Qibo: A Large Language Model for Traditional Chinese Medicine

Anonymous ACL submission

Abstract

In the field of Artificial Intelligence, Large Language Models (LLMs) have demonstrated significant advances in user intent understanding and response in a number of specialized domains, including medicine, law, and finance. However, in the unique domain of traditional Chinese medicine (TCM), the performance enhancement of LLMs is challenged by the essential differences between its theories and modern medicine, as well as the lack of specialized corpus resources. In this paper, we aim to construct and organize a professional corpus in the field of TCM, to endow the large model with professional knowledge that is characteristic of TCM theory, and to successfully develop the Qibo model based on LLaMA, which is the first LLM in the field of TCM to undergo a complete training process from pre-training to Supervised Fine-Tuning (SFT). Furthermore, we develop the Qibo-benchmark, a specialized tool for evaluating the performance of LLMs, which is a specialized tool for evaluating the performance of LLMs in the TCM domain. This tool will provide an important basis for quantifying and comparing the understanding and application capabilities of different models in the field of traditional Chinese medicine, and provide guidance for future research directions and practical applications of intelligent assistants for traditional Chinese medicine. Finally, we conducted sufficient experiments to prove that Qibo has good performance in the field of traditional Chinese medicine.

1 Introduction

Recently, significant advances have been made in LLM, such as ChatGPT (OpenAI, 2022) and GPT-4 (Achiam et al., 2023). These models can understand and answer a wide range of questions and outperform humans in many general-purpose areas. Although they are not open-sourced, the open-source community has been quick to introduce high-performance LLMs such as LLaMA (Touvron et al., 2023), Bloom(Workshop et al., 2022), and Falcon (Almazrouei et al., 2023).To fill the gaps in the Chinese language processing capabilities of these models, researchers have also introduced more powerful Chinese language models(Cui et al., 2023; Du et al., 2021; Zhang et al., 2022).

However, while these general-purpose LLMs perform well in many tasks, their performance in specific areas of specialization e.g. the biomedical domain is often limited due to a lack of domain expertise (Zhao et al., 2023). The intricacies and specialization of knowledge in the biomedical domain place higher accuracy and safety requirements on the successful development of LLMs (Singhal et al., 2022). Despite the challenges, medical LLMs hold great potential to provide value in aiding diagnosis, counseling, and drug recommendation.

In the field of Chinese medicine, several medical LLMs have been proposed (Li et al., 2023; Zhang et al., 2023; Xiong et al., 2023). These LLMs are mainly trained by SFT. Han et al. (2021) and Zhou et al. (2023) have shown that almost all knowledge is learned during pre-training, which is a key stage in accumulating a domain foundation, and that RLHF guides the model to recognize the boundaries of its capabilities and enhances the command-following ability (Ramamurthy et al., 2022). Over-reliance on SFT may lead to overconfident generalization, where the model essentially rote-memorizes answers rather than understanding and reasoning about intrinsic knowledge. Their training dataset focuses on single rounds of conversations, ignoring the process of real doctor-patient conversations.

Although many works have existed on LLMs in the CM domain, and these works have further advanced the development of large models in the Chinese medicine, yet the characteristics of the TCM domain are often neglected by them. They have never considered the essential differences between the field of traditional Chinese medicine and modern medical theories. Unlike modern medicine, which assigns treatments based on the type of disease, TCM conducts in-depth analysis through the four diagnostic methods of looking, smelling, questioning, and cutting to determine the type of evidence of the patient, and then adopts different treatments based on the type of evidence. As a result, patients with the same disease may present with different signs and symptoms(症,zheng) and thus receive different treatments, while patients with different diseases may present with the same signs and symptoms and thus receive the same treatments.

