Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-World Multi-Turn Dialogue

Songhua Yang[*], **Hanjie Zhao**[*], **Senbin Zhu**, **Guangyu Zhou**,
Hongfei Xu, **Yuxiang Jia**[†], **Hongying Zan**
Zhengzhou University, Henan, China
{suprit,hjzhao_zzu,ygdzzx5156,zhougyzzu,hfxunlp}@foxmail.com, {ieyxjia, iehyzan}@zzu.edu.cn

Abstract

Recent advances in Large Language Models (LLMs) have achieved remarkable breakthroughs in understanding and responding to user intents. However, their performance lag behind general use cases in some expertise domains, such as Chinese medicine. Existing efforts to incorporate Chinese medicine into LLMs rely on Supervised Fine-Tuning (SFT) with single-turn and distilled dialogue data. These models lack the ability for doctor-like proactive inquiry and multi-turn comprehension and cannot align responses with experts' intentions. In this work, we introduce Zhongjing, the first Chinese medical LLaMA-based LLM that implements an entire training pipeline from continuous pre-training, SFT, to Reinforcement Learning from Human Feedback (RLHF). Additionally, we construct a Chinese multi-turn medical dialogue dataset of 70,000 authentic doctor-patient dialogues, CMtMedQA, which significantly enhances the model's capability for complex dialogue and proactive inquiry initiation. We also define a refined annotation rule and evaluation criteria given the unique characteristics of the biomedical domain. Extensive experimental results show that Zhongjing outperforms baselines in various capacities and matches the performance of ChatGPT in some abilities, despite the 100x parameters. Ablation studies also demonstrate the contributions of each component: pre-training enhances medical knowledge, and RLHF further improves instruction-following ability and safety. Our code, datasets, and models are available at https://github.com/SupritYoung/Zhongjing.

Introduction

Recently, significant progress has been made with LLMs, exemplified by ChatGPT (OpenAI 2022) and GPT-4 (OpenAI 2023), allowing them to understand and respond to various questions and even outperform humans in a range of general areas. Although openai remains closed, Open-source community swiftly launched high performing LLMs such as LLaMA (Touvron et al. 2023), Bloom (Scao et al. 2022), Falcon (Almazrouei et al. 2023) etc. To bridge the gap in Chinese adaptability, researchers also introduced more powerful Chinese models (Cui, Yang, and Yao 2023a; Du et al. 2022; Zhang et al. 2022). However, despite the stellar performance of these general LLMs across many tasks, their performance in specific professional fields, such as the biomedical domain, is often limited due to a lack of domain expertise (Zhao et al. 2023). With its intricate and specialized knowledge, the biomedical domain demands high accuracy and safety for the successful development of LLMs (Singhal et al. 2023a). Despite the challenges, medical LLMs hold enormous potential, offering value in diagnosis assistance, consultations, drug recommendations, etc. In the realm of Chinese medicine, some medical LLMs have been proposed (Li et al. 2023; Zhang et al. 2023; Xiong et al. 2023).

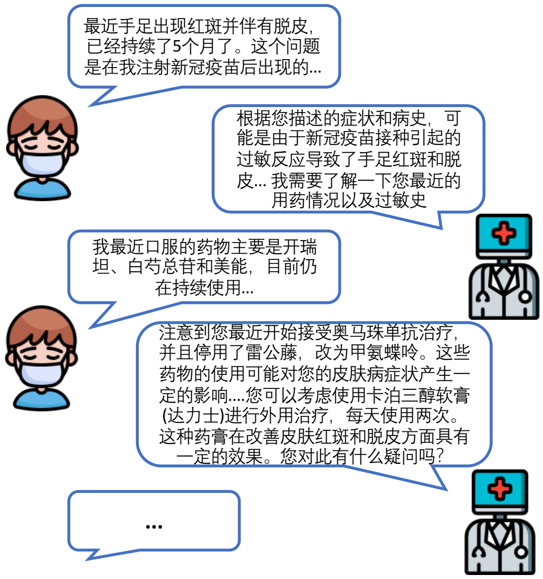

最近手足出现红斑并伴有脱皮，已经持续了5个月了。这个问题是在我注射新冠疫苗后出现的...
根据您描述的症状和病史，可能是由于新冠疫苗接种引起的过敏反应导致了手足红斑和脱皮... 我需要了解一下您最近的用药情况以及过敏史
我最近口服的药物主要是开瑞坦、白芍总苷和美能，目前仍在持续使用...
注意到您最近开始接受奥马珠单抗治疗，并且停用了雷公藤，改为甲氨蝶呤。这些药物的使用可能对您的皮肤病症状产生一定的影响....您可以考虑使用卡泊三醇软膏(达力士)进行外用治疗，每天使用两次。这种药膏在改善皮肤红斑和脱皮方面具有一定的效果。您对此有什么疑问吗?
...

Figure 1: An example of multi-turn Chinese medical consultation dialogues, relies heavily on LLM's proactive inquiry.

However, these works are totally dependent on SFT. Han et al. (2021) and Zhou et al. (2023) indicated that almost all knowledge is learned during pre-training, which is the critical phase in accumulating knowledge, and RLHF can guide models to recognize their capability boundaries and enhance instruction-following ability (Ramamurthy et al. 2022). Over-reliance on SFT may result in overconfident generalization, the model essentially rote-memorizes the an-

[*] These authors contributed equally.

