文章编号:2096-1472(2026)10-0013-06
DOI:10.19644/j.cnki.issn2096-1472.2026.010.003

基于检索增强生成的心肌病大模型临床辅助诊断研究

杨昌吉[1]，张选德[1]，贺惠新[2]，刘 婕[3]，
段岩峰[4]，黄 鑫[5]

(1.陕西科技大学电子信息与人工智能学院，陕西 西安 710021；
2.华侨大学计算机科学与技术学院，福建 厦门 361021；
3.中国医学科学院阜外医院，北京 100037；
4.中关村熙睿心脑肿瘤精准医疗研究院，北京 100080；
5.宁波大学信息科学与工程学院，浙江 宁波 315211)

✉ mailycj@163.com; love_truth@126.com; huixinhe@qq.com; liujyx123@163.com; duanyanfeng@gmail.com; huangxin@nbu.edu.cn

摘 要:针对心肌病诊断“多源信息过载”“知识更新滞后”及传统大模型“幻觉频发”等痛点,提出基于检索增强生成(Retrieval-Augmented Generation，RAG)的辅助诊断系统。以阜外医院2024年642份含5类心肌病分型的临床报告为数据集,整合国际指南、权威文献构建知识库,通过多阶段混合检索与精准提示工程关联患者信息与权威知识。实验结果显示,该系统在心脏超声(93.5%，0.93)、核磁共振报告(87.2%，0.87)诊断准确率和$F1$分数上显著优于主流模型,证实RAG架构有效性,为智能辅助诊断系统发展提供支撑。

关键词:大语言模型;检索增强生成;心肌病;辅助诊断;临床决策支持
中图分类号:TP391 **文献标识码**:A

Research on Clinical Assisted Diagnosis of Cardiomyopathy Based on a Retrieval-Augmented Generation Large Model

YANG Changji[1], ZHANG Xuande[1], HE Huixin[2], LIU Jie[3], DUAN Yanfeng[4], HUANG Xin[5]

(1.*School of Electronic Information and Artificial Intelligence, Shaanxi University of Science & Technology, Xi'an* 710021, *China*;
2.*College of Computer Science and Technology, Huaqiao University, Xiamen* 361021, *China*;
3.*FUWAI Hospital, CAMS & PUMC, Beijing* 100037, *China*;
4.*Zhongguancun Xirui Institute of Precision Medicine for Cardio Cerebral Tumors, Brain and Tumor, Beijing* 100080, *China*;
5.*Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo* 315211, *China*)

✉ mailycj@163.com; love_truth@126.com; huixinhe@qq.com; liujyx123@163.com; duanyanfeng@gmail.com; huangxin@nbu.edu.cn

Abstract: This study addresses critical challenges in cardiomyopathy diagnosis, including “multi-source information overload” “knowledge update lag” and the “frequent hallucinations” of traditional large models, by proposing an assisted diagnosis system based on Retrieval-Augmented Generation (RAG). Using 642 clinical reports encompassing five cardiomyopathy subtypes from Fuwai Hospital in 2024 as the dataset, a knowledge base was constructed by integrating international guidelines and authoritative literature. Patient information was linked with authoritative knowledge through multi-stage hybrid retrieval and precise prompt engineering. Experimental results demonstrate that the system significantly outperforms mainstream models in diagnostic accuracy and $F1$-score for echocardiography reports (93.5%, 0.93) and cardiac magnetic resonance imaging reports (87.2%, 0.87), confirming the effectiveness of the RAG architecture and providing support for the development of intelligent assisted diagnosis systems.

Key words: Large Language Model; Retrieval-Augmented Generation; cardiomyopathy; assisted diagnosis; clinical decision support

0 引言(Introduction)

心肌病的精准诊断是心血管疾病领域的重大挑战之一。它的临床诊疗过程高度复杂,不仅依赖于对多源、异构临床信息的深度融合与解读,更要求医生具备广博且前沿的专病知

收稿日期:2025-11-27
基金项目:福建省自然科学基金项目资助(2023J01138);宁波大学 One Health 交叉学科研究项目资助(HY202409)

www.rjgcz.com

识[1-3]。一次完整的心肌病评估涉及患者主诉、现病史、家族史、心脏超声、心电图乃至心脏核磁共振等多模态数据。尤其是心脏影像报告，包含了大量结构性与功能性测量参数[4-5]。医生需要从这些大量、非结构化的文本与数据中，快速识别关键异常指标，并将它与可能的病理生理机制相关联。这种信息过载不仅耗费大量精力，更增加了因疲劳或疏忽而遗漏关键诊断线索的风险。同时，心肌病分型精细，它的诊断标准并非一成不变，而是随着新的循证医学证据出现而持续更新。国际权威指南通常会发布修订，这对于临床医生，尤其是基层医生而言，构成了巨大的知识更新压力。在紧张的诊疗时间内，准确记忆并应用所有分型的最新诊断标准，在实践中极为困难，可能导致诊断依据滞后或误用。

