



**[专家简介]** 何昆仑,解放军总医院医学创新研究部某中心主任,主任医师、教授,主要从事医疗大数据、医学人工智能、智能卫勤装备和系统等研究。担任智能医学工程国家级重点实验室主任,承担重大项目等27项,制定国家行业标准29项,授权发明专利103项,获国家科技进步二等奖1项,省部级一、二等奖5项。担任中国研究型医院学会副会长、北京医师协会副会长等。

## 基于电子病历数据和知识增强的医疗大语言模型构建方法研究

王博<sup>1</sup>, 于志昊<sup>2</sup>, 张军雁<sup>2</sup>, 石戈<sup>3</sup>, 冯冲<sup>1</sup>, 庄严<sup>2</sup>, 何昆仑<sup>2</sup>

<sup>1</sup>北京理工大学计算机学院, 北京 100081; <sup>2</sup>解放军总医院医学创新研究部, 医疗大数据应用技术国家工程研究中心, 解放军总医院医学工程实验室, 国家药监局人工智能医疗器械研究与评价重点实验室, 北京 100853; <sup>3</sup>北京工业大学信息科学技术学院, 北京 100124

**摘要:** **背景** 电子病历数据在构建医疗领域大规模语言模型中具有关键作用。**目的** 研究一种基于通用大语言模型的三阶段训练范式, 以充分挖掘电子病历数据的价值。**方法** 第一阶段, 利用大规模电子病历文本对预训练的通用模型进行进一步训练, 增强其医疗领域的语言知识; 第二阶段, 利用标注的电子病历数据针对特定临床任务对模型进行微调, 从而赋予其专业的任务处理能力; 第三阶段, 通过结合医师的反馈优化模型输出, 进一步提升其决策的准确性和可解释性。**结果** 该方法显著提升了模型在临床任务中的表现, 减少了模型产生幻觉的现象, 并增强了输出的可信度。**结论** 该研究为构建规范化、可信赖的医疗大规模语言模型提供了有效的方法, 具有重要的实际应用价值。

**关键词:** 大规模语言模型; 医疗人工智能; 继续训练; 有监督指令微调; 强化学习

中图分类号: R197.323

文献标志码: A

文章编号: 2095-5227(2025)01-0096-08

DOI: 10.12435/j.issn.2095-5227.24070107

引用本文: 王博, 于志昊, 张军雁, 等. 基于电子病历数据和知识增强的医疗大语言模型构建方法研究 [J]. 解放军医学院学报, 2025, 46 (1): 96-103.

## A paradigm for constructing medical large language models based on electronic health record data and knowledge enhancement

WANG Bo<sup>1</sup>, YU Zhihao<sup>2</sup>, ZHANG Junyan<sup>2</sup>, SHI Ge<sup>3</sup>, FENG Chong<sup>1</sup>, ZHUANG Yan<sup>2</sup>, HE Kunlun<sup>2</sup>

<sup>1</sup>School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China; <sup>2</sup>Medical Innovation Research Department of PLA General Hospital, National Engineering Research Center for Medical Big Data Application Technology, Medical Engineering Laboratory of PLA General Hospital, Key Laboratory of Artificial Intelligence Medical Device Research and Evaluation of the National Medical Products Administration, Beijing 100853, China; <sup>3</sup>Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

Corresponding author: HE Kunlun. Email: kunlunhe@301hospital.com.cn

**Abstract: Background** Electronic Medical Records (EMRs) play a pivotal role in training large-scale language models (LLMs) within the medical domain. **Objective** To explore the value of electronic medical record data by studying a three-stage training paradigm based on a general large language model. **Methods** Firstly, in the continued training phase, extensive EMR texts were employed to further train the pre-existing general model, thereby enhancing its medical-specific linguistic knowledge. Secondly, during the supervised fine-tuning phase, annotated EMR data were utilized to modify the model for specific clinical tasks such as medical named entity recognition and clinical trial screening, enabling the model to acquire specialized task-oriented skills. Finally, in the reinforcement learning phase, feedback from doctors was integrated to optimize the model's outputs, improving the accuracy

收稿日期: 2024-07-01

基金项目: 新一代人工智能国家科技重大专项(2021ZD0140408); 省部级课题

第一作者: 王博, 硕士。Email: bwang@bit.edu.cn

通信作者: 何昆仑, 博士, 主任医师, 教授, 博士生导师。Email: kunlunhe@301hospital.com.cn

and interpretability of decision-making. **Results** Experimental results demonstrated that the model's performance in clinical tasks was significantly enhanced, the occurrence of hallucinations was mitigated, and the reliability of its outputs was improved. **Conclusion** This study provides an effective approach for constructing standardized and trustworthy medical LLMs, offering substantial practical application value.

**Keywords:** large language models; medical artificial intelligence; continue training; supervised instruction tuning; reinforcement learning

**Cited as:** Wang B, Yu ZHH, Zhang JY, et al. A paradigm for constructing medical large language models based on electronic health record data and knowledge enhancement[J]. Acad J Chin PLA Med Sch, 2025, 46(1): 96-103.

