CUMed-LLaMA: A Specialized and Human-aligned Chinese Medical Large Language Model with Full Training Pipeline

Anonymous ACL submission

Abstract

1

With the rapid advancement of artificial 2 intelligence, large language models (LLMs) 3 have made significant progress in capturing 4 and responding to user needs. However, 5 when applied to specific fields, particularly 6 the Chinese medical domain, these models 7 still face challenges. Existing medical 8 LLMs often rely on supervised fine-tuning 9 (SFT) for general medical tasks, but they 10 struggle with understanding complex 11 medical issues and aligning with expert 12 intentions. To address this, we introduce 13 CUMed-LLaMA, a Chinese medical LLM 14 specifically designed for urology. It has 15 undergone a complete training process, 16 including pre-training, supervised fine-17 tuning, and reinforcement learning from 18 human feedback (RLHF), ensuring its 19 strong performance in the urology domain. 20 We have also developed a dataset 21 containing various medical materials to 22 enhance the model's ability to handle 23 complex dialogues and proactive 24 questioning. multidimensional А 25 considering evaluation framework, 26 professionalism, and 27 relevance, user experience, was used to assess the model's 28 output. Experimental results show that 29 CUMed-LLaMA outperforms existing 30 baseline models in various medical tasks, 31 particularly in urology, demonstrating 32 capabilities on par with expert 33 professionals. Although CUMed-LLaMA 34 is focused on urology, it performs 35 impressively across other medical tasks, 36 supporting broader applications of LLM 37 technology in the medical field. This study 38 contributes to the development of Chinese 39 medical LLMs and provides a solid 40 foundation for practical applications in 41 urology. 42

43 **1** Introduction

44 With the rapid development of large language 45 models (LLMs) represented by ChatGPT, their 46 application prospects in the medical field have 47 become increasingly promising. LLMs have 48 already demonstrated significant potential in 49 medicine, not only for the rapid generation of 50 medical documents but also for assisting doctors in 51 clinical decision-making(Thirunavukarasu et al., 52 2023). Furthermore, in summarizing clinical text, 53 certain LLMs have even outperformed medical 54 experts(Van Veen et al., 2024). The application of 55 LLMs in medical question-answering systems has 56 also gained significant momentum(D. Wang & 57 Zhang, 2024). However, due to the complexity and 58 specialization of the biomedical knowledge system, ⁵⁹ developing high-precision, high-safety medical 60 question-answering systems presents a high 61 threshold(Singhal, Azizi, et al., 2023). Despite 62 these challenges, notable progress has been made 63 in relevant research, such as Google's development 64 of the Med PaLM2 model, which has achieved 65 leading results in multiple medical question-66 answering benchmarks(Singhal, Tu, et al., 2023). 67 Subsequently, the potential of large models in 68 medical dialogue has been widely recognized, and 69 an increasing number of research institutions have ⁷⁰ started developing related medical models. Models 71 such as PMC-LLaMA(Wu et al., 2023) and 72 MediQ(S. S. Li et al., 2024) have emerged. These 73 models often utilize the open-source large model 74 LLaMA(Touvron, Lavril, et al., 2023), which 75 performs comparably to closed-source GPT 76 models. Moreover, LLaMA offers various 77 parameter types, better meeting the practical 78 application requirements, especially in specific 79 domain tasks, where open-source models can be ⁸⁰ more efficiently optimized and deployed.