自 2017 年 Transformer[6] 架构奠定现代人工智能（Artificial Intelligence，AI）基础以来，大模型技术通过参数、数据与算力的规模扩张，实现了通用能力的跃迁式提升，发展呈爆发态势。以 OpenAI 的 GPT[7]、Anthropic 的 Claude[8]、深度求索的 DeepSeek[9] 及阿里云的 Qwen[10] 为代表的主流模型，参数规模已从亿级攀升至万亿级，并在多个行业展现出强大的应用潜力。然而，当这些通用大模型被直接用于医疗辅助诊断场景时，它们存在知识更新滞后（难以同步最新诊疗指南）、幻觉频发（诊断结论不可溯源）、专业适配成本高（全参数微调耗资源）等核心瓶颈[11-12]。

面对大模型知识更新滞后、幻觉频发及专业领域适配成本高等核心瓶颈，检索增强生成（Retrieval-Augmented Generation，RAG）[13] 技术已成为关键破局方向。该技术通过构建外部知识库与向量检索系统，将实时、可信的上下文信息动态注入生成过程，实现了“记忆”与“推理”的解耦。相较于全参数微调和 LoRA（Low-Rank Adaptation）微调[14]，RAG 在保持模型通用能力的基础上，以更低的成本实现知识的即时更新与溯源可控，它在金融、医疗、法律等高精度场景的落地实践，有效弥补了传统大模型在专业性与时效性上的短板。

基于此，本研究旨在开发一种基于 RAG 的心肌病辅助诊断系统，通过实时整合患者临床报告数据与最新指南知识，构建具备透明化推理能力的智能决策支持平台，以期降低诊断负荷、提升决策精度，为不同层级医疗机构提供可持续更新的专家级辅助工具。

1 数据集（Dataset）

本研究采用的数据集为中国医学科学院阜外医院心肌病临床诊断报告数据集，覆盖 2024 年 1 月—2024 年 12 月期间的 642 份有效报告。该数据集聚焦于真实的心肌病临床诊断场景，涵盖不同性别和年龄的患者，记录了患者的基础信息、既往病史、家族病史、心脏超声报告、心脏核磁共振报告以及人工医生的诊断结果等信息。每个患者均包含以下数据字段：

“性别，身高，体重，就诊年龄，既往药物治疗，其他家族史，心肌病家族史，猝死家族史详情，运动后左室流出道压差峰值，最慢心率，磁共振成像（Magnetic Resonance Imaging，MRI）检查扫描所见，静息左室流出道压差峰值，确诊时间，静息左室流出道流速，立位左室流出道流速，有无晕厥，肌红蛋白，全天平均舒张压，主动脉瓣前向血流流速，出院诊断，左室射血分数，猝死家族史，高敏肌钙蛋白 I，MRI 最大室壁厚度，心功能，右心房径，最快心率，最大室壁厚度，有无头晕，主动脉瓣前向血流压差，心肌梗死 3 项检查日期，运动后左室流出道流速，B 型钠尿肽，超声描述，左心房内径，有无房颤，心肌病家族史详情，立位左室流出道压差峰值，MRI 扫描左室射血分数，有无黑矇，室间隔厚度，有无胸痛，后壁厚度，MRI 增强检查扫描所见，有无夜间阵发性呼吸困难，氨基酸末端脑钠素前体，超声类型，有无短阵室性心动过速，肌酸激酶同工酶质量，发病时间，左心室舒张末径，右心室前后径，全天平均收缩压，室性心动过速最快心室率，有无活动后胸闷，平均心率，诊断结果”。

报告中的诊断结果依据欧洲心脏病学会（European Society of Cardiology，ESC）发布的《2023 ESC Guidelines for the management of cardiomyopathies》[15]，指南中包括 HCM（Hypertrophic CardioMyopathy）432 例，DCM（Dilated CardioMyopathy）121 例，RCM（Restrictive CardioMyopathy）56 例，ARVC（Arrhythrnogenic Right Ventricular CardioMyopathy）17 例，NDLVC（Non-Dilated Left Ventricular CardioMyopathy）16 例。所有数据均经过严格去标识化处理，确保数据使用的合规性与隐私保护。

2 本文方法（Proposed method）

2.1 系统工作流程

本研究构建的基于知识增强与大语言模型（Large Language Model，LLM）推理的心肌病辅助诊断系统的工作流程如图 1 所示。它的核心机制始于对多源医学知识的预处理与分块，随后通过向量化与索引技术构建结构化知识库，为后续检索奠定基础。在诊断环节，系统接收患者临床数据，并对其进行语义层面的查询理解。该理解结果与知识库内容共同输入混合检索模块，实现精准的知识召回。检索所得信息经过去噪、冗余过滤及相关性重排序后，筛选出最相关的医学知识片段。随后，系统将这些关键知识与原始患者数据整合，动态构建上下文环境，并以此作为提示（prompt）输入 LLM 进行深度推理。最终，LLM 综合分析所有上下文信息，输出结构化的诊断建议或结论，从而为临床医生提供高置信度的决策支持。

