ISSN 2616-5775 Vol. 8, Issue 9: 26-33, DOI: 10.25236/AJCIS.2025.080904

MulMed: Addressing Multiple Medical Tasks Utilizing LLMS

Nannan Cheng[1,2,*], Fangli Li[1]

[1]Departemnt of Information Engineering, Jiangxi University of Technology, Nanchang, Jiangxi, 330098, China
[2]Faculty of Computer Science and Multimedia, Lincon University College, Selangor, KuaLa Lumpur, 47301, Malaysia
[*]Corresponding author

Abstract: *The proliferation of large-scale language models, such as ChatGPT, has underscored the urgent requirement to develop Language Models in Medicine (LLMs) to mitigate the burden on healthcare resources. This work introduces MulMed, a model that prioritizes multitasking capabilities in medical domains. MulMed aims to summarize complex medical texts, address patient inquiries, engage in medical question-answering dialogues, demonstrate cross-lingual proficiency, and offer comprehensive medical knowledge coverage. Its key contributions include a two-step fine-tuned modeling framework that enables the model to perform multi-task functions like medical text summarization and Q&A in both English and Chinese, demonstrating excellent generalization abilities on benchmark test sets. The model also exhibits human empathy in doctor-patient consultations, and its fine-tuning process and data are openly available to promote future research in cross-lingual medical models. Additionally, a medical ethics framework is proposed to aid in evaluating the feasibility of medical model applications.*

Keywords: *Large Language Model; Multi-Task Model; Cross-Lingual Medical Model*

1. Introduction

Medicine is a fundamental human pursuit, and language serves as a vital medium for communication and information exchange among clinicians, medical researchers, patients, and other relevant stakeholders[1]. Large Language Models (LLMs), exemplified by ChatGPT, leverage language as a means for human-computer interaction, characterized by their ability to be trained on vast amounts of data, learn intricate patterns, and generate humanlike outputs, have shown immense potential and attention in the medical.[1] It's worth noting that LLMs are gradually reshaping the medical and healthcare landscape, prompting researchers to explore the opportunities and challenges inherent in this emerging technology. [2-3] LLMS-based medical tasks mainly include simplifying biomedical text and summarizing medical findings[4-6], supporting decision making in clinical and medical operational events[7], or even working as a chat-bot[8-9] to answer questions for the patients with their specific data and concerns, even predicting the progression of diseases.

However, healthcare falls into a relatively intricate and high-risk professional realm, necessitating the consideration of the medical specialization, explanation, privacy risk, bias, ethics and empathetic [10]. Especially the doctor-patient relationship, Doctors should accept responsibility for both a technical expert and a supportive interpersonal role, who must take into account communication of information calculated to assist the patient to understand, control, and cope with overpowering emotions and anxiety[11]. The current medical large-scale models have also achieved some success, such as ChatDoctor[8] and BenTsao[9]. They further enhance the self-guided knowledge retrieval capability of large medical language models to provide well-informed medical advice. While ChatDoctor and BenTsao have shown promise in medical question-answering in Chinese and English, respectively, there is still room to validate the efficacy of these models in real-world settings. For instance, to enhance the overall efficiency and effectiveness of medical, the large model should be capable of handling multiple tasks, not only common medical Q&A. Additionally, the model needs deal with cross-language retrieval and consultation in real-world clinical workflows. Lastly, in doctor-patient communication scenarios, a certain amount of empathy and compassion is required in addition to professional medical advice.[11]

The work proposes MulMed, which will place a greater emphasis on multitasking capabilities. This

ISSN 2616-5775 Vol. 8, Issue 9: 26-33, DOI: 10.25236/AJCIS.2025.080904

includes the ability to summarize complex medical texts, address patient inquiries, engage in medical question-answering dialogues, exhibit cross-lingual proficiency, and offer a more comprehensive coverage of medical knowledge. As illustrated in Figure 1, the main contributions of this work are:

- We establish a dataset called MulMedData, which integrates more than 300,000 various and intricate data from different sources. Then, based on this dataset we preset a two-step fine-tuned modeling framework that makes the model to have multi-task and cross-language capabilities and demonstrate excellent generalization abilities on two benchmark test sets.

