# MedQA 数据集整合与专项评测分析报告

核心议题: 本报告旨在总结将开源MedQA（选择题）数据集整合入RAG知识库的工作流程，展示其在专项评测任务上的性能表现，并分析该评测结果的意义。

## 1. 工作流程概述

为了扩充知识库的多样性并进行专项能力验证，执行了以下步骤将MedQA选择题数据集整合入现有系统：

1. **数据获取与分析**:
   * 获取了包含医学选择题的中文MedQA数据集（train.jsonl等文件），其格式为JSON Lines，每条包含问题（question）、选项（options）和正确答案字母（answer\_idx）及文本（answer）。
2. **数据格式化与整合**:
   * 开发并运行了 integrate\_new\_data.py 脚本（适配选择题格式 v2）。
   * 该脚本读取MedQA的 .jsonl 文件，提取其中的“问题”文本和“正确答案”文本。
   * 将提取出的“问题-答案”对转换为知识库所需的标准格式（问题作为name，答案作为sections中的“正确答案”），并赋予新的类型标识 MedQA选择题。
   * 将格式化后的MedQA数据与清洗后的 cleaned\_medical\_data\_v2.json (源自小荷医典) 合并，生成了 cleaned\_medical\_data\_v3\_combined\_with\_medqa.json。
3. **构建融合知识库**:
   * 使用 create\_model.py 脚本，以 cleaned\_medical\_data\_v3\_combined\_with\_medqa.json 作为输入数据。
   * 采用之前实验验证的最优配置（recursive切分, Chunk Size=512, BAAI/bge-large-zh-v1.5模型），创建了一个新的FAISS索引版本，命名为 v7\_combined\_medqa\_recursive\_bge\_large。
4. **开发专项评测脚本**:
   * 创建了 evaluate\_on\_medqa.py 脚本，专门用于在此 v7 知识库上测试MedQA选择题的回答能力。
   * 该脚本读取MedQA的 test.jsonl 文件作为问题来源，执行RAG流程（检索+生成），并使用一个特殊的Prompt指令，要求LLM（DeepSeek）**只输出**其判断的正确选项字母。
5. **执行专项评测**:
   * 运行 evaluate\_on\_medqa.py 脚本，对 test.jsonl 中的前50个问题进行了测试。
   * 脚本自动将LLM输出的字母与数据集中记录的正确答案字母（answer\_idx）进行比对，计算选择题的准确率。

## 2. 评测结果与效果影响

### 2.1 核心结果

在对MedQA test.jsonl 数据集的前50个选择题进行的专项评测中，基于 v7 融合知识库的RAG系统取得了 **86.00%** 的选择题回答准确率（答对43题）。

### 2.2 效果影响分析

这一高达86%的准确率显著高于之前在通用问题集上测得的检索命中率（最高86.67%，但仅代表检索阶段）和生成ROUGE分数（最高约0.3）。这种差异主要体现了以下几点关于整合MedQA数据的影响：

1. **验证了数据整合的有效性**: 高准确率首先证明了MedQA的问答对信息**已被成功整合、向量化并索引**到了 v7 知识库中。系统能够有效地“找到”这些信息。
2. **提升了对特定知识点的精准问答能力**: MedQA主要包含具体的医学事实和概念（例如，“正常脑脊液压力是多少？”）。将这些精确的“问题-答案”对加入知识库，使得系统在被问到**与这些知识点高度相似或完全相同**的问题时，能够进行非常精准的检索和回答。
3. **展示了RAG在“已知答案”场景下的上限**: 这次专项测试更像是一次检验系统**信息提取和复述能力**的“闭卷考试”。当知识库中**直接包含**问题的答案时，配合合适的Prompt（强制输出选项字母），现代RAG系统可以达到非常高的准确度。

## 3. 结果解释：为什么与通用评测差异巨大？

理解专项测试（86%准确率）与通用问题集自动化评测（~0.3 ROUGE）之间的巨大差异至关重要：

1. **任务性质不同**:
   * **MedQA专项测试**: 任务是**选择题判断**。输入（问题+选项）和输出（单个字母）都高度结构化，评判标准（是否等于正确字母）极其明确。检索目标是找到与问题最匹配的那个已知答案。
   * **通用问题集测试**: 任务是**开放式问答生成**。输入（自然语言问题）和输出（自然语言答案）都是非结构化的。检索目标是找到相关的**概念或段落**。评判标准（ROUGE）衡量的是**文本相似度**，而非绝对的对错。
2. **知识存在方式不同**:
   * **MedQA**: 知识以离散的、原子化的“问答对”形式存在。
   * **小荷数据/通用问题**: 知识以连续的、段落式的百科文章形式存在。LLM需要进行阅读理解、信息抽取和总结归纳才能生成答案。
3. **评估指标的侧重点不同**:
   * **选择题准确率**: 衡量系统**做出正确判断**的能力。
   * **ROUGE**: 衡量系统生成文本与参考文本在**字面上**的相似程度。