医疗问答系统项目报告

一、项目概述

本项目构建了一个基于 RAG 架构的医疗领域智能问答系统。系统以 792,099 条中文医疗 QA 数据为基础，通过 m3e-base 模型构建向量知识库，结合 Qwen2.5-1.5B 模型实现检索增强生成。该系统能够理解并回答多种医疗健康问题，在提供专业解答的同时显示参考来源，并包含医疗免责声明，旨在为用户提供可靠、可追溯的医疗信息辅助服务。

代码仓库：https://github.com/2758395517/01-NLP/tree/main

二、数据来源与处理

1. 数据来源：中文医疗对话数据集，数据集链接：https://github.com/Toyhom/Chinese-medical-dialogue-data
2. 数据量：792,099 条中文医疗 QA 数据
3. 处理流程：
   ① 编码检测与转换，通过自动检测机制识别并处理多种中文编码格式，包括 gbk、utf-8 等常见编码，确保所有文件可正确读取。
   ② 列名标准化，自动识别原始列名并映射为统一的 question、answer、department 字段，对于识别失败的采用启发式规则处理。
   ③ 随后进行数据清洗，包括删除空值行、去除文本空白字符、过滤短答案、基于问题去重等操作，确保数据质量。
   ④ 分块存储，将清洗后的数据转换为包含完整元数据的 JSON 格式，支持多种分块策略以适应不同检索需求。

三、方法

1. **向量检索**：系统采用 moka-ai/m3e-base 中文文本表示模型生成 768 维向量表示，该模型专门针对中文语义相似度任务优化。使用 FAISS 库构建高效的向量索引，支持快速近似最近邻搜索。检索过程结合向量相似度分数和文本匹配分数，设置可调节的相关性阈值过滤低质量结果，确保检索到的文档与用户查询高度相关。
2. **RAG 架构**：系统基于检索增强生成框架，用户查询首先转换为向量并在 FAISS 索引中检索最相关的文档片段。采用重排序机制筛选高质量结果，将检索到的 top-3 文档作为上下文信息提供给语言模型。生成阶段结合检索上下文和用户问题产生最终回答，同时将检索来源信息以结构化格式展示给用户，增强答案的可解释性和可信度。
3. **LLM 集成**：采用 Qwen2.5-1.5B-Instruct 作为生成模型，该模型专为中文问答优化并支持 32K 长上下文。通过 transformers 库加载模型和分词器，设置合理的生成参数控制输出质量。系统实现多层回退机制，当大模型不可用时自动切换到基于规则的回复，确保系统鲁棒性和可用性。

4. **提示工程**：设计简洁专业的医疗问答提示模板，明确 AI 角色为专业医疗助手。提示词要求基于检索信息回答，避免编造内容，回答需简洁专业且包含免责声明。上下文信息以结构化格式嵌入提示词中，确保语言模型能够有效利用检索到的医疗知识生成准确回答。
5. **网页实现**：使用 Gradio 框架构建直观的 Web 交互界面，采用垂直布局设计，上方展示对话历史，中部提供输入和控制区域，下方显示检索结果。界面支持多轮对话管理、RAG 功能开关和对话清空操作，响应速度快且用户体验良好。系统通过单命令启动，默认运行于本地 7860 端口，便于演示和测试。

四、实验结果

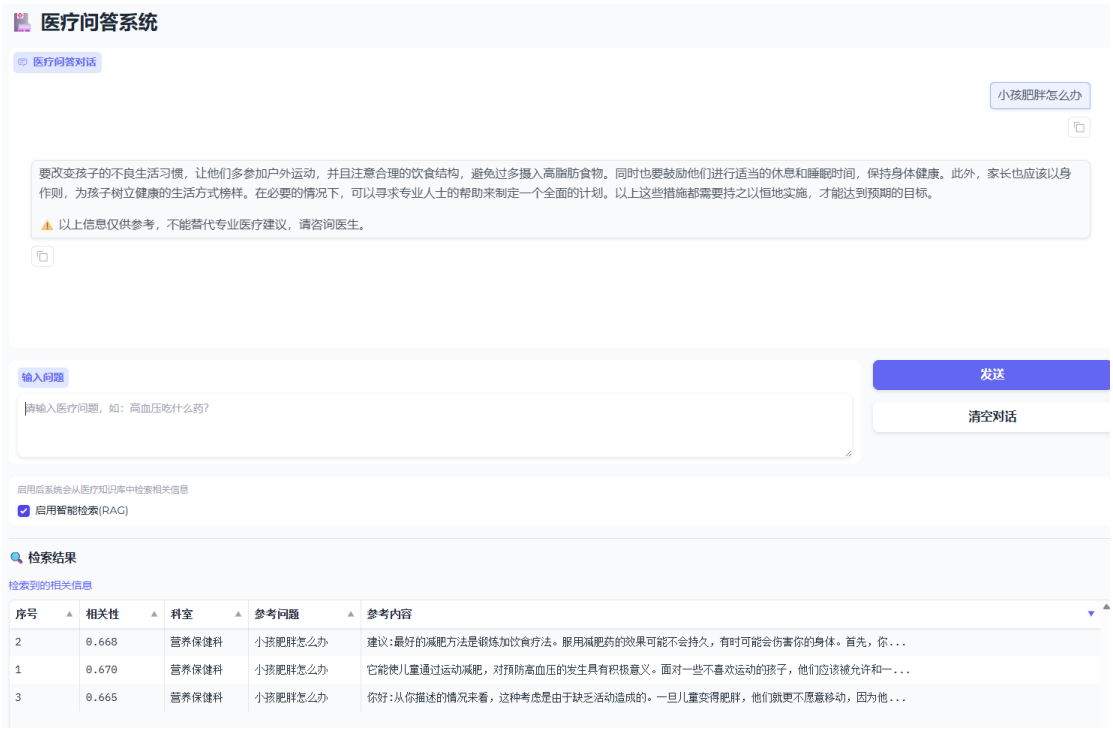
采用加权综合评分进行相关性评估：70%基于向量语义相似度（使用 m3e-base 模型计算余弦相似度），30%基于关键词匹配度。首先通过 FAISS 检索初步结果，然后计算每个结果的综合得分，过滤低于阈值的条目，最后按分数降序排列。

