融合知识图谱与向量数据库的二级检索及反馈优化医学问答模型*

张骏伟[1], 黄 江[2], 王 翔[1], 蒋晓红[3]
(1. 重庆交通大学 机电与车辆工程学院,重庆 400074; 2. 重庆交通大学 信息科学与工程学院,重庆 400074;
3. 重庆交通大学 交通运输学院,重庆 400074)

摘 要：随着AI的迅速发展,大语言模型在医疗领域的应用潜力日益凸显,但通用大模型在医学专业领域存在知识不足、易产生幻觉等问题. 本文以构建疾病医疗问答助手为核心任务,提出一种基于二级检索与反馈迭代机制的医学智能问答系统HQ-LLM(HealthQuery Large Language Model). 整合医学百科、公共卫生部门等数据,构建医学知识图谱和中文医学问答数据集(Chinese Medical Knowledge Graph and Question-Answering Dataset, CnMedKG-QA);通过LoRA技术实现ChatGLM3-6B模型的轻量化微调,单张32 GB vGPU即可支撑训练过程,兼顾成本控制与训练效率;针对传统RAG局限,借助paraphrase-multilingual-MiniLM-L12-v2模型进行文本向量化,构建高相关性、高检索效率的向量数据库,创新设计粗粒度筛选与细粒度推理融合的二级检索机制;建立涵盖专业术语准确率、逻辑连贯性等维度的医学专属置信度评估体系,当综合置信度低于阈值时,动态优化检索关键词、扩展知识图谱追溯层数或补充外部权威数据,实现"检索-生成-评估-再检索"反馈迭代机制. 结果表明,HQ-LLM在BLEU-4、Rouge-L、F1 Score和Accuracy指标上分别达到20.47%、42.12%、50.26%和79.20%,显著优于Qwen-UMLS-7B-Instruct等现有医学模型.

关键词：检索增强生成;知识图谱;微调;二级检索;反馈迭代
DOI：10.13568/j.cnki.651094.651316.2026.01.12.0002
中图分类号：TP391.1 **文献标识码**：A **文章编号**：2096-7675(2026)05-0562-015
引文格式：张骏伟,黄江,王翔,蒋晓红. 融合知识图谱与向量数据库的二级检索及反馈优化医学问答模型[J]. 新疆大学学报(自然科学版中英文),2026,43(5):562-576.
英文引文格式：Zhang Junwei, Huang Jiang, Wang Xiang, Jiang Xiaohong. Medical Q&A model integrating knowledge graph and vector database with two-stage retrieval and feedback optimization[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English),2026,43(5):562-576.

Medical Q&A Model Integrating Knowledge Graph and Vector Database with Two-Stage Retrieval and Feedback Optimization

Zhang Junwei[1], Huang Jiang[2], Wang Xiang[1], Jiang Xiaohong[3]
(1. School of Mechatronics and Vehicle Engineering, Chongqing Jiaotong University, Chongqing 400074, China; 2. School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China; 3. School of Traffic and Transportation, Chongqing Jiaotong University, Chongqing 400074, China)

Abstract: With the rapid advancement of AI, the application potential of large language models (LLMs) in the medical field has become increasingly prominent. However, general-purpose LLMs still suffer from insufficient professional knowledge and a tendency to generate hallucinations in medical domains. Focusing on the core task of constructing a disease-oriented medical question-answering assistant, this paper proposes a medical intelligent question-answering system named HQ-LLM

* **收稿日期**: 2026-01-12; **修回日期**: 2026-05-20; **录用日期**: 2026-05-21.
基金项目: 重庆市技术创新与应用发展专项重大项目"基于大模型的汽车网络安全测评技术服务平台研发"(CSTB2024TIAD-STX0027).
作者简介: 张骏伟(2002—),男,硕士生,从事深度学习、大语言模型的研究,E-mail:xiaozhang@mails.cqjtu.edu.cn.

(HealthQuery Large Language Model) based on a two-stage retrieval and feedback iteration mechanism. Multi-source data from medical encyclopedias and public health authorities are integrated to construct a Chinese Medical Knowledge Graph and Question-Answering Dataset (CnMedKG-QA). The LoRA (Low-Rank Adaptation) technique is adopted to implement lightweight fine-tuning of the ChatGLM3-6B model, which can be supported by a single 32 GB vGPU for the training process, balancing cost control and training efficiency. To address the limitations of traditional RAG (Retrieval-Augmented Generation) systems, the paraphrase-multilingual-MiniLM-L12-v2 model is utilized for text vectorization to build a vector database with high relevance and retrieval efficiency. An innovative two-stage retrieval mechanism combining coarse-grained filtering and fine-grained reasoning is designed, along with a "retrieval-generation-evaluation-re-retrieval" feedback iteration mechanism. A medical-specific confidence evaluation system covering dimensions such as professional terminology accuracy and logical coherence is established. When the comprehensive confidence is below the threshold, retrieval key words are dynamically optimized, the number of knowledge graph tracing layers is expanded, or external authoritative data is supplemented to continuously improve retrieval quality. Experimental results show that the HQ-LLM achieves 20.47%, 42.12%, 50.26% and 79.20% in BLEU-4, Rouge-L, F1 Score and Accuracy, respectively, significantly outperforming existing medical models such as Qwen-UMLS-7B-Instruct.

Key words：RAG; knowledge graph; fine-tuning; two-stage retrieval; feedback iteration

0　引 言

近年来，传统神经网络在图像识别、数据分析和结果预测等领域取得较大进展，但在生成和理解人类自然语言方面存在一定缺陷[1]，这一缺陷在医学领域尤为突出，医学文本的专业性、知识的动态更新特性及对回答准确性与可解释性的要求，使传统模型面对复杂医学问题时，出现实体召回不准、逻辑链断裂、回答无证据支撑等情况．同时，人工智能领域中大语言模型（Large Language Model，LLM）技术正在飞速发展，从BERT引入双向Transformer结构以提升语言理解能力[2]，到GPT系列采用自回归架构强化文本生成[3]，再到T5将多任务统一为"文本到文本"框架以增强泛化能力[4]，以及PaLM等超大规模模型通过参数扩展显著提升推理与上下文理解能力[5]．大语言模型强大的自然语言理解能力和生成能力推动人工智能在医学领域的发展．

目前，通用大模型在医学领域的应用仍面临局限．一方面，通用模型缺乏对医学专业术语、临床路径等知识的深度适配，生成不准确的医疗建议；另一方面，不同医疗场景存在差异化需求，通用模型的泛化能力难以满足应用需求[6]．故对模型进行微调，成为提升其应用价值的关键．

LLM在通用数据集上训练，其参数隐含大量知识[7]．为使其满足某专业领域的应用要求，通过检索增强生成（Retrieval-Augmented Generation，RAG）框架在知识图谱（Knowledge Graph，KG）中检索相关知识后，提供给LLM以生成答案成为一种有效的方法[8]．但检索增强技术存在缺点，一是检索机制单一，容易出现实体召回不准；二是缺乏置信度评估，无法识别低质量回答；三是无反馈机制，导致低置信度回答无法有效纠正．传统RAG仅做表层语义匹配，如图1所示，将咽肌痉挛泛化为普通咽喉不适，完全忽略了咽肌痉挛的核心病因多与神经系统、神经肌肉疾病相关．

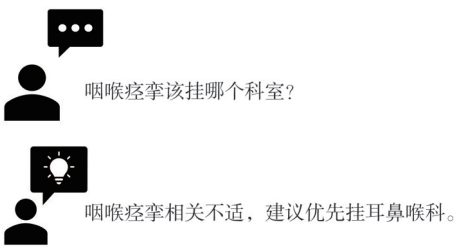


图 1　传统RAG示例

Figure 1　Traditional RAG example

学者们提出了多种改进方式．Xiong等[9]采用混合R2D-LLM方法的多源数据提取实体，构建动态知识图谱，设计了DR-RAG框架，在飞机设计相关的知识检索任务中，DR-RAG的表现优于其他方法．Shin等[10]集成教科书、指南数据库与LLM，构建基于向量数据库产生患者特定临床推理输出的Thyro-GenAI模型，该模型幻觉更少、可靠性更高．Xiong等[11]将低秩自适应（Low-Rank Adaptation，LoRA）应用于InterLM-20B模型，形成

InterLM-20B-LoRA,并与Langchain-Chatchat和本地知识库集成开发问答系统,该系统的回答准确率达97%. Kresevic等[12]结合RAG和提示词工程技术设计自定义的LLM框架,将引导信息转换为LLM能够有效处理的最佳结构化格式,以提供最准确的输出. 欧阳钞等[13]提出聚类感知与细粒度反馈驱动的多智能体系统自动优化方法(CAFG),在6个基准数据集上取得84.2的平均得分. Yuksekgonul等[14]提出TextGrad框架,通过反向传播LLM生成的反馈信号实现框架优化,通过在博士级别科学问题求解、放射治疗方案优化、特定属性分子设计、代码编写及智能体系统优化等场景开展研究,验证了TextGrad的通用性与有效性.

