个人简历

求职意向：大模型开发/算法工程师

基本信息

姓　　名：彭正隆　　政治面貌：团员

性　　别：男　　毕业院校：桂林电子科技大学信息科技学院

电　　话：13702435054　　学历 ： 本科

邮　　箱：pengzhenglong2016@163.com　　专业 ： 电子信息科学与技术

学习工作

- 2024. 04—2025. 04　深圳建广数字科技有限公司　　岗位：大模型开发工程师
- 2020. 06—2024. 03　环球资源广告（深圳）有限公司　　岗位：Java工程师
- 2016. 09—2020. 05　埃士信贸易（深圳）有限责任公司　　岗位：Java工程师
- 2012—2016　　桂林电子科技大学信息科技学院　　专业：电子信息科学与技术

求职意向

- 大模型开发工程师：基于大模型的应用开发与落地、原应用重构。
- 大模型训练工程师：大模型微调及垂类应用的开发部署。

专业技能

- 熟悉Dify + Ollama + Coze + Cursor等各种大模型应用工具的使用。能够基于Dify完整RAG和工作流的定制化
- 熟练使用远程服务器对 LLaMa3 进行 Lora 微调测试评估、模型合并与量化、部署量化后的模型
- 熟练使用数据分析可视化三剑客 NumPy、Pandas、Matplotlib 进行数据处理和数据可视化。能熟练掌握Python编程，具备良好的数据分析和问题解决能力
- 熟练使用 LangChain 构建本地向量数据库，实现存储与查询，掌握 LangGraph 核心组件。
- 熟悉 Pytorch 和TensorFlow 的网络架构模型，熟练使用里面的 API 进行模型搭建。
- 熟练使用 ModelScope 在线训练平台进行数据下载和模型调用，在线训练过 Llama 模型。
- 能够自定义 FunctionTool，实现工具的同异步调用。实现聊天、天气查询和网页搜索的 Agent 调用功能。
- 理解 LoRA 微调的基本原理，能够使用 LLaMA-Factory 基于自定义数据集进行微调训练。
- 理解主流机器学习算法原理：线性回归、逻辑回归、Softmax回归等。

- 熟练使用已有机器学习算法进行模型训练和预测，有炼丹经验，有实际项目经验。
- 理解 CNN、RNN、注意力机制、Embedding、Transformer、BERT、GPT 等模型原理和架构。
- 熟悉模型压缩方法、长文本技术、熟悉 SFT、RLHF 等调优技术。
- 熟悉 Python、JAVA、JS等编程语言和html、css、Linux
- 熟悉Kafka、Elasticsearch的使用和优势优点，熟悉kafka的组成结构、多副本备份机制的原理
- 熟练使用MySQL、PostgreSQL，熟悉InnoDB引擎下的索引、事务以及SQL优化。
- 熟练使用分布式缓存 Redis，熟悉Redis数据结构 、内存淘汰策略、持久化机制、集群。
- 熟练Kubernetes编排和docker容器化部署技术
- 了解大数据相关技术Hadoop、Flink、Spark
    -

项目经验

项目：国际物流SaaS系统-RAG企业知识库（深圳建广数字科技有限公司） 2024. 12–2025. 04

项目概述

国际物流客户需快速查询散落在内部文档、工单中的非结构化数据（如异常处理规则），传统搜索工具准确率<70%；构建一个结合国际海关法律/国际物流等专业领域的RAG系统，实现实时知识更新与可验证回答，支撑物流客户高效获取精准信息

方案设计

1、向量数据库在用弹性、分布式部署的Milvus数据库集群，采用Docker部署，可弹性增加节点

2、使用PyMuPDF解析PDF/PPT文档。通过Linux服务器的私有化部署UnstructuredLoader。使用UnstructuredLoader加载非结构化文档（如 PDF、Word、MD等）。设置 hi_res策略来提取章节层级、表格等。通过Tesseract （OCR）提取图片信息。 采用章节段落分块和语义分块相结合的方式来解决表格或者内容跨页断裂问题。

3、采用私有化部署的bge-large-zh-v1.5和BM25对Chunk进行密集向量化。创建密集向量和稀疏向量的索引。其中稀疏向量索引采用DAAT算法，控制词频饱和度参数BM25_k1，加速查询性能。稠密向量索引设置构建图时每个节点的最大邻接数M为16，efConstruction=64，提高搜索广度和精度。实现"向量+关键词"混合检索