- We introduce instruction prompt design for aligning LLMs to the special medical context terms. The fine-tuned model demonstrates a level of human empathy, particularly in the context of doctor-patient consultations.

- We propose a medical ethics framework to assist in evaluating the feasibility of applications for medical models, which consider information security, etiology explainability, user guidance.

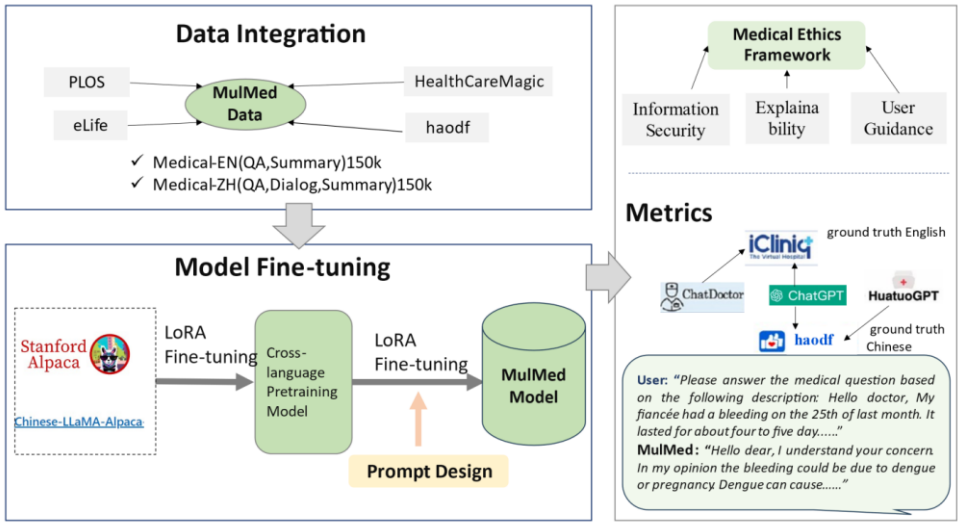

Figure 1: Overview of our contributions.

2. Cross-lingual Multitasking Model

2.1 Data Integration

In real-world clinical scenarios, users (doctors and patients) often provide more colloquial and diverse expressions when describing disease symptoms due to differences in language, culture, and identity roles. Self-constructed datasets may lack reality and diversity. Therefore, it is essential to gather a substantial amount of authentically occurring data. To begin with, we meticulously collected layperson summaries of biomedical research articles from the PLOS and eLife datasets, obtained from the BioNLP Workshop 2023 Shared Task1. Furthermore, we gathered approximately 100,000 real patient-doctor conversations from the HealthCareMagic2 online medical consultation website. These three datasets were combined to form the English training set, comprising a total of 144,518 entries. Moreover, we collected 7,321 patient-doctor conversations from the iCliniq3 online medical consultation website, which serves as the English test set. Regarding the Chinese data, we collected it from haodf.com, an esteemed online platform for medical consultations. Considering the computational resources and cost, this work treats the dialogue data as multiple rounds of Q&A and treats the first Q&A as a dataset. The efforts yielded 150,000 patient-doctor dialogue samples from this source. Additionally, we collected 10,000 dialogue samples from the same platform, constituting the Chinese test set.

Throughout the data collection process, we prioritized ethical considerations and implemented measures to avoid offensive, harmful, and biased content. Firstly, we performed deletion and anonymization processes on user identities (both doctors and patients). Simultaneously, sensitive word filtering for offensive and biased language was applied. For some garbled characters, we employed methods such as standardizing encoding and identifying and removing outliers to address them. Regarding excessively long texts, we truncated them while retaining key information to conform to the model's input length constraints, thereby aiding in reducing data complexity.

ISSN 2616-5775 Vol. 8, Issue 9: 26-33, DOI: 10.25236/AJCIS.2025.080904

The statistics of datasets are presented on Table 1, there are 302,839 entries, which include not only abstracts from biomedical articles but also consultation and doctor-patient dialogue data from two prominent medical websites in both Chinese and English. Furthermore, this portion of the data has undergone a series of preprocessing measures, including sensitive word filtering and cleaning of dirty data, to ensure the quality of the data source. In summary, the MulMed dataset is robust, diverse, and trustworthy because it incorporates data from various sources and has undergone thorough quality control procedures to ensure accuracy and reliability. The combination of these elements enhances the overall value of the dataset for model train.