本文核心在于以低成本实现高效的医学问答系统,整体结构如图2所示. 采用LoRA微调技术优化预训练模型,并结合提示词工程引导系统以专业视角解读用户需求,提升回答的适配性;同时,精准提取医学文本中的疾病、症状和病因等核心实体,识别治疗科室等实体关联关系,生成标准化三元组,借助Neo4j图数据库实现医学知识网络的可视化构建与结构化存储;创新设计粗粒度筛选与细粒度推理融合的二级检索机制,通过paraphrase-multilingual-MiniLM-L12-v2模型构建高相关性向量数据库,先通过粗粒度筛选快速锁定核心知识范围,再基于知识图谱的路径模板与子图扩展策略实现隐性关联的深度挖掘,有效解决传统RAG实体召回不准、推理深度不足的问题;引入反馈迭代机制,构建医学专属置信度评估体系,当综合置信度未达阈值时,动态优化检索关键词、扩展知识图谱追溯层数或补充外部权威数据,动态修正低置信度回答,有效抑制模型幻觉并提升答案可靠性. 通过两大核心创新机制的协同作用,显著提升医学问答的准确性、专业性与完整性.

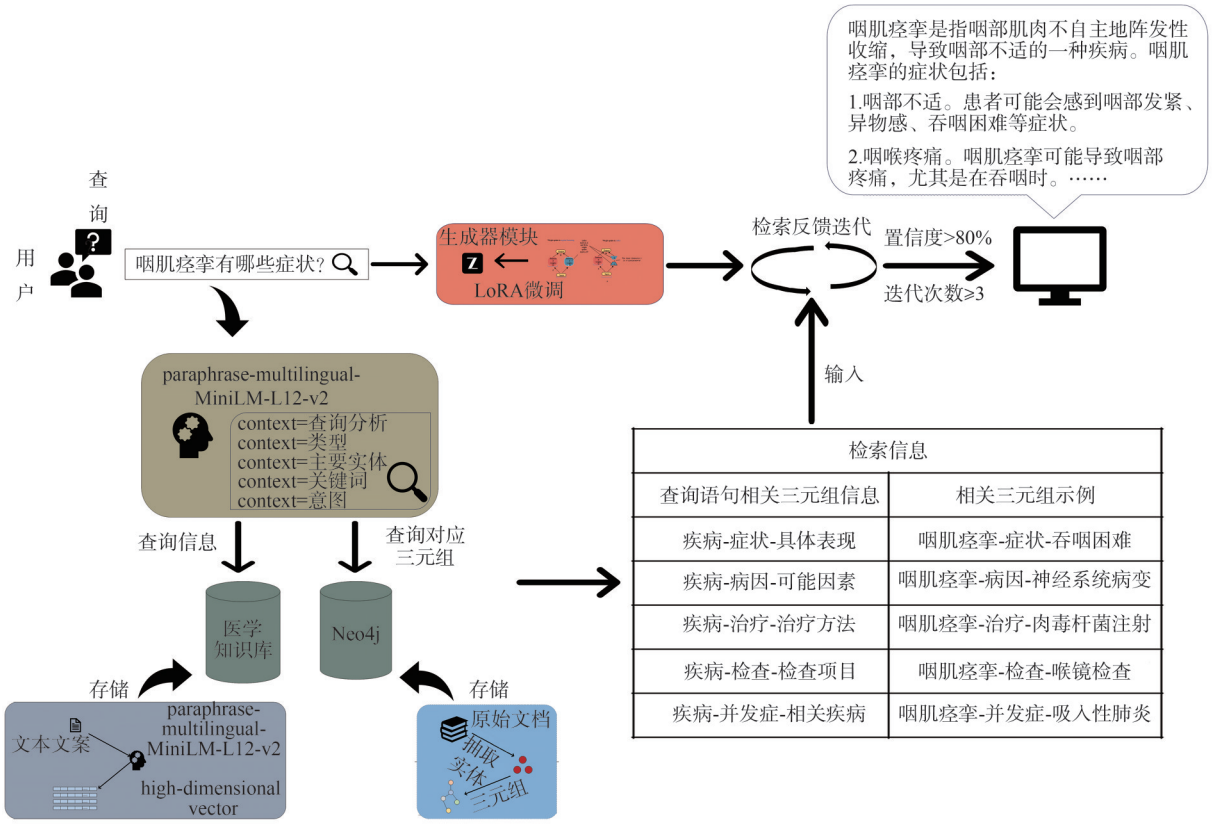


图 2 HQ-LLM结构示意图

Figure 2 Schematic of the HQ-LLM structure

1 数据集与方法

1.1 数据集

医学领域数据分散在临床诊疗、医学研究等多个独立环节,现有方案既缺乏统一数据结构,在数据格式标准化上也存在缺陷;同时,医学领域公共数据集多维度数据的完整性和全面性无法满足实际需求,如诊疗指南更新慢、病历信息不完整等关键方面. 值得关注的是公开数据集(如MIMIC系列数据集等)以英文为核心语言

载体．在医学场景中，这些现状不仅限制了中文医学相关任务效果，更阻碍了医学人工智能系统知识库发展．

为提升大模型在医学知识领域问答任务中的表现，本文构建一个医学领域专用的CnMedKG-QA数据集．原始数据来自国家级临床指南、权威医学教材和专业医学百科．对原始数据进行清洗、分句、去重与格式统一；构建医学术语标准化体系，基于国家卫生健康委标准术语等权威资源，构建标准医学术语词典，对原始文本中的口语化表达、俗称、缩写及异名实体，通过同义词匹配、规则映射、语义归一完成标准化，所有实体经过标准化后均绑定唯一标准名称与唯一实体ID，确保同一概念在知识图谱中仅存在一个标准节点，避免实体歧义、召回错误与知识冲突．

针对收集的结构化数据，本文采用封闭预定义实体类型和封闭预定义集合，覆盖诊疗全流程，定义Disease（疾病）、Symptom（症状）、Etiology（病因）、Department（科室）和Treatment（治疗方法）等7类基础核心实体，确保主体框架稳定；定义Has-Symptom（疾病-症状）、Has-Etiology（疾病-病因）、Belongs-to-Department（疾病-科室）和Has-Treatment（疾病-治疗方法）等10类核心关系，以保证推理路径稳定．

构建医学知识图谱时，先对原始医学文本数据进行预处理与结构化处理[15]，得到规范化格式文本．通过分析该文本，精准提取疾病名称、描述、分类和病因等实体，赋予每个实体唯一标识、医学术语和定义描述等标准化属性，并识别实体间的关联关系，包括治疗科室、治疗方法和治疗周期等．将提取结果输出为标准化三元组，并转化为可视化网络关系，构建完整表达医学知识关联的图谱体系，部分知识图谱结构如图3所示．