82 Chinese data, particularly in the field of Chinese 134 practical requirements. ⁸³ medical dialogue, research institutions in China ¹³⁵ ⁸⁴ have begun developing medical large models based ¹³⁶ we have drawn on valuable insights from cutting-85 on Chinese corpora. For example, ChatDoctor(Y. 137 edge research to develop CUMed-LLaMA, a 86 Li et al., 2023) and DoctorGLM(Xiong et al., 2023) 138 Chinese medical large model based on Ziya-87 are both striving to improve the model's 139 LLaMA-V2(Gan et al., 2024). CUMed-LLaMA ⁸⁸ performance in the Chinese medical context. ¹⁴⁰ implements the complete process from pre-training, ⁸⁹ However, despite these models being fine-tuned on ¹⁴¹ fine-tuning, to reinforcement learning. During both ⁹⁰ Chinese medical data, they still face a series of ¹⁴² the pre-training and fine-tuning phases, a 91 challenges. These models primarily rely on general 143 substantial corpus of Chinese medical professional ⁹² English data during the pre-training phase, with 144 texts was utilized, enabling the model to effectively ⁹³ Chinese medical domain knowledge only being 145 transfer the medical knowledge acquired during 94 incorporated during the fine-tuning phase. This 146 pre-training to the fine-tuning phase, thereby ⁹⁵ discrepancy between pre-training and fine-tuning ¹⁴⁷ significantly enhancing its understanding of ⁹⁶ data can lead to difficulties when transitioning from ¹⁴⁸ Chinese medical texts. Unlike Zhongjing and 97 general fields to highly specialized medical fields, 149 ChiMed-GPT, which aim to cover the entire ⁹⁸ as the models may struggle to adapt quickly(Hua et 150 medical field, CUMed-LLaMA focuses on a single ⁹⁹ al., 2024; L. Zhou et al., 2024). The language ¹⁵¹ specialty, urology. Particularly during the fine-100 patterns and knowledge structures involved in 152 tuning and reinforcement learning stages, the 101 English general-purpose data typically do not 153 model emphasizes urological dialogue data, ¹⁰² apply to the medical domain, resulting in logical ¹⁵⁴ ensuring more precise performance in the urology ¹⁰³ biases or information loss when the model is faced ¹⁵⁵ domain and better alignment with clinical needs. 104 with complex medical issues. Furthermore, the 156 To construct high-quality training data, we 105 knowledge of large language models is almost 157 collected a large amount of professional and 106 entirely derived from the pre-training phase(C. 158 authentic medical content, including medical 107 Zhou et al., 2023). If medical professional data is 159 encyclopedias, medical textbooks, and case studies 108 not included in the pre-training phase, the model's 160 for the pre-training phase, as well as urological 109 performance in medical tasks will be significantly 161 dialogue data for fine-tuning and reinforcement 110 compromised. Recently. models 111 Bianque(Chen et al., 2023), Zhongjing(Yang et al., 163 doctor-patient interactions in outpatient settings, 112 2023), and ChiMed-GPT(Tian et al., 2024) have 164 covering the diagnostic processes of common 113 partially addressed these issues by incorporating 165 urological conditions such as urinary tract 114 pre-training, fine-tuning. and 115 learning, significantly improving ¹¹⁶ performance in the Chinese medical context. The ¹⁶⁸ demonstrates robust professional capabilities in the 117 goal of these models is to develop a general 169 field of urology but also possesses a broad ¹¹⁸ medical model that covers knowledge across ¹⁷⁰ knowledge base in medicine, enabling it to ¹¹⁹ multiple departments and various medical tasks, ¹⁷¹ effectively address a wide range of medical issues. ¹²⁰ including clinical diagnosis, disease treatment, and ¹⁷² The overall process is illustrated in Figure 1. 121 health education. However, despite technological 122 breakthroughs, a parameter size of 13B is still 173 2 123 insufficient to handle the vast knowledge system of 124 the medical field. The content in the medical 174 2.1 125 domain is exceptionally complex, encompassing 175 High-quality pre-training datasets are crucial for 126 basic medicine (such as anatomy and physiology), 176 the performance of large models. The diversity, 127 multidisciplinary fields, as well as sub-specialties 177 richness, and professionalism of the data can 128 like cardiology and oncology. The depth of 178 significantly enhance the model's understanding 129 knowledge, expression styles, 130 diagnostic/treatment processes within each sub- 180 collected various types of data from multiple 131 specialty can vary significantly. Particularly in 181 sources, 132 interdisciplinary integration and handling complex 182 knowledge, medical knowledge graphs, and

In order to enable LLMs to better handle 133 cases, the model's performance often fails to meet

In response to the aforementioned challenges, like 162 learning. These dialogues simulate real-world reinforcement 166 infections, kidney stones, and prostate diseases. model 167 The resulting CUMed-LLaMA model not only

Methods

Continuous Pre-training

and 179 and generation capabilities. To this end, we including medical encyclopedic



184 185

Figure 1: An illustration of the overall framework of our study.

186 medical case data. In order to ensure the quality and 196 subfields, including basic medicine, clinical ¹⁸⁷ professionalism of the data, we applied rigorous ¹⁹⁷ medicine, and pharmacology. During the screening 188 standards 189 preprocessing stages. Each data point underwent 199 case records with high-quality clinical information, 190 manual review, with lower-quality or less-relevant 200 ensuring their representativeness and practical ¹⁹¹ data removed, leaving only high-quality data that ²⁰¹ value. ¹⁹² met the pre-training requirements. In the medical ²⁰² approximately 12,000 records that lacked valid 193 encyclopedic knowledge dataset, we ultimately 203 information or contained significant errors. For the ¹⁹⁴ retained 369,000 high-quality data points after ²⁰⁴ medical dialogue dataset, we filtered out 630,000 195 strict filtering. These data covered various 205 high-quality dialogues from the original data,

during the data screening and 198 of medical case data, we carefully selected 14,000 Simultaneously, we discarded

Stage	Dataset	Department	Token Number	Storage Size
Pre-training	Encyclopedia	Multiple	120,429,235	565.33MB
	Question-Answering	Multiple	97,666,379	444.88MB
	Knowledge Graph	Multiple	26,072,459	119.10MB
	Medical Cases	Multiple	12,799,680	55.90MBr
	Textbooks	Multiple	8,400,212	38.08MB
Fine-tune	Single-turn QA	Urology	9,825,615	15.49MB
	Multi-turn QA	Urology	9,595,993	15.58MB
Rejection Sampling	Reward QA	Urology	3,158,349	4.77MB

206 207 Table 1: A demonstration of datasets for different stages.