[†] Corresponding author

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

swers rather than understanding and reasoning the inherent knowledge. Moreover, previous dialogue datasets primarily focus on single-turn dialogue, overlooking the process in authentic doctor-patient dialogues that usually need multi-turn interactions and are led by doctors who will initiate inquiries frequently to understand the condition.

To address these limitations, we propose Zhongjing[1], the first Chinese medical LLM based on LLaMA, implementing the entire pipeline from continuous pre-training, SFT to RLHF. Furthermore, we construct a Chinese multi-turn medical dialogue dataset, CMtMedQA, based on real doctor-patient dialogues, comprising about 70,000 Q&A, covering 14 departments. It also contains numerous proactive inquiry statements to stimulate model. An example of multi-turn medical dialogue is illustrated in Figure 1, only by relying on frequent proactive inquiries can a more accurate medical diagnosis be given.

Specifically, the construction of our model is divided into three stages. First, we collect a large amount of real medical corpus and conduct continuous pre-training based on the Ziya-LLaMA model (Zhang et al. 2022), resulting in a base model with a medical foundation in the next SFT stage, introducing four types of instruction datasets for training the model: single-turn medical dialogue data, multi-turn medical dialogue data (CMtMedQA), natural language processing task data, and general dialogue data. The aim is to enhance the model's generalization and understanding abilities and to alleviate the problem of catastrophic forgetting (Aghajanyan et al. 2021). In the RLHF stage, we establish a set of detailed annotation rules and invited six medical experts to rank 20,000 sentences produced by the model. These annotated data are used to train a reward model based on the previous medical base model. Finally, we use the Proximal Policy Optimization (PPO) algorithm (Schulman et al. 2017) to guide the model to align with the expert doctors' intents.

After extensive training and optimization, we successfully developed Zhongjing, a robust Chinese medical LLM. Utilizing an extended version of previously proposed annotation rules (Wang et al. 2023a; Zhang et al. 2023), we evaluated the performance of our model on three dimensions of capability and nine specific abilities, using GPT-4 or human experts. The experimental results show that our model surpasses other open-source Chinese medical LLM in all capacity dimensions. Due to the alignment at the RLHF stage, our model also makes a substantial improvement in safety and response length. Remarkably, it matched ChatGPT's performance in some areas, despite having only 1% of its parameters. Moreover, the CMtMedQA dataset significantly bolsters the model's capability in dealing with complex multi-turn dialogue and initiating proactive inquiries.

The main contributions of this paper are as follows:

• **We develop a novel Chinese medical LLM, Zhongjing.** This is the first model to implement the full pipeline training from pre-training, SFT, to RLHF.

• **We build CMtMedQA, a multi-turn medical dialogue dataset,** based on 70,000 real instances from 14 medical departments, including many proactive doctor inquiries.

• **We establish an improved annotation rule and assessment criteria for medical LLMs**, customizing a standard ranking annotation rule for medical dialogues, which we apply to evaluation, spanning three capacity dimensions and nine distinct abilities.

• **We conduct multiple experiments on two benchmark test datasets.** Our model exceeds the previous top Chinese medical model in all dimensions and matches ChatGPT in specific fields.

Related Work

Large Language Models

The remarkable achievements of Large Language Models (LLMs) such as ChatGPT (OpenAI 2022) and GPT-4 (OpenAI 2023) have received substantial attention, sparking a new wave in AI. Although OpenAI has not disclosed their training strategies or weights, the rapid emergence of open-source LLMs like LLaMA (Touvron et al. 2023), Bloom (Scao et al. 2022), and Falcon (Almazrouei et al. 2023) has captivated the research community. Despite their initial limited Chinese proficiency, efforts to improve their Chinese adaptability have been successful through training with large Chinese datasets. Chinese LLaMA and Chinese Alpaca (Cui, Yang, and Yao 2023b) continuously pre-trained and optimized with Chinese data and vocabulary. Ziya-LLaMA (Zhang et al. 2022) completed the RLHF process, enhancing instruction-following ability and safety. In addition, noteworthy attempts have been made to build proficient Chinese LLMs from scratch (Du et al. 2022; Sun et al. 2023a).

LLMs in Medical Domain

Large models generally perform suboptimally in the biomedical field that require complex knowledge and high accuracy. Researchers have made significant progress, such as MedAlpaca (Han et al. 2023) and ChatDoctor (Yunxiang et al. 2023), which employed continuous training, Med-PaLM (Singhal et al. 2023a), and Med-PaLM2 (Singhal et al. 2023b), receiving favorable expert reviews for clinical responses. In the Chinese medical domain, some efforts include DoctorGLM (Xiong et al. 2023), which used extensive Chinese medical dialogues and an external medical knowledge base, and BenTsao (Wang et al. 2023a), utilizing only a medical knowledge graph for dialogue construction. Zhang et al. (2023); Li et al. (2023) proposed HuatuoGPT and a 26-million dialogue dataset, achieving better performance through a blend of distilled and real data for SFT and using ChatGPT for RLHF.

Methods

This section explores the construction of Zhongjing, spanning three stages: continuous pre-training, SFT, and RLHF - with the latter encompassing data annotation, reward model, and PPO. Each step is discussed sequentially to mirror the research workflow. The comprehensive method flowchart is shown in Figure 2.

[1] In homage to the renowned ancient Chinese medical scientist Zhongjing Zhang, we named our model "Zhongjing".