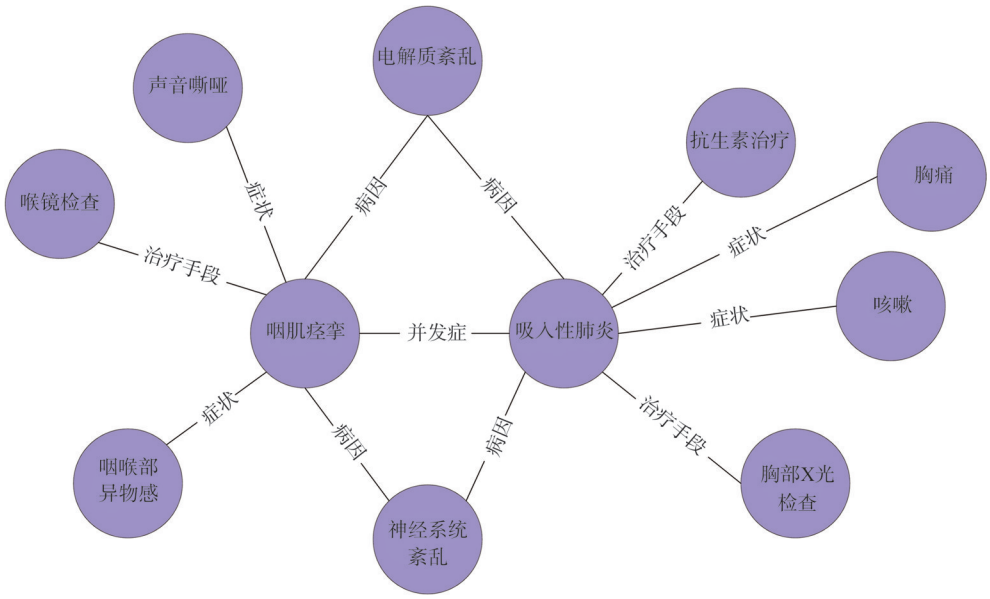


图 3　知识图谱

Figure 3　Knowledge graph

基于每个独立的疾病实体，通过知识图谱实体和关系组成的三元组中包含疾病介绍、病因分析和治疗方案等多个特征，构建医学知识问答数据集，专门用于训练和评估医学问答系统，每条数据均有instruction、input和output字段，分别对应问题、输入信息和答案内容．数据集的统计分布如表1所示．

表 1　数据集统计

Table 1　Dataset statistics

类别	项目	数量
医学知识图谱	实体	296个
	关系	583个
	三元组	597个
医学知识问答数据集	训练集	58 418条
	验证集	14 606条
	测试集	14 606条

对数据集进行统一的格式化处理[16],以“问题-答案”的结构存储,如图4所示. 数据集包含疾病介绍、病因分析和治疗方案等内容,保证了数据的多样性,使微调后的模型可以满足用户需求. 为确保CnMedKG-QA数据集具备临床可靠性与专业权威性,本文邀请重庆医科大学第一附属医院的医学专家对数据集进行审核,围绕医学术语规范性、临床逻辑合理性、内容完整性等维度,对每条记录进行独立验证,经批量初核、逐条细核、专家终审及随机抽检的规范流程完成验证,最终数据集所有记录均审核合格,为保证数据集质量提供了权威专业的支撑.

“instruction”:“你是一个专业的医学专家,请准确简洁地回答医学问题。只输出回答内容,不要重复问题。”
“input”:“脊髓小脑性共济失调的病因是什么?”
“output”:“本病的生化改变尚不明。有的病人有丙酮酸氧化的缺陷。脊髓小脑性共济失调(SCA)是遗传性共济失调的主要类型,其共同特征是中青年发病。血管病、占位病、常染色体显性遗传和共济失调,经CT扫描证实排除其他累及小脑及脑干的变化。常染色体显性遗传的脊髓小脑性共济失调具有遗传异质性,最具特征性的基因缺陷是扩增的CAG三核苷酸重复小脑中多聚谷氨酰胺通道,该通道在功能不明蛋白和神经末梢上发现的P/Q型钙通道α亚单位上;其他类型突变包括CTG三核苷酸(SCA8)和ATTCT五核苷酸(SCA10)重复序列扩增,在许多病例中这种扩增片断的大小与疾病严重性有关,且年龄愈小,病情愈重。SCAs基因突变改变蛋白的性质,实质无法被正常加工,异常加工的片断与一种参与非溶酶体降解的缺陷蛋白泛素结合,共同以蛋白酶体的型体移位式转运至核内,推测这种核内蛋白聚集可影响细胞核的功能。每种SCA亚型的基因位于不同的染色体,有不同的大小和基因突变部位。例如,SCA1基因位于8号染色体q22–23,基因组跨度450 kb,cDNA长11 kb,含9个外显子,编码816个氨基酸残基组成ataxia–1蛋白,该蛋白位于细胞核,CAG突变位于8号外显子,扩增拷贝数为40–83,正常人为6–38。SCA3(MJD)是我国最常见的SA亚型,基因位于14q24.3–32,至少含4个外显子,编码960个氨基酸残基组成ataxia–3蛋白,分布于细胞质中,CAG突变位于4号外显子,扩增拷贝数为61–89,正常人为12–41。尽管SCA有共同的基因突变机制,导致各型临床表现雷同,但仍有差异,如有的伴眼肌麻痹,有的伴视网膜色素变性,病理亚型临床和程度也不相同,提示除了多聚谷氨酰胺毒性作用外,其他因素可能也参与发病。”

图 4 医学知识问答数据集示例

Figure 4 Example of the medical question-answering dataset

1.2 基于LoRA的微调

虽然LLM经通用语料训练,已具备强大的自然语言理解与生成能力,但其应用受两大矛盾制约. 一是全量微调下显存需求与现有硬件资源供给之间的失衡;二是由于通用模型预训练语料的广泛性和非领域性导致专业能力不足与垂直领域知识更新之间的矛盾[17]. 指令遵循能力不足,如果直接用作生成器,效果不理想[18]. 若采用全参数策略对权重矩阵实施微调,则需对模型全部参数进行端到端优化,这一过程使计算成本增加,同时对硬件算力提出严苛要求[19]. 本文采用LoRA微调技术(图5),核心思想在于将预训练模型权重矩阵解构为两个低秩矩阵,借助对权重更新空间的限制,实现计算开销的显著降低[20].

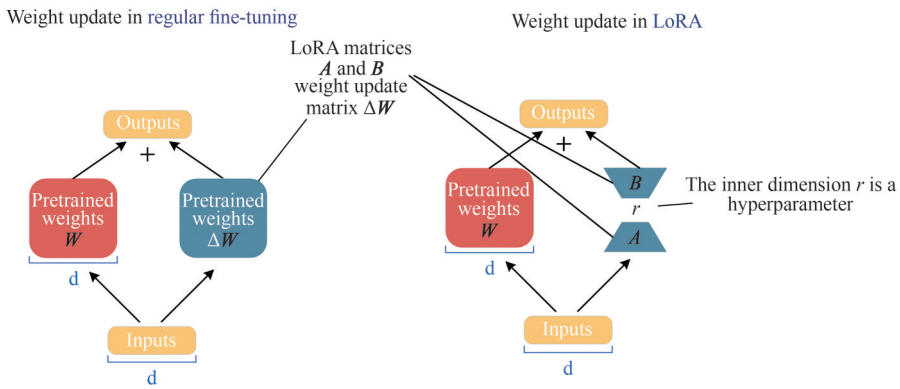


图 5 LoRA微调

Figure 5 LoRA fine-tuning

但LoRA微调的性能表现高度依赖基础模型的能力，本文针对中文医学知识科普场景，涵盖疾病介绍、病因分析、治疗方案和预防措施等内容，对模型的中文推理能力、长文本处理能力要求较高，故优先选择国内主流的开源预训练LLM，如Qwen、DeepSeek、Baichuan等．Qwen系列不仅拥有强大的通用问答能力，还具备出色的长文本理解与处理能力[21]．DeepSeek模型具有出色的长文本处理能力[22]．Baichuan系列支持十万长文本，计算效率较高[23]．GLM采用高效推理架构，适用于轻量级部署[24]．本文利用CnMedKG-QA数据集，通过问答任务对比实验筛选合适的生成器模型．