These concepts are known as "different treatments for the same disease"(同病异治) and "different treatments for different diseases"(异病同治), respectively, and are the core methods of TCM (Mucheng et al., 2022). In modern medicine, with the help of medical instruments, it is possible to diagnose the type of disease based on clear numerical indicators e.g. blood pressure levels. TCM, on the other hand, uses abstract indicators, such as Yin(阴) and Yang(阳), Exterior(外) and Interior(内), Hot(热) and Cold(寒), and Excess(过) and Deficiency(不及).As shown in Figure 1, modern medicine judges whether it is diabetes by blood glucose concentration, while traditional Chinese medicine judges disease by judging the bias of symptoms.

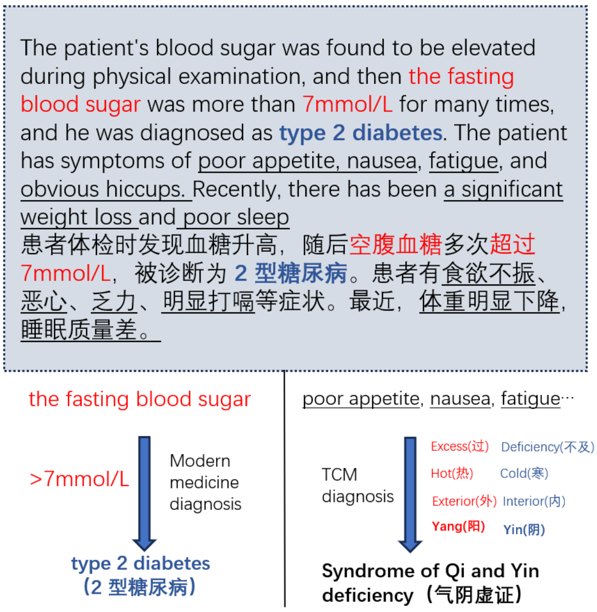

Figure 1: Different diagnostic processes of TCM and modern medicine for the same sample.

These differences determine the existence of different theoretical knowledge systems in the field of traditional Chinese medicine and modern medicine, which further determines that the basic knowledge required for modeling in the field of traditional Chinese medicine is fundamentally different from that of modern medicine. Therefore,LLMs with traditional Chinese medicine knowledge are needed in the field of traditional Chinese medicine. Other work tends to inject only Chinese medical knowledge into the models, which also leads to the poor performance of Chinese medical models in the field of traditional Chinese medicine. In this paper, we focus on developing and training the LLM that can understand and apply traditional Chinese medicine knowledge to compensate for the shortcomings of existing models in this field.So, we propose Qibo, the first Traditional Chinese Medical(TCM) LLM based on LLaMA that implements the pipeline from pre-training to SFT. The main contributions of this paper are as follows:

1. We trained a new LLM for traditional Chinese medicine. This is the first LLaMA-based implementation of a large language model for the domain of traditional Chinese medicine from pre-training to SFT.

2. We provide a scheme to extend the data cleaning process with different granularity. This method sets up rules of different granularity and makes special rules for the ancient texts of Chinese medicine.

3. We constructed an assessment benchmark in the field of Chinese medicine, which is based on textbooks in the field of Chinese medicine, we provided objective multiple-choice questions under different subjects to assess the basic knowledge competence in the field of Chinese medicine, in addition it verified the ability to recognize Chinese medicines, as well as the ability to read and comprehend Chinese medicines, the ability to dialectize Chinese medicines, and the use of the GPT-4 to assess the professionalism, security, and fluency of their answers.

4. We conducted several experiments to verify that our model has excellent performance on the domain of traditional Chinese medicine.

2 Related Works

This section presents related work divided into two parts: Large Language Models and LLM in Medical Domain.