大规模语言模型 (large language models, LLMs) 是近年来自然语言处理 (natural language processing, NLP) 领域的重大突破。不同于以往的 BERT<sup>[1]</sup> 等传统预训练语言模型, LLMs 通过在万亿级别的海量文本语料中进行预训练, 习得了丰富的语言知识和常识推理能力, 展现出了惊人的小样本学习乃至零样本泛化能力, 无需大量标注数据即可快速适应下游任务。此外, LLMs 还具备强大的语言生成能力, 可根据给定的上下文生成连贯、流畅、富有逻辑的文本。ChatGPT、PaLM、LLaMA 等代表性 LLMs 的出现, 标志着 NLP 正在从“特定任务导向”转向“通用人工智能”。

LLMs 为医疗 AI 注入了新的活力。医学知识的获取和应用高度依赖非结构化文本数据, 如电子病历 (electronic medical records, EMRs)、医学文献等。传统的医疗 AI 方法难以充分挖掘这些文本数据的价值。而 LLMs 恰恰擅长处理海量文本数据, 为医疗领域的知识表示和推理应用开辟了新的途径。一方面, LLMs 可通过在大规模医学文献语料中进行持续预训练, 学习医学领域的专业词汇、概念、实体关系等, 构建全面的医学知识库; 预训练好的医疗 LLMs 可辅助医师进行疾病诊断、药物推荐、病历分析、医患沟通等。另一方面, LLMs 还可应用于 EMRs 的自动化处理。原始 EMRs 通常格式不一、缺失值多、噪声大, 给后续分析带来困难。LLMs 可将非结构化的 EMRs 转化为结构化的信息, 提取患者的主诉、病史、检查、诊断、治疗等关键信息, 从而为疾病预测、药物警戒、智能问答等应用奠定基础。最新的研究成果表明, 在大规模医学文本数据的支持下, 医疗 LLMs 可在多项任务上达到甚至超越人类医师的水平。如谷歌研究团队提出的 MedPaLM<sup>[2]</sup> 模型在 PubMed 文献上进行预训练后, 在医学问答、临床决策支持等任务上表现出色。清华大学开发的 PMC-LLaMA<sup>[3]</sup> 模型则利用 PubMed Central (PMC) 全文数据进行训练, 在医疗命名实体识别、关系抽取等任务上取得了优异的效果。

尽管医疗 LLMs 取得了显著进展, 针对医疗场景构建 LLMs 仍面临诸多挑战。首先, 获取高质量且丰富的医学文本语料成本极高, 不同 LLMs 所使用的数据质量和覆盖范围差异较大, 影响模型的通用性和可靠性。医学文本中包含大量复杂的专业术语和领域特定的表达方式, 且存在多义性和上下文依赖性, 增加了模型理解和生成的难度。其次, 现有模型在医学领域的应用缺乏针对性的优化设计, 难以精准捕捉医疗专业知识的复杂性和细粒度特征。特别是在处理涉及多学科交叉的复杂医学问题时, 模型的推理能力和决策准确性亟需提升。最后, 随着医学知识的快速更新, 模型知识库的更新滞后现象突出, 导致模型易于出现“幻觉”, 生成不符合事实的回答, 威胁临床应用的安全性和可信度。此外, 医学数据的隐私性和敏感性限制了数据的共享和使用, 进一步增加了模型训练和优化的难度。

为应对上述挑战, 本文基于通用大语言模型, 提出了一种包含继续训练、有监督微调和强化学习的三阶段训练范式, 并进一步构建了基于检索增强的知识获取架构, 进一步强化模型在复杂医学问题上的推理能力。实验结果表明, 本方法在 C-EVAL 和 CMMLU 两个标准中文评测基准中的医疗领域问题中表现优异, 验证了其有效性。

## 1 医疗领域基础语言模型构建方法

本文针对医疗大语言模型在知识时效性和临床推理能力方面的局限, 提出了一种基于电子病历数据和知识增强的模型构建方法。如图 1 所示, 该方法包含训练和推理两个核心环节。在训练环节, 设计了一个仿生式三阶段训练框架, 分别对应基础医学教育、临床实习和正式医师 3 个阶段。具体而言, 首先通过医学文献和电子病历数据的继续训练构建领域知识表示, 然后利用标注数据进行有监督指令微调以获取特定临床任务能力, 最后基于专家反馈的强化学习提升模型决策水平。在推理环节, 设计了基于检索的知识增强机制,

通过实时调用外部医学知识库为模型推理提供动态知识支持。这种静态训练与动态知识增强相结合的范式设计，旨在构建一个具备持续学习能力和可靠临床决策能力的医疗大语言模型。下文将详细介绍训练框架和知识增强机制的具体设计。

### 1.1 任务定义

给定预训练语言模型  $M$ ，医疗文本语料集  $D_{med} = \{x_1, x_2, \dots, x_n\}$ ，医疗指令数据集  $D_{inst} = \{(q_i, a_i)\}$ ，以及医师反馈数据集  $D_{fb} = \{(c_j, r_j, s_j)\}$ 。其中， $x_i$  表示医学文献和病历等领域文本， $(q_i, a_i)$  为指令和目标响应对， $(c_j, r_j, s_j)$  分别表示临床场景、模型响应和专家评分。