1.3　二级检索机制：粗粒度筛选与细粒度推理融合

基于知识图谱构建的医学知识库，其文本信息以三元组结构进行组织，具备更精细的语义粒度．而传统检索策略受限于固有范式，难以从该类结构化知识库中精准提取与用户查询意图高度匹配的信息，导致复杂医学问题的检索精准度不足．本文提出结合粗粒度筛选与细粒度推理融合的二级检索机制，如图6所示．采用paraphrase-multilingual-MiniLM-L12-v2模型对用户查询语句进行解析，实现高效精准的检索与推理[25]，捕捉医学知识体系中的隐性关联，从而生成正确的检索结论．为实现这一目的，本文设计了提示词模板，结构如图7所示．

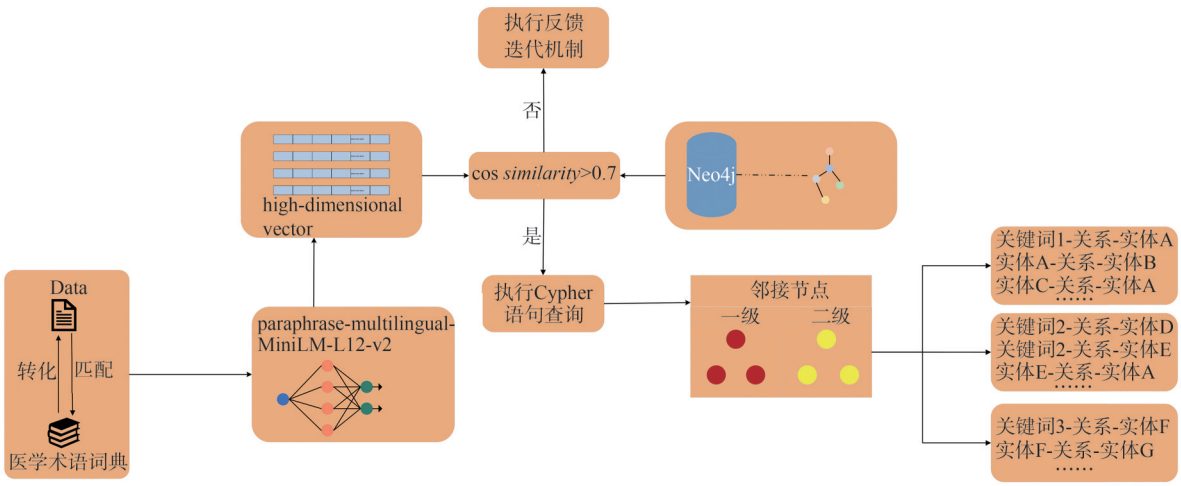


图 6　二级检索机制

Figure 6　Two-stage retrieval mechanism

```
你是一个医学信息分析专家，请分析以下医学问题，提取关键信息：
问题（question）
请按照以下格式回答：
1. 查询类型［诊断咨询/治疗方案/病因分析/症状查询/预防措施/其他］
2. 主要疾病/症状［提取主要的医学实体，用逗号分隔］
3. 关键词：［提取3~5个关键词，用逗号分隔］
4. 查询意图：［简要描述用户想要了解什么］
5. 知识粒度：［直接关联/间接关联，如“疾病-症状”为直接关联/“疾病-并发症-治疗”为间接关联］
6. 关联类型：［实体间核心关系，如“症状-病因”“并发症-治疗科室”等］

规则说明：
-优先提取具有明确医学意义的术语，避免通用词汇或日常用语
-对于可能有多种解释的词汇，提供消歧说明
  确保提取的实体、关键词及关系符合医学专业规范
-知识粒度需根据查询涉及的实体关联层数判断，单层关联为直接关联，
  两层及以上为间接关联
-保持输出格式的一致性，每点一行
```

图 7　提示词模板

Figure 7　Prompt template

粗粒度筛选的目标是快速缩小检索范围，基于医学知识库实现核心实体的高效定位．将解析得到的疾病实体与医学术语词典进行匹配，口语化表达转化为标准化术语，消除语义歧义；将CnMedKG-QA数据集按实体类型（如疾病、症状、科室等）分类存储，每个标准化实体通过paraphrase-multilingual-MiniLM-L12-v2模型编码

生成高维向量,并关联其在知识图谱中的唯一ID,构建分层向量索引;采用近似最近邻算法,将标准化核心实体向量输入向量数据库,匹配相关实体及其直接关联节点集合,设置余弦相似度阈值为0.7,低于阈值时自动补充相关实体检索,快速锁定与查询相关的核心知识范围.

在粗粒度筛选基础上,基于Neo4j构建的知识图谱进行细粒度推理,挖掘实体间的隐性关联,实现间接关联类的复杂查询. 先定义常见间接关联场景的路径模板,覆盖科室推荐、治疗方案推荐和并发症预防等核心医学问题类型,如"疾病→并发症→相关科室"对应科室推荐类查询、"疾病→症状→核心病因→推荐治疗方法"对应治疗方案推荐类查询;再基于解析得到的关联类型与知识粒度,采用Neo4j的Cypher语句自动匹配对应的路径模板执行多跳路径检索,如针对"咽肌痉挛应该挂哪个科室?"这一问题,通过匹配"疾病→并发症→相关科室"路径执行检索,返回科室名称及基于诊疗指南推荐等级设定的优先级,确保结果的临床权威性;若预设路径无匹配结果,则启动"子图扩展"策略,以核心实体为中心,遍历知识图谱中2~3层关联关系的所有节点,通过PageRank算法计算节点重要性,筛选与查询意图相关的隐性关联路径,避免遗漏潜在相关知识.

1.4 检索反馈迭代机制

为解决单次检索易出现的信息覆盖不全等问题,本文构建"检索-生成-评估-再检索"的闭环迭代机制(图8),通过生成结果置信度的动态评估调整检索策略,实现检索精度的持续优化.

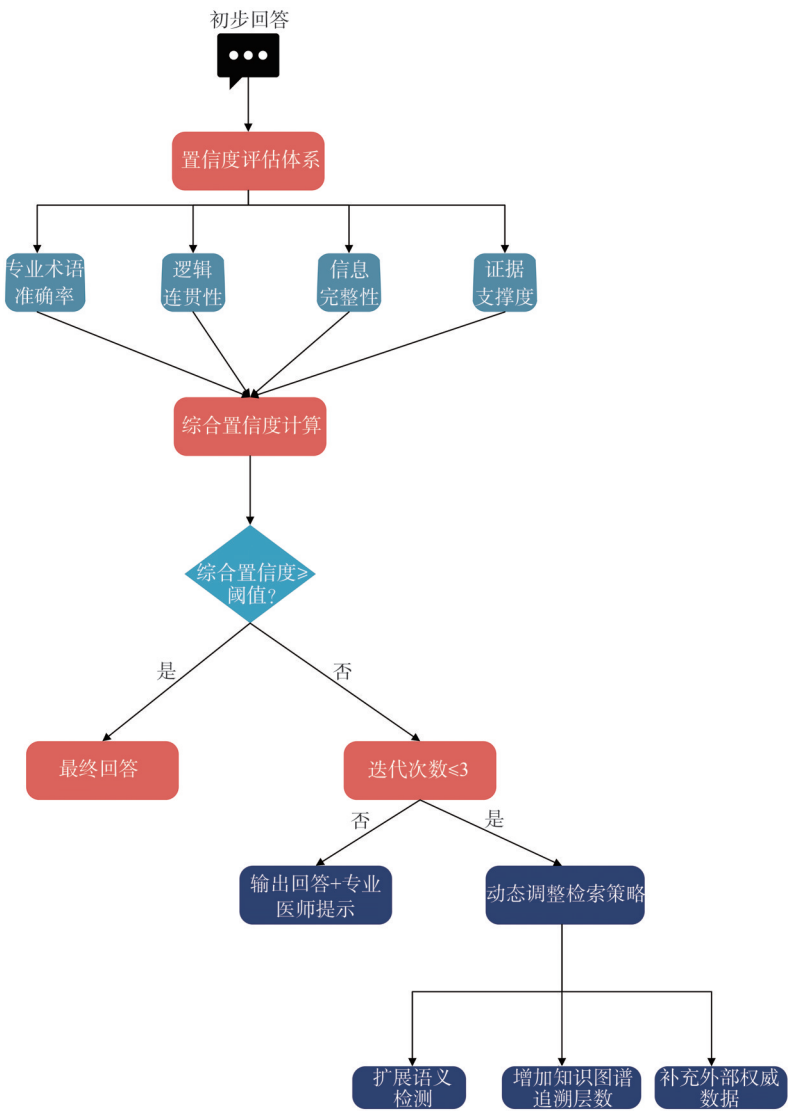


图 8 检索反馈迭代机制

Figure 8 Retrieval feedback iteration mechanism

为实现检索反馈迭代机制，本文构建面向医学知识的置信度评估体系，从专业术语准确率、逻辑连贯性、信息完整性和证据支撑度方面量化生成结果的质量.