208 removing around 160,000 dialogues of lower ²⁰⁹ quality. Through this meticulous data cleaning and ²⁴² **2.2** 210 selection process, we constructed a high-quality, 243 Supervised fine-tuning (SFT) is a crucial phase in 211 specialized medical dataset that provides a solid 244 enabling LLMs to acquire intelligence(Ouyang et ²¹² foundation for the pre-training of large models. The ²⁴⁵ al., 2022; Touvron, Martin, et al., 2023). Overspecific datasets used are shown in Table 1. 213 214

215 for Chinese, we chose Ziya-LLaMA-V2 for 248 performance(Gudibande et al., 2023; Shumailov et 216 subsequent development. Ziya-LLaMA is a version 249 al., 2024). Therefore, we collected over 20,000 217 developed by Chinese researchers based on 250 high-quality data points focused on urology. 218 LLaMA, significantly enhancing LLaMA's ability 251 Through these carefully curated doctor-patient 219 to process Chinese through training on a large 252 dialogue datasets, the model is able to effectively 220 volume of Chinese corpora. In the experiment, we 253 leverage the medical knowledge accumulated 221 selected AdamW(Loshchilov & Hutter, 2019) as 254 during pretraining, enabling it to better understand 222 the optimizer and set the initial learning rate to 255 and respond to user queries. Our dataset covers a 223 5×10-5. AdamW employs a decoupled weight 256 broad range of urological issues, rich in content and 224 decay regularization method that helps improve 257 highly representative. In the SFT process, the 225 training stability. To further enhance the stability of 258 model combines either single-turn or multi-turn 226 the training process and accelerate the model's 259 dialogue history with the patient's latest utterance 227 convergence, we applied a linear warmup strategy 260 to generate a doctor's response. In single-turn 228 to prevent gradient explosion caused by an 261 dialogues, the model generates a response based 229 excessively large initial learning rate. Additionally, 262 solely on the current patient query and relevant 230 we used the Megatron-LM framework (Shoeybi et 263 context. In multi-turn dialogues, the model 231 al., 2020) for efficient distributed training and 264 considers the 232 adopted **BF16** 233 training(Micikevicius et al., 2018) to optimize 266 responses, thereby generating more accurate and 234 GPU memory costs and accelerate computation. To 267 coherent doctor responses. During training, the 235 prevent overfitting, we set the weight decay to 0.01 268 model inputs the prompt, compares its generated 236 in the optimizer and applied Dropout(Srivastava et 269 response with the gold-standard answer, computes 237 al., 2014) with a dropout rate of 0.1 in some fully 270 the loss, and updates all parameters accordingly. To 238 connected layers. Through the comprehensive 271 efficiently utilize the model's maximum sequence 239 application of these strategies, we significantly 272 length, we concatenate short response pairs and 240 improved the training efficiency and performance 273 insert special tokens to mark the boundaries 241 of the model.

Supervised Fine-tuning

246 reliance on distilled data from GPT or other large Due to the relatively limited support of LLaMA 247 models can lead to a degradation of model entire conversation history. mixed-precision 265 integrating previous patient utterances and doctor 274 between different response pairs. To ensure optimal 275 training, we used LlamaFactory(Zheng et al., 2024), 276 a framework designed to enhance the training 277 process. The hyperparameters were set with a

Dimension	Ability	Explanation
Relevance	Comprehensiveness	The answer must include all relevant medical information, without omitting important details.
	Relevance	The answer must stay focused on the core of the question, ensuring that each part of the response directly addresses the inquiry.
	Logical Structuring	The answer should follow a clear logical structure, with well- organized content that is easy to follow.
Expertise	Safety	The answer must prioritize patient safety, ensuring that no misleading or potentially harmful advice is given.
	Professionalism	The answer must meet the standards and expectations of the medical industry, demonstrating appropriate professional expertise.
	Ethical Compliance	The answer must adhere to medical ethics, ensuring compliance with moral and professional standards.
Experience	Clarity	The answer should be concise and clear, avoiding overly technical terms to ensure patient understanding.
	Response Efficiency	The response should be timely, especially in urgent situations where rapid responses are necessary.
	Tone and Emotional	The answer should convey a friendly and warm tone, avoiding cold
	Resonance	or overly brief responses, ensuring that emotional empathy is effectively communicated.

Table 2: Medical question-answering ranking annotation criteria, divided into 3 capability dimensions and 9 specific abilities with explanations. The importance is ranked from high to low. 279

280

learning rate of 2×10^{-5} , weight decay of 0.1, and a 301 intermediate responses generated by these models 281 ²⁸² batch size of 16. Through these methods, the model ³⁰² to improve the model's performance. Specifically, 283 effectively learns from the training data and 303 we ranked the four responses in the following order: 284 generates high-quality, expected responses.