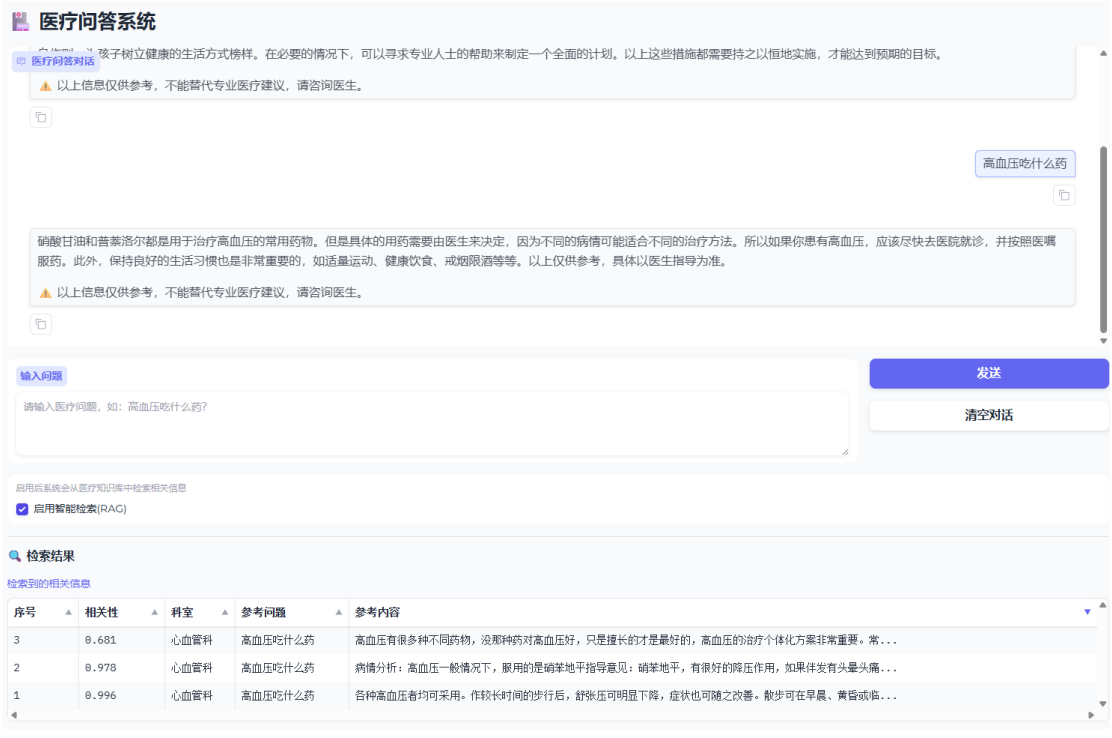
五、问题分析与创新点

遇到的问题：系统最初无法正确回答问题，检索结果只显示科室和相关性分数，但问题和答案内容为空，原因是数据清洗时字段映射不完整导致元数据缺失；LLM 模型下载受网络限制影响，通过配置国内镜像源解决；系统界面布局不够直观，我们调整了对话框和检索结果的显示位置以改善用户体验。

创新点：本项目的主要创新点在于构建了一个针对中文医疗领域的大规模 RAG 问答系统。系统处理了将近 80 万条医疗 QA 数据，并实现了完整的数据清洗、向量化与检索流水线；然后，针对中文医疗场景专门选用 m3e-base 中文 Embedding 模型与 Qwen2.5 中文优化 LLM，确保了语义理解与生成的专业性与准确性；最后，系统不仅完成了核心的检索增强生成功能，还部署了完整的 Gradio 交互界面，支持多轮对话、引用来源展示等实用特性，具备直接可运行的演示价值。

六、Demo 截图/链接





七、未来改进方向

在模型层面，可考虑引入领域适配微调以进一步提升回答的专业性和准确性；在数据层面，可扩充更多样化的医疗数据来源并优化数据质量；在系统功能上，可扩展症状追问、诊疗建议细化等实用特性；在评估与优化方面，需建立更全面的自动化评估体系，持续监控和提升回答质量与可信度；在工程部署上，可推动系统向容器化、服务化方向演进，以提升可维护性和扩展性。