### 1.2 基于继续训练的领域知识获取

本研究构建了一个大规模的医疗文本语料库用于模型继续训练。该语料库的构建过程包括数据采集和系统化预处理两个关键步骤。在数据采集阶段，我们从4个维度收集医疗领域文本：(1)专业医学教材与教辅材料；(2)医学期刊论文及会议文献；(3)标准化医学知识库；(4)电子病历数据集。其中，电子病历数据主要来自MIMIC-III重症监护数据库和经过脱敏处理的医院系统记录。为确保语料质量和标准化程度，本文设计了一套多阶段的数据预处理框架。该框架首先通过文本过滤算法剔除非医学相关内容，采用基于规则的方法识别并移除数据噪声。随后，利用医学文本分词模型进行语句切分和分词标注，该模型在医学领域语料上进行了特定优化。

在此基础上，我们基于 UMLS (Unified Medical Language System) 构建了医学实体标准化模块，将不同表述形式的医学概念统一映射至标准术语体系。之后，将处理后的数据转换为结构化格式，构建了包含 8.7 亿 Token 的医疗文本语料库。经统计分析，该语料库覆盖了常见医学概念。本阶段采用只有解码结构的模型架构进行继续训练。对于输入序列  $x = (x_1, x_2, \dots, x_T)$ ，模型通过最大化序列的联合概率来优化参数：

$$L_{pt} = - \sum_{i=1}^{|D_{med}|} \sum_{t=1}^T \log P(x_t^{(i)} | x_{<t}^{(i)}; \theta)$$

其中  $\theta$  为模型参数， $x_{<t}$  表示位置  $t$  之前的所有单元。

### 1.3 有监督医疗指令微调

本文构建了一个多任务导向的医疗指令数据集，包含4类核心子集：(1)症状-诊断数据集，将患者症状描述作为指令输入，相应诊断结果作为目标响应；(2)诊断-治疗数据集，以确诊结果为指令，对应治疗方案为响应；(3)拟诊讨论数据集，输入患者病情描述，输出对鉴别诊断及可能原因的系统性分析；(4)通用医疗指令数据集，基于临床医师实际工作指令，采用 Self-Instruct 方法进行数据扩充。

为规范化指令表达，针对不同任务设计了特定的模板格式。如在症状-诊断任务中，采用“根据以下症状进行诊断：[症状描述]”的标准化表达；在拟诊讨论任务中，使用“请对以下病例进行鉴别诊断分析：[病情描述]”的模板结构。同时

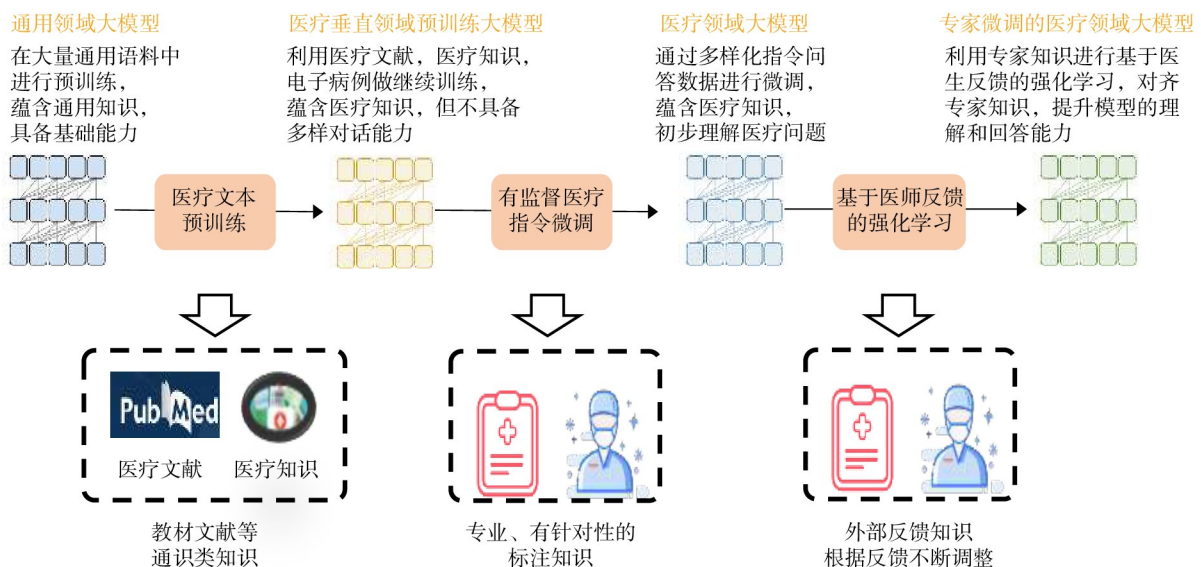


图1 医疗领域基础语言模型构建方法

Fig. 1 Construction of medical language models

对原始数据进行了系统化预处理，包括文本规范化、实体标准化和质量过滤等步骤。本阶段采用序列到序列的架构进行指令微调。将输入指令  $q$  和目标响应  $a$  拼接为完整序列  $[q; sep; a]$ ，通过最小化响应部分的负对数似然进行优化：

$$L_{ft} = - \sum_{i=1}^{|D_{inst}|} \sum_{t=1}^{|a^{(i)}|} \log P(a_t^{(i)} | a_{<t}^{(i)}, q^{(i)}; \theta)$$