专业术语准确率P_t的计算方式为：

$$P_t = \frac{N_{\text{std}}}{N_{\text{total}}} \times 100\%, \tag{1}$$

式中：N_{std}代表回答中符合医学术语规范的标准化术语数；N_{total}代表回答中所有术语数.

逻辑连贯性P_c的计算方式为：

$$P_c = \cos similarity\left(\boldsymbol{V}_{\text{gen}}, \bar{\boldsymbol{V}}_{\text{ret}}\right) \times 100\%, \tag{2}$$

式中：$\boldsymbol{V}_{\text{gen}}$代表paraphrase-multilingual-MiniLM-L12-v2模型编码的生成回答向量；$\bar{\boldsymbol{V}}_{\text{ret}}$代表检索结果经模型编码后的向量均值.

信息完整性P_i的计算方式为：

$$P_i = \frac{N_{\text{cov}}}{N_{\text{all}}}, \tag{3}$$

式中：N_{cov}代表回答中覆盖查询意图的核心要点数量；N_{all}代表检索结果中包含的查询意图核心要点总数.

证据支撑度P_e的计算方式为：

$$P_e = \frac{N_{\text{evi}}}{L_{\text{total}}}, \tag{4}$$

式中：N_{evi}代表回答中引用的知识库证据数量；L_{total}代表回答的总字符数.

综合置信度为各维度得分与对应权重的加权和，计算方式为：

$$Confidence = \omega_t \times P_t + \omega_c \times P_c + \omega_i \times P_i + \omega_e \times P_e, \tag{5}$$

式中：ω_t为专业术语准确率权重；ω_c为逻辑连贯性权重；ω_i为信息完整性权重；ω_e为证据支撑度权重.

本文设定综合置信度阈值，当置信度小于阈值时，自动触发检索反馈迭代. 针对不同原因生成的低置信度结果，采用不同的检索优化策略. 专业术语准确率得分小于60%时，利用paraphrase-multilingual-MiniLM-L12-v2模型对原始关键词进行语义扩展，提取向量数据库中相似度阈值降至0.6的同义实体与相关实体，补充检索关键词集合；信息完整性得分小于60%时，将知识图谱关联追溯层数从2扩展至3，同时基于模型计算查询意图与关联节点的语义相似度，优化子图扩展策略中的节点筛选逻辑，精准定位隐性关联路径；证据支撑度得分小于60%时，补充检索国家卫生健康委诊疗指南库、PubMed中文摘要库等外部权威数据源，通过定向抓取实时更新的临床研究数据与指南内容，丰富知识支撑；逻辑连贯性得分小于60%时，基于模型计算生成回答与各检索片段的语义关联度，过滤相似度低于0.5的无关信息，优先保留与查询意图高度相关的路径和文本片段用于重新生成. 若两个及以上维度得分低于阈值，则按实体修正、深度检索、噪声过滤和外部补充的顺序执行对应策略. 将迭代终止条件设为双重约束，当迭代后综合置信度超过阈值或累计迭代次数达到3时，终止检索流程.

2　实验与分析

2.1　实验设置

针对所提问答系统方案，为验证其技术可行性，本文采用随机采样的方式将医学知识问答数据集按8∶1∶1划分为训练集、验证集和测试集，利用测试集的问题，通过HQ-LLM生成结果与标准答案进行对比，量化系统的医学知识问答表现. 针对医学知识问答系统面临的领域专业术语密集、医学回答准确性要求高等挑战，本文从生成质量、内容覆盖度与专业准确性维度评估模型性能. 选取BLEU-4、Rouge-L、F1 Score和Accuracy指标，对HQ-LLM在任务中的表现进行分析.

BLEU系列指标通过计算生成文本与参考文本的*n*-gram重叠度评估生成质量. BLEU-4通过计算1-gram到4-gram(即连续4个单词的组合)的匹配精度，结合长度惩罚机制，综合评价生成文本与参考文本的相似度，计算方式为：

$$\text{BLEU-4} = BP \times \exp\left(\sum_{n=1}^{4} \omega_n \log p_n\right),$$

$$BP = \begin{cases} \exp\left(1 - \dfrac{lr}{lc}\right), & lc \leqslant lr, \\ 1, & lc > lr, \end{cases} \tag{6}$$

式中:BP为惩罚因子;ω_n为n-gram的权重;p_n为n-gram的精确度;lc为机器译文长度;lr为最短的参考翻译句子长度.

Rouge系列指标主要用于度量词汇、短语及最长公共子序列的召回率水平. Rouge-L的L指最长公共子序列(LCS),计算时需提取机器译文与参考译文的LCS,计算方式为:

$$R_{\mathrm{LCS}} = \frac{LCS(C,S)}{Len(S)},$$

$$P_{\mathrm{LCS}} = \frac{LCS(C,S)}{Len(C)},$$

$$F_{\mathrm{LCS}} = \frac{(1+\beta^2)R_{\mathrm{LCS}}P_{\mathrm{LCS}}}{R_{\mathrm{LCS}}+\beta^2 P_{\mathrm{LCS}}}, \tag{7}$$

式中:R_{LCS}为召回率;P_{LCS}为精确率;β为平衡召回率与精确率的参数;F_{LCS}为Rouge-L.

F1 Score则综合考虑精确率和召回率,通过计算生成文本与参考文本的词汇重叠度来评估相似性,计算方式为:

$$\text{F1 Score} = \frac{2PR}{P+R}, \tag{8}$$

式中:P为精确率;R为召回率. 计算方式为:

$$P = \frac{TP}{TP+FP}, \tag{9}$$

$$R = \frac{TP}{TP+FN}, \tag{10}$$

式中:TP为真正例;FP为假正例;FN为假负例.

为满足医学问答的专业正确性与临床安全性,本文邀请重庆医科大学第一附属医院的医学专家,对HQ-LLM、医学大模型及智能体在测试集上的问答(关于实体、病因和治疗方案等)是否完全正确进行审核,统计正确样本占比,即Accuracy,计算方式为:

$$\text{Accuracy} = \frac{N_{\mathrm{correct}}}{N_{\mathrm{total}}} \times 100\%, \tag{11}$$

式中:N_{correct}表示医学专家判定为正确的样本数;N_{total}为测试集样本数.

采用LoRA轻量化微调策略,学习率为1×10^{-4},训练轮数(epoch)为10,批大小(batch size)为4,LoRA Rank为8,LoRA Alpha为16. 基于Ubuntu操作系统搭建深度学习平台,硬件平台采用Intel (R) Xeon (R) Platinum 8352V处理器和vGPU-32 GB的高性能计算节点. 软件环境配置包括Python 3.8解释器和PyTorch 2.0.0深度学习框架,启用CUDA 11.8并行计算架构以加速模型训练. 所有实验均在同一平台上进行,确保结果的可比性和可重复性.

2.2 模型结果对比分析

为对比不同模型的医学对话任务表现,从CnMedKG-QA数据集中抽取10%的数据作为测试集,分别对Baichuan-7B-Chat、ChatGLM3-6B、Qwen-7B-Chat及DeepSeek-7B-Chat模型进行测试,结果如表2所示.