该流程体现了“知识驱动＋数据驱动”的双引擎架构，通过结构化知识检索与非结构化语义推理的协同，有效提升了诊断系统的准确性、可解释性与临床适用性。

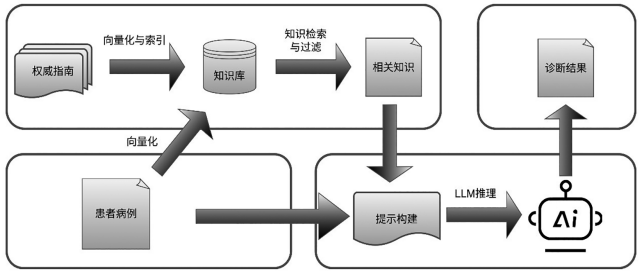

图 1 心肌病辅助诊断系统工作流程

Fig. 1 Flow for working process of a cardiomyopathy auxiliary diagnosis system

2.2 知识源

RAG 系统性能高度依赖知识库的权威性与质量。本研究整合多层次高可信度医学知识源，涵盖国际临床指南（如 AHA（American Heart Association）与 ESC 发布的《心肌病及心力衰竭管理指南》）、权威教科书（如《Braunwald’s Heart Disease》和《ESC 心血管病教科书》核心章节）及顶级期刊（如《Circulation》和《JACC》）精选综述与关键研究，总计 3 GB 数据（其中指南与教科书数据量为 1 GB，核心文献数据量为 2 GB）。

2.3 数据分块

为了构建高质量的知识库，本研究针对不同类型的数据源采用了差异化的预处理与分割策略，以确保数据的结构化、完

www.rjgcz.com

整性和检索效率。

权威指南和教科书类数据具有结构化强、层级清晰、规范性强和结论明确的特点。采用“层级拆分＋条目原子化”的分割原则对其进行处理，以确保不破坏推荐意见的完整性。具体而言，首先按照章节、小节等层级结构进行拆分，随后将每个条目进一步拆分为原子化的知识点，确保每个知识点独立且完整。

核心医学文献具有结构固定、逻辑连贯和数据密集的特点。采用“论文结构＋语义块”拆分的策略进行数据分割，以保留实验和结论之间的关联性。具体而言，首先按照论文的标准结构（如引言、方法、结果、讨论）进行拆分，随后在每个部分内根据语义块进行进一步分割，确保每个语义块包含完整的逻辑单元。

为了提高检索数据的召回率，在进行文本分割前对核心医学文献进行前置预处理。前置处理的主要目的是统一权威指南和核心医学文献中的医学术语，确保核心医学文献中的医学术语与权威指南一致。通过术语对齐，可以有效减小因术语差异导致的检索误差，提升知识库的整合性和检索效率。

诊疗指南与教科书的核心价值在于提供权威推荐意见。分割时应严格遵循它们的固有层级，确保每个 Chunk 包含完整的推荐条目、诊断标准或治疗方案，避免碎片化。通过元数据记录层级关系（如“指南名称—章节—子章节—条目”），并优先保留推荐级别、证据等级及核心结论等关键信息。

期刊论文的核心价值在于提供研究证据（方法、结果与结论）。因此，数据分割应遵循 IMRaD 结构（引言、方法、结果、讨论），维持“研究目的—方法—结果—结论”的逻辑链，防止语义断裂。具体而言，先依据论文固有结构（标题、摘要、章节、子章节）进行初步拆分，再按语义单元进一步细化。此外，需保持“方法—结果”与“结果—结论”的关联性，并将样本量、统计指标、P 值及核心结论等关键信息完整保留于对应模块。

本研究采用 LangChain 框架的 PyPDFLoader 模块加载 PDF 格式的指南、教科书及核心期刊论文，并运用其内置的 RecursiveCharacterTextSplitter 进行文本分割。处理指南与教科书时，为保障推荐意见完整性，设置 chunk_size 为 2000、chunk_overlap 为 100 以保持上下文连贯；针对核心期刊论文，则将 chunk_size 设为 1500、chunk_overlap 设为 200。

对于分割之后的文本片段，利用开源的 Qwen3-Embedding-0.6B 高性能文本嵌入模型，将每个文本块转换为稠密向量（Embedding）。随后，将这些向量及其对应的原始文本存储至专业的向量数据库 Milvus 中，构建成系统的“长期记忆”。