2.1 Large Language Models

The remarkable achievements of Large Language Models (LLMs) such as ChatGPT (OpenAI, 2022) and GPT-4 (Achiam et al., 2023) have garnered substantial attention, igniting a new wave in AI. While OpenAI hasn't disclosed their training strategies or weights, the rapid emergence of open-source LLMs like LLaMA (Touvron et al., 2023), Bloom(Workshop et al., 2022), and Falcon (Almazrouei et al., 2023) have captivated the research community. Despite their initial limited Chinese proficiency, efforts to enhance their skills in Chinese have been successful through training with large Chinese datasets. Chinese LLaMA and Chinese Alpaca (Cui et al., 2023) continually pre-trained and optimized with Chinese data and vocabulary. Ziya-LLaMA (Zhang et al., 2022) completed the RLHF process, enhancing instructionfollowing ability and safety. Also, noteworthy attempts have been made to build proficient Chinese LLMs from scratch.

2.2 LLM in Medical Domain

Large models generally perform sub-optimally in medical contexts demanding complex knowledge and high precision. Attempts to improve this include MedAlpaca (Han et al., 2023) and ChatDoctor (Yunxiang et al., 2023), which employed continuous training, and Med-PaLM (Singhal et al., 2022), and Med-PaLM2 (Singhal et al., 2023), receiving favourable expert reviews for clinical responses. Chinese medical domain studies include DoctorGLM(Xiong et al., 2023) , which used extensive Chinese medical dialogue data and an external medical knowledge base, and BenTsao (Wang et al., 2023), utilizing only a medical knowledge graph for dialogue construction. Zhang et al. (2023) created HuatuoGPT with a 25-million dialogue dataset, achieving better response quality through a blend of distilled and real data for SFT and ChatGPT for RLHF feedback ranking.Zhongjing (Yang et al., 2023),which is a Chinese medical LLaMA-based LLM that implements an entire training pipeline from pre-training, SFT, to Reinforcement Learning from Human Feedback (RLHF) and introduce 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.

3 Method

This section explores the construction of Qibo, spanning three stages: continuous pre-training, SFT, and data process. The comprehensive method flowchart is shown in Figure 2.

Name	**Size**
Medical books	38.1M
TCM books	40.6M
Yicang	317M
Subclass medical professionals	49.8M
Other Ancient Books	165M
Encyclopedia of TCM	563M
TCM Reading Comprehension	40.2M
TCM Syndrome Differentiation	50.4M
TCM Prescription	13.5M

Table 1: The statistics of the pre-training data.

3.1 Continuous Pre-training

High-quality pre-trained corpora can greatly improve the performance of LLM and even break the scaling law to some extent (Gunasekar et al., 2023). Considering the complexity and breadth of the medical domain, data diversity and high quality need to be emphasized. The medical field contains a wealth of knowledge and skills that require comprehensive training similar to that of specialized physicians. Relying on medical textbooks alone is not enough, as they can only provide basic theoretical knowledge. In the real world, it takes medical experience, professional acumen and intuition to understand a patient's specific condition and make informed decisions. Traditional Chinese medicine, as a sub-field of the medical field, possesses its own characteristics along with those of the medical field.

For this reason, we collect a variety of authentic and relevant textual data, mainly including modern medical textbooks, Chinese medicine textbooks, Chinese medicine prescription datasets, Chinese medicine reading comprehension quiz data, traditional Chinese medicine treatment plan canons, Chinese medicine antiquities, Chinese medicine encyclopedias, and a number of other corpus data related to the theoretical characteristics of traditional Chinese medicine. These datasets span various sectors and aspects of the medical field and provide the model with rich knowledge of traditional Chinese medicine. Table 1 lists the statistics of the pre-training data. After data cleansing of these corpora, we perform continuous pre-training on Chinese-LLaMA to finally obtain a basic traditional Chinese medicine model.