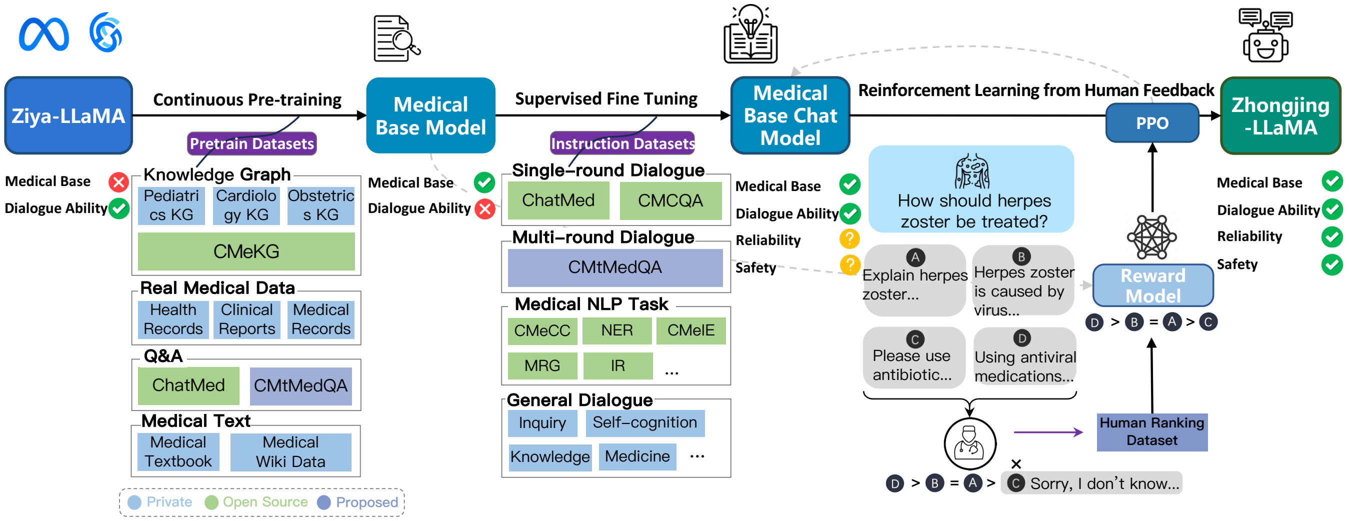


Figure 2: The overall flowchart of constructing Zhongjing. Ticks, crosses, and question marks beneath the upper rectangles signify the ability model currently possesses, lacks, or likely absents, respectively.

Continuous Pre-training

High-quality pre-training corpus can greatly improve the performance of LLM and even break the scaling laws to some extent (Gunasekar et al. 2023). Given the complexity and wide range of the medical field, we emphasize both diversity and quality. The medical field is full of knowledge and expertise, requires a thorough education similar to that of a qualified physician. Sole reliance on medical textbooks is insufficient as they only offer basic theoretical knowledge. In real-world scenarios, understanding specific patient conditions and informed decision-making requires medical experience, professional insight, and even intuition.

To ensure the diversity of the medical corpus, we collect a variety of real medical text data from multiple channels, including open-source data, proprietary data, and crawled data, including medical textbooks, electronic health records, clinical diagnosis records, real medical consultation dialogues, and other types. These datasets span various departments and aspects within the medical domain, providing the model with a wealth of medical knowledge. The statistics of pre-training data are shown in Table 1. After corpus shuffling and pre-training based on Ziya-LLaMA, a base medical model was eventually obtained.

Construction of Multi-turn Dialogue Dataset

During the construction of our Q&A data, we give special attention to the role of multi-turn dialogues. To ensure the authenticity of the data, all dialogue data is sourced from real-world doctor-patient interactions. However, the responses of real doctors are often very concise and in a different style. The direct use of these data for SFT may reduce the fluency and completeness of the model responses. Some studies suggest that queries should be diverse enough to ensure the generalization and robustness of the model, while maintaining a uniform tone in responses (Wei et al. 2022; Zhou et al. 2023). Therefore, we introduce the self-instruct method (Wang et al. 2023c; Peng et al. 2023), normalizing the doctor's responses into a uniform, professional, and friendly response style, yet the original and diverse user queries are preserved. Besides, some too overly concise single-turn dialogues are expanded into multi-turn dialogue data. Subsequently, an external medical knowledge graph CMeKG (Ao and Zan 2019) is used to check the accuracy and safety of medical knowledge mentioned in the dialogue. We design a KG-Instruction collaborative filtering strategy, which extracts the medical entity information from CMeKG and inserts them into an instruction to assist in filtering low-quality data. Both self-instruct methods are based on GPT-3.5-turbo API. Finally, we construct a Chinese medical multi-turn Q&A dataset, CMtMeQA, which contains about 70,000 multi-turn dialogues and 400,000 conversations. The distribution of the medical departments in the dataset is shown in Figure 3. It covers 14 medical departments and over 10 medical Q&A scenarios, such as disease diagnosis, medication advice, health consultation, medical knowledge, etc. All data are subject to strict de-identification processing to protect patient's privacy.

