湖北文理学院
HUBEI UNIVERSITY OF ARTS AND SCIENCE

基于RAG检索增强和多专家Agent的糖尿病智能管理系统

汇报人： 张聪颖　　导 师：吴钊教授

01 绪论

02 系统总体方案设计

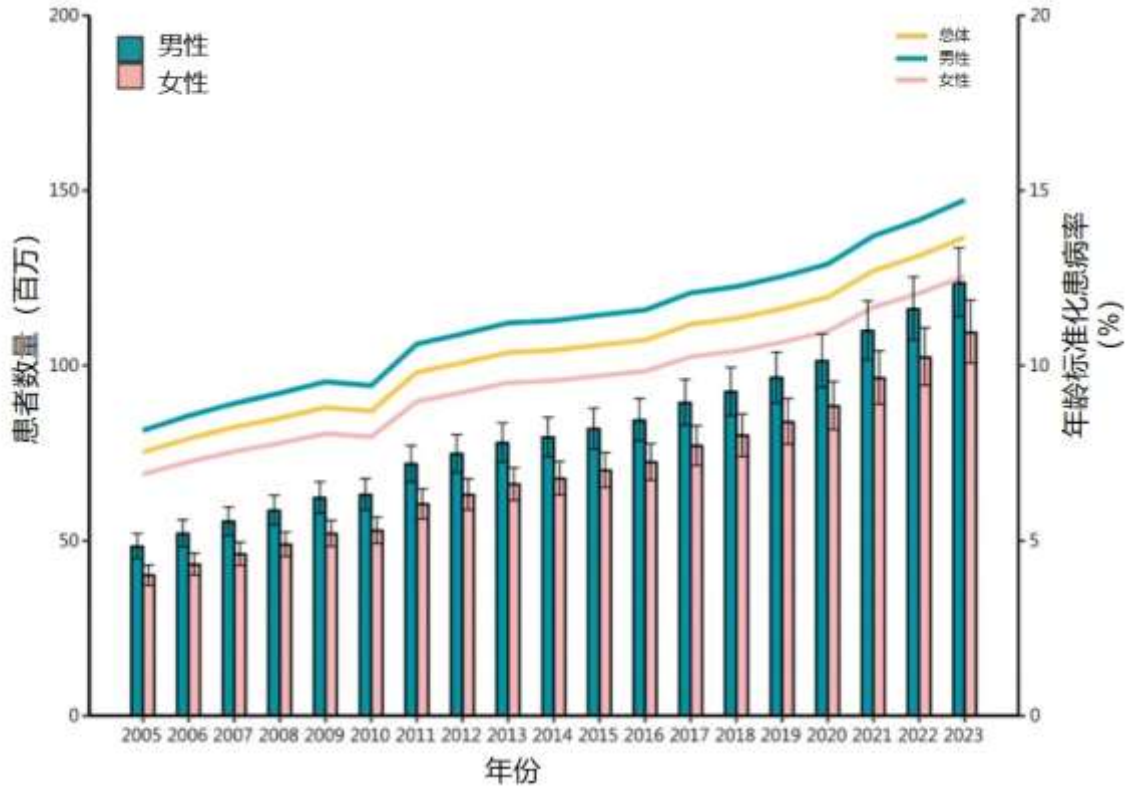
03 RAG引擎设计

04 MCP服务设计

05 智能体设计

06 总结与展望

背景及意义



现存问题

近年来我国糖尿病患病率呈持续**上升态势**，成人糖尿病患病比例不断攀升，且患者群体逐渐呈现**年轻化趋势**，已成为**威胁国民健康的主要慢性疾病之一，慢病管理的整体压力日益凸显。**

需求趋势

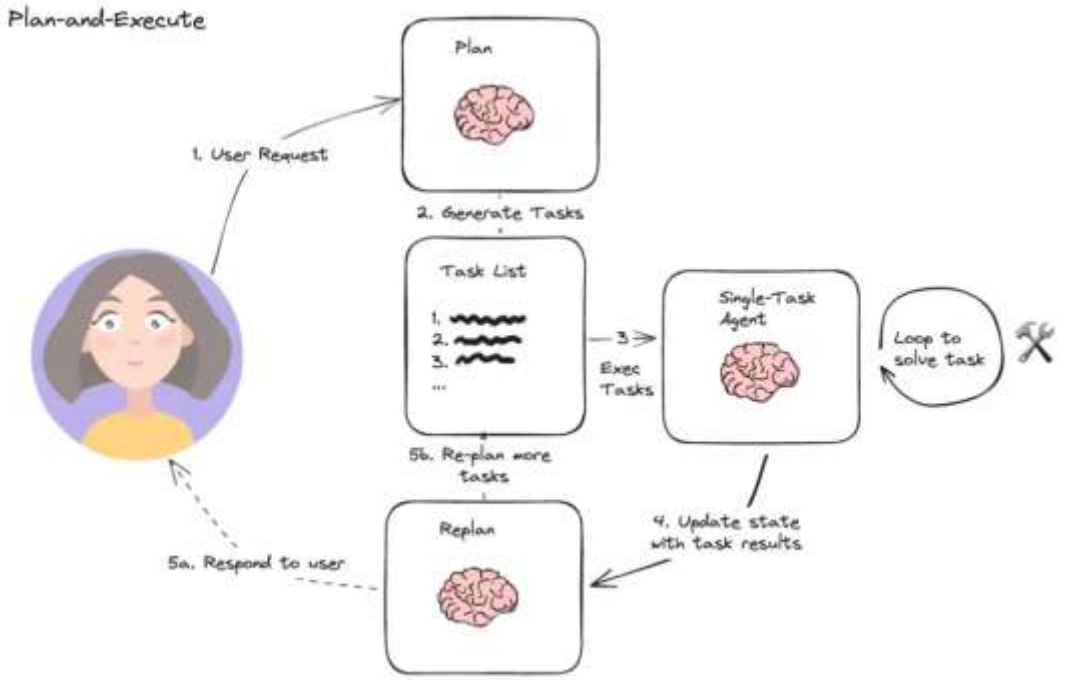
糖尿病智能管理向**智能化**发展的趋势，通过**多专家 Agent** 与 **RAG 技术**的创新融合，能够更好地适应糖尿病全周期精准管理的需求。

智能体研究现状

■ 核心范式

Plan-and-Execute 范式

- Plan-and-Execute（简称为P&E）是一种**面向复杂任务的AI任务处理范式**，其核心思想是将“完成复杂目标”这一整体任务，拆解为两个相互衔接的核心阶段——规划阶段（Planning）与执行阶段（Execution），并通过反馈机制实现两阶段的协同，确保任务在复杂环境或不确定条件下仍能高效推进。