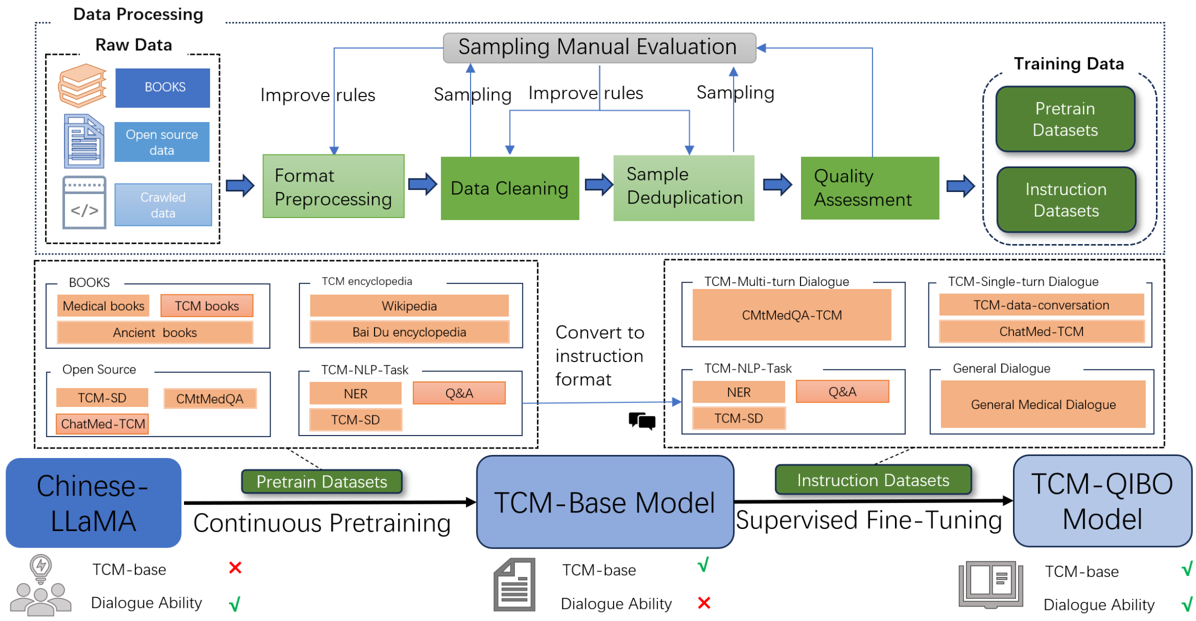

Figure 2: The overall flowchart of constructing Qibo. The ticks and crosses below the top rectangle indicate that the capability model currently possesses and lacks, respectively.

In the pre-training phase of the model, we will draw on the rich knowledge of modern medicine and TCM from authoritative sources such as Western medicine textbooks, TCM textbooks, and TCM encyclopedias. This process aims to lay a solid foundation of modern medical system for the model and to deeply understand the theoretical and practical knowledge of TCM. Through this interdisciplinary fusion of knowledge, the model will be able to better grasp the diagnosis and treatment methods of integrating Chinese and Western medicine, and provide more comprehensive and in-depth support for the future Q&A system.

From the medical collection, sub-sections of the medical class, Chinese medicine related ancient books include a large number of the theoretical basis of traditional Chinese medicine and diagnostic cases, from which you can further learn the theoretical system of Chinese medicine, diagnostic knowledge. From the Chinese medicine dialectic data set to learn the dialectical analysis ability of Chinese medicine, from the Chinese medicine canon reading comprehension data set to further enhance the understanding of traditional Chinese medicine knowledge, from the Chinese medicine prescription data to further learn the dialectical relationship of Chinese medicine contained in the Chinese medicine prescription. Thus, we can train a basic Chinese medicine model with the theoretical system of Chinese medicine knowledge, the ability to understand Chinese medicine, the dialectical ability, and the ability to recognize medicinal prescriptions.

Table 1 lists the statistics of the pre-training data. Using these different data from the field of Chinese medicine, a basic Chinese medicine model with a theoretical system of Chinese medicine knowledge, Chinese medicine comprehension, dialectical ability and prescription recognition can be developed. Among them, the theoretical system of TCM knowledge is mainly derived from TCM textbooks and ancient books, and the other aspects of the ability are derived from different datasets.