#### 1.4 基于医师反馈的强化学习方法

在图1中，经过有监督指令微调，语言模型已经具备执行特定医疗任务的能力。然而，为确保模型在实际应用中的安全性、可控性和符合伦理标准，需要进一步开展基于医师反馈的强化学习过程，对其进行人类价值对齐。

本研究采用强化学习方法对模型进行价值对齐优化。针对强化学习阶段的数据构建，首先从3个维度收集原始数据：(1)医疗伦理指南文献，包括赫尔辛基宣言、日内瓦宣言等标准规范；(2)临床案例分析，对实际医疗决策案例的系统评估；(3)行业标准与规范，反映医疗实践中的专业准则。随后进行标准化处理：构建情景-决策对，对文本进行规范化和质量过滤，通过分层采样确保不同评分区间样本分布均衡。

在此基础上，基于BERT架构训练Reward模型用于评估模型输出的合规性。Reward模型将临床场景  $c$  和模型响应  $r$  拼接后输入BERT，通过分类头输出评分  $s$ 。其损失函数定义为：

$$L_{reward} = \text{MSE}(f_{\phi}([CLS]), y)$$

其中  $f_{\phi}$  为分类头映射函数，将[CLS]标记的输出映射为标量评分， $y$  为专家标注分数。在强化学习阶段，采用PPO算法进行优化。对于给定的临床场景  $c$ ，模型生成响应  $r$  后由Reward模型给出评分  $s$ 。PPO的目标函数定义为：

$$L_{rl} = \min(\rho_t(\theta) A_t, \text{clip}(\rho_t(\theta), 1 - \varepsilon, 1 + \varepsilon) A_t)$$

其中  $\rho_t(\theta)$  表示新旧策略的概率比， $A_t$  为基于Reward评分计算的优势函数：

$$A_t = s_t + \gamma V(s_{t+1}) - V(s_t)$$

$\varepsilon$  为裁剪系数，设为0.2。为平衡探索与利用，采用自适应KL惩罚项：

$$L_{total} = L_{rl} + \beta \frac{1}{|D_{fb}|} \sum_{j=1}^{|D_{fb}|} \text{KL}(\pi_{\theta}^{(j)} \| \pi_{\theta_{old}}^{(j)})$$

其中  $\beta$  根据实际KL散度动态调整。通过此训练框架，模型能够基于专家反馈不断优化其决策

行为，提升输出的规范性。

#### 1.5 基于检索的知识增强推理

在图1中，外部反馈和知识动态融合阶段对应的是模型检索式知识增强过程。尽管语言模型通过前期的预训练和微调已掌握大量医学知识，但这些知识仍是静态的，无法实时更新和扩展。为进一步提升模型的医疗推理能力，使其能够应对不断变化的医疗需求和复杂的医学问题，本文引入基于检索的知识增强方法，类似于医师在诊疗过程中参考最新医学文献或指南，以确保其决策的准确性和时效性。

在知识库构建方面，本文开发了一个融合多源异构数据的医学知识库系统，该系统包含以下3方面。(1)利用医学教材、词典、百科等资源构建的大规模知识库，涵盖疾病、药品、治疗、解剖等多个医学子领域的核心概念；(2)临床指南和共识：来源于UpToDate、DynaMed等综合性临床决策支持系统，NGC、GIN等临床指南数据库的权威医学指南；(3)病例报告语料库：由脱敏后的EMRs提取而来的真实病例报告。

在知识检索阶段，本文采用BGE作为文本编码器。对于输入查询  $q$  和知识库文档  $d$ ，分别计算其向量表示：

$$e_q = \text{BGE}(q), e_d = \text{BGE}(d)$$

文档相关度通过余弦相似度计算：

$$s = \cos(e_q, e_d) = \frac{e_q \cdot e_d}{\|e_q\| \cdot \|e_d\|}$$

选取相似度最高的Top-k文档作为上下文知识。为了整合检索知识，我们采用基于Prompt的知识注入方式，将检索到的相关文档作为附加的上下文信息，输入查询为  $q$ ，检索到的Top-k相关文档集合为  $D_k = \{d_1, \dots, d_k\}$ 。知识融入采用序列拼接方式：

$$X = [q; d_1; d_2; \dots; d_k]$$

其中  $[; ]$  表示序列拼接操作。模型基于拼接后的完整序列  $X$  生成回答  $y = \text{LLM}(X)$ ，序列拼接方式保持了检索知识的完整语义信息，使模型能够在推理过程中充分利用外部知识。

## 2 模型验证

为了全面评估本文提出的医疗领域大模型训练范式的有效性，本文设计了一系列实验，分别从通用领域模型对比、电子病历数据增强、检索

式知识融合3个角度,展示本文方法在医疗问答任务上的优势。

## 2.1 数据描述

为了全面评估本文提出的医疗大语言模型在临床实践中的实用性,我们构建了多源异构的评测数据集。首先,从公开数据集方面,选取了C-Eval和CMMLU两个标准中文评测基准中的医疗领域问题。这些数据集包含了结构化的多选题形式,可以客观评估模型在医学知识掌握和推理能力方面的表现。