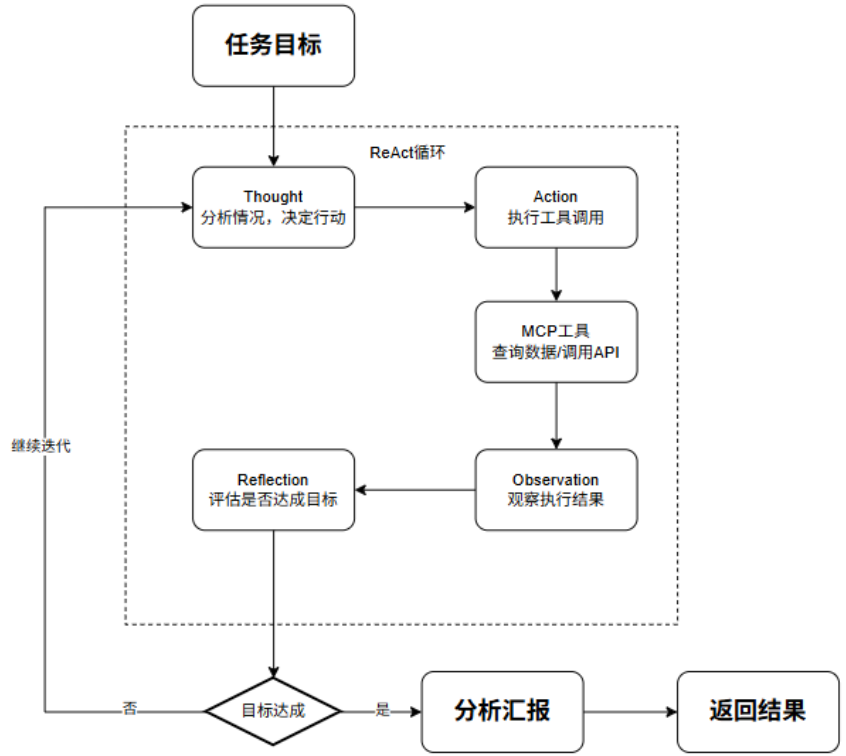
P&E 流程图

智能体研究现状

■ 核心范式

ReAct 范式

- ReAct 是一种智能体范式，它通过**结合推理和行动来增强大型语言模型在解决复杂任务时的能力。**ReAct框架强调在执行任务时，智能体不仅要理解上下文并进行逻辑推理，还要制定行动计划并执行行动，同时收集反馈以迭代优化自身行为。这种思考（Thought）→行动（Action）→观察（Observation）的循环使得Agent能够在复杂和动态的环境中更有效地工作。

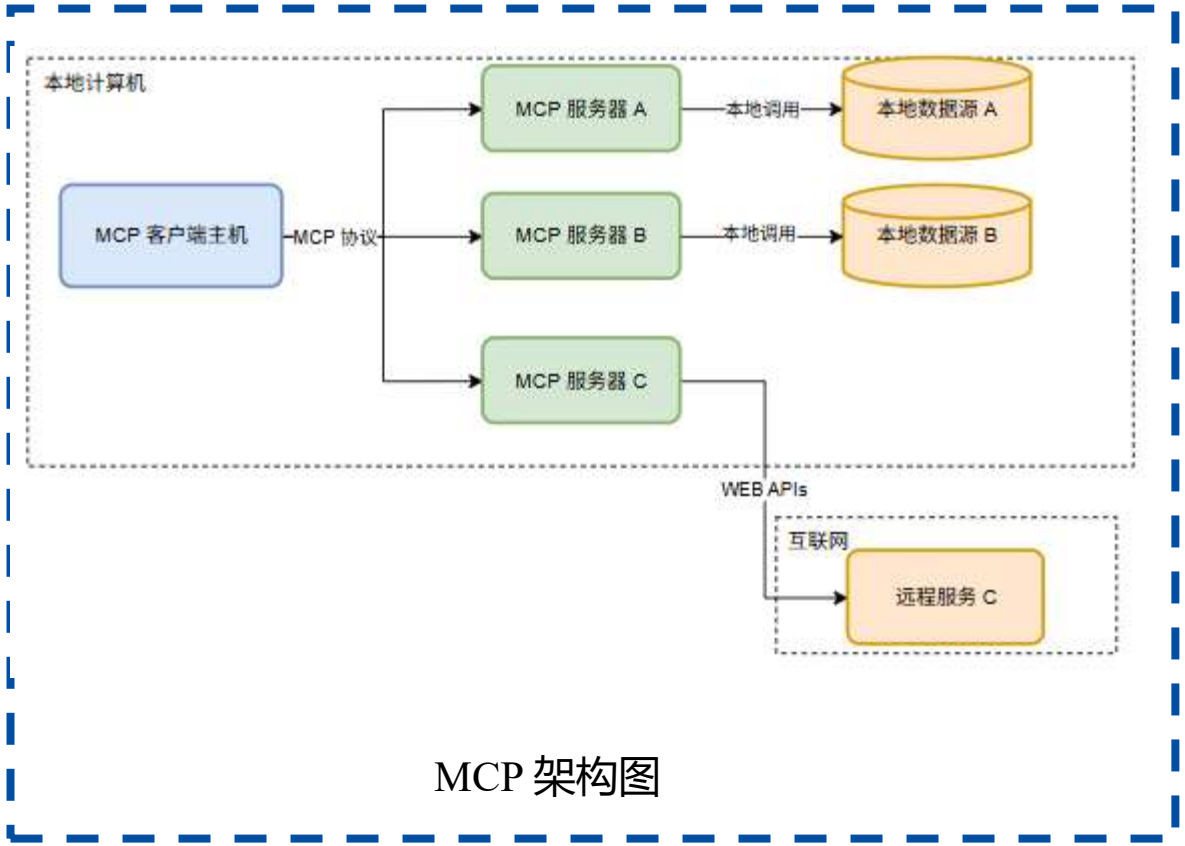


ReAct 流程图

智能体研究现状

■ MCP技术

- MCP 是 “模型上下文协议” （Model Context Protocol） 的缩写。它是一种开放标准协议，旨在**让大型语言模型（LLM）与外部工具和数据源无缝通信。**MCP 提供了一种标准化的方法，使 AI 模型能够高效、安全地访问和利用各种资源，类似于给 AI 模型装上了一个“万能接口”。
- MCP 的核心遵循客户端-服务器架构，其中主机应用程序可以连接到多个服务器：