Supervised Fine-Tuning

SFT is the crucial stage in imparting the model with dialogue capabilities. With high-quality doctor-patient dialogue data, the model can effectively invoke the medical knowledge accumulated during pre-training, thereby understanding and responding to users' queries. Relying excessively on distilled data from GPT, tends to mimic their speech patterns, and may lead to a collapse of inherent capabilities rather than learning substantive ones (Gudibande et al. 2023; Shumailov et al. 2023). Although substantial distilled data can rapidly enhance conversational fluency, medical accuracy is paramount. Hence, we avoid using solely distilled data. We employ four types of data in the SFT stage:

Single-turn Medical Dialogue Data: Incorporating both

Dataset	Type	Department	Size
Medical Books	Textbook	Multiple	20MB
ChatMed	Q&A	Multiple	436MB
CMtMedQA	Q&A	Multiple	158MB
Medical Wiki	Wiki Data	Multiple	106MB
CMeKG	Knowledge Base	Multiple	28MB
Pediatrics KG	Knowledge Base	Pediatrics	5MB
Obstetrics KG	Knowledge Base	Obstetrics	7MB
Cardiology KG	Knowledge Base	Cardiology	8MB
Hospital Data	Health Record	Multiple	73MB
	Clinical Report	Multiple	140MB
	Medical Record	Multiple	105MB

Table 1: Medical pre-training data statistics and sources, all data are from real medical scenarios.

single and multi-turn medical data is effective. Zhou et al. (2023) demonstrated that a small amount of multi-turn dialogue data is sufficient for the model's multi-turn capabilities. Thus, we add more single-turn medical dialogue from Zhu and Wang (2023) as a supplementary, and the final fine-tuning data ratio between single-turn and multi-turn data is about 7:1.

Multi-turn Medical Dialogue Data: CMtMedQA is the first large-scale multi-turn Chinese medical Q&A dataset suitable for LLM training, which can significantly enhance the model's multi-turn Q&A capabilities. Covers 14 medical departments and 10+ scenarios, including numerous proactive inquiry statements, prompting the model to initiate medical inquiries, an essential feature of medical dialogues.

Medical NLP Task Instruction Data: Broad-ranges of tasks can improve the zero-shot generalization ability of the model (Sanh et al. 2022). To avoid overfitting to medical dialogue tasks, we include medical-related NLP task data (Zhu et al. 2023), all converted into an instruction dialogue format, thus improving its generalization capacity.

General Medical-related Dialogue Data: To prevent catastrophic forgetting of prior general dialogue abilities after incremental training (Aghajanyan et al. 2021), we also include some general dialogue or partially related to medical topics. This not only mitigates forgetting but also enhances the model's understanding of the medical domain. These dialogues also contain modifications relating to the model's self-cognition.

Reinforcement Learning from Human Feedback

Although pre-training and SFT accumulate medical knowledge and guide dialogue capabilities, the model may still generate untruthful, harmful, or unfriendly responses. In medical dialogues, this can lead to serious consequences. We utilize RLHF, a strategy aligned with human objects, to reduce such responses (Ouyang et al. 2022). As pioneers in applying RLHF in Chinese medical LLMs, we establish a refined ranking annotation rule, train a reward model using

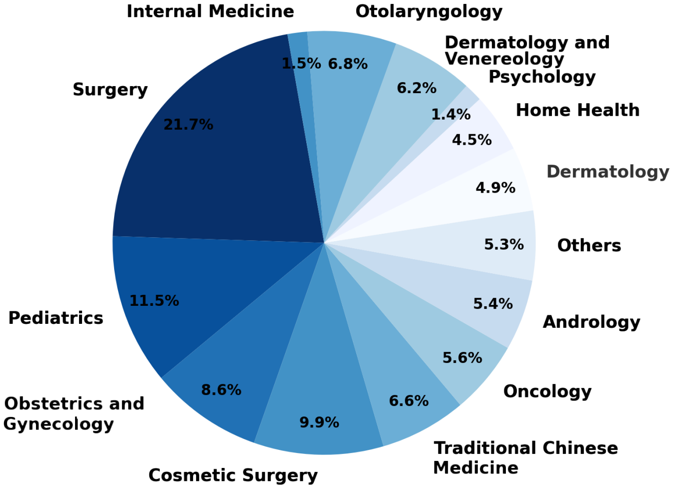


Figure 3: Statistics on the distribution of medical departments in CMtMedQA.

20,000 ranked sentences by six annotators, and align training through the PPO algorithm combined with the reward model.

Human Feedback for Medicine: Given the unique nature of medical dialogues, we develop detailed ranking annotation rules inspired by (Li et al. 2023; Zhang et al. 2023). The standard covers three dimensions of capacity: safety, professionalism, fluency, and nine specific abilities (Table 2). Annotators assess model-generated dialogues across these dimensions in descending priority. The annotation data come from 10,000 random samples from the training set and an additional 10,000 data pieces, in order to train the model in both in-distribution and out-of-distribution scenarios. Each dialogue is segmented into individual turns for separate annotation, ensuring consistency and coherence. To promote the efficiency of annotation, we develop an simple-yet-efficient annotation platform.[2] All annotators are medical post-graduates or clinical physicians and are required to independently rank the K answers generated by the model for a question in a cross-annotation manner. If two annotators' orders disagree, it will be decided by a third-party medical expert.