在真实临床场景数据方面,经过严格的脱敏处理和质量控制,所选数据包包含住院病历、手术记录、检查报告等多个阶段。此外,我们特别筛选出300份完整的典型病例作为评测集,这些病例涵盖了内科、外科、妇产科等多个专科,能够全面检验模型在不同临床场景下的表现。每份病例都包含完整的病史、体格检查、辅助检查、诊疗过程等信息,并由高年资主治医师进行标注和质量审核。

## 2.2 实现细节

本研究基于Baichuan2-7B进行实验。所有训练在8×NVIDIA A100 GPU上进行。在继续预训练阶段,Batch size设置为128,使用AdamW优化器( $\beta_1=0.9$ ,  $\beta_2=0.999$ ),初始学习率为 $1e-5$ ,采用线性预热策略(warmup\_steps=1 000)。训练持续100 K steps,使用混合精度训练,梯度裁剪阈值设为1.0。在指令微调阶段,采用LoRA进行参数高效微调,其中Rank=8,  $\alpha=32$ 。Batch size设为32,学习率为 $2e-4$ ,训练10个Epochs。在RLHF阶段,PPO算法的Clip ratio  $\epsilon$ 设为0.2,初始KL惩罚系数 $\beta=0.1$ 。采用Mini-batch训练,Batch size为8,每轮训练4个Epochs。Reward模型基于BERT-base构建,使用AdamW优化器,学习率设为 $1e-5$ 。推理检索增强每个输入Query通过检索获取相似度最高的知识作为上下文增强。

## 2.3 公开基准评测

为评估模型的整体性能,本文首先在中文公共数据集C-Eval和CMMLU中进行了系统测试。在评估过程中,我们专门提取了与医学相关的问题子集进行测试,以确保评估结果能够准确反映模型在医疗领域的实际表现。本文对比了几种现有的基线模型。

HuatuoGPT<sup>[4]</sup>, HuatuoGPT-II<sup>[5]</sup>: 基于通用大语言模型调整的医学专用模型,旨在提升在医学

文本任务中的表现。

Llama2-Chat (<https://huggingface.co/THUDM/chatglm2-6b>): 属于Meta的LLaMA系列,经过微调后适用于对话生成任务,并在多种语言理解任务中取得较好的效果。

ChatGLM2, ChatGLM3 (<https://huggingface.co/THUDM/chatglm3-6b>): 基于解码结构的通用对话模型,具备强大的中文处理能力,在中文语境中的表现尤为突出。

Baichuan2-Chat (<https://huggingface.co/baichuan-inc/Baichuan2-13B-Chat>): 百川智能推出的新一代开源大语言模型,主要应用于中文理解和生成。

Qwen-7B-Chat (<https://huggingface.co/Qwen/Qwen1.5-7B-Chat>): 面向多任务学习的模型,具备较强的推理能力和处理能力,在医学推理任务中有显著的优势。

ChatGPT (<https://chat.openai.com>): 由OpenAI开发的广泛应用的对话生成模型,在多个领域,包括医学领域,具备较好的通用性和推理能力。

如表1所示,通过与这些基线模型比较,验证了本文提出方法的有效性和优势,本模型在两个评测基准上均取得了最优成绩。具体而言,在CMMLU医疗相关任务上,本模型达到60.30%的准确率,显著优于通用大模型如ChatGPT(50.37%)和Baichuan2-Chat(52.95%),同时也超过了专门面向医疗领域的HuatuoGPT-II(59.08%)。在C-Eval医疗领域测试集上,本模型的表现更为突出,以65.43%的准确率领先第2名HuatuoGPT-II(62.40%)约3个百分点。同时,本模型也大幅超越了Llama2-Chat(32.64%, 31.47%)等通用基座模型的表现,表明针对医疗领域的特殊优化确实带来了实质性的性能提升。这些结果共同验证了本模型在医学知识掌握和专业推理能力方面的优越性。

## 2.4 临床场景评测

为评估基于检索的知识增强方法在实际医疗场景中的效果,我们设计了系统性的对比实验。实验采用两种方案:Base(不使用任何外部知识的原始模型)和Re-Aug(将检索到的医学文献、临床指南等知识通过简单拼接作为额外上下文输入)。测试数据源于300份真实病历,包含完整的病史、检查结果等信息。本实验设置中,评估指标包含了生成文本的平均长度(Len)、人工评估得分(Manual)、GPT-4自动评分(GPT-4)、ROUGE-L分数( $R_L$ )、F1分数( $F_1$ )、精确匹配率(EM)、命中率

表 1 不同模型在 C-Eval 和 CMMLU 医疗公共基准中心测试结果

Tab. 1 Test results of different models in C-Eval and CMMLU medical public benchmark dataset

模型	CMMLU/%	C-Eval/%
HuatuogPT	33.23	36.53
Llama2-Chat	32.64	31.47
ChatGLM2-	43.87	47.20
ChatGLM3	46.97	48.80
Baichuan2-Chat	52.95	58.67
Qwen-7B-Chat	54.65	52.80
ChatGPT	50.37	48.80
HuatuogPT- II	59.08	62.40
本文模型	60.30	65.43