2.3285 Feedback 286

287 We divide reinforcement learning with human 308 the lower-ranked response as a negative sample for 288 feedback (RLHF) into two main stages: reward 309 training the reward model. During training, we 289 model training and rejection sampling fine-tuning. 310 adopted a cosine decay strategy for the learning ²⁹⁰ In the reward model training stage, we constructed ³¹¹ rate, which started at 5×10^{-6} and gradually a dataset containing 5000 single-turn doctor- 312 decreased to 10% of the initial learning rate. The 291 ²⁹² patient Q&A instances, with 4500 instances in the ³¹³ training ran for two epochs with a batch size of 8, a ²⁹³ training set, and 250 instances each in the ³¹⁴ 3% warm-up phase, and a minimum step count of ²⁹⁴ validation and test sets. Each instance includes two ³¹⁵ 5. After training the reward model, we proceeded ²⁹⁵ answers: the "accepted answer" provided by the ³¹⁶ to the rejection sampling fine-tuning stage to 296 doctor and the "rejected answer" generated by 317 ensure the model's outputs better align with human 297 DeepSeek(DeepSeek-AI et al., 2024). To enhance 318 preferences. In this stage, we randomly sampled 298 the accuracy of the reward model, we did not 319 5000 prompts from the SFT dataset, and assigned a 299 directly use the answers generated by ChatGPT and 320 score to each generated output using the reward 300 GPT-3.5-Turbo but instead utilized

³⁰⁴ the doctor's positive answer, ChatGPT's response, 305 GPT-3.5-Turbo's response, and the rejected answer. Reinforcement Learning from Human 306 We then paired adjacent responses, using the 307 higher-ranked response as a positive sample and the 321 model. Based on these scores, we selected the top



326

327 328 329 $_{330}$ hyperparameters: β 1 set to 0.9, β 2 set to 0.98, ϵ set $_{354}$ in comprehensively evaluating critical factors like 331 332 333 ³³⁴ efficient tuning techniques to ensure the training ³⁵⁸ includes three key dimensions: 335 336 stable. As a result of these optimizations, the 360 assesses whether the response is closely related to 337 model's performance during rejection sampling 361 the question, ensuring that the answer does not 338 fine-tuning was significantly improved. The 362 deviate from the topic; professionalism evaluates ³³⁹ generated responses became more aligned with ³⁶³ the accuracy and correctness of the response; and ³⁴⁰ human preferences, while also enhancing the ³⁶⁴ user experience focuses on the appropriateness of ³⁴¹ model's applicability in the medical domain. This ³⁶⁵ the tone, ensuring the fluency of the dialogue and 342 approach draws inspiration from ChiMed-GPT's 366 the comfort of the user. For responses from ³⁴³ methodology and combines practical data with ³⁶⁷ different models to the same question, we classify 344 advanced technologies to further optimize the 368 the results and rate them as "win," "draw," or ³⁴⁵ conversational model in the medical field.

346 3 Results

Evaluation Metrics 347 3.1

³⁴⁸ Evaluating the quality of medical dialogues is a ³⁷⁴ terminology and information, as well as the ³⁴⁹ complex and multidimensional task. Traditional ³⁷⁵ adherence to medical ethics within the dialogue(P. ³⁵⁰ evaluation metrics, such as BLEU(Papineni et al., ³⁷⁶ Wang et al., 2023). However, relying on medical

k responses as the golden standard for further fine- 351 2002) and ROUGE(Lin, 2004), while capable of tuning. The rejection sampling fine-tuning phase 352 assessing certain aspects of lexical relevance and utilized the AdamW optimizer with the following 353 superficial similarity between dialogues, fall short to 1×10^{-2} , a learning rate of 5×10^{-5} , a weight decay 355 deep semantics, knowledge accuracy, safety, and coefficient of 0.2, a batch size of 64, and trained for 356 ethics. To address this, we propose a new 400 iterations. We applied distributed training and 357 evaluation scale (as shown in Table 2), which relevance, and fine-tuning processes were both effective and 359 professionalism, and user experience. Relevance ³⁶⁹ "loss," using this as a measure of each model's 370 overall performance in response quality. Given the ³⁷¹ highly specialized nature of the medical field, only 372 reviewers with professional medical knowledge 373 can accurately evaluate the correctness of the User:

大便颜色呈叶绿色异常半年有余。有胃炎伴有胃酸灼热感。大便无规律有时呈糊状。有时便秘 The stool has been abnormally dark green for over six months. I have gastritis with a sensation of acid reflux. The stool is irregular and sometimes mushy. Occasionally, I have constipation.

Zhongiing

你好,你这种情况需要做个肠镜检测。检测幽门杆菌,看看是否有胃溃疡。

Huatuo

大便呈叶绿异常可能与饮食有关,比如近期摄入了较多含叶绿色素的食物。此外,胃酸过量也可 能影响大便状态。建议注意饮食,并观察症状变化,如有持续不适,建议就医检查。 The abnormal green color of the stool may be related to diet, such as consuming more foods containing chlorophyll recently. Additionally, excess stomach acid may affect stool condition. It is recommended to pay attention to your diet and observe any changes in symptoms. If discomfort persists, a medical checkup is advised.