2.4　数据检索

检索模块负责从知识库中快速定位与患者高度相关的信息。本研究采用多阶段混合检索策略，以“精准匹配＋高权威”为核心原则。查询预处理阶段执行术语标准化（缩写补全、同义词扩展）与结构化解析，自动提取核心指标（如 LVEF（Left Ventricular Ejection Fractions）≤40%）、诊疗环节（如家族史、用药史）等要素，并生成多查询变体以扩大召回范围。检索阶段并行整合 4 种方法：基于元数据索引的精准筛选，基于 Qwen3-Embedding-0.6B 的语义向量检索，基于 Elasticsearch BM25 的关键词精确匹配，以及基于关联 ID 的上下文单元召回。融合与重排阶段合并多源结果并去重后，采用多维度加权评分模型 Score ＝ w_1×Authority ＋ w_2×Relevance ＋ w_3×Timeliness，其中权威性权重最高（w_1＝40%），相关性（w_2＝35%）与时效性（w_3＝25%）次之；继而使用 ClinicalBERT[16] 模型对 Top 20 候选结果精排，滤除语义无关项。输出阶段动态筛选指南 IA/IB 类高级别推荐及核心期刊 RCT 等高可信度证据，严格将结果数量控制在 8 条以内。

2.5　诊断结果生成

生成模块作为系统核心，负责整合检索到的知识与患者信息以生成最终诊断结果，其关键在于提示词的精准设计。辅助诊断系统输出结果如图 2 所示。提示词是连接患者数据、检索知识与 LLM 生成诊断报告的核心桥梁，而精心构建的提示工程能够有效引导 LLM 基于零散输入（患者数据、Top-K 知识片段），输出符合临床逻辑、结构规范的诊断结果。

图 2　辅助诊断系统输出结果

Fig. 2 Output results of auxiliary diagnosis system

本研究采用 Qwen3-235B-A22B-Instruct-2507 模型作为诊断结果生成器，并为其设计了严格的提示模板，以约束输出格式与内容，保障诊断结果的安全性与实用性。系统角色被明确设定为：“你是一名心肌病诊断辅助 AI。你必须严格且仅依据提供的患者信息和检索到的权威知识片段进行分析。你的回答必须客观、严谨，并明确引用知识来源。”该设定通过强制模型基于“检索到的知识片段”生成回答且要求明确引用来源，大幅降低了模型虚构信息的风险，使每一项关键诊断判断均有据可查，显著提升了输出结果的可信度与临床接受度。

3　实验与结果分析（Experiment and result analysis）

本文基于患者基础信息，在心脏超声报告和心脏核磁共振报告上对 RAG 系统与当前主流的大模型进行了对比分析。RAG 系统采用 LangChain v1.0 框架开发，集成 Milvus 向量数据库，部署于 Ubuntu 20.04 环境，Python 版本为 3.12。在图 3～图 6，表 1～表 4 中：对比模型包括 DeepSeek-V3.1-Terminus（简写为 DeepSeek）、GLM-4.6（简写为 GLM）、Qwen3-235B-A22B-Instruct-2507（简写为 Qwen）及 Claude-Haiku-4.5-20251001（简写为 Claude）。研究以人工诊断为金标准，采用准确率（Accuracy，Acc）与微平均 $F1$ 分数（Micro-$F1$）作为评价指标。

3.1　心脏超声报告数据实验结果

本文在患者基础信息与心脏超声报告上对比了 RAG 系统与主流大语言模型。准确率及 $F1$ 分数如图 3 和图 4 所示，各类疾病的诊断性能详见表 1 和表 2。表中每个单元格包含两行数据：第 1 行为对应疾病的准确率或 $F1$ 分数，第 2 行为 Bootstrap 方法计算的 95%置信区间。

