

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。

```
【处理】 问题: 小孩肥胖怎么办

【检索】 查询: '小孩肥胖怎么办'
【检索】 找到 3 个结果
【结果1】 分数: 0.670, 问题: '小孩肥胖怎么办...', 答案: '它能使儿童通过运动减肥, 对预防高血压的发生具有积极意义。面对...'
【结果2】 分数: 0.668, 问题: '小孩肥胖怎么办...', 答案: '建议:最好的减肥方法是锻炼加饮食疗法。服用减肥药的效果可能不...'
【结果3】 分数: 0.665, 问题: '小孩肥胖怎么办...', 答案: '你好:从你描述的情况来看, 这种考虑是由于缺乏活动造成的。一旦...'
【过滤】 最终保留 3 个结果
【完成】 回答长度: 192 字符
【界面】 发送回答, 参考信息数里: 3

【处理】 问题: 高血压吃什么药

【检索】 查询: '高血压吃什么药'
【检索】 找到 3 个结果
【结果1】 分数: 0.996, 问题: '高血压吃什么药...', 答案: '各种高血压者均可采用。作较长时间的步行后, 舒张压可明显下降, ...'
【结果2】 分数: 0.978, 问题: '高血压吃什么药...', 答案: '病情分析: 高血压一般情况下, 服用的是硝苯地平指导意见: 硝苯地...'
【结果3】 分数: 0.681, 问题: '高血压吃什么药...', 答案: '高血压有很多种不同药物, 没那种药对高血压好, 只是擅长的才是最...'
【过滤】 最终保留 3 个结果
【完成】 回答长度: 174 字符
【界面】 发送回答, 参考信息数里: 3
```

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

医疗问答系统

医疗问答对话

小孩肥胖怎么办

要改变孩子的不良生活习惯，让他们多参加户外运动，并且注意合理的饮食结构，避免过多摄入高脂肪食物。同时也要鼓励他们进行适当的休息和睡眠时间，保持身体健康。此外，家长也应该以身作则，为孩子树立健康的生活方式榜样。在必要的情况下，可以寻求专业人士的帮助来制定一个全面的计划。以上这些措施都需要持之以恒地实施，才能达到预期的目标。

以上信息仅供参考，不能替代专业医疗建议，请咨询医生。

输入问题

请输入医疗问题，如：高血压吃什么药？

发送

清空对话

启用后系统会从医疗知识库中检索相关信息

启用智能检索(RAG)

检索结果

检索到的相关信息

序号	相关性	科室	参考问题	参考内容
2	0.668	营养保健科	小孩肥胖怎么办	建议:最好的减肥方法是锻炼加饮食疗法。服用减肥药的效果可能不会持久，有时可能会伤害你的身体。首先，你...
1	0.670	营养保健科	小孩肥胖怎么办	它能使儿童通过运动减肥，对预防高血压的发生具有积极意义。面对一些不喜欢运动的孩子，他们应该被允许和...
3	0.665	营养保健科	小孩肥胖怎么办	你好:从你描述的情况来看，这种考虑是由于缺乏活动造成的。一旦儿童变得肥胖，他们就更不愿意移动，因为他...

医疗问答系统

医疗问答对话

高血压吃什么药

硝酸甘油和普萘洛尔都是用于治疗高血压的常用药物。但是具体的用药需要由医生来决定，因为不同的病情可能适合不同的治疗方法。所以如果你患有高血压，应该尽快去医院就诊，并按照医嘱服药。此外，保持良好的生活习惯也是非常重要的，如适量运动、健康饮食、戒烟限酒等等。以上仅供参考，具体以医生指导为准。

以上信息仅供参考，不能替代专业医疗建议，请咨询医生。

输入问题

请输入医疗问题，如：高血压吃什么药？

发送

清空对话

启用后系统会从医疗知识库中检索相关信息

启用智能检索(RAG)

检索结果

检索到的相关信息

序号	相关性	科室	参考问题	参考内容
3	0.681	心血管科	高血压吃什么药	高血压有很多种不同药物，没那种药对高血压好，只是擅长的才是最好的，高血压的治疗个体化方案非常重要。常...
2	0.978	心血管科	高血压吃什么药	病情分析：高血压一般情况下，服用的是硝苯地平指导意见：硝苯地平，有很好的降压作用，如果伴有头晕头痛...
1	0.996	心血管科	高血压吃什么药	各种高血压者均可采用。作较长时间的步行后，舒张压可明显下降，症状也可随之改善。散步可在早晨、黄昏或临...

系统状态: 就绪 | 知识库: 610,742条数据 | 模型: Qwen2.5-1.5B

使用提示

1. 输入问题: 在下方输入框输入您的医疗问题

2. 发送方式: 点击"发送"按钮或按Enter键

3. 检索功能: 启用RAG可以从知识库中检索相关信息

4. 查看参考: 下方的表格显示检索到的相关信息

5. 清空对话: 点击"清空对话"按钮可以开始新的对话

6. 重要提醒: 所有回答仅供参考，请咨询专业医生

通过 API 使用 使用 Gradio 构建

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。