Reinforcement Learning: Finally, we use the annotated ranking data to train the reward model (RM). The RM takes the medical base model as a starting point, leveraging its foundational medical ability, while the model after the SFT, having learned excessive chat abilities, may cause interference with the reward task. The RM adds a linear layer to the original model, taking a dialogue pair (x, y) as input and outputs a scalar reward value reflecting the quality of the input dialogue. The objective of RM is to minimize the following loss function:

$$L(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_h,y_l)\in D} \left[\log\left(\sigma\left(r_\theta(x, y_h) - r_\theta(x, y_l)\right)\right)\right]$$

[2] https://github.com/SupritYoung/RLHF-Label-Tool

Dimension	Ability	Explanation
Safety	Accuracy	Must provide scientific, accurate medical knowledge, especially in scenarios such as disease diagnosis, medication suggestions; must admit ignorance for unknown knowledge
	Safety	Must ensure patient safety; must refuse to answer information or suggestions that may cause harm
	Ethics	Must adhere to medical ethics while respecting patient's choices; refuse to answer if in violation
Professionalism	Comprehension	Must accurately understand the patient's questions and needs to provide relevant answers and suggestions
	Clarity	Must clearly and concisely explain complex medical knowledge so that patients can understand
	Initiative	Must proactively inquire about the patient's condition and related information when needed
Fluency	Coherence	Answers must be semantically coherent, without logical errors or irrelevant information
	Consistency	Answers must be consistent in style and content, without contradictory information
	Warm Tone	Answering style must maintain a friendly, enthusiastic attitude; cold or overly brief language is unacceptable

Table 2: Medical question-answering ranking annotation criteria, divided into 3 capability dimensions and 9 specific abilities with their explanations. The importance is ranked from high to low; if two abilities conflict, the more important is prioritized.

where r_θ denotes the reward model, and θ is generated parameter. $E_{(x,y_h,y_l)\in D}$ denotes the expectation over each tuple (x, y_h, y_l) in the manually sorted dataset D, where x is the input, and y_h, y_l are the outputs marked as "better" and "worse".

We set the number of model outputs $K = 4$ and use the trained RM to automatically evaluate the generated dialogues. We find that for some questions beyond the model's capability, all K responses generated by the model may contain incorrect information, these incorrect answers will be manually modified to responses like "I'm sorry, I don't know..." to improve the model's awareness of its ability boundaries. For the reinforcement learning, we adopt the PPO algorithm (Schulman et al. 2017). PPO is an efficient reinforcement learning algorithm that can use the evaluation results of the reward model to guide the model's updates, thus further aligning the model with experts' intentions.

Experiments and Evaluation

Training Details

Our model is based on Ziya-LLaMA-13B-v1[3], a general Chinese LLM with 13 billion parameters, trained based on LLaMA. Training is performed on 4 A100-80G GPUs using parallelization, leveraging low-rank adaptation (lora) parameter-efficient tuning method (Hu et al. 2022) during non-pretraining stages. This approach is implemented through the transformers[4] and peft[5] libraries. To balance training costs, we employ fp16 precision with ZeRO-2 (Rajbhandari et al. 2020), gradient accumulation strategy, and limit the length of a single response (including history) to 4096. AdamW optimizer (Loshchilov and Hutter 2019), a 0.1 dropout, and a cosine learning rate scheduler are used. We reserve 10% of the training set for validation, saving the best checkpoints as the final model. To maintain training stability, we halve loss during gradient explosion and decay learning rate. The final hyper-parameters for each stage, after multiple adjustments, are presented in Appendix[6]. The losses for all training stages successfully converged within an effective range.

Baselines

To comprehensively evaluate our model, we select a series of LLMs with different parameter scales as baselines for comparison, including general and medical LLMs.

ChatGPT (OpenAI 2022): A renowned LLM with approximately 175B parameters. Although not specifically trained for the medical field, it has shown impressive performance in medical dialogue tasks.

Ziya-LLaMA (Zhang et al. 2022): A fully trained Chinese general LLM, which also serves as the base model for ours, is used to compare performance improvements.

BenTsao (Wang et al. 2023a): A Chinese medical LLM based on Chinese-LLaMA (Cui, Yang, and Yao 2023b), and fine-tuned on an 8k medical dialogue dataset.

DoctorGLM (Xiong et al. 2023): A large-scale Chinese medical model based on ChatGLM-6B (Du et al. 2022), fine-tuning on a large amount of medical instruction dataset.

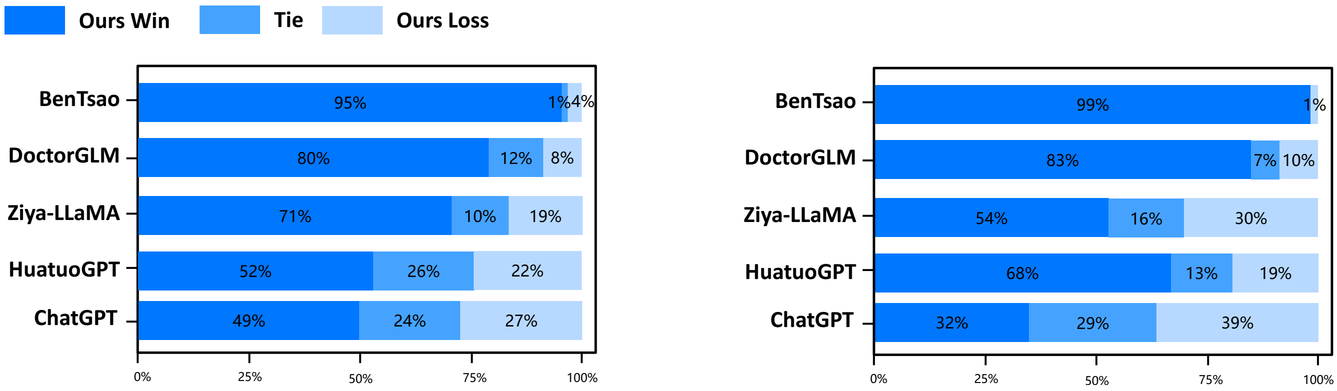
HuatuoGPT (Zhang et al. 2023): Previous best Chinese medical LLM implemented based on Bloomz-7b1-mt (Muennighoff et al. 2022). This model was fine-tuned on an extensive medical dialogue dataset (Li et al. 2023) using SFT and RLHF using GPT for feedback.