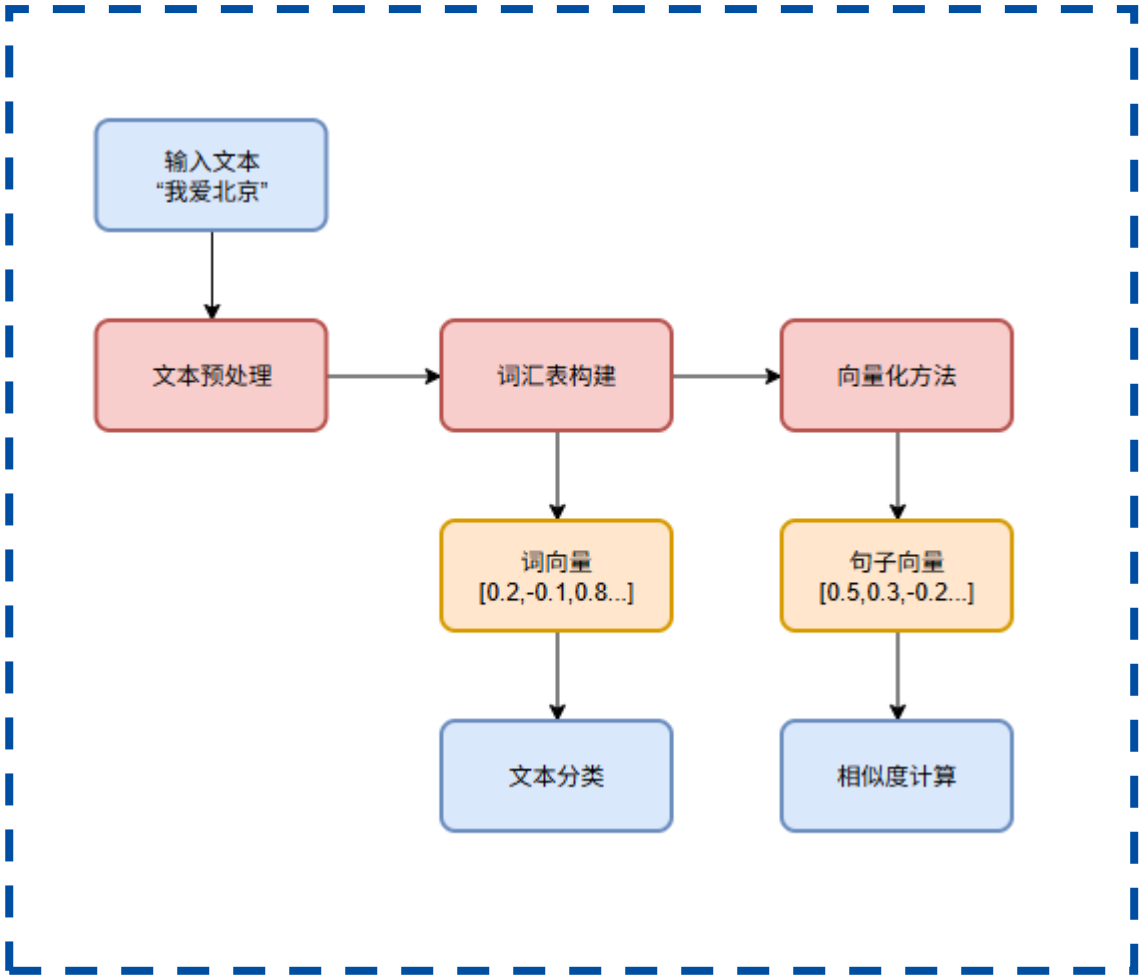


MCP 架构图

智能体研究现状

■ 文本向量化

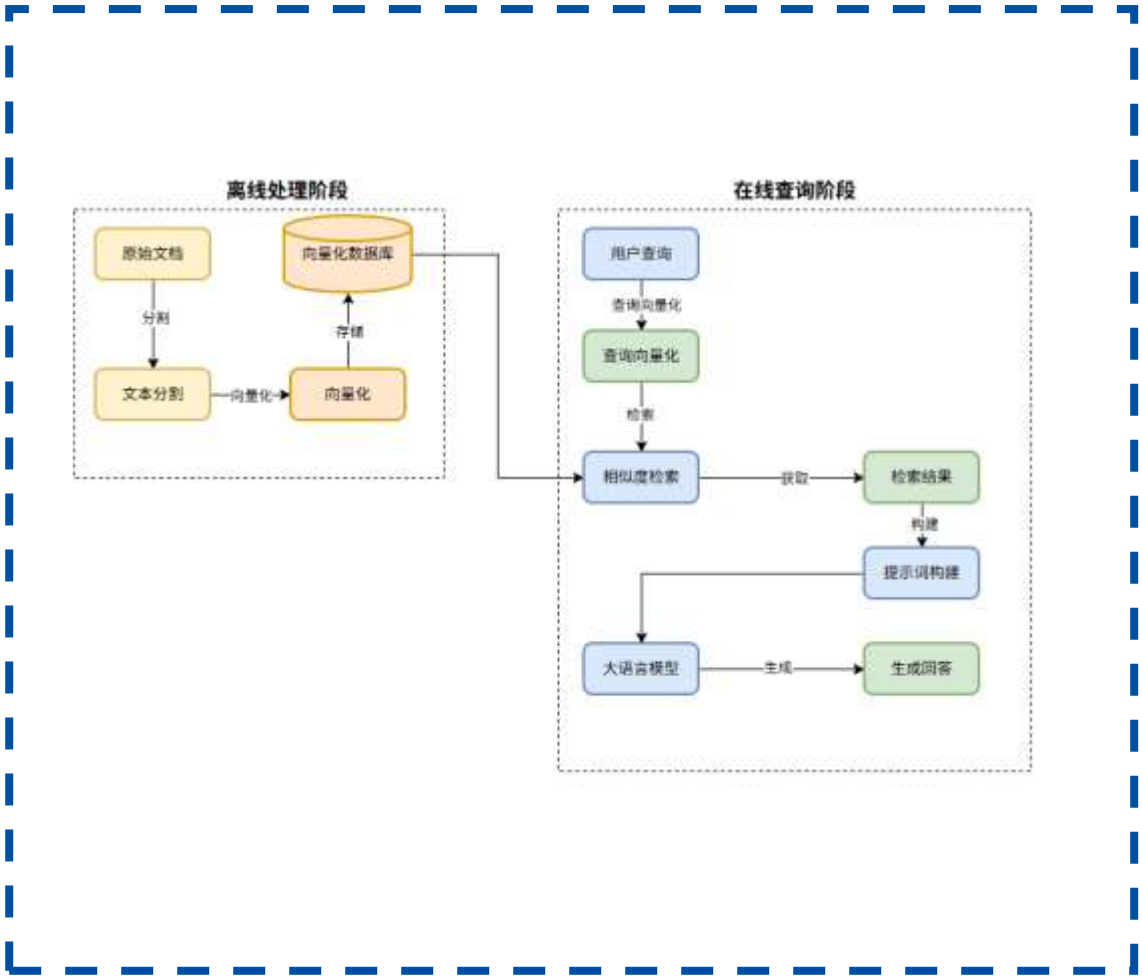
- 文本向量化是将自然语言文本转换为数值向量表示的核心技术，在本系统中采用预训练的向量化模型进行实现。系统通过调用专门的向量化模型（如BERT、Sentence-BERT、BGE等）将用户输入的文本、文档片段转换为高维稠密向量。**这些向量能够捕捉文本的语义信息，支持后续的相似度计算和检索任务。**相比传统的TF-IDF等稀疏表示方法，向量化模型生成的稠密向量具有更强的语义理解能力，能够识别语义相似但词汇不同的文本，为RAG系统的检索增强提供了重要基础。



