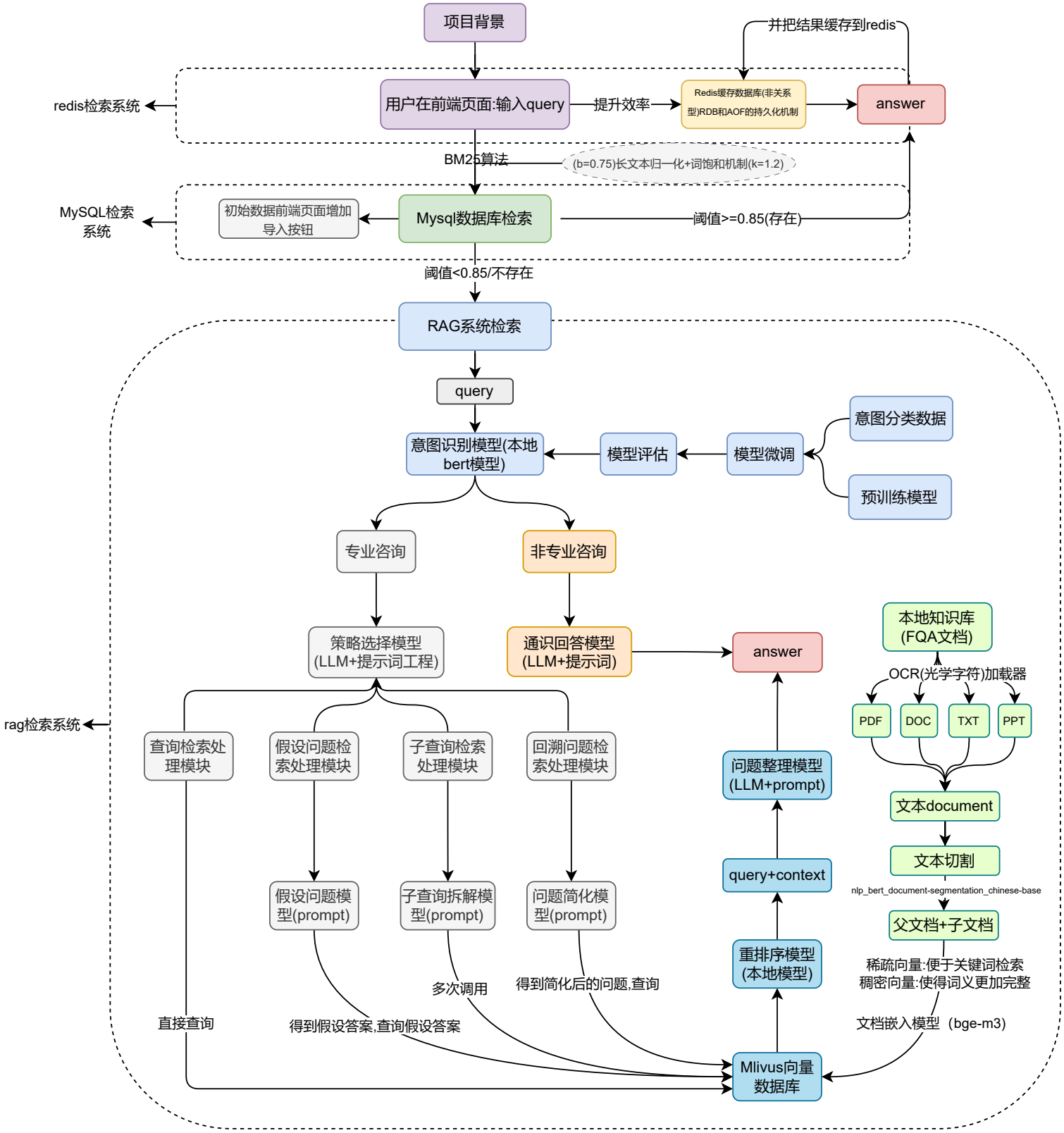

项目背景
并把结果缓存到redis
redis检索系统
用户在前端页面:输入query
提升效率
Redis缓存数据库(非关系型)RDB和AOF的持久化机制
answer
BM25算法
(b=0.75)长文本归一化+词饱和机制(k=1.2)
MySQL检索系统
初始数据前端页面增加导入按钮
Mysql数据库检索
阈值>=0.85(存在)
阈值<0.85/不存在
RAG系统检索
query
意图识别模型(本地bert模型)
模型评估
模型微调
意图分类数据
预训练模型
专业咨询
非专业咨询
策略选择模型(LLM+提示词工程)
通识回答模型(LLM+提示词)
answer
本地知识库(FQA文档)
OCR(光学字符)加载器
PDF
DOC
TXT
PPT
rag检索系统
查询检索处理模块
假设问题检索处理模块
子查询检索处理模块
回溯问题检索处理模块
问题整理模型(LLM+prompt)
文本document
文本切割
nlp_bert_document-segmentation_chinese-base
query+context
父文档+子文档
假设问题模型(prompt)
子查询拆解模型(prompt)
问题简化模型(prompt)
重排序模型(本地模型)
稀疏向量:便于关键词检索
稠密向量:使得词义更加完整
多次调用
得到简化后的问题,查询
直接查询
得到假设答案,查询假设答案
文档嵌入模型（bge-m3）
Mlivus向量数据库