(Hit)、平均效用(MAU)和多样性(Dis)。

实验分别在 3 种数据处理场景下进行, 即原始 EMR、压缩后的 EMR 以及压缩后的 EMR 与实验室数据组合。在原始 EMR 场景中, 检索增强方法使人工评分从 2.84 提升至 3.02, GPT-4 评分提升 0.15(3.31→3.46), ROUGE-L 和 F1 分别提升 2.16% 和 1.42%, 同时生成文本的多样性提升 2.20%。在压缩 EMR 场景下, 尽管基线表现相对较低, 检索增强方法仍带来了显著改善, 特别是在 ROUGE-L (提升 3.41%) 和 F1 (提升 5.05%) 等客观指标上。在最复杂的 EMR 与实验室数据组合场景中, 检索增强方法表现出最强的改进效果, 人工评分提升 0.31 (2.97→3.28), GPT-4 评分提升 0.22, 精确匹配率提升 4.89%(表 2)。从结果来看, 实验室结果缺乏足够的临床上下文支持, 可能导致模型产生片面理解, 完整的 EMR 和实验室数据组合会导致输入序列过长, 影响模型处理效率。通过对实验室数据

进行必要压缩, 我们在保持关键信息的同时优化了模型输入。实验结果充分证明了检索增强方法在提升模型临床决策支持能力方面的显著效果。该方法不仅提高了模型在各项指标上的表现, 更重要的是展现出了在处理不同类型医疗数据时的稳健性能。特别是在复杂病例分析中, 检索增强机制帮助模型提供了更全面、更准确的临床建议, 这对于提升医疗大语言模型的实际应用价值具有重要意义。

## 2.5 典型案例评测

本文首先将医疗领域大模型与 ChatGLM2-6B 和 GPT-4 进行人工对比, 以评估领域适应训练的必要性。从定性分析来看, ChatGLM2-6B 在医学知识和表达专业性上存在明显不足。以“女, 51 岁, 已婚, 山西省籍。主因进食吞咽困难 6 年于入院。”的病例为例, ChatGLM2-6B 虽然给出了一些初步的诊断思路, 如考虑食管痉挛、进一步检查等, 但推断欠缺深入的依据, 治疗建议也比较笼统, 整体专业性有限。相比之下, 本模型基于电子病历数据训练, 在医学专业知识和语言组织上都更接近真实医师。它能够通过系统的鉴别诊断分析, 逐一讨论食管良性狭窄、食管贲门癌、Chagas 病等可能性及依据, 体现了扎实的专业知识功底和缜密的临床思维。GPT-4 在医学知识的广度上则表现出强大的优势, 从贲门失弛缓、胃底黏膜下肿物、慢性胃炎等多个角度, 对患者的病情进行了全面梳理, 但 GPT-4 的表述相对较为口语化和冗长, 而本模型更加简洁凝练, 与医师的书写习惯更为贴近。综合来看, 尽管 GPT-4 展现了惊人的医学洞见, 但本模型在专业性和实用性上更胜一筹, 特别是在与

表 2 拟诊讨论任务在检索增强实验中的效果

Tab. 2 Outcomes of retrieval enhancement on diagnostic discussion tasks

实验	长度	Manual	GPT-4	$R_L$ /%	$F_1$ /%	EM/%	Hit	MAU/%	Dis/%
EMR									
Base	656.91	2.84	3.31	35.92	60.04	70.22	75.00	62.20	75.98
Re-Aug	631.60	3.02	3.46	38.08	61.46	67.62	77.00	68.66	78.18
改进	/	+0.18	+0.15	+2.16	+1.42	-2.60	+2.00	+6.46	+2.20
压缩 EMR									
Base	682.85	2.53	3.18	30.69	54.09	48.80	34.00	60.17	73.39
Re-Aug	645.12	2.60	3.22	34.10	59.14	47.48	27.00	67.50	75.19
改进	/	+0.07	+0.04	+3.41	+5.05	-1.32	-7.00	+7.33	+1.80
压缩 EMR & 实验检查结果									
Base	556.83	2.97	3.17	29.61	51.32	42.97	25.00	49.48	74.47
Re-Aug	610.80	3.28	3.39	32.91	56.44	47.86	29.00	47.13	75.55
改进	/	+0.31	+0.22	+3.30	+5.12	+4.89	+4.00	-2.35	+1.08

医师的日常工作流程结合时,结果见表3。

为了直观展示电子病历数据的价值,将本模型与医疗领域顶级模型 HuatuoGPT 进行了对比。以“心衰患者伴低血压应如何处理”的问题为例, HuatuoGPT 给出的建议虽然涵盖了一些基本的处理原则,如药物治疗、生活方式干预等,但总体比较笼统,缺乏针对性。而本模型则提供了更加具体、操作性更强的方案,如增加液体摄入、使用血管活性药物、给予氧疗和抗心律失常治疗等。电子病历数据帮助模型形成了更加实践化的思维模式,能够给出更符合临床实际的建议,见表4。

### 3 总结与展望

近年来,LLMs的发展推动了医疗领域专用模型的研究。通过进一步训练和任务特定优化,这些医疗大模型在知识表示和推理能力方面有了显著进步<sup>[6]</sup>。尽管这种方法在微调评估和知识探测测