智能体研究现状

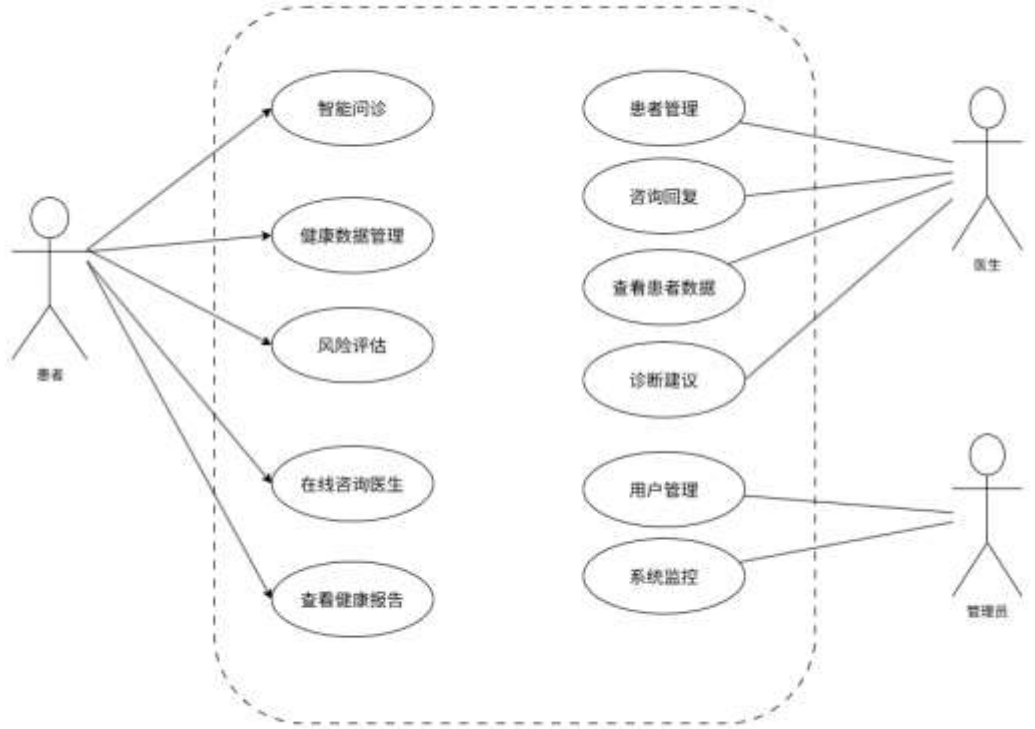
■ RAG技术

- RAG 是一种将**文本检索模块与文本生成模块结合的框架**，旨在提升知识密集型任务中生成回复的质量。从形式化定义来看，RAG模型通过检索器为序列到序列（seq2seq）生成器提供外部文本语料的访问能力。



系统功能分析

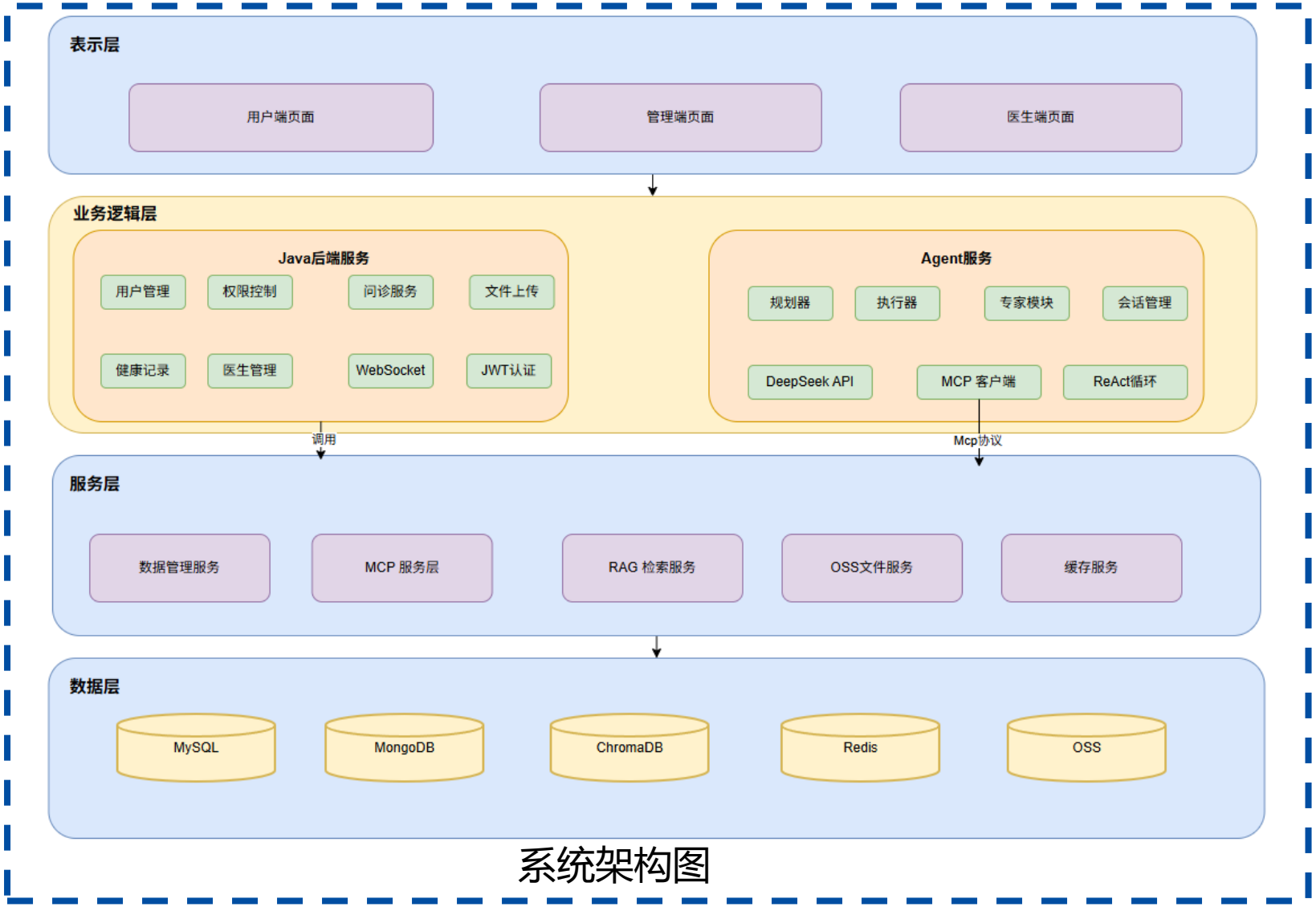
- 患者作为**系统的核心服务对象**，可以**通过自然语言与AI智能体进行交互式问诊**，系统会自动记录血糖、血压等健康数据，并由AI智能体基于历史数据进行糖尿病风险评估和预警。当患者需要专业医疗意见时，可发起在线咨询，与医生进行实时沟通。
- 医生端侧重于患者服务和数据查阅，医生可以查看患者的完整健康档案和监测数据，在线回复患者咨询并给出专业诊断建议，实现远程医疗服务。
- 管理员负责整个平台的运维工作，包括用户账号管理、医生资质审核以及系统运行状态的监控，确保平台稳定运行。