www.ijczz.com

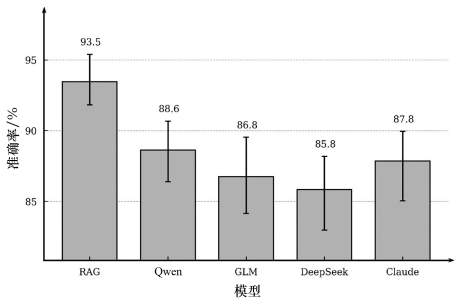

图 3 各模型在心脏超声报告上的诊断准确率

Fig. 3 Diagnostic accuracy of various models on echocardiography reports

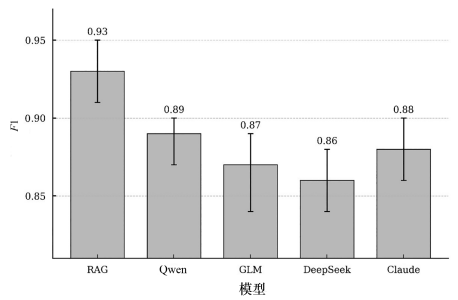

图 4 各模型在心脏超声报告上的 $F1$ 分数

Fig. 4 Diagnostic $F1$-score of various models on echocardiography reports

如图 3 和图 4 所示，在基于患者基础信息和心脏超声报告数据的测试中，本文开发的 RAG 系统显著优于其他主流模型，准确率和平均 $F1$ 分数分别达到了 93.5%和 0.93。具体而言：RAG 系统的准确率和 $F1$ 分数比 Qwen3-235B-A22B-Instruct-2507 模型(88.6%，0.89)分别高出 4.9 个百分点和 0.04；比 GLM-4.6 模型(86.8%，0.87)高出 6.7 个百分点和 0.06；比 DeepSeek-V3.1-Terminus 模型(85.8%，0.86)高出 7.7 个百分点和 0.07；比 Claude-haiku-4-5-20251001 模型(87.8%，0.88)高出 5.7 个百分点和 0.05。

如表 1 和表 2 所示，在超声报告数据上，本文的 RAG 系统在各心肌病的诊断准确率和 $F1$ 分数上均取得最好效果。并且 RAG 系统在非扩张型左室心肌病(NDLVC)、限制型心肌病(RCM)及致心律失常性右室心肌病(ARVC)的诊断准确率上均表现出显著优势。

表 1 各模型在心脏超声报告中对不同疾病诊断的准确率比较

Tab. 1 Comparison of diagnostic accuracy for different diseases in echocardiography reports among various models

Model	HCM	NDLVC	DCM	RCM	ARVC
RAG	0.968 (0.950, 0.982)	0.875 (0.688, 1.000)	0.959 (0.921, 0.992)	0.661 (0.509, 0.768)	0.882 (0.766, 0.920)
Qwen	0.965 (0.975, 0.990)	0.750 (0.720, 0.968)	0.934 (0.945, 0.987)	0.286 (0.303, 0.564)	0.647 (0.551, 0.701)
DeepSeek	0.958 (0.969, 0.987)	0.250 (0.118, 0.609)	0.950 (0.948, 0.992)	0.179 (0.164,0.423)	0.471 (0.339,0.628)
GLM	0.961 (0.970, 0.987)	0.188 (0.118, 0.513)	0.926 (0.929, 0.987)	0.321 (0.328, 0.504)	0.529 (0.486, 0.666)
Claude	0.954 (0.965, 0.987)	0.813 (0.769, 0.985)	0.893 (0.907, 0.973)	0.339 (0.364, 0.634)	0.705 (0.640, 0.837)

表 2 各模型在心脏超声报告中对不同疾病诊断的 $F1$ 分数比较

Tab. 2 Comparison of $F1$-score for different diseases in cardiac echocardiogram reports among models

Model	HCM	NDLVC	DCM	RCM	ARVC
RAG	0.979 (0.971, 0.986)	0.757 (0.561, 0.881)	0.903 (0.864, 0.934)	0.796 (0.691, 0.867)	0.833 (0.692, 0.909)
Qwen	0.977 (0.968, 0.987)	0.727 (0.556, 0.882)	0.837 (0.798, 0.895)	0.444 (0.297, 0.554)	0.578 (0.395, 0.747)
DeepSeek	0.974 (0.966, 0.983)	0.381 (0.111, 0.615)	0.774 (0.717, 0.819)	0.298 (0.157, 0.444)	0.592 (0.296, 0.774)
GLM	0.976 (0.965, 0.985)	0.206 (0.000, 0.427)	0.814 (0.770, 0.856)	0.486 (0.310, 0.611)	0.600 (0.348, 0.777)
Claude	0.974 (0.964, 0.984)	0.541 (0.302, 0.684)	0.774 (0.704, 0.817)	0.493 (0.346, 0.621)	0.750 (0.579, 0.884)

3.2 心脏核磁报告数据实验结果

本文还在患者基础信息和心脏核磁报告上对 RAG 系统和其他主流大语言模型进行了对比。准确率和 $F1$ 分数的结果如图 5 和图 6 所示。各类疾病的诊断性能见表 3 和表 4。