表 2 模型的比较结果

Table 2 Comparison results of the models

模型	BLEU-4/%	Rouge-L/%	F1 Score/%
Baichuan-7B-Chat	6.30	7.74	9.29
Qwen-7B-Chat	1.67	6.00	3.69
DeepSeek-7B-Chat	7.22	8.60	11.39
ChatGLM3-6B	11.42	8.08	15.00

ChatGLM3-6B在医学知识领域问答任务上表现良好,BLEU-4和Rouge-L分别达到11.42%和8.08%,较高的BLEU-4说明ChatGLM3-6B对医学领域专业术语具有良好的理解与生成能力. 虽然DeepSeek-7B-Chat获

得较高的Rouge-L，但其F1 Score相对略低．综合考虑，ChatGLM3-6B在医学知识问答任务中更具优势．

2.3　损失函数

使用交叉熵损失函数对模型生成的输出文本与目标的参考文本进行评估，计算方式为：

$$\text{Loss} = \sum_{i=1}^{N} i \sum_{t=1}^{T} t y_{i,t} \log y'_{i,t}, \tag{12}$$

式中：N为样本数量；T为每个样本的序列长度；$y_{i,t}$为目标分布；$y'_{i,t}$为模型预测概率分布．

图9展示了ChatGLM3-6B模型的训练损失曲线．训练初期损失值迅速下降至2.0，训练后期损失值缓慢下降，逐步收敛至1.1附近并趋于稳定，这一现象直观反映模型具备持续迭代优化的能力．整个训练过程损失变化稳定，说明微调后的ChatGLM3-6B模型具备良好的医学知识问答任务适配性，能够准确解析医学专业术语，并生成满足医学领域要求的回答．

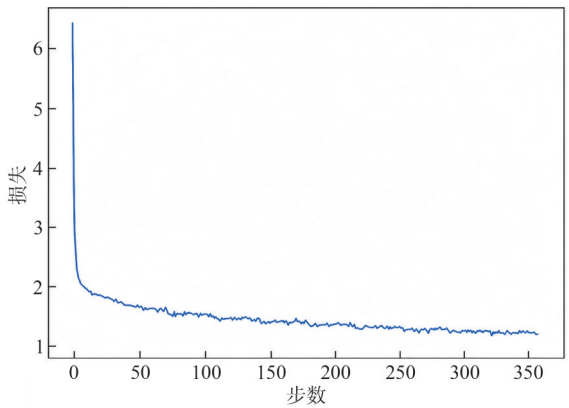


图 9　训练损失曲线

Figure 9　Training loss curve

2.4　检索模型评估

选择paraphrase-multilingual-MiniLM-L12-v2作为系统的检索模型，其是专为多语言文本语义理解设计的预训练模型．该模型通过将自然语言文本编码为高维向量表示，实现文本的语义化处理．在系统中，先对医学文档进行向量化处理，构建高质量的嵌入向量数据库．这些向量不仅保留原始文本的语义信息，还具备良好的数学性质，使相似度计算更加准确．

由表3可知，自相似度均值达到1.000 0，表明模型能够准确捕捉文本的内在一致性；平均相似度为0.368 6，说明不同文本间的语义区分度良好．聚类质量方面，轮廓系数高达0.999 7，且形成了3 007个聚类，反映向量空间具有优秀的结构组织性．在检索性能上，Recall达到1.000 0，MRR为0.900 0，证明该模型能够高效、准确检索相关文档．此外，医学相关性均值为0.766 5，表明生成的向量与医学领域高度相关．平均响应时间仅为0.012 4秒，满足实时应用的需求．

表 3　嵌入模型性能评估

Table 3　Performance evaluation of embedding models

指标类别	具体指标	数值
内部质量	自相似度均值	1.000 0
	平均相似度	0.368 6
聚类质量	轮廓系数	0.999 7
	聚类数量	3 007
检索性能	Recall	1.000 0
	MRR	0.900 0
医学相关性	平均相关性	0.766 5
效率	平均响应时间/秒	0.012 4

综上,paraphrase-multilingual-MiniLM-L12-v2支持高效的相似度计算和文档检索,为LLM提供精准的上下文信息,从而显著提升知识问答系统的准确性和专业性.

2.5 最优参数实验

为确定医学知识问答系统综合置信度计算的最优参数配置,明确权重分配与置信度阈值对系统性能的影响,基于测试集,通过控制变量法分别进行权重对比和置信度阈值对比实验. 权重对比实验以综合置信度计算的多维度权重分配为变量,将置信度阈值固定为80%. 综合置信度由专业术语准确率、逻辑连贯性、信息完整性、证据支撑度加权构成,某一维度权重过高或过低都会导致评估失衡,影响回答质量. 当权重在0.05范围内进行调整时,各项指标没有明显变化. 故将权重范围设定为[0.2,0.4],避免某一维度权重过高或过低,步长设定为0.1. 设计5组差异化权重组合(表4),明确各权重的核心作用及最优分配比例,实验结果如表5所示. B组在各指标上取得较好表现,BLEU-4为14.31%,Rouge-L为20.83%,F1 Score为33.59%,优于其他4个组合,较高的专业术语准确率权重保障医学回答的专业性,同时均衡其他维度权重,兼顾逻辑、完整性与可追溯性,可提高系统性能.

表 4 综合置信度计算权重组合

Table 4 Weight combinations for comprehensive confidence calculation

权重组合	专业术语准确率权重	逻辑连贯性权重	信息完整性权重	证据支撑度权重
A	0.25	0.25	0.25	0.25
B	0.40	0.20	0.20	0.20
C	0.20	0.40	0.20	0.20
D	0.20	0.20	0.40	0.20
E	0.20	0.20	0.20	0.40

表 5 综合置信度计算权重组合对比

Table 5 Weight combination comparison for comprehensive confidence calculation

权重组合	BLEU-4/%	Rouge-L/%	F1 Score/%
A	10.70	13.61	21.06
B	14.31	20.83	33.59
C	12.45	18.27	28.14
D	11.68	17.53	26.74
E	9.45	11.59	19.32

置信度阈值的作用是筛选低置信度的回答输出,避免存在错误、歧义或信息缺失的医学内容流向用户. 阈值过低时,大量低置信度回答未触发迭代,导致回答专业性与完整性缺失;阈值过高时,因频繁迭代引入的计算开销与信息冗余,导致整体性能下降. 阈值对比实验将权重固定为最优的权重分配B组,为评估不同置信度阈值对问答准确性、迭代触发频率的影响,设置4组梯度置信度阈值进行对比实验,结果如表6所示.

表 6 不同阈值的性能差异

Table 6 Performance differences at different thresholds

置信度阈值/%	BLEU-4/%	Rouge-L/%	F1 Score/%
60	9.79	11.46	17.09
70	10.70	12.53	19.04
80	16.24	22.45	32.68
90	12.87	18.73	29.42

将置信度阈值设为80%,BLEU-4达16.24%、Rouge-L达22.45%、F1 Score达32.68%,均优于其他阈值,能够有效拦截低置信度的错误、歧义或信息缺失的医学回答,又可避免高阈值引发频繁迭代导致的性能下降问题.

综上，本文在最优权重组合与阈值80%的协同作用下，构建完整的检索反馈迭代体系，既能有效过滤低置信度的低质回答、降低医学信息错误传播风险，又可避免过度迭代引发的性能下降，提升了系统的专业准确性与可靠性.

2.6　性能分析

为全面评估HQ-LLM性能，选取经医学数据训练的医学通用LLM和智能体（包括Qwen-UMLS-7B-Instruct、medicine-chat等医学大模型，MedGraphRAG智能医疗问诊系统，MDAgents智能体），在CnMedKG-QA数据集上进行微调，并在测试集上开展横向对比，结果如表7所示.