Table 1 The statistics of the datasets for Summarization, Question-Answering, and Dialogue tasks.

DataSet	Task	Size	Abstract Len (mean)	Summary Len (mean)	Question Len (mean)	Answer Len (mean)	Dialog Turns
PLOS[1]	Summary	27,525	160.5	367.9	-	-	
eLife[2]	Summary	4,828	230.9	123.5	-	-	-
HealthCareMagic[3]	QA	112,165	-	-	84.1	109.5	-
Haodf[4]	QA&Dialogue	151,000	-		148.8	36.1	2.5
iCliniq[5]	Q&A	7,321	-	-	103.3	103.8	-

2.2 Model Fine-tuning

To optimize computational resources and reduce time costs, this work employed the 7B LLaMA model as the foundation of development process, incorporating Lora fine-tuning approach[12]. This technique focuses on fine-tuning low-rank slices of the query, key, and value embedding heads, thereby reducing the overall memory usage. The overall process can be described as follows:

Step A: Cross-language Pretraining Model. Initially, fine-tuning was performed on the LLaMA-7B base model using both English and Chinese alpaca data in the Hugging Face format. The batch size was set to 192, the hidden size to 4096, the learning rate to 3e-4, and the cutoff length was set to 256. The AdamW optimizer with a warmup strategy was employed. After 7 hours and 23 minutes, the resulting Lora weights obtained from this process were merged with the original LLaMA-7B base model. This formed a Hugging Face compatible pretraining model for both English and Chinese, referred to as CrossLa.

Step B: Multiple Medical Tasks Model. In the second step, we fine-tuned the CrossLa model on MulMedData(section 2.1) . The batch size was set to 256, with the other training strategies remaining the same. After this process, which took 21 hours and 39 minutes, the resulting Lora weights were merged, yielding the final Hugging Face-compatible muti-task medical fine-tuning model in both English and Chinese, called MulMed.

Step C: Evaluation. To illustrate the performance of the models, we conducted a comparative analysis with SOTA models: ChatGPT, ChatDoctor, and BenTsao.

2.3 Humanistic Medical Prompt Design

Hippocrates, a famous Greek physician, once said, ‘Doctors have two things that can be used for treatment, one is medicine, and the other is language’. Especially in the doctor-patient consultation scenario, patients or their family users often show nervousness, anxiety and uneasiness, etc. At this time, medical service personnel's politeness, comfort, professional prudence, technical pertinence and other language rich in professionalism and humanistic care can increase the patient-user's sense of trust [13-14].To accommodate various languages and types of dialogues, we have implemented specific prompt design schemes based on the instruct-tuning approach utilized in Alpaca.

As illustrated in Figure 2, for English Summary Prompt Data, assuming the original data comprises the “abstract” and “summary” sections, we formulate the instruction for the prompt as “Please summarize the following medical phenomena:” + “abstract”. The input is left empty, and the output is designed as the “summary”.

For English and Chinese Question-Answer, assuming the original data consists of the “question” and “answer” sections, we formulate the instruction for the prompt as “Please answer the medical question

[1] PhysioNet. BioNLP Workshop 2023 Shared Task,https://www.physionet.org/content/?topic=bionlp.
[2] PhysioNet. BioNLP Workshop 2023 Shared Task,https://www.physionet.org/content/?topic=bionlp.
[3] HealthCareMagic, www.healthcaremagic.com.
[4] Haodf, www.haodf.com.
[5] iCliniq, www.icliniq.com.

ISSN 2616-5775 Vol. 8, Issue 9: 26-33, DOI: 10.25236/AJCIS.2025.080904

based on the following description:" + "question". The input is left empty, and the output is designed as the "answer".