试中表现出一定的提升,但在某些任务中可能会导致性能的显著下降。这反映了领域知识嵌入与通用能力保持之间的权衡问题。此外,研究表明,通过精心挑选的小规模数据集训练,有可能使模型性能接近更大模型<sup>[7-9]</sup>。医疗大模型通用大语言模型的成功推动了医疗领域专用模型的研究。如 MedPaLM 系列<sup>[2]</sup>在 MedQA 基准测试中达到60%以上的准确率,据报道已超越人类专家。现有医疗大模型主要采用两种方法:(1)利用通用大模型的提示<sup>[10-14]</sup>,通过设计任务特定的提示语直接应用通用模型,虽然无需训练但受限于原始模型的能力;(2)基于医疗数据的进一步训练<sup>[15-17]</sup>,通过医疗语料微调模型,嵌入领域知识,但需要大量计算资源,如 HuatuoGPT 等<sup>[4-5,18]</sup>。前者尽管无需训练,但受限于原始模型的能力;后者通过领域知识嵌入显著提升性能,但面临计算资源需求较高的挑战。

表3 EMR LLMs与通用领域大模型的对比实验

Tab. 3 Comparison between EMR LLMs and general domain LLMs

问题	患者一般情况:女,51岁,已婚,山西省籍,主因进食吞咽困难6年入院,病例特点如下。(1)中年女性,51岁;(2)病史特点:慢性病程,患者于2013年无明显诱因出现进食吞咽困难症状,每次进食需用水送服,夜间食物反流,间断呛咳,夜间偶有胸部针扎样疼痛,持续1~2 min缓解,胃底黏膜下肿物,慢性非萎缩性胃炎伴糜烂,未进行特殊治疗,就诊于我院,我院肠胃造影显示贲门失弛缓症,于我院行高分辨率食管测压检查诊断为贲门失弛缓症(Chicago II型),于我院行食管pH-阻抗监测,结果未见病理性胃食管反流。针对病例特点,医师可以推断出患者有什么样的问题?
Chatglm2-6B	针对病例特点,医师可以推断患者可能存在食管痉挛,导致夜间反流症状加重,同时出现胸部疼痛,另外患者还有体质量下降、慢性非萎缩性胃炎等病史,增加了食管痉挛的可能性,对于患者可能患病的问题,医师可以考虑食管痉挛的可能性同时需要进一步检查以明确病因。可以进行食管痉挛的辅助检查,如食管pH2-阻抗监测等,以明确是否存在食管痉挛。可使用胃酸分泌抑制剂、抗酸药等,以缓解症状。同时,需要关注患者的饮食和生活习惯,进行相关疾病的治疗,避免进食过快、过于油腻等,以减轻症状。
GPT-4	根据您提供的病例特点,患者的主要问题可能包括以下几点。 (1)贲门失弛缓症:这是一种食管运动障碍,表现为食管下段和贲门肌肉不能正常放松,导致食物难以进入胃部。患者的吞咽困难、食物反流、夜间呛咳等症状,以及高分辨率食管测压检查的结果,都支持这一诊断。 (2)胃底黏膜下肿物:这可能是良性或恶性的,需要进一步的检查和诊断。 (3)慢性非萎缩性胃炎伴糜烂:这是一种胃炎,表现为胃黏膜的炎症和糜烂。这可能是患者胸部针扎样疼痛的原因。 (4)体质量下降:可能是由于吞咽困难和食物反流导致的营养不良,需要注意的是,这些都是可能的问题,具体的诊断需要根据患者的详细病史、体格检查、实验室检查和影像学检查结果来确定。
本模型	拟诊讨论:患者为进食吞咽困难6年入院,初步诊断为贲门失弛缓症,现就其诊断讨论如下。 (1)食管良性狭窄:如创伤性食管炎及Barrett食管等疾病,患者可因瘢痕形成而导致食管狭窄,引起进食吞咽不畅等症状,患者既往多有食管创伤长期反酸、烧心等病史。该患者既往无食管创伤等病史,根据已完善胃镜检查,考虑该病可能性小。 (2)食管贲门癌:食管贲门癌造成的狭窄是由于癌组织浸润管壁所致,黏膜有破坏,可形成溃疡、肿块等改变,病变多以管壁的一侧为主,狭窄被动扩张性差,内镜通过阻力较大,狭窄严重者,常无法通过,强力插镜易造成穿孔,此患者为中年女性已完善胃镜检查,未见有食管占位等病变,考虑该病可能性小。 (3)Chagas病:可能患有巨食管症,为南美局部流行的锥虫寄生所致,并同时累及全身器官,其临床表现与失弛缓症不易区别。由于继发于寄生虫感染使肠肌丛退化,在生理学、药理学及治疗反应上与原发性失弛缓症相似,Chagas病除食管病变外,尚有其他内脏的改变。

表 4 EMR LLMs与医疗领域大模型(无病历数据训练)的对比

Tab. 4 Comparison of EMR LLMs and large models in the medical field (trained without medical record data)