ChiMed-GPT

你好,根据你的咨询,属于慢性胃炎,消化功能紊乱所致的症状,处理,一是不吃辛辣刺激食 物,二是不喝汤水,三是服用一段时间三九胃泰,整肠生胶囊和乳酸菌素片,保和丸治疗 Hello, based on your inquiry, this is likely chronic gastritis caused by digestive dysfunction. Treatment includes: 1) Avoid spicy and irritating foods, 2) Avoid soups and watery foods, and 3) Take medications such as Sanjiu Weitai, Zhengchangsheng capsules, and Lactic Acid Bacteria pills for a period of time, along with Baohe pills for treatment.

CUMed-LLaMA

如果是脾胃虚弱的话可以用健脾益气的药物,如果不是的话用力泻药注意调节饮食,使脾胃得到 恢复。建议你注意:1、注意饮食卫生,避免食用生冷食物,避开生冷,注意腹部保暖。2、避免 过度劳累,注重精神调理,增强体质。3、保持良好的情绪,学会自我调节,注射情绪。

If you have spleen and stomach weakness, you may consider using medications to strengthen the spleen and invigorate qi. If that is not the case, using strong purgative medications might be appropriate, but be sure to adjust your diet to help restore the function of the spleen and stomach. Recommendations: 1. Pay attention to food hygiene, avoid consuming cold or raw foods, and keep your abdomen warm. 2. Avoid overexertion, focus on mental relaxation, and strengthen your overall health. 3. Maintain a positive emotional state, learn how to manage stress, and practice emotional regulation.

 Table 3: Responses generated by various models, accompanied by their English translations. 377 378

379 professionals for manual evaluation is both time- 399 particularly when handling tasks requiring deep 380 consuming and labor-intensive. To improve 400 expertise, where the models' accuracy and depth efficiency and reduce evaluation 381 382 consumption, we introduce ChatGPT as an 402 CUMed-LLaMA exhibits certain gaps compared to ³⁸³ auxiliary scoring tool to assess relevance and user ⁴⁰³ ChatGPT, particularly when handling tasks related 384 experience.

385 3.2 **Results**

387 evaluates the performance in the general medical 408 models in the experiential dimension, without 388 domain, while RJUA-QA(Lyu et al., 2024) focuses 409 demonstrating a distinct advantage. In the urology-389 on the urology-specific evaluation. In the context 410 specific evaluation, CUMed-LLaMA again excels 390 of the overall medical domain, CUMed-LLaMA 411 in both relevance and expertise dimensions, further 391 outperforms other models in both relevance and 412 validating its strong capabilities in specialized 392 expertise dimensions, demonstrating its ability to 413 medical fields. This indicates that CUMed-LLaMA ³⁹³ precisely capture core information in the medical ⁴¹⁴ is highly effective at handling urology-related 394 field and generate high-quality professional 415 professional issues and generating accurate, in-395 responses. In comparison, ChatGPT and ChiMed- 416 depth responses. In contrast, other models show 396 GPT show similar performance to CUMed- 417 weaker performance in urology tasks, primarily 397 LLaMA, whereas Zhongjing and Huatuo perform 418 due to insufficient domain knowledge, resulting in ³⁹⁸ relatively weaker in these two dimensions,

resource 401 are insufficient. In terms of user experience, 404 to interactivity and optimization of user experience, ⁴⁰⁵ with its output quality and user satisfaction slightly 406 lower than that of ChatGPT. Overall, CUMed-386 As shown in Figures 2 and 3, Huatuo26M 407 LLaMA shows comparable performance to other

420 responses.

421 422 responses generated by different models. A 473 computational resources but also significantly ⁴²³ comparative analysis reveals that CUMed-LLaMA ⁴⁷⁴ enhances the model's performance and flexibility ⁴²⁴ provides the most comprehensive and specific ⁴⁷⁵ when addressing various specialized medical tasks. ⁴²⁵ answers. It not only recommends further diagnostic ⁴⁷⁶ We believe that by incorporating this strategy, 426 tests (such as gastroscopy and routine stool 477 CUMed-LLaMA will be better equipped to meet ⁴²⁷ analysis) but also incorporates advice on dietary ⁴⁷⁸ the increasingly complex demands of medical 428 adjustments, medication. and 429 management, reflecting a thorough and holistic 480 and reliable medical support. 430 approach to the patient's condition. In contrast, 431 Zhongjing offers useful recommendations for 481 5 Limitations 432 testing but lacks guidance on treatment and 433 lifestyle modifications. Huatuo provides dietary 482 Although CUMed-LLaMA has demonstrated 434 suggestions but lacks in-depth medical analysis. 483 significant potential in the medical field, it still has 435 ChiMed-GPT, while focusing on gastritis and 484 certain limitations. Currently, the model primarily overlooks 436 pharmacological treatment, 437 recommendation for additional diagnostic tests and 486 multimodal capabilities, which limits its ability to 438 exhibits a lower level of fluency in its Chinese 487 integrate diverse forms of medical data such as 439 responses. Overall, CUMed-LLaMA outperforms 488 imaging and diagnostic results. While CUMedother models 440 the in 441 professionalism and thereby strongly supporting the high quality of our ⁴⁹¹ generated responses. Given that medical data may 443 work and dataset and providing compelling 492 inherently be incomplete or inaccurate, we advise 444 evidence for the potential applications of this 493 users to exercise caution when making important 445 model in the medical field.