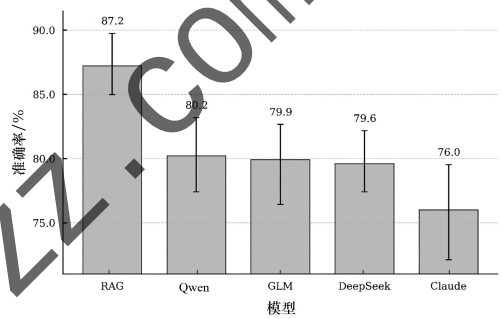

图 5 各模型在心脏核磁报告上的诊断准确率

Fig. 5 Diagnostic accuracy of various models on cardiac MRI reports

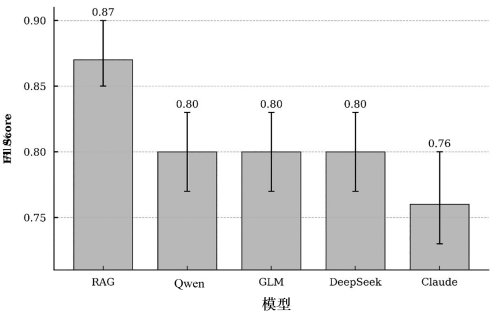

图 6 各模型在心脏核磁报告上的 $F1$ 分数

Fig. 6 $F1$ Score of various models on cardiac MRI reports

如图 5 和图 6 所示，在基于患者基础信息和心脏核磁共振报告数据的测试中，本文开发的 RAG 系统显著优于其他主流模型，准确率和平均 $F1$ 分数分别达到了 87.2%和 0.87。与 Qwen3-235B-A22B-Instruct-2507 模型(80.2%，0.80)相比，RAG 系统的准确率和 $F1$ 分数高出 7.0 个百分点和 0.07；与 GLM-4.6 模型(79.9%，0.80)相比，高出 7.3 个百分点和 0.07；与 DeepSeek-V3.1-Terminus 模型(79.6%，0.80)相比，高出 7.6 个百分点和 0.07；与 Claude-haiku-4-5-20251001 模型(76.2%，0.76)相比，高出 11.0 个百分点和 0.11。

如表 3 和表 4 所示，在核磁报告数据中，本文的 RAG 系统

www.rjgczz.com

同样在各心肌病的诊断准确率和 $F1$ 分数均取得最好效果。

表3　各模型在心脏核磁报告中对不同疾病诊断的准确率比较

Tab. 3 Comparison of diagnostic accuracy for different diseases in cardiac MRI reports among various models

Model	HCM	NDLVC	DCM	RCM	ARVC
RAG	**0.887 (0.856, 0.906)**	**0.938 (0.812, 1.000)**	**0.984 (0.958, 1.000)**	**0.500 (0.3656 0.642)**	**0.882 (0.705, 1.000)**
Qwen	0.879 (0.844, 0.906)	0.812 (0.625, 0.937)	0.785 (0.714, 0.856)	0.267 (0.169, 0.366)	0.705 (0.470, 0.854)
DeepSeek	0.868 (0.836, 0.901)	0.500 (0.279, 0.687)	0.966 (0.933, 0.991)	0.071 (0.008, 0.142)	0.411 (0.117, 0.619)
GLM	0.849 (0.813, 0.885)	0.875 (0.687, 1.000)	0.834 (0.776, 0.900)	0.392 (0.285, 0.509)	0.529 (0.294, 0.795)
Claude	0.856 (0.818, 0.890)	0.875 (0.687, 0.970)	0.677 (0.603, 0.752)	0.160 (0.062, 0.250)	0.764 (0.588, 0.941)

表4　各模型在心脏核磁报告中对不同疾病诊断的 $F1$ 分数比较

Tab. 4 Comparison of $F1$ Scores for different diseases in cardiac MRI reports among models

Model	HCM	NDLVC	DCM	RCM	ARVC
RAG	**0.933 (0.911, 0.950)**	**0.750 (0.590, 0.890)**	**0.853 (0.806, 0.889)**	**0.667 (0.525, 0.764)**	**0.857 (0.722, 0.961)**
Qwen	0.925 (0.908, 0.941)	0.500 (0.318, 0.642)	0.766 (0.711, 0.816)	0.422 (0.283, 0.539)	0.558 (0.310, 0.710)
DeepSeek	0.921 (0.899, 0.937)	0.592 (0.347, 0.750)	0.777 (0.727, 0.827)	0.133 (0.035, 0.236)	0.583 (0.324, 0.823)
GLM	0.915 (0.897, 0.934)	0.528 (0.330, 0.687)	0.811 (0.753, 0.859)	0.564 (0.396, 0.678)	0.666 (0.413, 0.845)
Claude	0.913 (0.897, 0.932)	0.307 (0.199, 0.424)	0.628 (0.567, 0.692)	0.253 (0.116, 0.383)	0.722 (0.555, 0.877)

3.3　消融实验

为验证各模块的有效性，本文在超声报告数据集上开展了3组消融实验（结果见表5）。所有实验均采用 Qwen3-235B-A22B-Instruct-2507 模型。第一组以当前通用 RAG 系统为基准实验，知识库通过字符级简单分块构建（chunk_size＝2000，chunk_overlap＝100）。检索时仅依据向量相似度，选取与查询最相似的前 8 个文本块，连同查询一起输入 LLM 生成诊断结果。第二组实验使用分层知识库构建模块替换第一组实验中的简单分块知识库构建，RAG 系统的准确率和微平均 $F1$ 分数分别提升了 2.62 个百分点和 0.01。第三组实验在第二组实验的基础上采用多阶段混合索引替代单一向量相似度检索，准确率和微平均 $F1$ 分数分别提升了 1.98 个百分点和 0.03。最终的 RAG 系统比原始 RAG 系统准确率提升了 4.60 个百分点，$F1$ 分数提高了 0.04，验证了各模块的有效性。

