深圳大学考试答题纸

(以论文、报告等形式考核专用)  
二○二四～二○二五学年度第二学期

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程编号 | 1501990049 | 课程名称 | | 大模型技术及开发 | | 主讲教师 | | 陈小军 | 评分 |  |
| 学 号 | 2022150221 | 姓名 | 何泽锋 | | 专业年级 | | 2022级计算机科学与技术（创新班） | | | |
|  | | | | | | | | | | |
| 教师评语： | | | | | | | | | | |
| 项目名称（体现领域）： | | 宋史探微 - 宋代历史RAG系统 | | | | | | | | |

摘 要：

[简单列出已经实现的功能]

一、 数据处理与知识库

* 支持处理用户上传的多种格式本地文档（如PDF, DOCX, TXT） 。
* 实现了围绕特定主题（宋代历史）的定向网络数据爬取 。
* 对获取的原始文本进行深度清洗和安全过滤 。
* 对长文本进行智能分块处理 。
* 为知识库构建了FAISS向量索引和BM25关键词双重索引 。
* 实现了处理后文件及其中间产物（如Embeddings）的缓存机制 。

二、 模型优化

* 搭建了利用大模型从原始文本自动化生成“问题-答案”对数据集的流程 。
* 在自建的宋史数据集上，对Bi-encoder嵌入模型（bge-large-zh-v1.5）进行了微调 。
* 使用LoRA方法对生成式大模型（deepseek-r1-7b）进行了指令微调 。
* 设计了包含难负样本的测试集，并对微调后的模型进行了多维度性能评测 。

三、 大模型分析与应用

* 实现了对用户输入的安全检测与意图识别（区分问答与闲聊） 。
* 实现了基于检索上下文的流式答案生成 。
* 实现了对多源信息的事实冲突检测功能 。

四、 检索与排序

* 实现了知识库检索、实时网络检索、用户指定网页检索的多路召回功能 。
* 对向量检索与关键词检索的结果进行加权混合 。
* 使用Cross-Encoder模型对召回结果进行二次精排序（Re-ranking） 。
* 实现了基于大模型的查询词递归优化与迭代检索 。

五、 用户交互与服务

* 开发了包含对话窗口、知识库管理、历史记录管理等功能的Web用户界面 。
* 实现了实时展示RAG流程（检索、生成、检测）的可视化进度条 。
* 基于Flask框架搭建了完整的后端API服务。
* 使用SQLite数据库对用户对话历史进行持久化存储 。

评分明细：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 评价维度 | 1 | 2 | 3 | 4 | 5 |
| 分项分数 |  |  |  |  |  |