湖北文理学院

毕业设计（论文）开题报告

毕业论文题目	基于 RAG 检索增强和多专家 Agent 的糖尿病智能管理系统
学 生 姓 名	张聪颖
学　　号	2022117353
所 在 学 院	计算机工程学院
所 学 专 业	计算机科学与技术
班　　级	计科 2211
指 导 教 师	吴钊

2025 年 11 月 29 日

一、课题研究的目的与意义

1.1 研究目的

随着我国糖尿病患病率持续上升，慢性病管理需求与传统医疗服务模式之间的矛盾日益突出。本研究旨在设计并实现一套基于 RAG 检索增强和多专家 Agent 的糖尿病智能管理系统，融合 Plan-Execute 规划模式与 ReAct 循环推理的多专家 Agent 协作架构，集成 RAG 医疗知识检索技术，通过 MCP（Model Context Protocol）协议实现标准化工具调用。系统采用 MySQL+MongoDB 混合数据库架构，为糖尿病患者提供智能问诊、健康数据管理、风险评估和个性化建议等服务，同时为医生提供患者管理工具，为管理员提供系统运维功能，构建三端协同的智能健康管理平台。

1.2 研究意义

本课题的研究和应用系统的研发，对推动智能医疗与健康管理、提升慢性病管理水平具有重要的研究意义和实用价值。

首先，该系统创新性地将 Plan-Execute 规划模式与 ReAct 循环推理相结合，设计了层次化的多专家 Agent 协作架构，能够处理复杂的医疗问诊场景，实现全局性规划与精确性推理的有机统一，为糖尿病患者提供专业、准确的健康管理服务。

其次，系统引入 RAG 检索增强技术，构建糖尿病诊疗知识库，通过语义检索动态获取医学专业知识，提升 AI 系统回答的专业性和准确性，弥补传统大模型在专业医疗领域知识不足的问题。

第三，通过 MCP 协议实现 AI Agent 与异构数据源的标准化接口调用，提高系统的模块化程度和可扩展性，为医疗 AI 应用提供了可复用的技术方案。

最后，系统支持用户端、医生端、管理员端三个业务端口，实现健康数据可视化、智能风险评估、医患互动等功能，为糖尿病患者提供全方位的健康管理支持，对促进慢性病管理模式创新具有实践意义。

二、课题国内外研究现状

糖尿病智能管理系统是结合人工智能、大数据和医疗信息化技术的创新应用。随着 RAG 检索增强技术和多专家 Agent 系统的发展，智能医疗领域出现了许多创新应用。国内外在糖尿病管理、RAG 技术和多专家 Agent 系统方面开展了广泛研究和实践。

2.1 国内现状

在糖尿病管理方面，国内部分互联网医疗平台（如平安好医生、丁香医生等）已开始提供糖尿病在线咨询和健康管理服务，但大多以简单的问答和提醒功能为主，缺乏深度的智能分析和个性化建议。近年来，国内研究机构如医渡云、零氪科技等开始探索知识图谱在医疗领域的应用，构建医疗知识库并应用于辅助诊疗。

在AI技术方面，国内大模型如DeepSeek、百川、通义千问等快速发展，但在医疗垂直领域的应用还处于起步阶段。RAG技术在国内也得到关注，智谱AI等团队开发了BGE系列嵌入模型，为中文语义检索提供了强大支持。

关于多专家 Agent 系统，国内研究主要集中在理论层面，实际医疗应用较少。清华大学、中国科学院等机构开展了 Agent 协作机制的研究，但在医疗场景下的应用还不够深入。

2.2 国外现状

国外在糖尿病智能管理方面发展较为成熟。代表性的工作主要包括：

在RAG技术方面，Facebook AI Research于2020年提出RAG框架，将检索增强与生成式模型结合，显著提升了知识密集型NLP任务的性能。OpenAI、Anthropic等公司在其大模型产品中广泛应用RAG技术，提升了回答的准确性和可信度。

在多专家 Agent 方面，2023 年普林斯顿大学提出 ReAct 框架，将推理与行动相结合，实现了更可解释的智能体行为。MIT、Stanford 等机构探索了 Plan-Execute 模式在复杂任务规划中的应用。Anthropic 开发的 Model Context Protocol (MCP)提供了标准化的 AI Agent 工具调用协议，促进了 Agent 系统的开发。

在医疗应用方面，Google Health 开发了基于 Med-PaLM 的医疗对话系统，IBM Watson Health 提供了智能诊疗辅助系统，但大多为单一模型应用，缺乏多专家协作机制。近年来，一些研究开始探索将 RAG 和多专家 Agent 结合应用于医疗场景，但在糖尿病管理领域的应用还不够深入和系统化。