4、通过LangGraph定制工作流，实现动态路由能力，根据上下文语音去不同的知识库或者网络进行检索

5、引入Corrective RAG和Adaptive RAG构建两层评估体系。第一层：从知识库中检索出来的doc进行相关性评估。评估不达标通过Transform_Query触发重新检索。第二层：在输出答案之前进行输入的关联性评估。全部评估通过最终输出答案。

项目技术

LangChain、LangGraph、Milvus、UnstructuredLoader、PyMuPDF、LLaMa3.1

项目经验

项目：安链云SaaS平台AI智能助手（深圳建广数字科技有限公司） 2024.07–2024.12

项目概述

围绕物流业务场景，结合SaaS服务流程，数据涵盖订单信息、配送信息、仓储信息等核心领域，替代客服和销售人员，通过交互的方式，完成物流SaaS系统中各个业务。由AI大模型自主完成：营销，配送，仓储查询和订单管理，金融服务产品销售和预定等

项目实现

主要包括业务工具模块、身份信息和权限控制模块、多Agent调度块、状态管理模块。

1、工具函数模块：将安链云已有的业务，提取整合成为LangChain工具; 依据功能风险将工具划分为“只读工具”与“敏感操作工具”，将所有工具安装业务分类（营销业务，订单业务，配送业务，仓储业务，公共业务），并与大模型完成交互绑定，使大模型能够动态触发工具调用

2、身份信息和权限控制模块：主助手智能体一开始就能通过节点从请求参数中获取用户ID，并调用Tools查询完整的物流信息，并保存在状态中，当整个工作流执行到任何敏感工具节点之前都会进行中断，并且保存上下文信息，即使中断的情况下Agent也能通过Confg中的thread_id和用户ID来保证上下文统一

3、多Agent调度块： 整个模块主要是把业务流程细分为四个子工作流，由一个主Agent进行居中调度，每个子工作流作为一个单独的Agent负责完成各自的业务，分为：营销业务Agent、订单业务Agent、仓储业务Agent，公共业务Agent。每个子工作流包含五个节点，分别是入口节点、决策节点、只读工具、敏感操作工具和离开当前任务节点，从主工作流到子工作流的调度是通过动态路由条件来实现，从子工作流到主工作流是通过控制状态中的来管理记录

4、状态管理模块：整个模块是State模块，就是保存工作流中的messages（历史消息列表）、用户信息(user_info)、以及工作流状态栈(dialog_stack)

算法技术

- 模型选择：本次使用的是 llama3.1 模型。
- 代码框架：使用 langgraph 创建工作流图，结合多个 agent 一起完成整个项目需求。
- 模块组成：Agent、Prompt、RAG、LLM

项目亮点

本项目基于 LangChain + LangGraph 框架开发多个Agent + WorkFlow实现了一套智能服务助手，聚焦于多轮对话管理、多任务链路执行和工具调用等能力。

项目经验

项目：基于 Llama3 的训练与微调（深圳建广数字科技有限公司） 2024. 04–2024. 07

项目概述:

Llama3 作为开源大模型，基于超过 15Ttoken 训练，相比 GPT，它具有更强的定制性和安全性。可应用于针对 B 端企业的自然语言处理、机器翻译、文本生成、问答系统、聊天机器人等场景。本次项目通过训练和微调 Llama3，为定制部署垂类应用做技术准备。

微调目的:

提升语言理解，帮助模型更深入地理解中文的语法和语义结构，从而在理解上下文、抽象概念和复杂句式方面表现得更好。通过在中文文本上微调，LLaMA3 能更准确地捕捉到中文使用中的地域性表达和文化特征，提高其在地化应用的效果。在法律、物流等专业领域中，中文数据集的微调能够让型更好地适应行业特定的术语和表达方式，增强其在特定领域的实用性和精准度。偏差减少:通过对包含多样化中文文本的数据集进行微调，可以减少模型在处理中文输入时的偏差，使其生成的内容更加公正和无偏见。

平台设备:

阿里云 DSW，NVIDIA A10 24G 显存。

数据集:

使用 Alpaca 数据集，已进行清洗优化。

模型:

llama-3-8b-bnb-4bit

模型训练:

- 使用 4 位量化训练，加快训练速度
- 保存模型
- 加载模型用于微调
- 使用 4 位 Q-LoRA/LoRA 微调

本地微调:

- 使用正则表达式和 langchain 工具清洗 PDF 文档，并使用 Agent 工具生成 json 格式的指令数据集。
- 执行微调，并将微调后的模型保存成 GGUF 格式，方便本地使用。
- 使用 Ollama 和 LM Studio 加载微调好的模型，测试生成效果。

微调后效果:

- 对比后，可以看出微调后的模型对中文支持度更好，泛化性提高，回答精准，不易出现幻觉。

项目经验

项目名称： 搜索平台组任务可视化系统 （环球资源） 2022.06-2024.03

核心技术： Spring，SpringBoot，Nacos，PostgreSQL，Redis

项目概述：

为解决以下痛点：

1、任务无法查看执行进度

2、服务器重启导致任务中断

项目职责：

1、 在架构师的指导下负责整个系统的方案设计。

2、 实现该系统高可用方案的设计与落地。

3、将部分定时任务接入该系统。

4、提供该系统对任务的新建、编辑参数、启动、停止等API的支持。

技术描述：

1、定义注解用于任务执行类的标记与任务相关信息的声明，用模版方法封装主流程。

2、使用 PostgreSQL 存储任务相关信息。

3、使用Redission的延迟队列来实现任务发布与任务执行的松耦合。

4、使用Redis的发布订阅来修改任务的中断标识实现任务的中断。

5、使用Nacos的服务监听机制实现任务的自动恢复。

项目名称： GSOL关键词系统 （环球资源） 2020.06-2022.06

核心技术： Spring，SpringBoot，Nacos，Fegin，Dubbo，PostgreSQL，Redis，XXL-Job

项目概述：

从内部（Global Sources）、外部（阿里国际站、敦煌网、中国制造网等）收集大量B2B外贸平台的搜索关键词，将这些关键词通过一系列的过滤、清洗、合并提炼出一批精准、热门的优质关键词来提供给网站内部使用。

项目职责：

1、 参与关键词需求分析，提出对应设计方案。

2、 编写并维护关键词接口代码，并编写单元测试，确保代码的质量和稳定性。

3、 编写关键词拉取、过滤、清洗、合并等相关定时任务，并保证代码的高扩展，高可读性。

4、 编写并维护关键词文档。

技术描述：

1、抽象关键词过滤器接口，利用DI一次性获取所有过滤器后循环过滤，从而实现高扩展性，符合开闭原则。

2、使用线程池提升关键词任务的处理速度。

3、自定义缓存注解实现零侵入数据缓存。

项目经验

项目名称：机器人管理平台 （埃士信） 2018.06-2020.05

核心技术： Spring, SpringBoot, Dubbo, Redis, RabbitMQ

项目概述：

该模块作用于如果用户与业务机器人之间的触发了兜底话术,用户的问题会被记录下来,次日由标注人员手动匹配更精确的答案,最终目的,是如果同样的问题,用户会得到更满意的答复。

项目职责：

1、参与需求分析、业务合理性讨论

2、维度管理和技能管理模板的开发;

3、负责智能教育的审核功能设计与开发,一个问题会由两个标注师同时进行标注,并且由业务方负责人,去确定最 终答案,并把答案归纳到知识库进行管理;

技术描述：

1、使用SpringBoot 进行项目构建，使用 Dubbo+Zookeeper 实现远程调用以及服务治理

2、使用 RabbitMQ 进行审核成功后的业务处理

3、使用Redis+定时任务实现今日进店、今日点击数据统计

项目名称：数据库可视化工具 （埃士信） 2016.10-2018.05

核心技术： Spring, SpringBoot, XXL-Job, Redis, RabbitMQ

项目概述：

云化可视化数据库设计工具，支持数据库物理模型设计、变更、审批、自动生成变更脚本、脚本上库、设 计规范管控、版本管理与分支管理等E2E协同作业。支持多种数据库（Mysql、Oracle、PostgreSql、GaussDB系列、Cassandra、Hive等）、支持逻辑模型设计。

项目职责：

1、参与需求分析、表结构设计的讨论。

2、设计规范管控/自定义规范模块

3、版本管理与分支管理

4、负责数据同步公共日志模块的开发

技术描述：

1、控制台SQL执行建模;

2、使用 RabbitMQ 进行数据异步同步与消费

3、使用本地消息表+定时任务保证数据的最终一致，使用 Redisson 对消息 Id 加锁，防止手动重试与定时任务重试导致重复消费的问题

4、使用单例模式与模板方法模式进行消息的统一分发

5、使用模板方法模式对于不同类型的商品进行业务调整

自我评价

- 喜欢探索钻研，具有较强的学习能力、适应能力和抗压力。
- 喜欢玩各种框架、开源项目，综合能力、动手实践能力强。
- 自学能力强，擅长学习官方文档，使用新特性进行应用开发。
- 对待工作认真严谨、积极向上。
- 具有较强的责任心、团队协作能力。
- 对 AI技术方面兴趣浓烈，跟随先进技术，不断学习。
- 有较强的自我学习和解决问题的能力