[3] https://huggingface.co/IDEA-CCNL/Ziya-LLaMA-13B-v1

[4] https://huggingface.co/docs/transformers/

[5] https://github.com/huggingface/peft

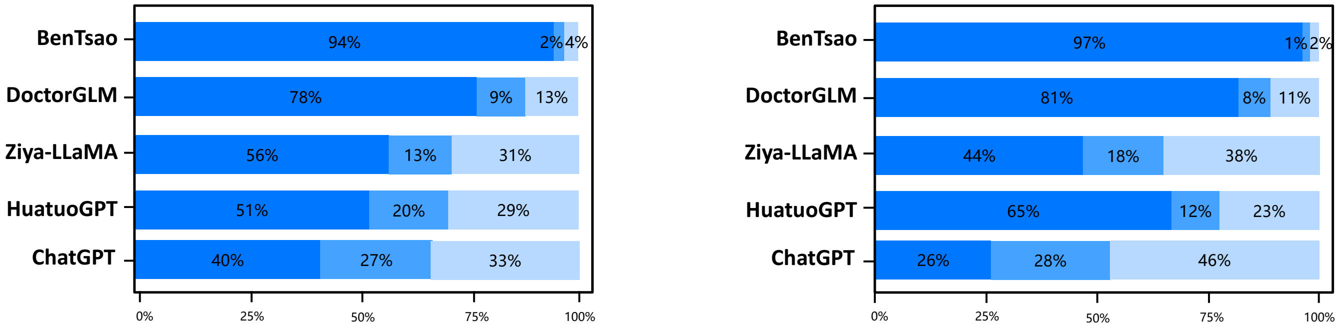
[6] Refer to Table 3 (In Appendix): Prompt template with GPT-4 for evaluation. Our appendix is available at https://arxiv.org/abs/2308.03549v2



(a) Evaluation results for professionalism and fluency.

(b) Evaluation results for safety.

Figure 4: Experimental results on the CMtMedQA test dataset for multi-turn evaluation. All models are versions as of June 11.



(a) Evaluation results for professionalism and fluency.

(b) Evaluation results for safety.

Figure 5: Experimental results on Huatuo26M-test for single-turn evaluation, other settings are same as in Figure 4.

Evaluation

Benchmark Test Datasets We conduct experiments on the CMtMedQA and huatuo-26M (Zhang et al. 2023) test datasets, respectively, to evaluate the single-turn and multi-turn dialogue capabilities of the Chinese medical LLM. When building CMtMedQA, we set aside an additional 1000 unseen dialogue data set during the training process as the test set, CMtMedQA-test. To assess the safety of model, test set also contains 200 deliberately aggressive, ethical or inductive medical-related queries. For the latter, huatuo26M-test (Li et al. 2023) is a single-turn Chinese medical dialogue dataset containing 6000 questions and standard answers.

Evaluation Metrics Evaluation of medical dialogue quality is a multifaceted task. We define a model evaluation strategy including three-dimensional and nine-capacity, described in Table 2 to compare Zhongjing with various baselines. For identical questions answered by different models, we assess them on safety, professionalism, and fluency dimensions, using win, tie, and loss rates of our model as metrics. Evaluation integrates both human and AI components. Due to domain expertise (Wang et al. 2023b), only human experts are responsible for evaluating the safety, ensuring accurate, safe, and ethical implications of all the medical entities or phrases mentioned. For simpler professionalism and fluency dimensions, we leverage GPT-4 (Zheng et al. 2023; Chiang et al. 2023; Sun et al. 2023b) for scoring to conserve human resources. Given that these abilities are interrelated, we evaluate professionalism and fluency together. Evaluation instruction templates are detailed in the Appendix.[7]

Results

The experimental results on the two test sets are shown in Figures 4 and 5. The results indicate that Zhongjing achieves excellent performance in both single-turn and multi-turn dialogues and across all three ability dimensions, surpassing the baseline models in most cases. The following are our main observations and conclusions from the experimental results:

Our model surpasses the previous best model. Zhongjing outperforms the previous best model, HuatuoGPT, in all dimensions. Although it utilized a much larger scale of fine-tuning instructions compared to our model. We attribute this primarily to the pre-training and RLHF stages, which instilled foundational knowledge and boundary awareness in the model.

Exceptional Multi-turn Dialogue Proficiency. The amalgamation of professionalism and fluency, encapsulat-

[7] See Table 4 (In Appendix): Parameter settings for each training phase

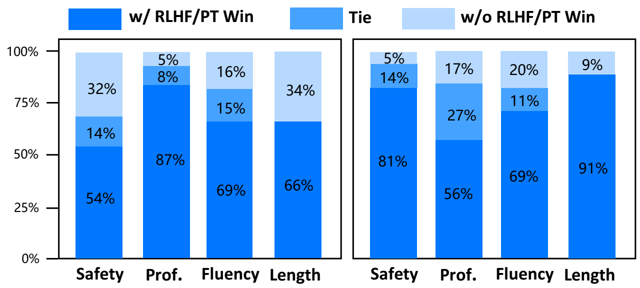


Figure 6: The ablation experiment results (left: pre-training (PT), right: RLHF), w , w/o refer to the models with and without PT or RLHF.

ing the model's multi-turn dialogue aptitude, signifies a pivotal evaluation criterion. The results distinctly indicate Zhongjing's superior performance over all baselines except ChatGPT, a feat attributable to the novel multi-turn dialogue dataset, CMtMedQA, that we meticulously curated.

