刘梓鑫

17860765547 ZixinLiu@stu.xjtu.edu.cn http://blog.yours.me

教育经历

中山大学 985 211 2023年09月 – 2026年06月

交通运输 硕士

研究方向：基于大模型的交通事故预测

已经立第一作者身份投稿《SoftwareQualityJournal》(CCF-C)发表论文为《ReAPR: Automatic Program Repair via Retrieval-Augmented Large Language Models》的学术论文（under review）

中山大学 985 211 2020年09月 – 2024年06月

交通工程 本科

成绩荣誉：本科综测排名5/108（前5%），连续4年获学习奖学金，学习标兵荣誉称号，山东省2023年度优秀毕业生。

实习经历

北京小米移动软件有限公司 2024年10月 – 至 今

算法工程师小爱音法策略组-北京NLP组 北京

1.多轮意图解析模型迭代

项目介绍：该模型位于小爱productAgent框架内的Agent Parser模块，主要是根据用户问题以及系统回复等上下文信息，从中抽取关键词，属性等；比如：品牌、价格、设备名等等。还可以在Box中写入code内容；利用用户需求结构化能力接收用户的意图信息，用户声音分类(Amazon Compare,Recommend,OutOfDomain)，以及查询修复等问题。

主要内容：

- 利用Qwen-Max做few-shot构建训练数据150k条，并对构建测试数据20k条(其中正样本500条(覆盖率200条，无覆盖率400条)，负样本500条(属于OutOfDomain))。
- 以Qwen2.5-4B做配方基座，使用SFT微调基座模型，使得function code的准确率由86.2%->90.2%。
- 通过用户实际query数据进行评测，以及对公测的10k真实进行评测，Function code准确率由90.2%->92.4%。
- 基于规则优化了子问流程，能够输出function code的错误类型(比如：差误召回，设置现场大学)以及准确率等。

2.多轮query改写模型建设

项目介绍：该模型位于RAG系统的前处理模块，主要针对用户多轮query中存的信息缺失，消歧，同时也结合用户近期输出的设备意图补全，以增RAG后续的覆盖召回；测试指标：产品数据召回重排序(31%->71%)，商品(23%->83%)。

主要内容：

- 从公司的数据厂广泛抽取关于重城的线上用户多轮问答，同时利用Qwen-Max做few-shot生成对应的改写query，这样构造好的训练一批次训练数据以及测试数据。
- 以Qwen2.5-1.5B做配方基座，使用SFT微调基座模型，使得微调后的改写准确率由75%->81.6%。
- 针对每轮测试集中的badcase，分析badcase，补充对应的训练数据，不断迭代模型，改写准确率从81.6%->89.2%。
- 针对SFT数据不了的badcase(比如:多设备清晰代指)，在SFT微调的基础上使用DPO对大人类偏好，使得改写准确率从89.2%->92%。
- 除了准确率外，还调研了各种模型架构评估指标，如：BLEU、ROUGE、PPL等等，最终选取ROUGE以及BERTScore作为辅助评价指标。

项目经历

ReAPR:Automatic Program Repair via Retrieval-Augmented Large Language Models 2024年02月 – 2024年06月

项目介绍: 针对完整包含bug的Java函数，采用检索增强生成技术（RAG），构建外部知识库，并分别采用稀疏检索（BM25算法）和密集检索（DPR算法）检索出与待修复的Java函数相似的bug-fix-pair来辅助LLM进行Java函数的修复。

主要内容:

- 结合APR(Automatic Program Repair)领域深入人使用的数据集，构建针对Java完整函数级别的bug-fix-pair语料库，以供检索器使用并优化。
- 采用PyTorch框架和比较学习并使用InfoNCE损失函数在GPU环境下分布式训练微调GraphCodeBERT，将其作为密集检索系统;
- 同时使用Faiss检索库构建向量数据库加速引，优化索检。
- PromptE利用LLMs的few-shot learning能力，整合检索到的相关函数级Prompt，以辅助LLMs进行修复能力。
- LLMs采用Huggingface上开源的高性能开源模型(GPT-Neo, Phi-2, CodeLlama等)，大小从125M-7B，包括自然语言模型以及专门用于代码任务的模型。

Retrieval-Augmented Intelligent Agent for Enhancing Developer Answers 2024年08月 – 2024年10月

项目介绍: 针对开发者提出的问题，采用LangChain框架的ReAct Agent框架，设计一个Agent，并集成对应的检索Stack Overflow等工具，用Agent自动规划解决问题的步骤，通过结合检索内容更好的帮助LLM进行开发问题回答。

主要内容:

- 使用LangChain提供的ReAct Agent框架，并改进ReAct显性模板，构建一个Agent。
- 使用Ollama最新提供的ollama3.1（SOTA模型）作为Agent的大脑，基本了覆盖了OpenAI的用例，完全实现了无网化。
- 嵌入Agent的工具函数，如模拟用户问题的关键词检索Stack0verflow，并将搜索到的结果使用Ollama提供的nomic-embed-text提取特征向量放入向量数据库中，或者保存在Pinecone向量数据库中。
- 使用Maximum Marginal Relevance(MMR)将Pinecone搜索结果推理到相关新信息上，并检索出与用户问题相关的前15个结果，并检验其与用户向量人工相关性增大。
- 检索索引的结果作为llama3.1模型已有知识的补充，以便更好的回答用户问题。

⚙ 相关技能

- 专业技能：熟练掌握Python，PyTorch框架，熟悉Linux系统，掌握Linux系统常用命令及相关工具的使用（vim，git等），在Linux系统下进行开发经验。
- 大模型技能：使用过Huggingface以及Ollama等开源模型，以及LangChain等大模型应用开发框架，掌握大模型架构相关技术，如检索，RAG，Agent等。
- 论文技能：阅读过几十件有关AAAI,ACL等顶级会议论文，了解良好的论文阅读能力并且具有独立复现论文实验以及追踪前沿技术的能力。

ℹ 个人总结

- 对大模型相技术术关注深刻，有较好的学习能力，乐于扩展大模型应用技术。
- 乐于团队协作，善于沟通能够独立思考，善于自我驱动，对待工作认真负责。
- 乐意公司运作开发方式，剑了解环境能够快速上手从事研发工作。
- 性格较好，与同事相处融洽，工作勤勉诚实。