系统用例图

系统总体架构

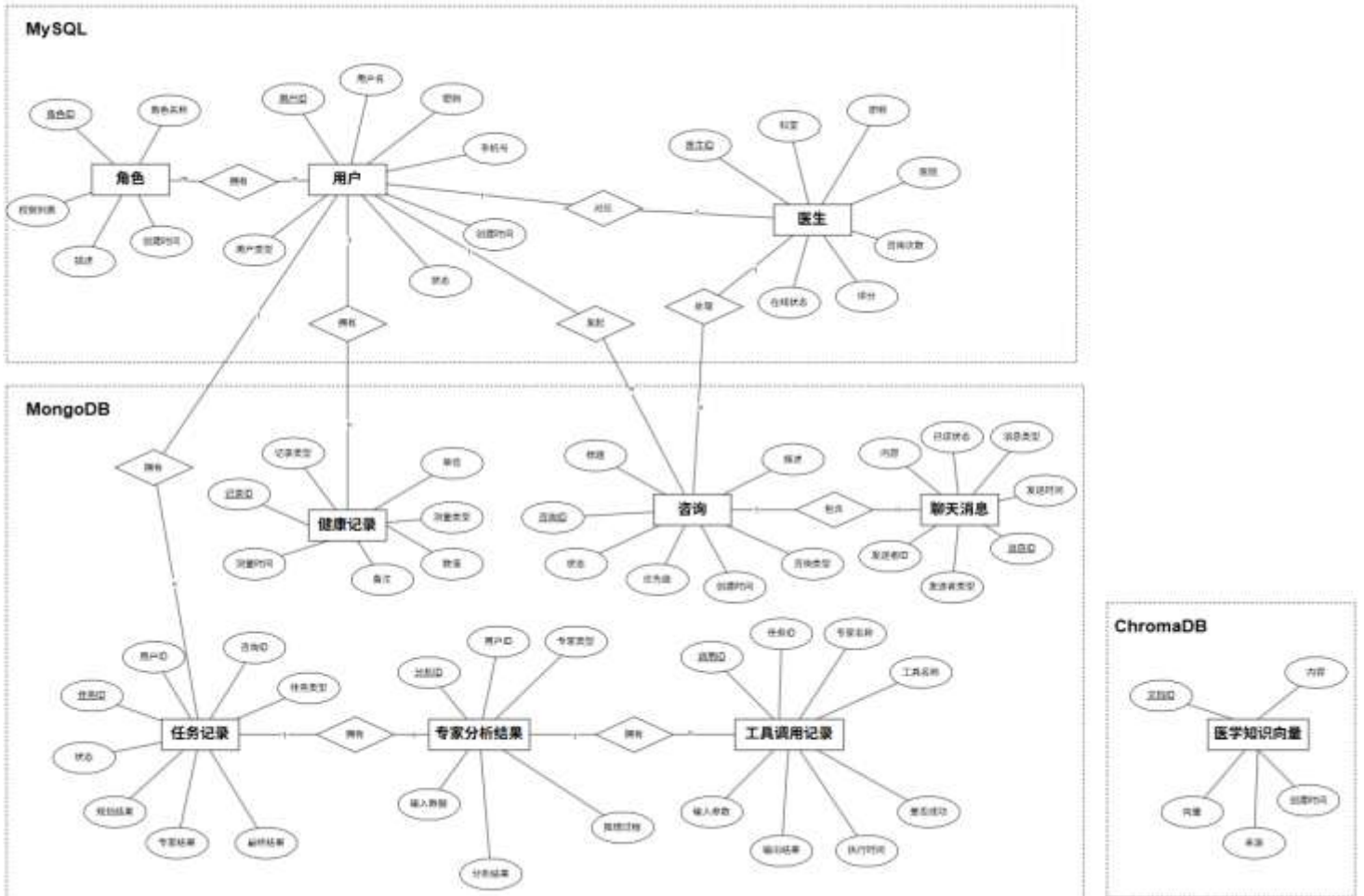
- 本系统为实现糖尿病智能服务管理系统的完整功能，我们设计了如图所示的**四层架构**：表示层负责用户交互，业务逻辑层包含Java后端和Python Agent两大服务，服务层提供文件存储、缓存、MCP工具和RAG检索等支撑服务，数据层采用MySQL、MongoDB、ChromaDB、Redis多数据库协同存储。



系统架构图

核心技术架构 -数据库

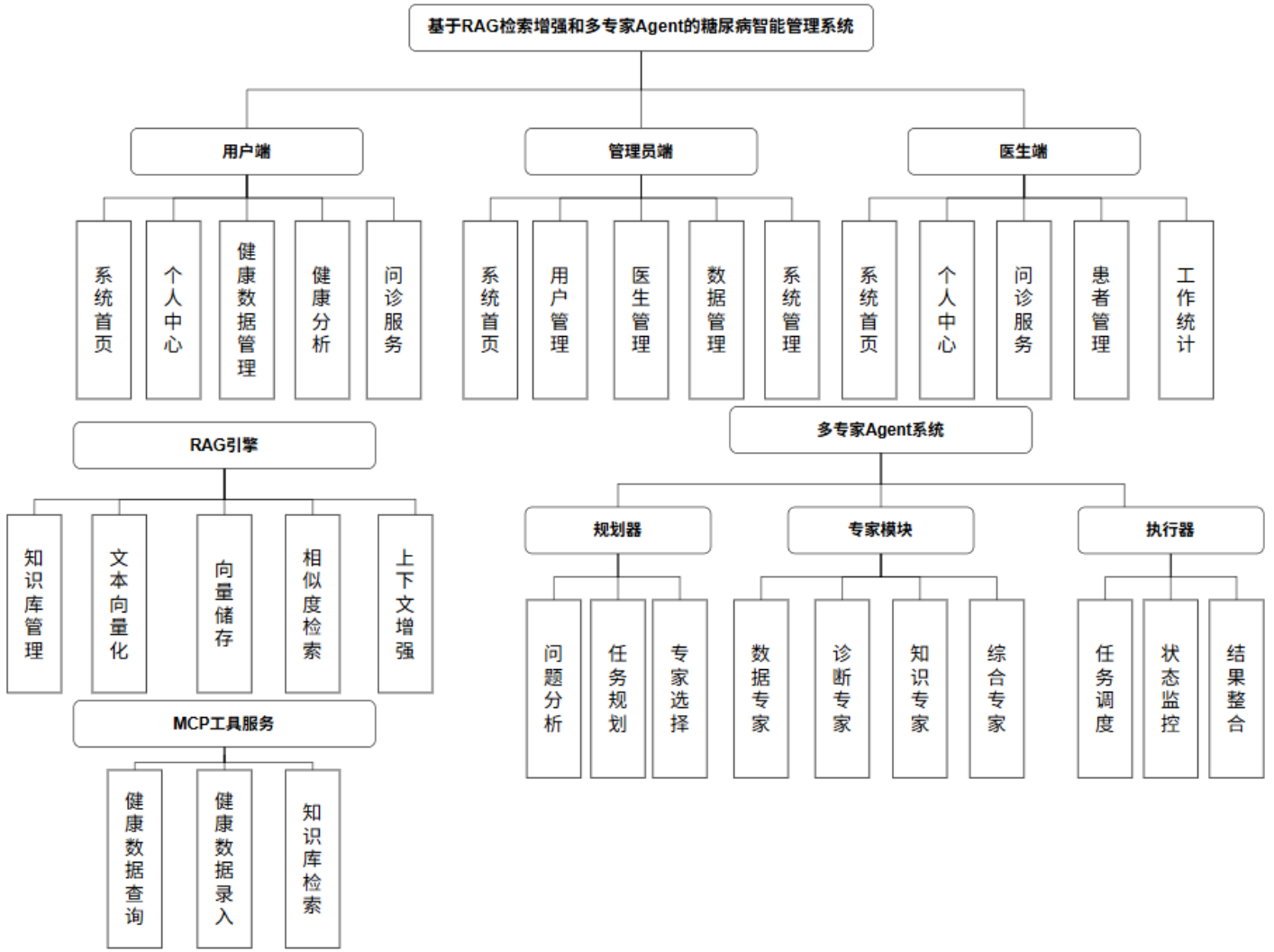
- 系统采用**MySQL+MongoDB+ChromaDB+Redis多数据库混合架构**，MySQL负责存储用户信息、医生资料等结构化业务数据，MongoDB用于存储Agent任务日志、MCP调用记录、咨询记录、健康数据录等半结构化数据，向量数据库支持医学知识的语义检索，Redis提供高性能缓存服务