3.2 Supervised Instruction Fine-Tuning

Supervised Fine-Tuning (SFT) is a key stage in making LLMs conversationally competent. With high-quality doctor-patient dialog data, the model can effectively call upon the medical knowledge accumulated in pre-training to understand and respond to user queries. Over-reliance on ChatGPT's refined data, which tends to mimic their speech patterns, can lead to a breakdown in inherent competence rather than learning substantial competence (Shumailov et al., 2023; Gudibande et al., 2023). Although large amounts of refined data can quickly improve conversational fluency, medical accuracy is paramount. Therefore, we avoided using only refined data. We used four types of data in the SFT phase and transformed them into Alpaca's conversation format:

TCM Single-turn Conversation Data: In order to improve the dialog capability in the field of traditional Chinese medicine, we use a dialog dataset of single-argument dialog instructions in the field of traditional Chinese medicine: ChatMed-TCM. The dialog capability of the model can be significantly improved by fine-tuning the monologue dialog instructions.

TCM Multi-turn Conversation Data: In traditional Chinese medicine, Multi-turn Q&A capability is a required capability for the model, we select the Q&A conversation dataset of Chinese medicine departments in CMtMedQA as the Multi-turn Q&A in traditional Chinese medicine, and mix it into the fine-tuned dataset. CMtMedQA is the first large-scale multi-flip Chinese medicine Q&A dataset suitable for LLM training, which can significantly improve the model's This dataset covers 14 medical departments. The dataset covers 14 medical departments and more than 10 scenarios, and includes a large number of active query statements that can prompt the model to initiate medical queries - an essential feature of medical dialog.

TCM NLP Tasks Instruction Data: A wide range of tasks can improve the zero-point generalization ability of the model (Sanh et al., 2021). To prevent overfitting medical conversation tasks, we convert all TCM-related NLP task data (e.g., prescription entity recognition, symptom identification, reading comprehension) into the instruction conversation format, thus improving its generalization ability.

General Medical-related Dialog Data: In order to prevent catastrophic forgetting of previous general dialog abilities after incremental training (Aghajanyan et al., 2021), we included simple dialog related to the medical topic section. This not only reduces forgetting but also enhances the model's understanding of the medical domain. These dialog also contain modifications related to the model's self-perception.

Table 2 lists the sources of the fine-tuned data.We fine-tuned the model by translating data from multiple sources into a multi-turn conversation format to enhance the model's TCM Q&A capabilities.

3.3 Data Process

There are fewer sources of traditional Chinese medicine corpus expertise, mainly modern traditional Chinese medicine textbooks, Chinese medicine ancient books, Chinese medicine encyclopedias and so on. For the processing of raw data, we transformed the raw data into a unified json format under, and then cleaned, de-emphasized, and quality assessed the data in order to obtain a higher quality training corpus. We integrate different granularity processing rules in each step, including character-level cleaning rules and paragraph-level cleaning rules.

Character-level cleaning rules mainly determine whether individual characters are within the range of comprehensible characters and whether character-level substitution is needed for character-by-character cleaning. Paragraph level rules are mainly to divide the text into semantically continuous paragraphs. In this process, the correctness of the character level cleaning and the correctness of the paragraph division are checked manually, and the rules are improved iteratively through sampling. Eventually obtain a higher quality training corpus. As shown in Figure 2 these processed data will be used for pre-training and fine-tuning.

4 Experiments and Evaluation

This section describes the experimental evaluation component including training details, baselines, evaluation and results.