三、研究内容

3.1 研究内容

本课题的研究旨在设计并实现一套基于 RAG 检索增强和多专家 Agent 的糖尿病智能管理系统。系统融合 Plan-Execute 规划模式与 ReAct 循环推理，集成 RAG 医疗知识检索，通过 MCP 协议实现标准化工具调用，为糖尿病患者提供智能问诊、健康管理、风险评估等服务。课题的研究内容包括：

(1) 需求分析

通过调研糖尿病患者、医生和管理员对健康管理和医疗服务的需求，明确系统的核心功能模块。确定系统性能、智能化水平、数据安全性和用户体验等非功能性需求，设计用户端、医生端、管理员端三个业务端口的功能需求。

(2) 系统设计

采用前后端分离架构，后端基于 Spring Boot 3 框架提供 RESTful API，AI服务基于 FastAPI 框架实现高性能异步服务，前端使用 Vue 3 + Element Plus 构建现代化用户界面。设计MySQL-MongoDB混合数据库架构：MySQL存储用户信息、权限管

理等结构化数据，MongoDB 存储健康数据、血糖血压记录、AI 问诊会话记录等时序数据。设计基于 Plan-Execute+ReAct 的多专家 Agent 协作架构，包括 Plan 规划器、数据专家（ReAct 循环推理）、诊断专家、知识专家、数据记录专家等。构建基于 ChromaDB 的 RAG 医疗知识检索系统，使用 BGE-Large-zh-v1.5 模型实现语义检索。通过 MCP 协议实现 AI Agent 与异构数据源之间的标准化接口调用。

(3) 前端开发

使用 Vue 3 + Element Plus 开发用户端、医生端和管理员端三个独立前端应用。上述三个前端应用的主要功能包括：

用户端功能：用户注册登录、个人中心、健康数据管理（血糖、血压、体重记录）、健康数据统计与可视化、智能问诊（AI 对话）、医生列表浏览、在线咨询等。

医生端功能：医生工作台、个人信息管理、咨询管理、咨询用户列表等。

管理员端功能：数据看板、用户管理、医生管理、咨询管理、权限管理、系统设置等。

使用 Vite 构建工具和 Pinia 状态管理，集成 WebSocket 实现实时通信，基于角色的访问控制确保不同用户类型的权限隔离。

(4) 后端开发

使用 Spring Boot 3 开发主要后端服务，提供用户认证、健康数据管理、权限控制等功能。上述后端服务的功能主要包括：

用户认证：使用 RESTful API 接口实现健康数据增删改查、用户信息管理、医生信息管理、权限管理等功能。

健康数据管理：使用 Spring Data MongoDB 管理健康记录时序数据，通过 MongoDB 存储血糖、血压、体重等健康数据，支持高效的时间范围查询和数据统计。

权限控制：集成 JWT 实现用户认证和授权。

(5) AI 智能问诊系统开发

基于 FastAPI 开发 AI 问诊服务，集成 DeepSeek 大模型提供智能对话能力。主要功能包括：

实现多专家 Agent 协作架构：Plan 规划器负责任务分解和专家调度，数据专家（ReAct 循环推理）执行多轮健康数据检索与趋势分析，诊断专家基于分析结果评估糖尿病相关风险，知识专家结合 RAG 检索提供医学依据，数据记录专家解析用户输入并生成结构化健康记录。

开发 MCP 工具调用层，提供标准化的数据库操作（MongoDB、MySQL 查询）、知识检索（RAG）、健康数据分析等功能接口。

实现 WebSocket 实时推送，支持 AI 推理过程的实时展示。

(6) RAG 检索增强系统开发

构建糖尿病医疗知识库，收集整理糖尿病诊疗指南、用药规范等专业医学知识。使用 BGE-Large-zh-v1.5 Embedding 模型将医疗文档转换为向量表示。部署 ChromaDB 向量数据库实现高效的语义检索和相似度匹配。开发 FastAPI 检索服务提供知识检索 API 接口。开发检索增强模块，在 AI 推理过程中通过 MCP 协议动态检索相关医疗知识，提升回答的专业性和准确性。