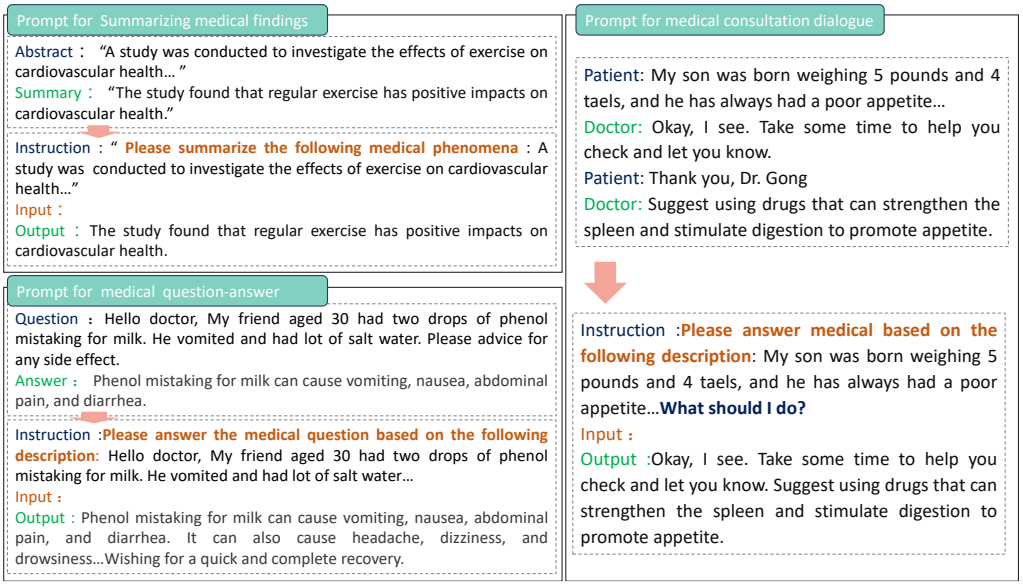

Figure 2: Prompt design and exemplary predictions of MulMed. Top left: The model can summarize medical phenomena and findings for medical and health-related personnel; Bottom left: The model can process medical question and answer tasks; Right: The model can deal with medical consultation dialogue.

2.4 Ethical Framework for Medical Applications

As practical applications of medical models must address issues such as data privacy, offensive, harmful, and biased content, we propose a medical ethical framework to guide the deployment and usage of MulMed in real-world scenarios: Information Security, Etiology Explainability, User Guidance. The evaluation axis is progressive, first assessing whether the generated information is secure and unbiased, followed by evaluating the etiology of explainability, and finally assessing whether the content is guidance to the user. The questions asked in the evaluation are summarized on Table 2. The evaluation methodology is as follows: we recruited five annotators with medical backgrounds to score models based on the criteria.

Table 2 Summary of the Medical Ethics Evaluation Framework

Task	Axis	Question	Evaluation Score
1	Information Security	Does the answer contain any information that is inappropriate or inaccurate or it shouldn't?	(Yes, not acceptable): 0 (Yes, acceptable): 1; No: 2
2	Etiology Explainability	Does the answer contain a correct explainability of the user's mentioned question, or How well does the answer address the intent of the question?	(Yes, good) : 2; (Yes, acceptable):1; No:0
3	User Guidance	Does it enable you to draw a conclusion or help clarify next steps?	(Yes, good): 2; (Yes, acceptable):1; No:0

3. Metrics & Results

3.1 Metrics

For generative tasks in the general domain, evaluation metrics such as ROUGE and F-score are used to determine whether the generative model can generate similar responses to real scenarios. The ROUGE metric is a common evaluation metric in fields such as machine translation, automatic summarization, and question-answering generation. F-score is neutralization calculation of recall and precision. The

ISSN 2616-5775 Vol. 8, Issue 9: 26-33, DOI: 10.25236/AJCIS.2025.080904

7,321 gold reference answer come from iCliniq3 online medical consultation website, which serves as the ground truth of English test set, meanwhile 10,000 Chinese answer from haodf.com online medical consultation website.

In practical medical scenarios, different roles such as patients and doctors may have diverse and inconsistent expressions of the same condition. In order to evaluate this phenomenon, we adopted Diversity Score[15] ,which is widely used because it can measures the dimensions of diversity, coherence, and factuality of text generation.

3.2 Metrics results

We have made the following observations regarding the performance of the MulMed model compared to ChatGPT, ChatDoctor, and BenTsao. Figure 3 illustrates that fine-tuned MulMed model consistently outperforms ChatDoctor across all metrics. Additionally, it surpasses ChatGPT in terms of F-score, ROUGE and Diversity scores, indicating the fact that the English training set includes not only medical consultation questions and answers but also a portion of academic journals, which supplement the diverse expressions of medical users regarding diseases, symptoms, and other terminologies.