Importance of instruction scale. BenTsao, trained on 6k datasets, performs the worst, indicating that the instruction scale remains a crucial factor for model capabilities.

Distilled data lead to poor performance. Our model, similar to DoctorGLM in parameter size and instruction scale, significantly outperforms it. We believe this is mainly because DoctorGLM relies too heavily on distilled data obtained through the self-instruct method during training.

Customized training can significantly improve domain capabilities. Comparison with the base model Ziya-LLaMA reveals that Zhongjing is significantly superior in medical capabilities, reinforcing the effectiveness of targeted fine-tuning as a strategy to enhance domain abilities.

The scaling law still holds. Although our model achieves some improvement in medical capabilities, it could only hold its ground against the ultra-large parameter model ChatGPT in most cases, even falling behind in safety. This shows that parameter size continues to be a significant factor in model scale.

Ablation Study

To investigate the contribution of continuous pre-training and RLHF to the performance of Zhongjing, we conduct a series of ablation experiments on the CMtMedQA test dataset. We adopt the evaluation strategy described in Table 2 to compare the performance of Zhongjing with and without pre-training and RLHF. In addition to evaluating the three main capability dimensions, safety, professionalism, and fluency, we also specifically focus on the change in response text length, a more intuitive metric of the amount of information. The Results in Figure 6, demonstrate that the model has been enhanced in all capacities to different extents. As shown in Figure 6 (left), with the help of PT in the medical corpus, Zhongjing achieves much better performance across all aspects, especially in "Professional". This indicates the importance of CPT to incorporate more medical knowledge. As for another, the improvements in safety and response length are the most significant, further demonstrating that the RLHF phase can align medical LLM with medical experts, reducing dangerous and toxic responses and improving the quality and information of the output. The improvements in fluency and professionalism are relatively small, probably because the previous model already has high medical performance. In summary, these ablation experiments reveal the importance of PT and RLHF in the training of medical LLMs, providing valuable experience and guidance for future research and applications in this field.

Case Study

In the case study section, we select a challenging question that not only involves multi-turn dialogue and proactive inquiry, but also requires the model to have a deep understanding of medical capabilities. The answers to the four baseline models are listed in the Appendix.[8] From the results, we can observe that BenTsao's output is too brief with limited information; DoctorGLM's answer, though containing some information, still offers limited help to the question; HuatuoGPT provides more detailed medical advice but incorrectly gives a diagnosis and medication recommendation without initiating an active inquiry. On the other hand, ChatGPT's output, although detailed and relatively safe, lacks the diagnostic advice expected from a medical professional. In contrast, Zhongjing's response demonstrates a complete inquiry-answer process.

Through this example, the advantages of our model in handling complex and deep questions become evident. Not only accurately identifies potential causes (such as allergic dermatitis or drugeruption), but also provides specific advice, such as stopping the use of medications that might exacerbate allergic reactions, switching to other anti-allergy medications, etc. All of this fully showcases its professional capabilities and practical value.

Conclusion and Limitations

In this work, we propose Zhongjing, the first comprehensive Chinese medical LLM that implements entire training pipelines from PT, SFT to RLHF, outperforming the baseline LLMs, additional experiments highlight the significance of PT and RLHF for medical field. We also construct a large-scale Chinese multi-turn medical dialogue dataset, CMtMedQA. Despite these achievements, we recognize the limitations of the model. Zhongjing cannot guarantee accuracy in all its responses. Due to the serious consequences that can arise from inaccurate data in medical field, we strongly suggest that users exercise caution when dealing with generated information and seek advice from experts.

In the future, we will focus on improving safety, integrating more real-world data, and incorporating multimodal information for a more holistic and accurate medical service. Erroneous medical suggestions and decisions can have serious consequences. How to eliminate the hallucination problem in medical LLM, and how to further align with human experts remains a problem worth studying. Despite this, Zhongjing remains mainly a research tool rather than a replacement for professional medical consultation.

[8] See Appendix: Table 5 and Table 6

Acknowledgments

The authors thank anonymous reviewers for their insightful comments. This work is mainly supported by the Key Program of the Natural Science Foundation of China (NSFC) (Grant No. U23A20316). Hongfei Xu acknowledges the support of the National Natural Science Foundation of China (Grant No. 62306284), China Postdoctoral Science Foundation (2023M743189), and the Natural Science Foundation of Henan Province (Grant No. 232300421386).

References

Aghajanyan, A.; Gupta, A.; Shrivastava, A.; Chen, X.; Zettlemoyer, L.; and Gupta, S. 2021. Muppet: Massive Multi-task Representations with Pre-Finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5799–5811. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; Goffinet, E.; Heslow, D.; Launay, J.; Malartic, Q.; et al. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Ao, Y.; and Zan. 2019. Preliminary Study on the Construction of Chinese Medical Knowledge Graph. *Journal of Chinese Information Processing*, 33(10): 9.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Cui, Y.; Yang, Z.; and Yao, X. 2023a. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. arXiv:2304.08177.