系统多数据库数据模型架构示意图

核心技术架构 - 数据库

- 为实现糖尿病智能服务管理系统，我们设计了如图所示的功能模块：**用户端、管理端、医生端三个前端界面，RAG引擎提供知识检索能力，多专家Agent系统实现智能问诊，MCP工具服务支撑Agent数据交互。**各模块协同配合，共同完成健康管理、智能问诊、在线咨询等核心功能。



系统功能模块图

RAG引擎模块

■ 文本向量化

- 文本本系统选择预训练向量化模型，因其能够识别语义相似但词汇不同的医学文本，满足RAG检索的精准性要求。

方法	优点	缺点
TF-IDF	简单高效，计算速度快	无法理解语义，仅基于词频统计
Word2Vec	能捕捉词级语义关系	无法处理一词多义，缺乏上下文理解
Doc2Vec	支持文档级向量表示	训练不稳定，效果依赖语料质量
预训练向量化模型	深度语义理解，支持上下文	计算资源需求较高

RAG引擎模块

■ 文本向量化

- 本系统选择预训练向量化模型BGE-Large-ZH-V1.5
- BGE-Large-ZH-V1.5 是由智源研究院（BAAI）发布的中文文本嵌入模型，在中文大规模文本嵌入基准（C-MTEB）中排名第一。

特点	具体说明
中文优化	专为中文语义理解设计，在 C-MTEB 基准排名第一，能准确捕捉医学知识语义
部署成本低	模型体积适中，可在普通 GPU 甚至 CPU 上运行，无需高昂硬件投入
检索性能优异	采用指令微调和硬负样本挖掘技术，提升知识检索准确性
开源免费	可本地部署，保障医疗数据隐私安全

RAG引擎模块

■ 向量化储存

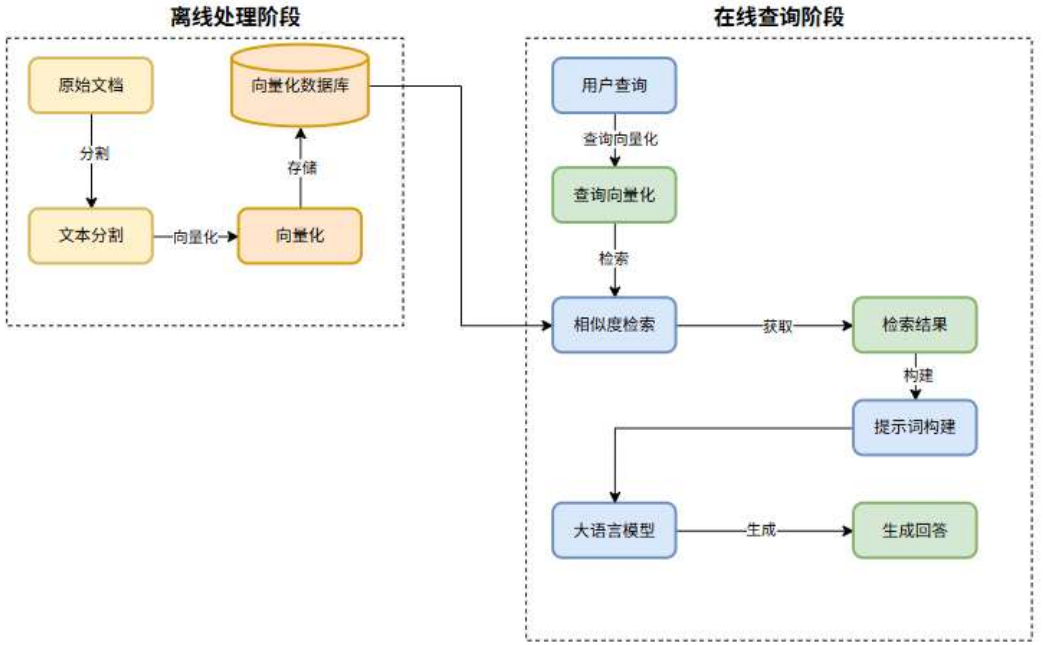
- 文本我们对比了市面上常见的向量数据库方案。Pinecone等云服务存在数据隐私风险且需付费；Milvus性能强但部署复杂、资源占用高；Faiss检索快但不支持持久化。综合考虑部署成本、数据安全和实际需求，我们选择了**ChromaDB**作为向量存储方案。

数据库	优点	缺点
Pinecone	全托管服务，扩展性强	需付费，数据存储在云端，存在隐私风险
Milvus	性能优异，支持大规模数据	部署复杂，资源占用较高
Faiss	检索速度快，Facebook 开源	仅支持内存存储，无持久化功能
ChromaDB	**轻量易用，支持持久化**	面对大规模数据时，性能表现一般