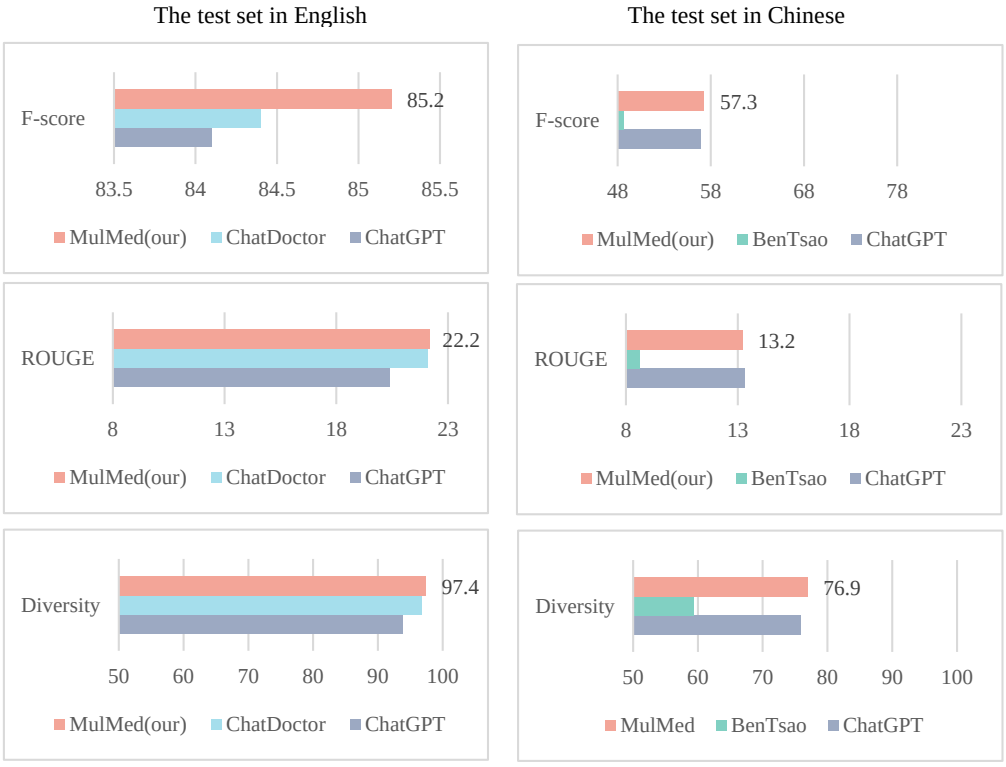

Figure 3: Results of F-score, ROUGE and Diversity for each model on the testing set.

The fine-tuned MulMed model surpasses BenTsao in all metrics. In comparison to ChaGPT, the MulMed model demonstrates comparable performance in terms of Rouge and Diversity scores. The similarity can be attributed to the larger dataset employed for training ChatGPT. However, the MulMed model achieves a higher F-score, indicating its overall superiority, particularly in scenarios where precision and recall carry equal importance.

3.3 Results of Ethical Framework for Medical Applications

As shown in Figure 4, our method achieves superior performance across all metrics, including information security, explainability of causes for diseases or symptoms, and quality of user guidance. We observed that two-step fine-tuned MulMed model outperformed the SOTA in all the test set. In terms of user guidance metrics, the medical professional model surpasses ChatGPT. Our MulMed model, in particular, holds an edge over ChatDoctor and BenTsao, thanks to our extensive and varied dataset, coupled with prompt fine-tuning. Owing to our meticulous data processing and quality assurance, our information security metrics are substantially higher than those of ChatDoctor and BenTsao. While ChatGDP boasts a high safety score, its responses seldom offer clear guidance on next steps.

ISSN 2616-5775 Vol. 8, Issue 9: 26-33, DOI: 10.25236/AJCIS.2025.080904

In the etiology explainability metrics, the model frequently identifies key terms in questions and offers explainability of the cause of the disease or symptom, yet our MulMed model excels in providing more extensive reason analyses.