Conlusion 446 4

448 model, which focuses on the medical field, 499 nuanced communication techniques that human 449 particularly demonstrating 450 performance in the application of urology. Through 501 or provide emotional support, particularly when 451 a comprehensive training process that includes 502 dealing with sensitive health issues. Patients often 452 pretraining, supervised fine-tuning (SFT), and 503 require more personalized care and empathy, which 453 reinforcement learning with human feedback 504 the model is unable to offer. Therefore, while 454 (RLHF), CUMed-LLaMA outperforms other 505 CUMed-LLaMA can provide medical knowledge 455 open-source Chinese medical large language 506 and 456 models at each stage. This achievement not only 507 comprehensive care and 457 highlights the importance of pretraining and RLHF 508 provided by a physician. Future research will focus 458 in the medical domain, especially in medical 509 on enhancing the model's safety, reducing 459 dialogue generation and related tasks, but also 510 hallucinations, and exploring the integration of 460 validates the critical role of the meticulously 511 multimodal 461 designed datasets used during the training process. 512 comprehensive and accurate medical services. 462 These datasets have played a pivotal role across 513 However, CUMed-LLaMA currently remains a 463 pretraining, fine-tuning, and 464 learning stages, ensuring the model's accuracy and 515 professional medical consultation. Its effective 465 reliability when handling complex medical tasks. 516 collaboration with human experts remains a crucial 466 Looking ahead, we plan to build several mixture- 517 direction for future development. 467 of-experts (MoE) models based on the CUMed-468 LLaMA architecture. The MoE model dynamically 469 activates specialized expert networks based on the

⁴¹⁹ lower accuracy and professionalism in their ⁴⁷⁰ requirements of different tasks, thereby improving 471 both the efficiency and accuracy of the model. This As demonstrated in **Table 3**, we present example 472 approach not only optimizes the use of emotional 479 applications in the future, offering more precise

the 485 focuses on dialogue generation and lacks terms of medical 489 LLaMA performs well in medical-related tasks, it patient communication, 490 cannot guarantee the complete accuracy of all 494 decisions and to seek professional medical advice ⁴⁹⁵ when necessary. Another key limitation lies in the ⁴⁹⁶ model's language and communication skills. 497 Although CUMed-LLaMA is capable of 447 In this study, we introduce the CUMed-LLaMA 498 generating fluent dialogue, it cannot employ the exceptional 500 doctors use to reassure patients, address concerns, suggestions, it cannot replace the clinical judgment information to offer more reinforcement 514 research tool rather than a substitute for

518 6 References

		572
519	Chen, Y., Wang, Z., Xing, X., zheng, huimin, Xu,	573
520	Z., Fang, K., Wang, J., Li, S., Wu, J., Liu,	574
521	Q., & Xu, X. (2023). <i>BianQue: Balancing</i>	575
522	the questioning and suggestion ability of	576
523	health LLMs with multi-turn health	577
524	conversations polished by ChatGPT (No.	578
525	arXiv:2310.15896). arXiv.	579 L
526	https://doi.org/10.48550/arXiv.2310.158	580
527	96 D G LALL' A E D X D W	581
528	DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang,	582
529	B., WU, B., LU, C., Zhao, C., Deng, C., Zhang, C., Duan, C., Dai, D., Cua, D.	583 L
530	Zhang, C., Kuan, C., Dai, D., Guo, D., Yong D. Chen D. I. D. Li E. Lin F.	584
531	I alig, D., Chen, D., Ji, D., Li, E., Lili, F., Dai, F., Dan, \overline{Z} (2024), Deen Seek V2	585
532	Dal, Γ ., Pan, Z. (2024). DeepSeek-VS	586
533	arViv	587
534	alalv. https://doi.org/10.48550/arXiv.2412.104	588
535	27	589
536	Gan P. Wu Z. Sun P. Lu I. Wu Y. Zhang	590
537	D Pan K He I Tian V Vang P	591
538	$V_{ang} \cap W_{ang} H$ Zhang I & Song	592
539	Y (2024) Ziva2: Data-centric learning	593
540	is all LLMs need (No. arXiv:2311.03301)	594
5/2	arXiv	595
543	https://doi.org/10.48550/arXiv.2311.033	596 IN
544	01	597
545	Gudibande, A., Wallace, E., Snell, C., Geng, X.,	598
546	Liu, H., Abbeel, P., Levine, S., & Song,	599
547	D. (2023). The false promise of imitating	600
548	proprietary LLMs (No.	601
549	arXiv:2305.15717). arXiv.	603
550	https://doi.org/10.48550/arXiv.2305.157	604
551	17	605
552	Hua, R., Dong, X., Wei, Y., Shu, Z., Yang, P., Hu,	606
553	Y., Zhou, S., Sun, H., Yan, K., Yan, X.,	607
554	Chang, K., Li, X., Bai, Y., Zhang, R.,	608
555	Wang, W., & Zhou, X. (2024). Lingdan:	609
556	Enhancing encoding of traditional	610
557	Chinese medicine knowledge for clinical	611
558	reasoning tasks with large language	612
559	models. Journal of the American Medical	613
560	Informatics Association, 31(9), 2019–	614 P
561	2029.	615
562	https://doi.org/10.1093/jamia/ocae087	616
563	Li, S. S., Balachandran, V., Feng, S., Ilgen, J. S.,	617
564	Pierson, E., Koh, P. W., & Tsvetkov, Y.	618
565	(2024). MediQ: Question-asking LLMs	619
566	and a benchmark for reliable interactive	620
567	<i>clinical reasoning</i> (No.	621
568	arXiv:2406.00922). arXiv.	622
569	https://doi.org/10.48550/arXiv.2406.009	623
570	22	624 S

