曹剑沣

13814943087 | jianfeng_cao@qq.com | 上海
在职 | 大模型开发工程师

教育经历

常熟理工学院 2007年09月 - 2011年06月

电子信息工程 本科

技术栈

- 熟悉 Python 以及 Pytorch 框架，对常用 API 都有深入了解，并在项目中熟练使用
- 了解 RAG、Agent 等LLM 应用技术，以及多模态RAG、GraphRAG、DeepResearch等技术
- 熟悉 LoRA、QLoRA等LLM 参数高效微调方法，了解 LlaMA、Qwen、Deepseek 等常见开源模型，了解RLHF、DPO等偏好对齐方法
- 熟练掌握Linux操作系统及shell、python编写脚本，能够高效地进行系统管理和自动化任务
- 熟悉K8S，Docker技术栈，并将大数据平台，其他应用和自研系统进行容器化改造
- 精通使用Ambari/Cloudera Manager搭建和维护企业级Hadoop平台（CDH和HDP）

工作经历

上海深众信息技术有限公司 2024年10月 - 至今

大模型应用开发工程师

背景: 应对企业服务管理中工单处理效率低、响应慢等痛点，通过开发问答助手精准解决问题，显著提升公司内部运营效率。
职责:

- 基于LlamaFactory利用LoRA技术微调Qwen大模型，集成DeepSpeed加速多卡训练，结合测试数据集评估模型性能，提升训练效率30%。
- 基于LlamaIndex框架开发企业知识库智能助手，支持高效检索与知识管理，优化用户查询体验。
- 实现多模态RAG管道，集成CLIP、Faiss及智谱大模型，支持文本与图像数据的索引、检索及生成，满足多模态查询需求。

CLPS（PayPal） 2021年03月 - 2024年09月

运维开发工程师

背景：为湖仓一体大数据分析平台实现全服务容器化，优化PayPal Hadoop集群升级，解决系统扩展性与运维效率痛点，提升数据处理性能与服务稳定性。

职责：

- 设计并搭建Rancher平台，部署Kubernetes（K8s）集群，Ray集群，针对大数据组件与自研系统重新构建镜像，编写Docker file，YAML文件确保兼容性，实现大数据组件与系统无缝联动，提升部署效率与稳定性。
- 负责PayPal大数据平台运维（数据中心建设支持），包括节点扩容、新用户接入（创建账号、Hive表、HBase表及授权），灾备环境搭建，漏洞修复，提升系统可扩展性与用户满意度。
- 搭建Zabbix监控系统，集成大数据组件至Zabbix与Grafana实现监控与告警，显著提高系统响应速度与运维效率。
- 支持Control-M系统，与全球团队对接完成新Job上线、系统维护及灾备环境搭建，定期进行灾备切换演练，确保业务连续性与系统稳定性。

浩鲸科技有限公司（汇丰） 2020年07月 - 2021年03月

运维开发工程师

背景: 为汇丰银行系统接入阿里云平台，优化云服务部署与运维流程，确保系统稳定性和用户体验，提升平台性能与客户满意度。
职责:

- 负责阿里云服务机房验收，全面检查相关云服务功能，整理问题清单并与实施团队协作解决问题，确保达到验收标准。
- 负责汇丰银行阿里云平台运维（DataWorks、MaxCompute），针对用户问题进行验证与分析，提供清晰解释与操作指导；对于平台层面问题，与阿里云研发团队紧密合作，通过Hotfix或版本升级快速解决，提升客户满意度。

上海浦东发展银行信用卡中心 2017年06月 - 2020年07月

大数据工程师

背景: 构建稳定的大数据平台，部署欺诈养卡防控模型，为各个组件添加监控，保障生产环境的稳定性和可靠性
职责:

- 主导200+节点大数据集群的规模化部署与运维，基于Ambari/Cloudera Manager实现集群自动化部署、监控及性能调优，保障日均4T级数据处理稳定性
- 部署欺诈养卡风控模型，集成至大数据平台实现分布式训练，利用Spark ML优化特征工程，模型迭代效率提升60%
- 基于Jenkins构建应用系统的自动化发布，实现十多套应用系统的改造，极大提升了系统发布的效率

项目经历

企业知识库智能问答助手 2025年03月 - 至今

大模型应用开发工程师

项目概述

针对企业服务管理中工单处理效率低、响应延迟的痛点，开发基于大语言模型的智能知识库问答助手，取代传统工单流转，基础工单下降30%。结合运维经验，优化系统稳定性与高并发处理能力，确保企业运营效率显著提升。

技术亮点

- 高效模型微调与运维优化: 基于 LlamaFactory，利用 2000 条工单问答对，采用 Qwen1.5-14B-Chat 进行 LoRA 微调，合并模型后使用 int8 量化，降低 32% 内存占用，领域问答准确率达 95%。结合运维经验，通过自动化脚本优化微调流程，并实现模型部署的自动化监控与故障恢复。
- 混合 RAG 检索与重排序: 利用 LlamaIndex 注入知识库文档，存储于 Chroma 向量数据库，构建语义搜索与关键词匹配的混合索引，召回率提升 40%。开发基于 bge-reranker的重排序模块，通过上下文相关性分析优化答案排序，整体问答准确率达 93%。优化 Chroma 数据库配置，减少 20% 查询延迟，确保高负载下检索稳定性。
- 高性能推理与系统架构: 基于 vLLM 搭建异步批处理管道，首字节响应时间降至 0.32 秒，较基线提升 48%。

湖仓一体大数据分析平台与系统升级 2021年03月 - 2024年09月

运维开发工程师

项目概述：湖仓一体大数据分析平台所有服务容器化

技术亮点

- 安装rancher，部署管理k8s集群，以及根据特定版本重新build镜像将大数据组件以及相关联应用容器化，实现组件之间的联合访问
- Airflow指定版本的容器化，并经过重新build镜像后和hdfs，hive，sqoop的联合操作，直接在airflow部署任务就可以实现调度hadoop的资源
- 安装本地镜像仓库Harbor，将使用到的镜像推送并存储在本地镜像仓库，方便调用，也避免服务器资源不足，自动删除镜像导致容器服务挂掉，维持服务的稳定运行
- 测试环境测试升级版本，测试hdfs读写,yarn测试job,hive内外部表读写,hbase读写等功能

汇丰系统对接阿里云平台 2020年07月 - 2021年03月

运维开发工程师

项目概述

汇丰系统接入阿里云平台

技术亮点

- 按照验收标准，测试系统的可用性，和实施反馈问题，确保服务达到验收标准

- 负责用户侧使用dataworks和maxcompute遇到的各种问题，进行测试验证，并将结果反馈给用户，指导用户正确使用
- 针对阿里云平台侧的问题，收集并反馈给阿里研发，后续通过hotfix或者大版本升级来解决问题
- 每天运行系统巡检系统，及时发现问题，并处理问题

欺诈养卡防控模型 2018年06月 - 2019年11月

大数据工程师

项目概述

利用AI高维机器学习算法和计算架构，充分挖掘欺诈风险客户在不同场景行为表现中的海量特征数据，依托高性能预估引擎，在确保模型预测精确度的同时，扩大欺诈养卡客群的监控范围，推进实现风险行为监控范围的无缝覆盖，模型识别准确率达95%以上。通过机器自主学习，完成风险指标参数化更新，并在数据引入，模型训练，发布上线闭环中实现自我迭代，保证模型长期监测效果

技术亮点

- 搭建建模环境hadoop集群
- 利用F5负载均衡技术保障高可用性，通过防火墙策略隔离生产与验证环境，确保数据安全部
- 署欺诈养卡防控模型，启动相关进程
- 更新配置文件，对接hadoop集群