L.; Ma, L.; Wang, M.; Liu, M.; Lin, M.; Nie, N.; Guo, P.; Sun, R.; Zhang, T.; Li, T.; Li, T.; Cheng, W.; Chen, W.; Zeng, X.; Wang, X.; Chen, X.; Men, X.; Yu, X.; Pan, X.; Shen, Y.; Wang, Y.; Li, Y.; Jiang, Y.; Gao, Y.; Zhang, Y.; Zhou, Z.; and Wu, Z. 2023a. Baichuan 2: Open Large-scale Language Models. arXiv:2309.10305.

Yang, S.; Zhao, H.; Zhu, S.; Zhou, G.; Xu, H.; Jia, Y.; and Zan, H. 2023b. Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-world Multi-turn Dialogue. arXiv:2308.03549.

Zeng, G.; Yang, W.; Ju, Z.; Yang, Y.; Wang, S.; Zhang, R.; Zhou, M.; Zeng, J.; Dong, X.; Zhang, R.; Fang, H.; Zhu, P.; Chen, S.; and Xie, P. 2020. MedDialog: Large-scale Medical Dialogue Datasets. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9241–9250. Online: Association for Computational Linguistics.

Zeng, W.; Xu, C.; Zhao, Y.; Lou, J.-G.; and Chen, W. 2024. Automatic Instruction Evolving for Large Language Models. *arXiv preprint arXiv:2406.00770*.

Zhang, H.; Chen, J.; Jiang, F.; Yu, F.; Chen, Z.; Li, J.; Chen, G.; Wu, X.; Zhang, Z.; Xiao, Q.; Wan, X.; Wang, B.; and Li, H. 2023. HuatuoGPT, towards Taming Language Model to Be a Doctor. arXiv:2305.15075.

Zhang, K.; Zeng, S.; Hua, E.; Ding, N.; Chen, Z.-R.; Ma, Z.; Li, H.; Cui, G.; Qi, B.; Zhu, X.; Lv, X.; Jinfang, H.; Liu, Z.; and Zhou, B. 2024a. UltraMedical: Building Specialized Generalists in Biomedicine. arXiv:2406.03949.

Zhang, X.; Tian, C.; Yang, X.; Chen, L.; Li, Z.; and Petzold, L. R. 2024b. AlpaCare:Instruction-tuned Large Language Models for Medical Application. arXiv:2310.14558.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; YU, L.; Zhang, S.; Ghosh, G.; Lewis, M.; Zettlemoyer, L.; and Levy, O. 2023. LIMA: Less Is More for Alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.