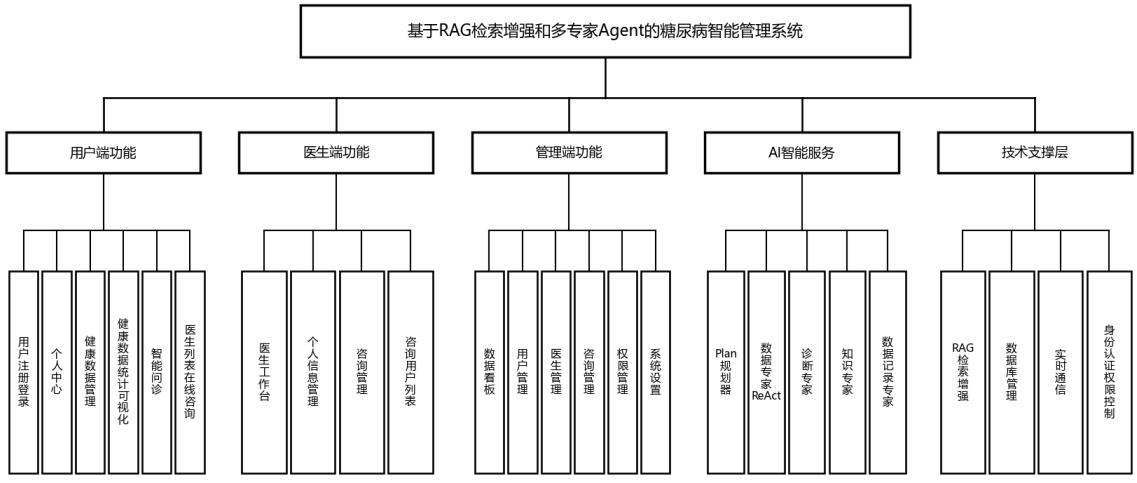
(7) 实时通信开发

集成 WebSocket 实现用户与 AI 的实时对话功能。实现 AI 推理过程的流式输出，支持多专家协作过程的实时展示。开发会话管理功能，支持会话创建、切换、历史记录查询等。将问诊聊天记录存储到 MongoDB，记录完整的多专家协作过程和推理链路。

(8) 系统集成与测试

将前端界面、后端服务、AI 智能体和 RAG 检索系统集成，确保数据交互的准确性和稳定性。测试各功能模块的正确性和完整性，验证多专家 Agent 协作的有效性和 RAG 检索的准确性。优化健康数据统计算法和图表展示性能。确保用户数据的安全性和隐私保护，实现基于 JWT 的权限控制。

3.2 核心功能模块



功能模块图

四、研究方法及技术途径

研究方法和技术路线涉及前后端技术、AI 技术、数据库技术等多方面的综合应用，主要包括：

(1) 前端框架：系统采用 Vue 3 作为前端框架，结合 Element Plus UI 组件库构建

现代化的用户界面。使用 Vite 作为构建工具，提供快速的开发体验。采用 Pinia 进行状态管理，实现组件间数据共享。集成 ECharts 实现健康数据的可视化展示。通过 WebSocket 实现与 AI 服务的实时通信，展示多专家协作过程。

(2) 后端框架：后端采用 Spring Boot 3 框架，提供 RESTful API 服务。使用 Spring Data MongoDB 管理健康数据的时序存储，使用 Spring Data JPA 管理用户信息和权限数据。集成 JWT 实现用户认证和授权，基于角色的访问控制确保系统安全。支持文件上传到阿里云 OSS 存储。

(3) AI 服务层：基于 FastAPI 框架构建 AI 服务，使用 Uvicorn 作为 ASGI 服务器，实现高性能异步处理。集成 DeepSeek 大模型提供智能对话能力。实现 Plan 规划器进行任务分解和专家调度，数据专家采用 ReAct 循环推理模式检索与分析健康数据，诊断专家负责医疗风险判断，知识专家提供医学知识支撑，数据记录专家完成结构化数据登记。通过 MCP 协议实现标准化的工具调用，包括数据库操作、RAG 检索等。

(4) RAG 检索增强：使用 BGE-Large-zh-v1.5 Embedding 模型将医疗文档转换为向量表示。部署 ChromaDB 向量数据库实现高效的语义检索。开发 FastAPI 检索服务提供知识检索 API 接口。在 AI 推理过程中动态检索相关医疗知识，提升回答的专业性和准确性。

(5) 数据库技术：系统采用 MySQL+MongoDB 混合数据库架构。MySQL 存储结构化数据，包括用户信息、医生信息、权限数据等，保证数据一致性。MongoDB 存储半结构化数据，包括健康记录（血糖、血压、体重）、AI 问诊会话记录等时序数据，支持高效的时间范围查询和聚合分析。这种混合架构充分发挥了不同数据库的优势。