问题	患者一般情况:女,51岁,已婚,山西省籍,主因进食吞咽困难6年入院,病例特点如下。(1)中年女性,51岁;(2)病史特点:慢性病程,患者于2013年无明显诱因出现进食吞咽困难症状,每次进食需用水送服,夜间食物反流,间断呛咳,夜间偶有胸部针扎样疼痛,持续1~2 min缓解,胃底黏膜下肿物,慢性非萎缩性胃炎伴糜烂,未进行特殊治疗,就诊于我院,我院肠胃造影示贲门失弛缓症,于我院行高分辨率食管测压检查诊断为贲门失弛缓症(ChicagoII型),于我院行食管pH-阻抗监测,结果未见病理性食管反流。针对病例特点,医师可以推断出患者有什么样的问题?
HuatuoGPT	基于你提供的信息,这个患者的主要问题是吞咽困难、反流以及胸骨后持续性疼痛,这些症状可能与她的胃肠道疾病有关,包括幽门螺杆菌感染或胃动力障碍等,此外根据她的医疗记录,她之前曾在多家医院就诊并接受过多种检查,这表明她在治疗和管理自己的健康方面已经采取了一些措施,然而由于没有更多的信息和测试结果,我们无法确定病情的具体原因。
本模型	拟诊讨论:患者为进食吞咽困难6年入院,初步诊断为贲门失弛缓症,现就其诊断讨论如下。 (1)食管良性狭窄:如创伤性食管炎及Barrett食管等疾病,患者可因瘢痕形成而导致食管狭窄,引起进食吞咽不畅等症状,患者既往多有食管创伤长期反酸、烧心等病史。该患者既往无食管创伤等病史,根据已完善胃镜检查,考虑该病可能性小。 (2)食管贲门癌:食管贲门癌造成的狭窄是由于癌组织浸润管壁所致,黏膜有破坏,可形成溃疡、肿块等改变,病变多以管壁的一侧为主,狭窄被动扩张性差,内镜通过阻力较大,狭窄严重者,常无法通过,强力插镜易造成穿孔,此患者为中年女性已完善胃镜检查,未见有食管占位等病变,考虑该病可能性小。 (3)Chagas病:可能患有巨食管症,为南美局部流行的锥虫寄生所致,并同时累及全身器官,其临床表现与失弛缓症不易区别。由于继发于寄生虫感染使肠肌丛退化,在生理学、药理学及治疗反应上与原发性失弛缓症相似,Chagas病除食管病变外,尚有其他内脏的改变。

## 4 总结与展望

本文提出了一个基于电子病历数据的三阶段医疗LLMs训练范式,包括继续训练、有监督微调与强化学习,并在此基础上探索了基于检索的知识增强技术。结果表明,该方法可在医学知识获取、任务执行与价值对齐方面显著提升模型性能,缓解知识滞后问题,并增强模型输出的准确性、全面性和可解释性。与现有模型相比,该方法在适配医疗领域的同时,显著提升了推理能力与知识更新效率,并针对临床任务的实际需求进行了精细化优化。

本研究为构建高效、可信的医疗领域大规模语言模型提供了有力支撑和重要参考,但医疗LLMs在技术、数据等方面仍有很大的优化空间。在技术层面,需要探索更高效的模型架构和训练范式,进一步提升LLMs的表示能力和泛化能力。在数据层面,未来需要拓展电子病历的采集渠道,获取更加全面、多样、有代表性的医疗数据,同时严格遵循隐私保护等相关法规。

**作者贡献** 王博:进行实验验证,理论分析;于志昊、张军雁、石戈:辅助实验验证和数据处理;冯冲、庄严:审读和修订;何昆仑:监督指导。

**利益冲突** 所有作者声明无利益冲突。

**数据共享声明** 本论文相关数据可依据合理

理由从作者处获取, Email: kunlunhe@plagh.org。

## 参考文献

- Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. <https://arxiv.org/abs/1810.04805v2>.
- Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models [J/OL]. <https://doi.org/10.1038/s41591-024-03423-7>.
- Wu C, Lin W, Zhang X, et al. PMC-LLaMA: toward building open-source language models for medicine [J]. *J Am Med Inform Assoc*, 2024, 31 (9): 1833-1843.
- Zhang HB, Chen JY, Jiang F, et al. HuatuoGPT, towards taming language model to be a doctor [C] //Findings of the Association for Computational Linguistics; EMNLP 2023. Stroudsburg: ACL, 2023: 10859-10885.
- Chen J, Wang X, Ji K, et al. HuatuoGPT-II, One-stage Training for Medical Adaption of LLMs [C/OL]. <https://doi.org/10.48550/arXiv.2311.09774>.
- Zhang N, Liu YC, Zhao XJ, et al. Pruning as a domain-specific LLM extractor [C] //Findings of the Association for Computational Linguistics; NAACL 2024. Stroudsburg: ACL, 2024: 1417-1428.
- Bețianu M, Mălan A, Aldinucci M, et al. DALLMi: domain adaptation for LLM-based multi-label classifier [C] //Advances in Knowledge Discovery and Data Mining. Singapore: Springer Nature Singapore, 2024: 277-289.
- Ge Y, Hua W, Mei K, et al. Openagi: When llm meets domain experts [J/OL]. <https://doi.org/10.48550/arXiv.2304.04370>.
- Aycock S, Bawden R. Topic-guided Example Selection for Domain Adaptation in LLM-based Machine Translation [C/OL]. <https://aclanthology.org/2024.eacl-srw.13.pdf>.

(下转封三)