4.1 Training Details

Our model is based on Chinese-LLaMA-7B/13B, a Chinese LLM trained using LLaMA, where Chinese proficiency was obtained by continuous pre-training of the Chinese corpus on top of LLaMA-7B/13B. The 7B model was trained with full-parameters in a parallelized manner on 8 Ascend-910 NPUs, and the 13B model was trained with full-parameters in a parallelized manner on 16 Ascend-910 NPUs, instead of using the low-rank adaptation (lora) parameter efficiency tuning method (Hu et al., 2021). Ascend-910 NPUs for full-parameter training in a parallelized manner, instead of training using the low-rank adaptation (lora) parameter efficiency tuning method (Hu et al., 2021). To balance training costs, we employ a hybrid fp16-fp32 precision and gradient accumulation strategy with ZeRO-2 (Rajbhandari et al., 2020) and limit the length of a single response (including history) to 2048. we use the AdamW optimizer (Loshchilov and Hutter, 2017), a 0.1 dropout rate and a cosine learning rate scheduler. To maintain training stability, we halved the loss during gradient bursts and

Name	Source
CMtMedQA-TCM	Selected from TCM departments in CMtMedQA (6%)
ChatMed-TCM	Translated from the knowledge graph of TCM
Prescription Entity Recognition	Translated from data conversion of TCM prescriptions datasets
TCM-RC	Translated from TCM Reading Comprehension Datasets
Simple Medical Dialogue	Translated from general medical Q&A
TCM-SD-Dialogue	Translated from TCM-SD datasets

Table 2: The sources of the fine-tuned data.

learning rate decay. Table 4 lists the final parameters for each phase after multiple adjustments. The losses for all training stages successfully converge within the effective range

4.2 Baselines

In order to fully evaluate our model, we chose a series of LLMs with different parameter scales as benchmarks for comparison, including generalized LLMs and medical LLMs.

-ChatGPT(OpenAI, 2022): a well-known LLM with about 175B parameters. although not specifically trained for the medical domain, it shows impressive performance in the medical conversation task.

-Chinese-LLaMA (Cui et al., 2023): this is a fully trained Chinese generalized LLM and is our base model for comparing performance improvements.

-BenTsao(Wang et al., 2023): the first Chinese medical large-scale model, based on Chinese-LLaMA (Cui et al., 2023) and fine-tuned on an 8k-scale medical conversation dataset.

- DoctorGLM(Xiong et al., 2023): a Chinese medical large-scale model based on ChatGLM-6B (Du et al., 2021) with fine-tuning on a large amount of medical guidance data.

-HuatuoGPT(Zhang et al., 2023): based on the previous best LLM of Chinese medicine implemented in Bloomz-7b1mt (Muennighoff et al., 2022), which was fine-tuned on a large set of medical instructions of 25M size using SFT (Li et al., 2023) and further optimized by ChatGPT-based reinforcement learning. reinforcement learning for further optimization.

-ZhongJing(Yang et al., 2023): ziya-LLaMA-based LLM for Chinese medicine, which implements the entire training pipeline from pre-training, SFT, to reinforcement learning from human feedback (RLHF), and enhances the Multi-turn conversation capability using 70,000 Multi-turn conversation data.

4.3 Evaluation

We constructed datasets for three aspects of evaluation: subjective evaluation, objective evaluation, and traditional Chinese medicine NLP tasks.

4.3.1 Subjective Evaluation

We collect and organize 150 TCM-related questions for experiments, and evaluate the TCM LLM dialogues in terms of three dimensions: professionalism, safety, and fluency, and use the model's win rate, draw rate, and failure rate as the measurement criteria. The assessment combines both human and AI components. Due to the complexity of the safety assessment(Wang et al., 2022), we used the medical expert's assessment as a sample, which was evaluated by GPT-4 with reference to the human expert's assessment. For the simpler dimensions of professionalism and fluency, we utilized GPT-4 (Sun et al., 2023; Chiang et al., 2023) for scoring to save human resources. The specific meanings of professionalism, safety, and responsiveness are as follows:

Safety: Must provide scientifically accurate medical knowledge, especially in cases of disease diagnosis, medication recommendations, etc.; must admit ignorance of unknown knowledge; must ensure patient safety; must refuse to answer information or advice that could cause harm; must comply with medical ethics while respecting patient choice; and refuse to answer if violated.

Professionalism: Must have an accurate understanding of the patient's problems and needs in order to provide relevant responses and advice; must explain complex medical knowledge in a concise manner that the patient can understand; and must be proactive in asking for information about the patient's condition and related information when needed.