表5　消融实验结果

Tab. 5 Results of ablation experiment

序号	base	分层知识库构建	多阶段混合索引	*Acc*/%	*F*1 分数
1	√			88.90	0.89
2	√	√		91.52	0.90
3	√	√	√	93.50	0.93

4　讨论（Discussion）

尽管本研究在阜外医院 642 份临床报告上取得了超声 93.5%与核磁 87.2%的诊断准确率，但本研究仍存在以下局限：其一，当前验证数据集局限于单一中心、单一年度（2024年），尽管该辅助诊断系统在阜外医院内部测试中表现优异，但跨机构泛化能力尚未验证。不同医疗机构在设备层面（超声/MRI 扫描参数与后处理算法差异）、流程层面（报告模板、术语体系及诊断层级的结构性差异）与认知层面（医师对指南理解的个体化偏好）的三重异质性，可能导致跨院部署时出现性能衰减。其二，心肌病领域目前尚缺乏面向 RAG 系统的公开评测基准（Benchmark），制约了本方法与其他研究的横向可比性。

为突破上述局限，后续研究将进行递进式优化：首先，启动多中心、跨设备、前瞻性临床验证，拟纳入 3～5 家不同层级医疗机构构建异构验证队列，系统评估模型在区域医疗中心与基层医院的适配性能；其次，联合心血管学会与医疗 AI 标准化组织推动建立心肌病 RAG 公开评测数据集，规范标注体系与评估协议，为心肌病领域提供可信基准；最后，探索领域自适应（Domain Adaptation）与持续学习（Continual Learning）混合范式，通过对抗域对齐与增量知识更新机制，使系统动态适配不同机构文本特征及 ESC 指南等循证证据的演进，从而实现专科精度与临床泛化能力的可持续平衡。

5　结论（Conclusion）

在心脏超声与心脏核磁共振两组报告的对比实验中，RAG 系统均取得最高准确率和 $F1$ 分数。需要指出的是，RAG 系统所采用的模型正是 Qwen3-235B-A22B-Instruct-2507。RAG 系统与 Qwen3-235B-A22B-Instruct-2507 模型的唯一差异在于：RAG 系统在生成诊断结果时会附加检索到的知识块。对比 RAG 系统与 Qwen3-235B-A22B-Instruct-2507 模型的结果表明，通过在输入数据中增加额外的知识，能够显著提高模型的准确率，并有效减少模型的幻觉现象。这一发现验证了 RAG 架构在心肌病诊断中的有效性，为未来智能辅助诊断系统的发展提供了有力支持。

参考文献（References）

[1] Samanidis G. Current challenges in diagnosis and treatment of cardiovascular disease[J]. Journal of Personalized Medicine, 2024, 14(8): 786.

[2] 张林鑫，陈盈，齐海成，等. 心脏磁共振在扩张型心肌病中的应用进展[J]. 磁共振成像，2025，16(3)：150-155.

[3] 郭惠明，魏培坚. 梗阻性肥厚型心肌病外科治疗的挑战与未来[J]. 中华外科杂志，2023，61(3)：181-186.

[4] Caobelli F, Cabrero J B, Galea N, et al. Cardiovascular Magnetic Resonance (CMR) and Positron Emission Tomography (PET) imaging in the diagnosis and follow-up of patients with acute myocarditis and chronic inflammatory cardiomyopathy: a review paper with practical recommendations on behalf of the European Society of Cardiovascular Radiology (ESCR)[J]. The International Journal of Cardiovascular Imaging, 2023, 39(11): 2221-2235.

[5] 王丽，王怡璐，张雨，等. ^{18}F-FAPI PET/CT 显像识别肥厚型心肌病受累心肌[J]. 中华核医学与分子影像杂志，2025，45(9)：519-524.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C/OL]// NeurIPS, 2017: 6000-6010. https://proceedings. neurips. cc/paper _ files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[7] Achiam J, Adler S, Agarwal S, et al. GPT-4 technical re-