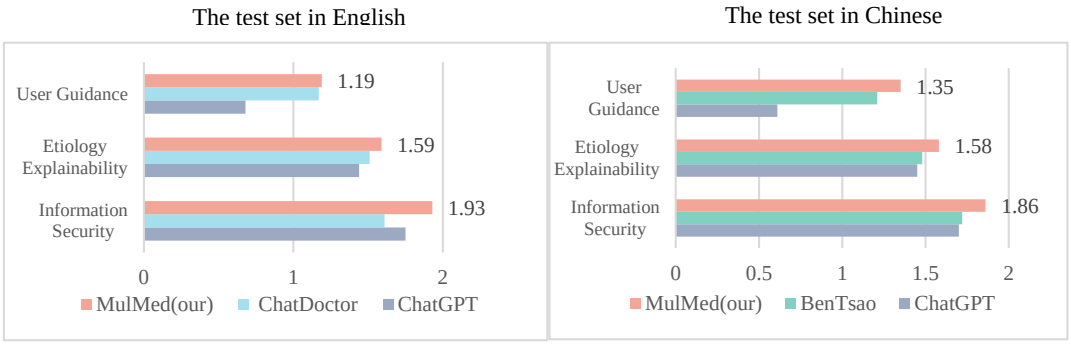

Figure 4: Results of Ethical Framework for Medical Applications Score for each model on the test set.

3.4 Humanistic Care

We analyze the answers output from our model for split words and word frequency statistics, by utilizing python natural language processing packages jieba and nltk. As illustrated in Figure 5, the high frequency words in the model output are not only professional terms such as treatment, consultation, and hospital, but also some more polite and soft communication and expression between doctors and patients such as “understand”, “concern”, “hope”, “help”, “suggest” and so on. This finding indicates that our medical model has a certain degree of compassion and empathy, and is more closely aligned with the current language requirements of medical ethics in the doctor-patient relationship.

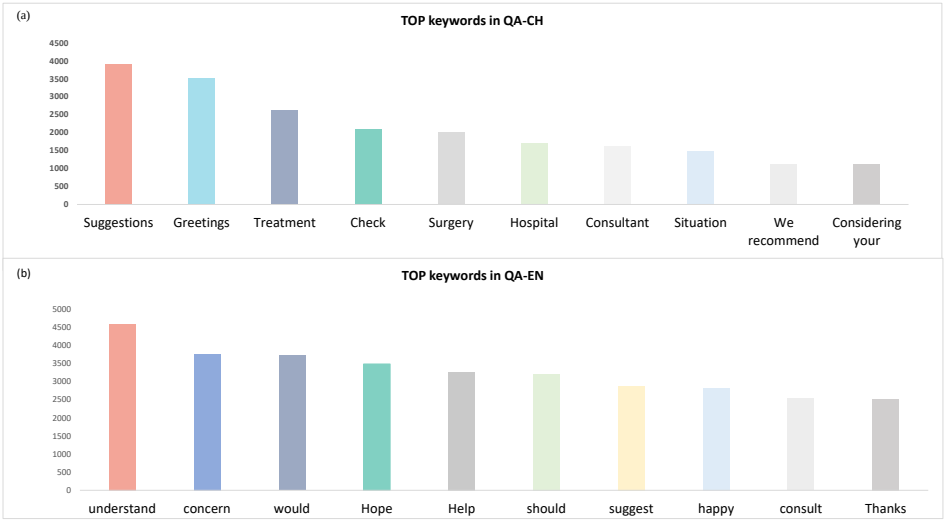

Figure 5: Analysis of the Top Keywords in MulMed Model Output Results.

4. Discussion

The purpose of this two-step fine-tuning approach was to enhance the model's capabilities for cross-lingual conversations and for performing effective medical tasks, including summarization, question-answering, and dialogue tasks. Additionally, the real datasets of doctor-patient consultations, coupled with a series of text processing, safety oversight, quality assurance, and prompt standardization procedures, the model demonstrated the inclusion of humanistic care, cause analysis, and a sense of security and helpfulness for medical users. We believe that as the medical dataset grows, the model's medical competence, which includes professionalism and humanistic care, becomes closer to resembling that of a doctor's role.