表 7　HQ-LLM与专业模型对比结果

Table 7　Comparison results between HQ-LLM and professional models

模型	BLEU-4/%	Rouge-L/%	F1 Score/%	Accuracy/%
Qwen-UMLS-7B-Instruct	9.99	13.62	19.06	45.40
medicine-chat	10.70	13.61	21.06	47.00
Medical_expert	4.12	20.94	30.40	51.30
Medical_Consultation_Assistant	14.31	20.83	33.59	57.20
MedGraphRAG	17.72	35.21	42.18	71.20
MDAgents	19.17	40.54	47.86	75.30
HQ-LLM	20.47	42.12	50.26	79.20

Qwen-UMLS-7B-Instruct基于通义千问7B基础模型，使用UMLS医学术语集及基础医学问答数据微调，侧重医学术语的标准化映射；medicine-chat基于开源7B模型微调，训练数据为中文医学对话数据集，覆盖常见病诊疗场景；Medical_expert聚焦临床专科知识，基于7B模型微调内科病例与专家问答数据；Medical_Consultation_Assistant基于通用对话7B模型微调，训练数据为多科室医学咨询对话，涵盖全科诊疗场景. MedGraphRAG是基于LangGraph状态机的智能医疗问诊Agent，支持多轮对话、RAG证据检索、风险分层与就医建议. MDAgents为基于医学知识训练后进行疾病诊疗的会诊系统.

HQ-LLM在各指标上表现均优于其他医学模型. 具体而言，HQ-LLM模型的BLEU-4、Rouge-L、F1 Score及Accuracy分别达到20.47%、42.12%、50.26%、79.20%，BLEU-4相较Qwen-UMLS-7B-Instruct、medicine-chat、Medical_expert、Medical_Consultation_Assistant、MedGraphRAG、MDAgents分别领先10.48%、9.77%、16.35%、6.16%、2.75%、1.30%，Rouge-L比次优模型提升1.58%，F1 Score显著优于其他模型，Accuracy比MDAgents提升3.90%. 证明HQ-LLM在医学问答领域文本生成质量方面具有优势，能够准确表达疾病名称、描述、分类和病因等基础信息，可更加清晰精确地理解分析用户提出的问题.

2.7　消融实验

为进一步评估HQ-LLM的有效性，对不同组件的有效性进行分析，包括原始的ChatGLM3-6B模型（a）、仅通过微调的模型（b）、仅使用传统RAG的模型（c）、仅使用二级检索机制的模型（d）、仅使用检索反馈迭代机制的模型（e）及最终的HQ-LLM，结果如表8所示.

表 8　HQ-LLM消融实验结果

Table 8　Ablation experiment results of HQ-LLM

模型	BLEU-4/%	Rouge-L/%	F1 Score/%
a	11.42	8.08	15.00
b	12.18	13.48	22.52
c	15.46	24.25	30.38
d	17.82	35.64	42.75
e	16.95	32.76	37.92
HQ-LLM	20.47	42.12	50.26

作为核心优化模块,仅使用二级检索机制(d)与仅使用检索反馈迭代机制(e)的各项指标均大幅提升. 模型d的BLEU-4、Rouge-L与F1 Score分别达到17.82%、35.64%和42.75%,与传统RAG(c)相比,分别提升2.36%、11.39%和12.37%. 模型e的BLEU-4、Rouge-L与F1 Score分别为16.95%、32.76%和37.92%,说明在无检索增强的前提下,也可实现生成质量的提升.

HQ-LLM的BLEU-4、Rouge-L及F1 Score分别达到20.47%、42.12%、50.26%,比ChatGLM3-6B(a)分别提升9.05%、34.04%和35.26%. 基于以上数据,说明HQ-LLM在医学问答领域生成的文本更准确且信息更完整,二级检索机制与检索反馈迭代机制的协同作用,可有效提升模型对医学专业知识的理解与应用水平.

2.8 可视化界面

基于Gradio框架构建医学问答系统直观易用的Web界面,用户可通过浏览器访问并与系统进行交互. 以查询问题"咽肌痉挛有哪些症状?"为例(图10),虽然回答涉及病例症状,但缺乏病例症状的解读,未提供额外信息,专业表达不够详细.

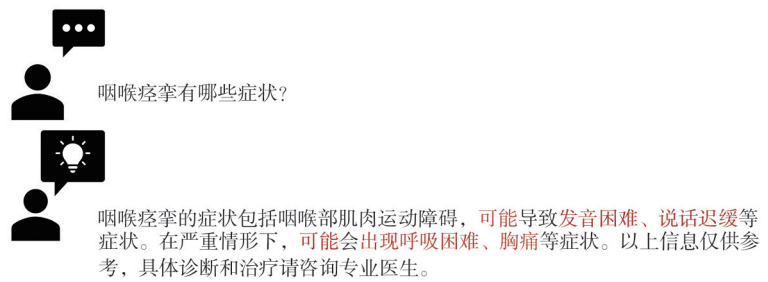


图 10 原始模型问答演示

Figure 10 Demonstration of the original model's Q&A

如图11所示,HQ-LLM回答出病例症状的同时,还提供描述症状更加详细的信息,通过二级检索与检索反馈迭代的结合,有效提升医学问答的准确性、专业性和完整性.

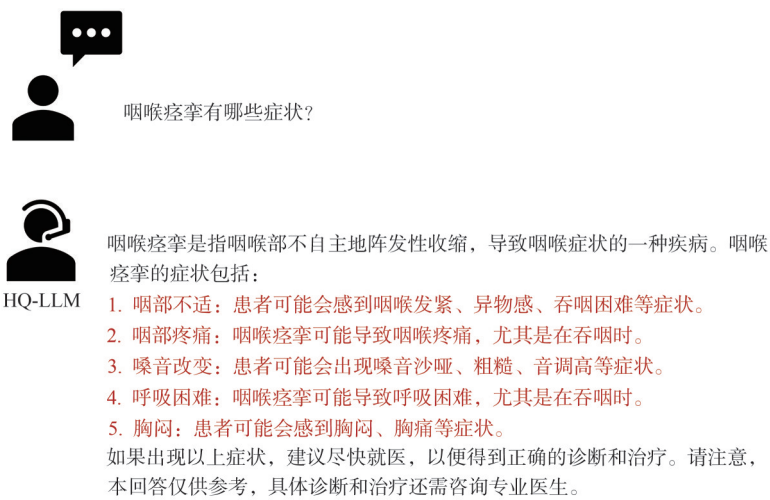


图 11 HQ-LLM问答演示

Figure 11 Demonstration of the HQ-LLM Q&A

3 结 论

针对医学问答任务中专业信息输出的精准性问题,本文提出一个融合检索增强生成架构的智能问答系统HQ-LLM,以提升系统在疾病名称、描述、分类和症状等方面的回答质量. 与现有方法相比,HQ-LLM以paraphrase-multilingual-MiniLM-L12-v2模型为检索核心,通过精细化查询、解析、提取用户意图与关键医学实体,并采用粗粒度筛选与细粒度推理协同作用的二级检索机制,先通过向量数据库实现核心实体的快速匹配,再基于医学知识图谱的预设路径模板与子图扩展策略,将知识图谱关联追溯层数从2扩展至3,同时对检索结

果进行相关性排序，确保高权威、高匹配度的医学信息优先返回；为消除单次检索的信息欠缺，HQ-LLM引入检索反馈迭代机制，构建医学专属置信度评估体系，当综合置信度低于阈值时，动态调整检索策略，包括扩展语义关键词、增加知识图谱追溯层数、补充外部权威诊疗数据等，通过闭环优化，持续提升检索信息的精准性与全面性．

针对生成器ChatGLM3-6B，HQ-LLM引入LoRA微调技术实施指令优化训练，在单张vGPU-32 GB上即可实现高效低耗的模型训练，有效保障资源受限条件下的模型适应性．实验结果表明，HQ-LLM的BLEU-4、Rouge-L和F1 Score指标分别达到20.47%、42.12%和50.26%，证实本文方法在医学问答任务中的可行性．

本研究存在一定局限性．如图12所示，面对“血液肿瘤化疗后危重并发症的治疗方案”这类多系统并发症叠加的问题，模型只输出泛化的对症治疗，并没有覆盖核心诊疗措施．反映出模型在复杂临床场景下无法整合多维度决策逻辑、知识覆盖不足且临床精准度有限的核心缺陷，也印证其在重症、罕见病场景下的劣势．未来工作可从以下方面进行优化：数据方面，实现医学文本、图像、视频等多模态数据的深度融合，通过挖掘不同模态数据间的关联信息，提升模型对复杂临床问题的分析与判断能力；检索精度还有提升空间，通过优化查询反馈中的评分算法提高结果相关性；知识库方面，构建动态更新的医学知识库，确保模型知识与最新临床共识同步，进一步缩小模型输出与真实临床实践的差距；建立科学合理的临床效用评估基准，从适用性、准确性、效率等多个方面对模型效果进行评估，以提升系统在临床医学中的实际应用价值．

