段嘉旋

女 | 2年工作经验 | 24 | 18973579750 | 2720381329@qq.com

项目经历

智答RAG大模型 2024.06 - 2024.12

背景

智答大模型旨在服务集团内部员工从海量内部文档信息（如章程、流程手册、FAQ等）中精准定位 、筛选出所需信息，并提供快速可交互的咨询服务，避免繁琐的信息查找和整理流程，显著提高工作效率。

数据收集和清洗

- 收集并归类资源档案、管理条例、操作指令等公司内部文档数据，**部署MinerU**统一文档至markdown格式
- 设计基于规则和词典的**自动化数据清洗和校验pipeline**，确保数据的准确性和合规性
- **多线程**实现**minhash**算法用于去除海量数据中的内容高度相似部分

向量化落库

- 基于**递归字符切分策略**对清洗后的数据进行分块，落入精准数据库
- 设计并构建小批量测试集，**测试embedding模型**向量化知识块后的命中率，向量检索召回Top30，知识命中率为69.7%

系统框架构建

- 设计**多级索引**，构建基于向量检索、BM25稀疏检索的**混合检索系统**，召回Top30知识块，知识命中率为78.3%，下游任务准确率为63.7%
- 集成**bge-reranker-large**模型，设计"Top30粗排→Top10精排"的**级联召回策略**，下游任务准确率提升至79.6%

系统优化

- **微调bge-reranker-large模型**，提升难负例知识块重排准确率，下游回答准确率提升至92.3%
- 针对用户高频问题设计**动态缓存机制**，高频问题的平均响应速度提升80%
- 采用**异步处理+分段返回**的方式，缩短用户等待感知，提升用户满意度20%

冷链标书大模型 2024.12-2025.06

背景

为满足招标方对冷链合规性、季节性供应稳定性等垂直领域的需求，构建专注于撰写生鲜配送场景投标书的大模型，实现招标文件关键点自动解析，关键板块撰写及风险条款智能标注等功能。

数据统筹和制作

- 分析、沟通并制定数据标注规范，明确重点数据和高度定制化数据的制作细则
- 获取公司内部近两年共2400+专业招标、投标文件，批量解析成结构化文档设
- 计并实现自动化数据去重、打分、和校验pipeline，得到共**22000+**数据

确定测评指标

- 针对用户需求和目标，设计合规性、专业性、全面性三大数据评估板块，人工构建共1200条高质量评估数据，数据类别比例为1:3:1
- 设计自动化测评和人工测评结合的方式计算模型回答合格率

模型微调

- 测试不同base model领域能力，选取Qwen2.5-32B作为base model，测试合格率为**78.8%**
- 基于Qwen2.5-0.5B进行多组实验，得到通用数据 : 私有数据的比例为 **2:3**
- 使用**llama-Factory**对Qwen2.5-32B模型进行**Lora微调**，模型合格率为**89.4%**
- 通过修改训练框架源码，增加数据ID追踪功能，定位异常数据来解决loss跳增问题，模型合格率提升至**92.1%**

模型部署

- 基于4*NVIDIA RTX 4090 GPU集群，使用vLLM框架部署Qwen2.5-32B模型，通过FastAPI封装问答、SQL操作等接口，使用Docker Compose容器化部署，实现依赖隔离和一键服务启停

工作经历

乐禾食品集团股份有限公司　　AI算法工程师　　2023.09 - 2025.07

- 大规模数据处理与文本预处理
- RAG系统构建，优化和维护
- 模型训练和优化，选择合适模型完成项目功能模块，优化模型效果
- 针对 NLP 相关项目中的问题提供解决方案，进行算法的实现及改进

教育背景

湖南农业大学　　会计学　　2020.06 - 2024.06

专业技能

- 熟悉 PyTorch 框架与 Llama-Factory 工具链
- 熟悉 CPT / SFT / PPO / DPO / GRPO 算法 和 Lora / QLora 高效微调算法
- 熟悉使用 DeepSpeed 进行多机多卡分布式训练
- 能够复现 Transformer / BERT 等模型架构
- 熟悉 vLLM、Docker、Linux 等模型部署相关技术