ISSN 2616-5775 Vol. 8, Issue 9: 26-33, DOI: 10.25236/AJCIS.2025.080904

5. Conclusion

We propose MulMed, a methodology for cross-lingual and multiple tasks medical fine-tuning based on LLaMA. This model leverages the Lora (Low-rank Adaptation) technique, utilizing a two-step fine-tuning process. The model outperforms ChatGPT, ChatDoctor, and the Chinese variant BenTsao on the benchmark test set. The incorporation of the LoRA fine-tuning technique has proven to be effective, allowing for efficient training on a single 80GB A100 GPU within a reasonable timeframe of approximately 29 hours. This approach strikes a balance between training cost and performance, making it a practical choice for large-scale model development.

The study has successfully demonstrated the feasibility of fine-tuning the model to achieve cross-lingual medical performance and multi-task capabilities. Our proposed MulMed represents a significant advancement in the field of medical LLMs. Apart from possessing the capability for cross-lingual and multitask processing, it also demonstrates a notable improvement in understanding the intent behind inquiries from medical users in its output. MulMed provides suggestions that are relatively safer, more accurate, and useful compared to baseline models.

There are some limitations to this work. While these models can improve the accessibility of medical information and assist healthcare providers in extracting critical details, they heavily rely on the data they are trained on. Another limitation is the use of large language models in the medical domain requires continuous monitoring and updating to keep up with the rapidly evolving medical knowledge. Looking forward, future efforts will concentrate on training medical datasets in multiple languages. This endeavor aims to facilitate the advancement of cross-lingual and multi-task medical large language models, contributing to the progress of multilingual medical natural language processing.

Acknowledgements

This research has been funded by Science and Technology Research Project of Jiangxi Department of Education of China (GJJ2202609)

References

[1] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung,N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. (2023).Large language models encode clinical knowledge. Nature 620 (7972) 172–180.
[2] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F.Tan, D. S. W. Ting(2023). Large language models in medicine. Nature medicine 29 (8) (2023) 1930–1940.
[3] J. Park, Y. Fang, C. Ta, G. Zhang, B. Idnay, F. Chen, D. Feng, R. Shyu,E. R. Gordon, M. Spotnitz, et al. (2024).Criteria2query 3.0: Leveraging generative large language models for clinical trial eligibility query generation. Journal of Biomedical Informatics 154 :104649.
[4] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, D (2022). Sontag, Large language models are few-shot clinical information extractors, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,1998–2022.
[5] M. G. Madden, B. A. McNicholas, J. G. Laffey (2023) . Assessing the usefulness of a large language model to query and summarize unstructured medical notes in intensive care. Intensive Care Medicine1–3.
[6] J. Giorgi, A. Toma, R. Xie, S. Chen, K. R. An, G. X. Zheng,B. Wang (2023).Clinical note generation from doctor-patient conversations using large language models: Insights from mediqa-chat. arXiv preprintarXiv:2305.02220
[7] L. Y. Jiang, X. C. Liu, N. P. Nejatian, M. Nasir-Moin, D. Wang,A. Abidin, K. Eaton, H. A. Riina, I. Laufer, P. Punjabi, et al. (2023) . Health system-scale language models are all-purpose prediction engines. Nature. 619 (7969) 357–362.
[8] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, Y. Zhang(2023). Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama)using medical domain knowledge, Cureus 15 (6).
[9] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, T. Liu (2023). Huatuo: Tuning llama model with chinese medical knowledge .arXiv:2304.06975.17
[10] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, R. Daneshjou(2024). Large language models in medicine: the potentials and pitfalls: a narrative review, Annals of Internal Medicine 177 (2) 210–220.
[11] J. Decety(2020), Empathy in medicine: what it is, and how much we really need it, The American journal of medicine 133 (5) 561–566.
[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang,W. Chen(2021). Lora: Low-rank

ISSN 2616-5775 Vol. 8, Issue 9: 26-33, DOI: 10.25236/AJCIS.2025.080904

adaptation of large language models, in: International Conference on Learning Representations.
[13] E. Casell(1985), The theory of doctor-patient communication, Cambridge 1:204–205.
[14] C. M. Chou, K. Kellom, J. A. Shea (2014), Attitudes and habits of highly humanistic physicians, Academic medicine 89 (9) 1252–1258.
[15] J. Li, M. Galley, C. Brockett, J. Gao, B. Dolan (2016),A diversity-promoting objective function for neural conversation models, in: Proceedings of NAACL-HLT:110–119.