- 571 Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., &
 572 Zhang, Y. (2023). ChatDoctor: A medical
 573 chat model fine-tuned on a large
 574 language model meta-AI (LLaMA) using
 575 medical domain knowledge (No.
 576 arXiv:2303.14070). arXiv.
 577 https://doi.org/10.48550/arXiv.2303.140
 578 70
 - Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 74–81. https://aclanthology.org/W04-1013/
 - Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization (No. arXiv:1711.05101). arXiv. https://doi.org/10.48550/arXiv.1711.051 01
 - Lyu, S., Chi, C., Cai, H., Shi, L., Yang, X., Liu,
 L., Chen, X., Zhao, D., Zhang, Z., Lyu, X.,
 Zhang, M., Li, F., Ma, X., Shen, Y., Gu,
 J., Xue, W., & Huang, Y. (2024). *RJUA-QA: A comprehensive QA dataset for urology* (No. arXiv:2312.09785). arXiv.
 https://doi.org/10.48550/arXiv.2312.097
 85
 - Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., & Wu, H. (2018). *Mixed precision* training (No. arXiv:1710.03740). arXiv. https://doi.org/10.48550/arXiv.1710.037 40
 - Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback (No. arXiv:2203.02155). arXiv. https://doi.org/10.48550/arXiv.2203.021 55
 - Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.107313 5
- ⁶²⁴ Shoeybi, M., Patwary, M., Puri, R., LeGresley, P.,
 ⁶²⁵ Casper, J., & Catanzaro, B. (2020).

626 parameter language models using model 682 627 *parallelism* (No. arXiv:1909.08053). 683 628 arXiv. 629 684 https://doi.org/10.48550/arXiv.1909.080 685 630 53 631 686 Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., 687 632 Papernot, N., & Anderson, R. (2024). The 688 633 curse of recursion: Training on generated 689 634 models forget data makes 635 arXiv:2305.17493). arXiv. 691 636 https://doi.org/10.48550/arXiv.2305.174 692 637 93 638 639 Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, 694 J., Chung, H. W., Scales, N., Tanwani, A., 695 640 Cole-Lewis, H., Pfohl, S., Payne, P., 696 641 Seneviratne, M., Gamble, P., Kelly, C., 697 642 Babiker, A., Schärli, N., Chowdhery, A., 698 643 Mansfield, P., Demner-Fushman, D., ... 699 644 Natarajan, V. (2023). Large language 700 645 models encode clinical 646 620(7972), 172-180. 702 Nature. 647 https://doi.org/10.1038/s41586-023-703 648 06291-2 649 704 650 Singhal, K., Tu, T., Gottweis, J., Sayres, R., 705 Wulczyn, E., Hou, L., Clark, K., Pfohl, S., 706 651 Cole-Lewis, H., Neal, D., Schaekermann, 707 652 M., Wang, A., Amin, M., Lachgar, S., 708 653 Mansfield, P., Prakash, S., Green, B., 709 654 Dominowska, E., Arcas, B. A. y, ... 710 655 Natarajan, V. (2023). Towards expert- 711 656 level medical question answering with 712 657 large language models 658 arXiv:2305.09617). arXiv. 714 659 https://doi.org/10.48550/arXiv.2305.096 715 660 17 661 716 662 Srivastava, N., Hinton, G., Krizhevsky, A., 717 Sutskever, I., & Salakhutdinov, R. (2014). 718 663 Dropout: A simple way to prevent neural 719 664 networks from overfitting. J. Mach. 720 665 Learn. Res., 15(1), 1929–1958. 721 666 Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, 722 667 K., Gutierrez, L., Tan, T. F., & Ting, D. S. 723 668 W. (2023). Large language models in 724 669 medicine. Nature Medicine, 29(8), 1930-725 670 https://doi.org/10.1038/s41591-726 1940. 671 023-02448-8 672 727 Tian, Y., Gan, R., Song, Y., Zhang, J., & Zhang, 728 673 Y. (2024). ChiMed-GPT: A Chinese 729 674 medical large language model with full 730 675 training regime and better alignment to 731 676 human preferences (No. 732 677 arXiv:2311.06025). 678 https://doi.org/10.48550/arXiv.2311.060 734 679 25 735 680