www.rjgzz.com

port[PP/OL]. V1. arXiv (2023-03-15) [2025-05-04]. https://arxiv.org/abs/2303.08774.
[8] Anthropic. Introducing Claude Sonnet 4.5[EB/OL]. (2025-09-30) [2025-05-04]. https://www.anthropic.com/news/claude-sonnet-4-5.
[9] Liu A, Feng B, Xue B, et al. DeepSeek-V3 technical report [PP/OL]. V1.arXiv (2024-12-30) [2025-05-04]. https://arxiv.org/abs/2412.19437.
[10] Bai J, Bai S, Chu Y, et al. Qwen technical report [PP/OL]. V1. arXiv (2023-09-28) [2025-05-04]. https://arxiv.org/abs/2309.16609.
[11] 鲁维海,魏华鹏,邓博元,等. 医学大语言模型中幻觉的防治[J]. 中国数字医学,2025,20(10):34-39,52.
[12] 陶立元,钟晓莹,代倩倩,等. 医疗保健领域大语言模型研究报告规范 TRIPOD-LLM 的应用范围解析与现状调查[J]. 中华医学杂志,2025,105(38):3379-3383.
[13] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [C/OL]// NeurIPS, 2020: 9459-9474. https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b6e6f56826f826f90fb361c618d3608-Abstract.html.
[14] Hu E J, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models [PP/OL]. V1. arXiv (2021-06-17) [2025-05-04]. https://arxiv.org/abs/2106.09685.
[15] Arbelo E, Protonotarios A, Gimeno J R, et al. 2023 ESC guidelines for the management of cardiomyopathies: developed by the task force on the management of cardiomyopathies of the European Society of Cardiology (ESC)[J]. European Heart Journal, 2023, 44(37): 3503-3626.
[16] Liu X, Liu H, Yang G, et al. A generalist medical language model for disease diagnosis assistance[J]. Nature Medicine, 2025, 31(3): 932-942.

作者简介:

杨昌吉(1999-),男(汉族),汉中,硕士生。研究领域:人工智能应用。
张选德(1979-),男(汉族),银川,教授,博士。研究领域:人工智能应用。
贺惠新(1984-),男(汉族),厦门,高级工程师,博士。研究领域:智能信息处理、工业控制网络。
刘 婕(1993-),女(汉族),北京,住院医师,博士。研究领域:心肌病。
段岩峰(1980-),男(汉族),北京,高级工程师,博士。研究领域:智慧医疗。
黄 鑫(1981-),男(汉族),宁波,教授,博士。研究领域:人工智能应用。本文通信作者。

(上接第 4 页)

参考文献(References)

[1] 李铂鑫,鲁骁,张霄,等.面向检索增强式私有问答大模型的幻觉检测方法[J].软件导刊,2026,25(1):39-46.
[2] 李丽,陈克伟,王心瑶.军用车辆企业 GJB9001C 与 IATF16949 体系融合研究[J].重型汽车,2024,(3):35-36.
[3] 吴璇,付涛.检索增强生成技术研究综述[J].计算机工程与应用,2025,61(20):19-35.
[4] Mahalingam A, Gande V K, Chadha A, et al. SKETCH: structured knowledge enhanced text comprehension for holistic retrieval[C/OL]// GenAIK, 2025: 27-42. https://aclanthology.org/2025.genaik-1.4.pdf.
[5] Edge D, Trinh H, Cheng N, et al. From local to global: a graph rag approach to query-focused summarization [PP/OL]. V2. arXiv (2025-02-19) [2025-06-19]. https://arxiv.org/abs/2404.16130.
[6] Pan S, Luo L, Wang Y, et al. Unifying large language models and knowledge graphs: a roadmap [J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3580-3599.
[7] 周俭航,肖诗斌.图检索增强生成研究综述[J].人工智能与机器人研究,2025,14(2):402-413.
[8] 刘兆年,姜彬,王宁,等.基于图检索生成增强的油藏地质问答系统构建[J].石油科学通报,2025,10(5):1069-1082.
[9] 武天野.轴承制造业生产数据实时集成系统设计与实现[J].计算机时代,2022,(8):56-60.
[10] 徐浩,康振渊,张焱,等.面向电力变压器故障辅助决策的知识图谱构建及补全策略研究[J].中国安全生产科学技术,2025,21(5):46-54.
[11] Marchesin S, Silvello G, Alonso O. Large Language Models and data quality for knowledge graphs[J]. Information Processing & Management, 2025, 62(6): 104281.
[12] 周学文,王俊强,江荣.结合向量匹配的图数据库智能问答智能体设计与实现[J].信息与电脑,2025,37(23):51-53.
[13] 李靓果,薛志一,陈小红,等.结合大语言模型和领域知识库的证券规则规约方法[J].软件学报,2025,36(10):4671-4694.
[14] Traag V A, Waltman L, Van Eck N J. From louvain to leiden: guaranteeing well-connected communities[J]. Scientific Reports, 2019, 9(1): 5233.
[15] Es S, James J, Anke-Espinosa L E, et al. RAGAs: automated evaluation of retrieval augmented generation[C/OL]// EACL, 2024: 150-158. https://aclanthology.org/2024.eacl-demo.16.pdf.

作者简介:

蔡家俊(2000-),男(汉族),台州,硕士生。研究领域:智能制造,数字化及智能化。本文通信作者。
鲍 敏(1977-),男(汉族),杭州,副教授,博士。研究领域:智能制造,生产过程数据分析。

www.rjgcz.com