(6) 实时通信：系统集成 WebSocket 实现用户与 AI 的实时对话功能。前端使用 STOMP 协议进行消息订阅和发布，后端支持 WebSocket 连接管理。实现 AI 推理过程的流式输出，支持多专家协作过程的实时展示。开发会话管理功能，支持会话创建、切换、历史记录查询等。

(7) MCP 协议：系统引入 Model Context Protocol (MCP)作为 AI Agent 与异构数据源之间的标准化接口层。通过 MCP 协议实现对 MySQL、MongoDB 等多种数据源，以及 RAG 检索系统的统一调用规范。使用 MCP Inspector 工具进行协议调试和测试，确保工具调用的正确性和稳定性。

(8) 开发工具：集成开发环境包括 IntelliJ IDEA、PyCharm、VS Code。使用 Git 进行版本控制，托管平台 GitHub。数据库管理工具包括 Navicat、Studio 3T。使用 Apifox 进行接口测试。

五、实施计划：

上学期：

第 19 周：课题的选择

第 20 周：查阅相关资料，收集有关文献

下学期：

第 1-2 周：课题选题与文献调研，需求分析，学习 RAG 和多专家 Agent 技术

第 3-4 周：系统总体架构设计，多专家 Agent 协作架构设计，技术选型，数据库设计

第 5-6 周：后端基础功能开发（用户认证、权限管理、健康数据管理）

第 7-8 周：RAG 检索增强系统开发（医疗知识库构建、文档向量化、语义检索）

第 9-10 周：多专家 Agent 开发（Plan-Execute 规划器、ReAct 推理、MCP 工具调用层）

第 11-12 周：前端三端界面开发（Vue.js 实现，展示专家协作过程和推理链路）

第 13-14 周：问诊服务开发、WebSocket 实时通信、系统集成测试

第 15-16 周：系统功能测试与性能优化，验证创新点有效性，论文撰写

第 17 周：答辩准备与毕业设计答辩

六、参考文献

[1]中华医学会糖尿病学分会. 中国 2 型糖尿病防治指南(2020 年版)[J]. 中华内分泌代谢杂志,2021,37(04):311-398.

[2]Lewis P., Perez E., Piktus A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks[C]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[3]Yao S., Zhao J., Yu D., et al. ReAct: Synergizing Reasoning and Acting in Language Models[C]. International Conference on Learning Representations, 2023.

[4]刘知远,孙茂松,林衍凯,等. 知识表示学习研究进展[J]. 计算机研究与发展,2016,53(02):247-261.

[5]国家卫生健康委疾病预防控制局. 中国居民营养与慢性病状况报告(2020 年)[J]. 营养学报,2020,42(06):521.

[6]王淇,吴浩,魏学娟,等. 方庄社区卫生服务中心应用移动终端 APP 进行糖尿病管理的效果评价研究[J]. 中国全科医学,2020,23(07):844-848.

[7]Guan Z.Y.,Li H.T.,Liu R.H.,et al. Artificial intelligence in diabetes management:Advancements,opportunities,and challenges[J]. Cell Rep Med,2023,4(10):101213.

[8]Xiao M., Liu Z., Fu Y., et al. FlagEmbedding: A Toolkit for Embedding-Based

Retrieval[EB/OL]. https://github.com/FlagOpen/FlagEmbedding ,2025-02-07/2025-11-29.

[9]金春林,何达. 人工智能在医疗健康领域的应用及挑战[J]. 卫生经济研究,2018,(11):3-6.

[10]姚裕忠,马晓骏,宋懽,等. 基于“全专精准管理”的糖尿病“1358 模式”对社区糖尿病患者的管理效果研究[J]. 中国全科医学, 2023,26(34):4308-4314.

[11]侯梦薇,卫荣,陆亮,等. 知识图谱研究综述及其在医疗领域的应用[J]. 计算机研究与发展,2018,55(12):2587-2599.

[12]张智琪,杨四涛,刘冬瑞,等. 基于大语言模型结合 RAG 技术的慢性肾脏病人工智能体应用模型构建[J]. 中国现代应用药学,2025,42(17):2936-2942.

[13]Tiangolo S. FastAPI Framework[EB/OL]. https://fastapi.tiangolo.com,2025-07-11/2025-11-29.

[14]The Chroma Team. Chroma: The AI-native open-source embedding database[EB/OL]. https://www.trychroma.com,2024/2025-11-29.

[15]周志华. 机器学习[M]. 北京:清华大学出版社,2016.

指导教师审核意见：

指导教师签名：

2025 年 月 日