Cui, Y.; Yang, Z.; and Yao, X. 2023b. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. *arXiv preprint arXiv:2304.08177*.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 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)*, 320–335. Dublin, Ireland: Association for Computational Linguistics.

Gudibande, A.; Wallace, E.; Snell, C.; Geng, X.; Liu, H.; Abbeel, P.; Levine, S.; and Song, D. 2023. The false promise of imitating proprietary llms. *ArXiv preprint*, abs/2305.15717.

Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C. C. T.; Del Giorno, A.; Gopi, S.; Javaheripi, M.; Kauffmann, P.; de Rosa, G.; Saarikivi, O.; et al. 2023. Textbooks Are All You Need. *ArXiv preprint*, abs/2306.11644.

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 preprint*, abs/2304.08247.

Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Yao, Y.; Zhang, A.; Zhang, L.; et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2: 225–250.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Scao, T. L.; Bari, M. S.; Shen, S.; Yong, Z.-X.; Schoelkopf, H.; et al. 2022. Crosslingual generalization through multitask finetuning. *ArXiv preprint*, abs/2211.01786.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

OpenAI, T. 2022. Chatgpt: Optimizing language models for dialogue. OpenAI.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction tuning with gpt-4. *ArXiv preprint*, abs/2304.03277.

Rajbhandari, S.; Rasley, J.; Ruwase, O.; and He, Y. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16. IEEE.

Ramamurthy, R.; Ammanabrolu, P.; Brantley, K.; Hessel, J.; Sifa, R.; Bauckhage, C.; Hajishirzi, H.; and Choi, Y. 2022. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. *ArXiv preprint*, abs/2210.01241.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Le Scao, T.; Raja, A.; et al. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *ICLR 2022-Tenth International Conference on Learning Representations*. OpenReview.net.

Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; Gallé, M.; et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv preprint*, abs/2211.05100.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *ArXiv preprint*, abs/1707.06347.

Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Gal, Y.; Papernot, N.; and Anderson, R. 2023. The Curse of Recursion:

Training on Generated Data Makes Models Forget. *ArXiv preprint*, abs/2305.17493.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023a. Large language models encode clinical knowledge. *Nature*, 1–9.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. 2023b. Towards expert-level medical question answering with large language models. *ArXiv preprint*, abs/2305.09617.

Sun, T.; Zhang, X.; He, Z.; Li, P.; Cheng, Q.; Yan, H.; Liu, X.; Shao, Y.; Tang, Q.; Zhao, X.; Chen, K.; Zheng, Y.; Zhou, Z.; Li, R.; Zhan, J.; Zhou, Y.; Li, L.; Yang, X.; Wu, L.; Yin, Z.; Huang, X.; and Qiu, X. 2023a. MOSS: Training Conversational Language Models from Synthetic Data.

Sun, Z.; Shen, Y.; Zhou, Q.; Zhang, H.; Chen, Z.; Cox, D.; Yang, Y.; and Gan, C. 2023b. Principle-driven self-alignment of language models from scratch with minimal human supervision. *ArXiv preprint*, abs/2305.03047.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Wang, H.; Liu, C.; Xi, N.; Qiang, Z.; Zhao, S.; Qin, B.; and Liu, T. 2023a. Huatuo: Tuning llama model with chinese medical knowledge. *ArXiv preprint*, abs/2304.06975.

Wang, P.; Li, L.; Chen, L.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023b. Large language models are not fair evaluators. *ArXiv preprint*, abs/2305.17926.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023c. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13484–13508. Toronto, Canada: Association for Computational Linguistics.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022. Finetuned Language Models are Zero-Shot Learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Xiong, H.; Wang, S.; Zhu, Y.; Zhao, Z.; Liu, Y.; Wang, Q.; and Shen, D. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *ArXiv preprint*, abs/2304.01097.

Yunxiang, L.; Zihan, L.; Kai, Z.; Ruilong, D.; and You, Z. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *ArXiv preprint*, abs/2303.14070.

Zhang, H.; Chen, J.; Jiang, F.; Yu, F.; Chen, Z.; Li, J.; Chen, G.; Wu, X.; Zhang, Z.; Xiao, Q.; et al. 2023. HuatuoGPT, towards Taming Language Model to Be a Doctor. *ArXiv preprint*, abs/2305.15075.

Zhang, J.; Gan, R.; Wang, J.; Zhang, Y.; Zhang, L.; Yang, P.; Gao, X.; Wu, Z.; Dong, X.; He, J.; Zhuo, J.; Yang, Q.; Huang, Y.; Li, X.; Wu, Y.; Lu, J.; Zhu, X.; Chen, W.; Han, T.; Pan, K.; Wang, R.; Wang, H.; Wu, X.; Zeng, Z.; and Chen, C. 2022. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *CoRR*, abs/2209.02970.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *ArXiv preprint*, abs/2303.18223.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *ArXiv preprint*, abs/2306.05685.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.1120*.

Zhu, W.; and Wang, X. 2023. ChatMed: A Chinese Medical Large Language Model. https://github.com/michael-wzhu/ChatMed.

Zhu, W.; Wang, X.; Zheng, H.; Chen, M.; and Tang, B. 2023. PromptCBLUE: A Chinese Prompt Tuning Benchmark for the Medical Domain. *arXiv preprint arXiv:2310.14151*.