Megatron-LM:Trainingmulti-billion681Touvron, H., Lavril, T., Izacard, G., Martinet, X.,parameter language models using model682Lachaux, M.-A., Lacroix, T., Rozière, B.,parallelism(No.arXiv:1909.08053).683Goyal, N., Hambro, E., Azhar, F.,arXiv.684Rodriguez, A., Joulin, A., Grave, E., &https://doi.org/10.48550/arXiv.1909.080685Lample, G. (2023). LLaMA: Open and53686efficient foundation language models (No.lov, I., Shumaylov, Z., Zhao, Y., Gal, Y.,687arXiv:2302.13971).Papernot, N., & Anderson, R. (2024). The688rXiv:2302.13971).curse of recursion: Training on generated68971

 (No. 690 Touvron, H., Martin, L., Stone, K., Albert, P.,

 rXiv. 691
 Almahairi, A., Babaei, Y., Bashlykov, N.,

 .174
 692
 Batra, S., Bhargava, P., Bhosale, S., Bikel,

 693
 D., Blecher, L., Ferrer, C. C., Chen, M.,

 Wei, 694
 Cucurull, G., Esiobu, D., Fernandes, J.,

 ii, A., 695
 Fu, J., Fu, W., ... Scialom, T. (2023).

 e, P., 696
 Llama 2: Open foundation and fine-tuned

 v, C., 697
 chat models (No. arXiv:2307.09288).

 v, A., 698
 arXiv.

 v, A., 699
 https://doi.org/10.48550/arXiv.2307.092

knowledge. 701VanVeen, D., Van Uden, C., Blankemeier, L.,
172–180. 702172–180. 702Delbrouck, J.-B., Aali, A., Bluethgen, C.,
Delbrouck, J.-B., Aali, A., Bluethgen, C.,
Pareek, A., Polacin, M., Reis, E. P.,
Seehofnerová, A., Rohatgi, N., Hosamani,
P., Collins, W., Ahuja, N., Langlotz, C. P.,
Hom, J., Gatidis, S., Pauly, J., &
Chaudhari, A. S. (2024). Adapted large
language models can outperform medical
experts in clinical text summarization.
Nature Medicine, 30(4), 1134–1142.
https://doi.org/10.1038/s41591-024-
02855-5

(No. 713 Wang, D., & Zhang, S. (2024). Large languagerXiv. 714models in medical and healthcare fields:.096715Applications, advances, and challenges..716Artificial Intelligence Review, 57(11),A., 717299. https://doi.org/10.1007/s10462-024-.014). 71810921-0

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). Large language models are not fair evaluators (No. arXiv:2305.17926). arXiv. https://doi.org/10.48550/arXiv.2305.179

Wu, C., Lin, W., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2023). PMC-LLaMA: Towards building open-source language models for medicine (No. arXiv:2304.14454). arXiv. https://doi.org/10.48550/arXiv.2304.144 54

arXiv. 733 Xiong, H., Wang, S., Zhu, Y., Zhao, Z., Liu, Y., 1.060 734 Huang, L., Wang, Q., & Shen, D. (2023). 735 DoctorGLM: Fine-tuning your Chinese

26

- *doctor is not a herculean task* (No. arXiv:2304.01097).
- https://doi.org/10.48550/arXiv.2304.010
 97
- Yang, S., Zhao, H., Zhu, S., Zhou, G., Xu, H., Jia, 740 Y., & Zan, H. (2023). Zhongjing: 741 Chinese Enhancing the medical 742 capabilities of large language model 743 through expert feedback and real-world 744 multi-turn dialogue (No. 745 arXiv:2308.03549). arXiv. 746 https://doi.org/10.48550/arXiv.2308.035 747 49 748
- 749 Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., & Ma, Y. (2024). LlamaFactory: 750 Unified efficient fine-tuning of 100+ 751 language models (No. 752 arXiv:2403.13372). arXiv. 753 https://doi.org/10.48550/arXiv.2403.133 754 72 755
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y.,
 Ma, X., Efrat, A., Yu, P., Yu, L., Zhang,
 S., Ghosh, G., Lewis, M., Zettlemoyer, L.,
 & Levy, O. (2023). LIMA: Less is more
 for alignment. Advances in Neural
 Information Processing Systems, 36,
 55006–55021.
- Zhou, L., Chen, Y., Li, X., Li, Y., Li, N., Wang, 763 X., & Zhang, R. (2024). A new adapter 764 tuning of large language model for 765 Chinese medical named entity 766 recognition. Artificial Applied 767 Intelligence, 38(1). 2385268. 768 https://doi.org/10.1080/08839514.2024.2 769 385268 770
- 771