RAG引擎模块

■ 处理流程

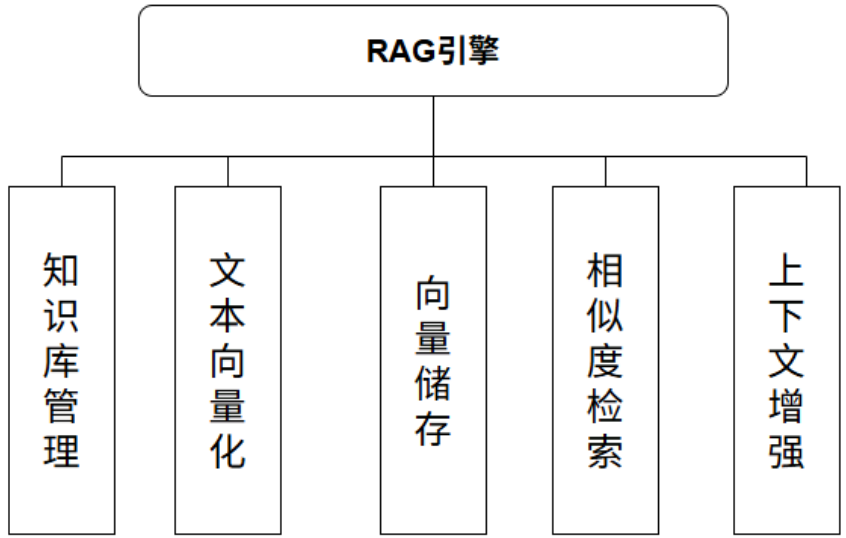
- RAG处理流程分为**离线处理阶段**和**在线查询阶段**两部分。
- 离线处理阶段：将糖尿病知识文档进行文本分割，切分为适当大小的知识片段；通过向量化模型将每个片段转换为高维向量；最后将向量存入向量数据库，构建知识索引。
- 在线查询阶段：用户提问后，首先将问题向量化；在向量数据库中进行相似度检索，召回最相关的知识片段；将检索结果与用户问题一起构建提示词，输入大模型生成最终回答。



RAG 处理流程

RAG引擎模块

- RAG为实现上述RAG处理流程，系统设计了以下功能模块：
- 知识库管理：负责糖尿病医学知识的导入、分割和维护，支持知识的增删改查。
- 文本向量化：调用BGE-Large-ZH-V1.5模型，将文本转换为高维语义向量。
- 向量储存：基于ChromaDB向量数据库，实现向量的持久化存储和索引管理。
- 相似度检索：根据用户问题向量，快速检索语义最相关的知识片段。
- 上下文增强：将检索结果注入提示词，增强大模型回答的专业性和准确性。

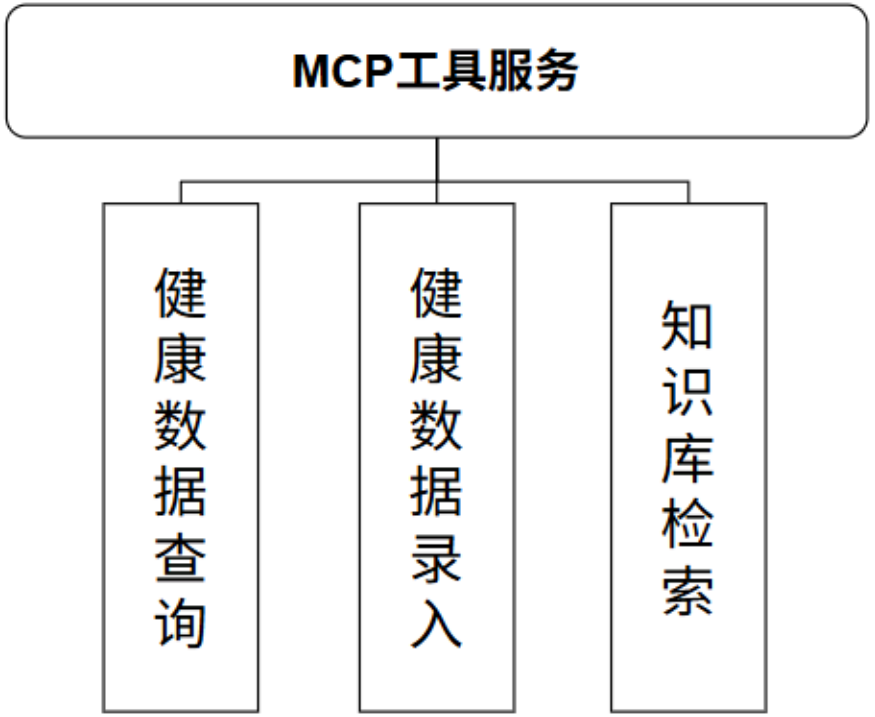


RAG引擎功能模块图

功能模块

■ MCP技术

- 为实现智能体与系统数据的交互，我们基于MCP协议设计了如图所示的工具服务模块。
- 健康数据查询：支持查询用户的血糖、血压、体重等历史健康记录，为智能体提供分析数据源。
- 健康数据录入：支持通过自然语言解析并记录用户的健康数据，实现对话式数据采集。
- 知识库检索：对接RAG检索服务，从糖尿病知识库中检索相关医学知识，为知识专家提供专业信息支撑。



MCP 功能模块图

简述

- 系统在设计智能问诊系统时，我们分析了两种主流Agent架构的特点，并针对各自的优缺点进行了融合设计。

基于ReAct实现的Agent架构

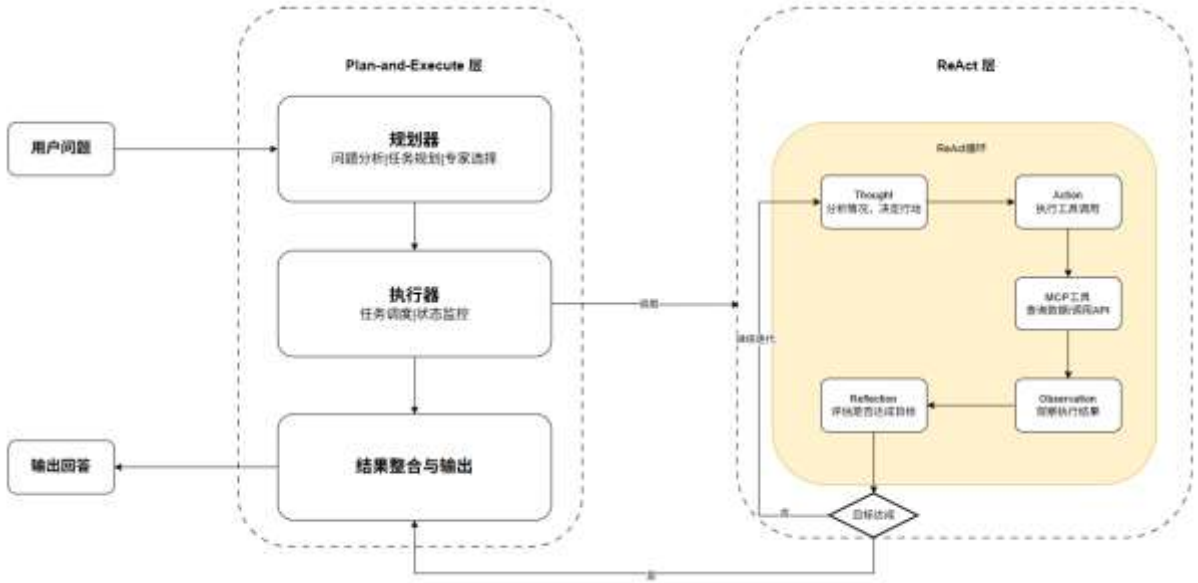
- 优点：具备强大的动态推理能力，能够根据每一步的执行结果灵活调整后续行动；通过“思考-行动-观察-反思”的迭代过程，来处理需要多步探索、信息不完整的复杂任务；能够自主判断何时停止，避免过度执行。
- 缺点：缺乏全局视角，容易陷入局部最优；对于需要多模块协作的任务，单一ReAct循环难以有效协调；每次迭代都需要LLM推理，在简单任务上效率较低。