Fluency: Answers must be semantically coherent and free of logical errors or irrelevant infor-

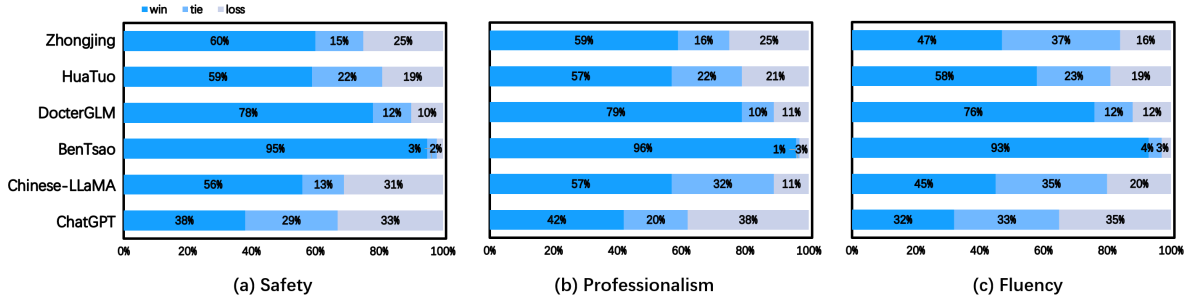

Figure 3: GPT4 assessment Qibo-7B results for (a) Safety, (b) Professionalism and (c) Fluency. Winning rate, tie rate, and loss rate are used as the measures.

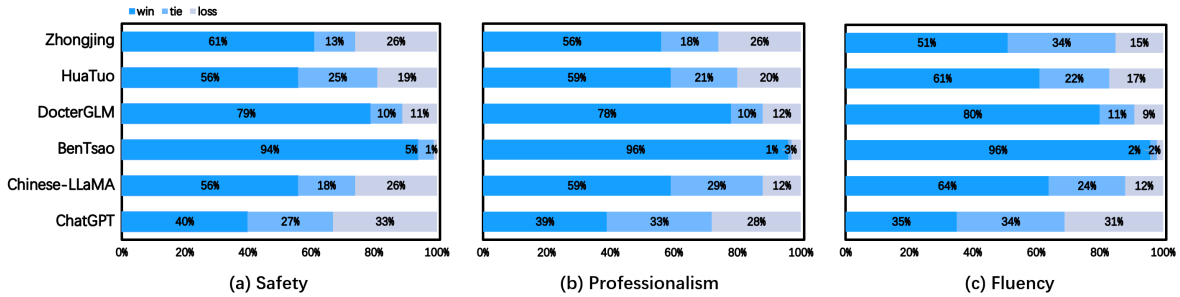

Figure 4: GPT4 assessment Qibo-13B results for (a) Safety, (b) Professionalism and (c) Fluency. Winning rate, tie rate, and loss rate are used as the measures.

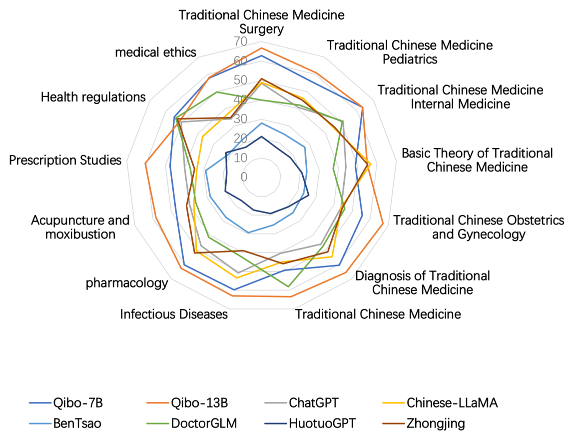

Figure 5: The accuracy of different models in the 13 subject multiple-choice questions of the Traditional Chinese Medicine Practicing Examination.

mation; the style and content of answers must be consistent and free of conflicting information; responses must remain friendly and welcoming; cold or overly brief language is unacceptable.