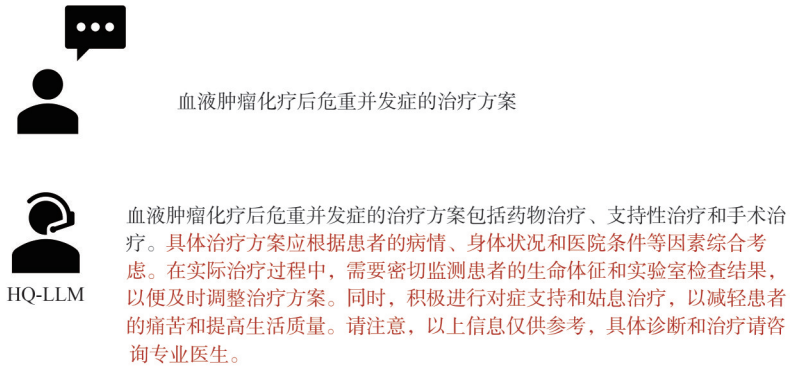


图 12　HQ-LLM的局限性

Figure 12　Limitations of HQ-LLM

参考文献：

[1] 张玉晨，王传涛，夏海梁，等．基于大语言模型的医疗智能助手应用研究进展[J]．协和医学杂志，2025，16(6)：1511-1518．
Zhang Y C，Wang C T，Xia H L，et al．Research progress on the application of large language model-based intelligent medical assistants[J]．Medical Journal of Peking Union Medical College Hospital，2025，16(6)：1511-1518．(in Chinese)

[2] Devlin J，Chang M W，Lee K，et al．BERT：Pre-training of deep bidirectional Transformers for language understanding[PP/OL]．V1．arXiv (2018-10-11) [2025-09-15]．https://arxiv.org/pdf/1810.04805v1．

[3] Radford A，Wu J，Child R，et al．Language models are unsupervised multitask learners[EB/OL]．[2025-09-15]．https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf．

[4] Raffel C，Shazeer N，Roberts A，et al．Exploring the limits of transfer learning with a unified text-to-text Transformer[J]．Journal of Machine Learning Research，2020，21：1-67．

[5] Chowdhery A，Narang S，Devlin J，et al．PaLM：Scaling language modeling with pathways[J]．Journal of Machine Learning Research，2023，24：240．

[6] 钱波，李富江，郑常乐，等．医疗大模型发展现状与展望[J]．数据采集与处理，2025，40(3)：562-584．
Qian B，Li F J，Zheng C L，et al．A review of development and future directions of medical foundation models[J]．Journal of Data Acquisition and Processing，2025，40(3)：562-584．(in Chinese)

[7] Li M C，Kilicoglu H，Xu H，et al．BiomedRAG：A retrieval augmented large language model for biomedicine[J]．Journal of Biomedical Informatics，2025，162：104769．

[8] 赵海燕,高东昇,曹健,等. 融合知识图谱和大模型的医疗智能问答方法研究[J]. 小型微型计算机系统,2026,47(5):1127-1133.
Zhao H Y,Gao D S,Cao J,et al. Research on medical intelligent question answering method combining knowledge graph and big model[J]. Journal of Chinese Computer Systems,2026,47(5):1127-1133.(in Chinese)
[9] Xiong X R,Cai H M,Shen B Q,et al. DR-RAG:Domain-rule-based retrieval-augmented generation for aviation digital model design[J]. Advanced Engineering Informatics,2025,68:103688.
[10] Shin M,Song J,Kim M G,et al. Thyro-GenAI:A chatbot using retrieval-augmented generative models for personalized thyroid disease management[J]. Journal of Clinical Medicine,2025,14(7):2450.
[11] Xiong J,Pan L M,Liu Y J,et al. Enhancing plant protection knowledge with large language models:A fine-tuned question-answering system using LoRA[J]. Applied Sciences,2025,15(7):3850.
[12] Kresevic S,Giuffrè M,Ajcevic M,et al. Optimization of hepatological clinical guidelines interpretation by large language models:A retrieval augmented generation-based framework[J]. npj Digital Medicine,2024,7(1):102.
[13] 欧阳钞,张捷,白昱阳,等. 聚类感知与细粒度反馈驱动的大模型智能体系统优化方法[J/OL]. 计算机工程与应用,2026-01-06. https://link.cnki.net/urlid/11.2127.TP.20260106.1426.006.
Ouyang C,Zhang J,Bai Y Y,et al. Cluster-awareness and fine-grained feedback driven optimization method for LLM-based multi-agent systems[J/OL]. Computer Engineering and Applications,2026-01-06. https://link.cnki.net/urlid/11.2127.TP.20260106.1426.006.(in Chinese)
[14] Yuksekgonul M,Bianchi F,Boen J,et al. Optimizing generative AI by backpropagating language model feedback[J]. Nature,2025,639(8055):609-616.
[15] 赵恩毅. 大数据中的数据清洗与预处理技术研究[J]. 信息记录材料,2024,25(3):195-197.
Zhao E Y. Research on data cleaning and preprocessing technology in big data[J]. Information Recording Materials,2024,25(3):195-197.(in Chinese)
[16] Wang Y Z,Hu W B,Dong Y P,et al. Align is not enough:Multimodal universal jailbreak attack against multimodal large language models[J]. IEEE Transactions on Circuits and Systems for Video Technology,2025,35(6):5475-5488.
[17] 林晓淅. LoRA/QLoRA微调技术在DeepSeek大模型问答系统中的应用[J]. 信息记录材料,2025,26(10):205-207.
Lin X X. Application of LoRA/QLoRA fine-tuning technology in DeepSeek large model question answering system[J]. Information Recording Materials,2025,26(10):205-207.(in Chinese)
[18] Kaddour J,Harris J,Mozes M,et al. Challenges and applications of large language models[PP/OL]. V1. arXiv(2023-07-19)[2025-09-15]. https://arxiv.org/pdf/2307.10169.
[19] Zhang D,Feng T,Xue L L,et al. Parameter-efficient fine-tuning for foundation models[PP/OL]. V1. arXiv(2025-01-23)[2025-09-15]. https://arxiv.org/pdf/2501.13787.
[20] Hu E,Shen Y L,Wallis P,et al. LoRA:Low-rank adaptation of large language models[PP/OL]. V2. arXiv(2021-10-16)[2025-09-15]. https://arxiv.org/pdf/2106.09685.
[21] Qwen Team. Qwen3 technical report[PP/OL]. V1. arXiv(2025-05-14)[2025-09-15]. https://arxiv.org/pdf/2505.09388.
[22] Bi X,Chen D L,Chen G T,et al. DeepSeek LLM:Scaling open-source language models with longtermism[PP/OL]. V1. arXiv(2024-01-05)[2025-09-15]. https://arxiv.org/pdf/2401.02954.
[23] Yang A Y,Xiao B,Wang B N,et al. Baichuan 2:Open large-scale language models[PP/OL]. V4. arXiv(2025-04-17)[2025-09-15]. https://arxiv.org/pdf/2309.10305.
[24] Du Z X,Qian Y J,Liu X,et al. GLM:General language model pretraining with autoregressive blank infilling[C/OL]//ACL,2022:320-335. https://aclanthology.org/2022.acl-long.26.pdf.
[25] 贾子硕,张俭鸽,贺浩峰,等. 大模型与知识图谱互增技术与应用综述[J]. 计算机科学,2026,53(8):219-228.
Jia Z S,Zhang J G,He H F,et al. Survey on mutually augmenting technologies and applications of large models and knowledge graphs[J]. Computer Science,2026,53(8):219-228.(in Chinese)

责任编辑:张自强