基于P&E实现的Agent架构

- 优点：具备全局规划能力，能够将复杂问题分解为结构化的子任务序列；支持多专家并行协作，适合处理涉及多个领域的综合性问题；任务分配明确，执行效率高。
- 缺点：计划一旦制定，灵活性不足；对于需要动态探索、结果不可预知的任务，预先规划可能不够准确；难以应对执行过程中的意外情况。

简述

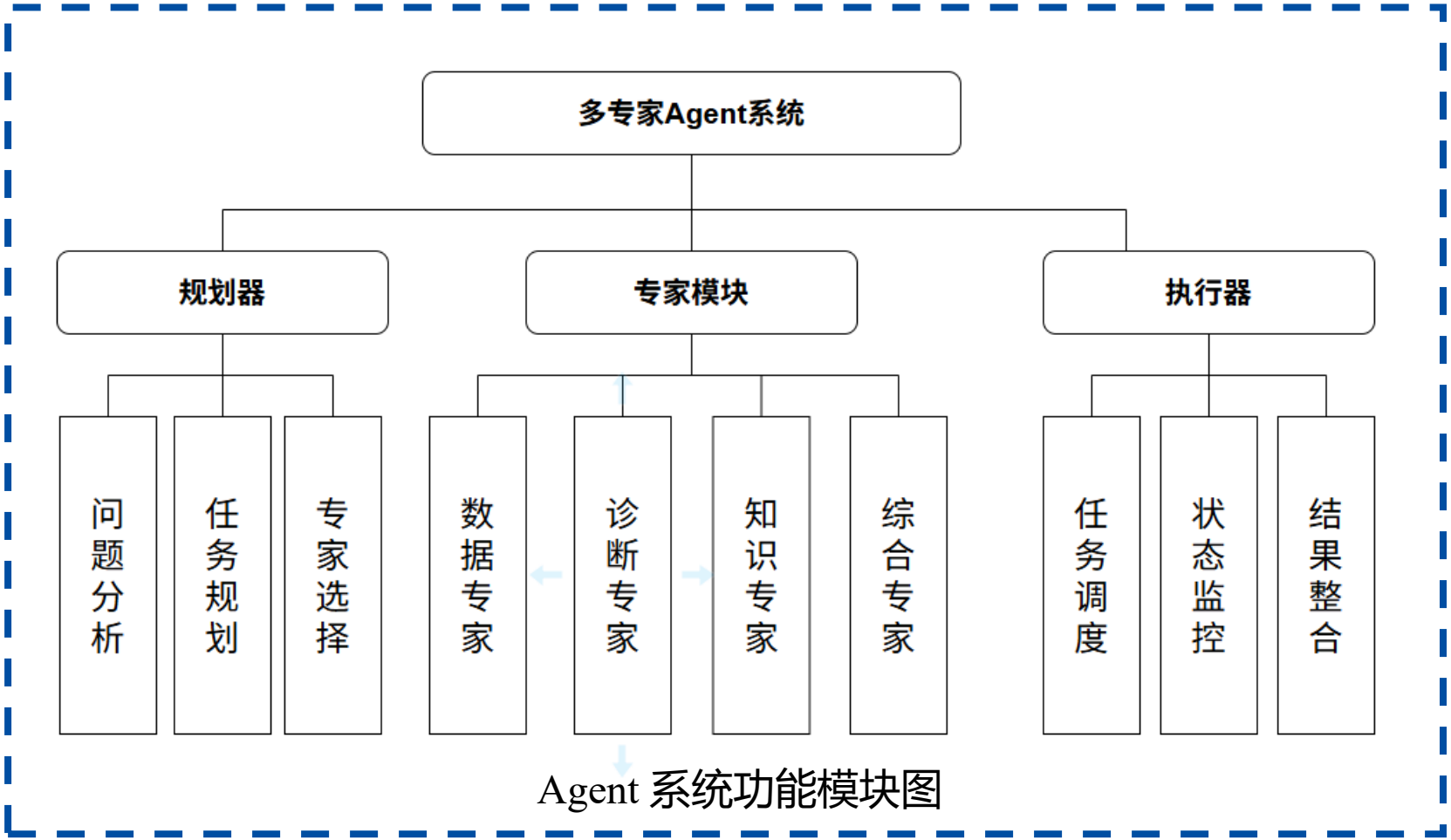
- 针对上述分析，本系统采用**Plan-and-Execute与ReAct相结合的混合架构**。宏观层面使用Plan-and-Execute模式，由规划器进行全局任务分解和专家分配，执行器协调诊断专家、知识专家、综合专家等多模块协作；微观层面在数据专家内部采用ReAct模式，通过迭代循环实现健康数据的多步查询和深度分析。这种设计既保证了复杂任务的全局协调能力，又赋予了数据分析环节的动态推理灵活性，实现了两种模式的优势互补。



结合框架流程图

功能模块

- 为实现上述混合架构的设计构想，我们设计了如图所示的功能模块。



Agent 系统功能模块图

总结与展望

总结

- 基于糖尿病健康管理需求，本文设计了智能服务管理系统，采用Vue+Spring Boot+Python的前后端分离架构。
- 智能问诊方面，采用Plan-and-Execute与ReAct混合架构，实现多专家协同的智能问诊服务，通过MCP协议实现Agent与数据库的标准化交互。
- RAG引擎方面，选用BGE-Large-ZH-V1.5向量化模型和ChromaDB向量数据库，构建糖尿病知识检索系统，为智能问诊提供专业知识支撑。
- 系统实现了用户健康数据管理、AI风险评估、在线医生问诊等核心功能，完成了患者、医生、管理员三端的完整业务流程。

展望

- ◆ 可引入更多专家模块，如用药专家、饮食专家等，提供更全面的健康管理建议。
- ◆ 可扩展知识库规模，接入更多权威医学数据源，提升RAG检索的覆盖面和准确性。
- ◆ 可结合可穿戴设备，实现血糖、血压等数据的自动采集，减少用户手动录入负担。
- ◆ 可引入多模态能力，支持医学影像分析，拓展系统的诊断辅助功能。

湖北文理学院
HUBEI UNIVERSITY OF ARTS AND SCIENCE

感谢各位的聆听与指导

THANK YOU FOR LISTENING AND GUIDING

汇报人：张聪颖　　导 师：吴钊教授