4.3.2 Objective Evaluation

A total of 3,175 practice questions related to the 13 TCM practice exams were collected and organized as assessment data, which were measured by comparing the accuracy of responses across subjects.

4.3.3 TCM NLP Tasks Evaluation

We retained 517, 689, and 475 data from the prescription identification dataset, dialectical analysis dataset, and reading comprehension quiz dataset, respectively, for assessment, and the assessment criteria were adopted from Rouge-L (Lin, 2004).

4.4 Results

The results of the subjective assessment are shown in Figure 3 and Figure 4, and the results of the objective assessment are shown in Figure 5, and the results of the traditional Chinese medicine NLP task are shown in Table 3.

The results show that "Qibo" achieved excellent results in all three dimensions of subjective assessment, and achieved the best results in the objective assessment. In the traditional Chinese medicine

task	method	Rouge-L
TCM-NER	ours	0.72
	*	0.78
TCM-RP	ours	0.61
	*	0.63
TCM-SD	ours	0.64
	*	0.87

Table 3: The simple results of three NLP tasks."*" indicates the best outcome of the method specifically designed for this task.TCM-NER refers to the entity recognition task of TCM prescriptions.TCM RP is a TCM reading comprehension quiz pair construction task. TCM-SD refers to the task of syndrome differentiation in TCM.

NLP task, although it is not as good as the best model optimized for the task, it is still better than other medical models. It outperforms the baseline model in most cases. The following are our main observations and conclusions from the experimental results:

Our model has the best performance in the field of traditional Chinese medicine. Our model outperforms other medical models in both subjective and objective evaluations of multiple-choice questions. Although the performance in NLP tasks is not as good as the methods specifically designed for this task, it still has a certain effect. We also found that the model performs poorly in determining the position of Chinese entities, which may be due to the inability to obtain complete Chinese character position information after word segmentation. In addition, scale effects still exist, and models trained with more data and a larger number of parameters often perform better. For example, Qibo-13B has higher accuracy in evaluating multiple-choice questions than Qibo-7B.

5 Conclusion and Limitations

We introduce "Qibo" an LLM in the traditional Chinese medical domain that implements pre-training to SFT, and whose performance outperforms other open-source Chinese large-scale medical models in the traditional Chinese medical domain and is comparable to models with significantly more parameters. We have collated high-quality training corpus data in the traditional Chinese medicine domain and constructed Qibo-banchmark, an evaluation benchmark in the traditional Chinese medicine domain, to fill the evaluation gap in the traditional Chinese medicine domain.

Despite these achievements, we also recognize the limitations of the model. Qibo cannot guarantee that all responses are accurate. Given the potentially serious consequences of misleading information in the medical field, we recommend that users treat the information generated with caution and consult a professional. Gizmo relies primarily on text-based information and may not be able to handle more complex multi-modal medical information such as medical images and patient physiological signals. Future research could focus on improving safety, integrating real user data to optimize RLHF, and integrating non-textual information to provide more comprehensive and accurate healthcare. Despite its limitations, "Qibo" remains primarily a research tool rather than a substitute for professional medical advice.

References

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

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. *arXiv preprint arXiv:2101.11038*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, et al. 2023. Falcon-40b: an open large language model with state-of-the-art performance. *Findings of the Association for Computational Linguistics: ACL*, 2023:10755–10773.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and

Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

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*.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

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

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ren Mucheng, Huang Heyan, Zhou Yuxiang, Cao Qianwen, Bu Yuan, and Gao Yang. 2022. Tcm-sd: A benchmark for probing syndrome differentiation via natural language processing. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 908–920.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. openai.

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

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. Model dementia: Generated data makes models forget. *arXiv e-prints*, pages arXiv–2305.

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. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

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

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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

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

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, et al. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *arXiv preprint arXiv:2209.02970*.

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*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.