王润龙

男 | 年龄：24岁 | 1347203568@qq.com
3年工作经验 | 求职意向：大模型算法 | 期望薪资：28-35K | 期望城市：上海

个人优势

本人性格随和，真诚谦虚，热爱代码。对待工作认真热情，肯吃苦，具备较强的学习能力和抗压能力。具有团队精神和协作意愿。经过多年的代码训练养成了良好的代码风格和思维逻辑，比较注重代码质量，解决问题的能力比较强。愿意迎接及应对工作中的各种挑战。

工作经历

智子熹源（上海）科技有限公司 agent 2024.03-至今

内容:

1. 负责企业内部Agent框架开发
2. 负责Agent意图分类的NLP及NER部分算法实现
3. 负责对Agent项目的项目管理
4. 负责RAG在线算法服务的设计&开发
5. 负责Infra离线数据平台的后台搭建和文件处理服务搭建
6. 负责对新技术的验证复现(例如Deepseek论文复现...)
7. 对私有化交付算法部署落地的支持(NPU/其他算力一体机的模型&服务部署)

业绩:

1. 实现自主研发的Agent框架，支持动态添加某个问题域的工作流
2. 实现问题分类，现已实现【6个域】的问题分类及后续的意图识别
3. 自主研发的RAG服务，MRR比百度的千帆高出5个点，首条召回相关率比千帆高出7个点。比RagFlow的MRR高出3个点，收招呼召回相关率大致持平，约差RagFlow 2个点左右。
4.Infra平台完成企业内部2TB垂类数据的导入和存储，实现从数据到训练预料的一条龙处理。
5. 对企业私有化交付提供支持，实现在NPU（昇腾）平台及l4t（英伟达算力一体机）平台的模型搭建
6.实现Deepseek论文的复现，使用对应思路成功对NLU进行优化。

信华信（大连）数字技术有限公司 Java 2022.03-2024.01

内容:

java后端业务开发和项目框架开发。
优化公司后端框架

业绩:

获得过2023年公司年度优秀员工

项目经历

能碳大模型多智能体协作平台 Agent算法端开发负责人 2024.03-至今

内容：

该项目主要适用于能源领域的multi-agent协作场景。针对能碳领域特点，完成从底层框架到上层多Agent应用的完整商业用项目。项目以能源领域数据为牵引，具体应用场景为导向，
意在解决用户在能源领域 源网荷储 四大应用场景的实际需求。项目实现从Infra（数据准备，模型训练）-> RAG -> Agent -> Multi-Agent整套数据流转和业务闭环。

主要技术：
工程架构：采用 Python、Fastapi、aiohttp 搭建高性能服务框架，集成 Jeager 与 PushGateway 实现全链路监控与性能优化
算法体系：基于 VLLM/SGLang/XInference 完成模型高效部署，基于 Modelscope-Agent 多智能体框架开发企业内部框架、LlamaFactory 模型训练体系实现业务逻辑开发
数据存储：使用 ES、MySQL、Redis、Milvus(向量数据库)、Kafka 支持服务高性能高可用

业绩：

1. 框架创新：主导 Agent 框架选型与二次开发，设计企业级多智能体协作架构，支撑日均1万+,并发100+次智能体交互
2. 业务落地：完成企业知识问答、文档生成等核心业务智能体开发，自主研发 RAG 检索增强、实时联网搜索、MCP等工具，基于企业专属数据完成模型微调。
3. 技术引领：统筹组内技术选型与评审，推动多次关键技术改造，主导 2 场企业内部智能体框架培训，赋能团队快速掌握多智能体开发能力
4. 生态整合：牵头完成与 Java 服务端、数据平台的深度对接，构建以 SSE 通信为核心的高效数据交互机制，实现多系统应用协同与智能体任务协同

能源企业知识库管理平台 RAG-算法端开发负责人 2024.05-至今

内容：

针对能源行业专业文献、企业规章、技术手册等知识的高效问答需求，主导开发基于检索增强生成（RAG）技术的智能知识问答系统。通过精准知识检索与智能内容生成，赋能企业员工快速获取专业知识，显著提升能源生产、运维及管理效率。

主要技术：
数据处理：使用 Python 结合 Pandas、OCR、VLM(视觉语言模型)进行企业文档处理
数据增强：使用LLM、ES、Milvus进行检索数据的增强和存储
模型优化：使用LlamaFactory&sentence_transformer进行模型的二次调优

业绩：

1. 向量数据库架构优化：搭建 Milvus 向量数据库存储企业知识图谱，采用 HNSW 空间图向量索引技术，在 10B token 数据规模下，检索响应速度提升 70%
2. RAG 系统全链路开发：独立设计并开发 RAG 后端服务，构建「文档解析→纯文本处理→语义切片→多路召回」的完整技术链路
3. 多模态检索创新：设计支持图片、表格、公式的多模态召回逻辑，功能覆盖度超出行业竞品 20%
4. 模型性能调优：基于能源领域专业问答场景，对 Embedding 与 Rerank 模型进行定向优化，将平均倒数排名（MRR）指标从 0.79 提升至 0.85

企业应用国产化&一体机改造 开发负责人 2024.10-2025.02

内容：

1. 针对企业核心业务系统中模型算法模块依赖国外技术栈带来的安全隐患与成本问题，主导完成模型算法领域的国产化改造项目。通过替换国外主流模型相关框架，优化模型推理的全流程国产化替代，保障数据安全与技术自主可控，同时降低系统运行与维护成本。
2. 同一时间还展开了对一体机部署服务&模型的调研改造，支持NVIDIA-ORIN机器的部署

主要技术：

大模型部署框架：MindIE, Xinference

小模型部署框架：Xinference

模型训练框架：Llamafactory(NPU), MindSpore

中间件：Milvus, PolarDB, EasySearch...

业绩：

1. 国产化适配调研: 负责调研并实施国产化算法&模型部分改造，已经支持 华为昇腾&沐曦 国产化的改造
2. 国产化应用服务改造: 负责主导组内成员进行国产化适配改造，包括但不限于服务框架改造、模型部署改造、服务适配...
3. 一体机调研: 对不同模型在不同规格的一体机上进行benchmark

dodoAPI底层SDK　需求&开发　2023.11-2024.03

内容：

基于 Dodo 语音平台提供的开发者 API ，开发便于开发者使用的专用 Dodo-Python-SDK ，主要功能包含客户端与服务端的通信建立、具有鉴权功能的通信接口封装、通用日志模块的封装、ws 服务器信息的过滤、路由及各个事件的处理器功能。后续还会开发基于 Pymysql 的 Dodo 专用 ORM 框架。

以及装饰器支持作为独立后端路由功能的开发。

技术栈：Python、WebSocket、AIOHTTP、logging、ABC

项目职责：负责全部 SDK 的架构设计及底层架构开发，负责部分接口、事件封装的开发

项目地址 https //github.com/omoidesu/dodo.py

业绩：

已经实现 WS 的异步连接，同时支持异步运行 WS 的心跳请求维持 WS 连接。

已经实现接收指令到装饰器被修饰函数的方法路由，同时有一套完备的基于消息队列的处理系统。

实现 logging、http 请求功能的封装，以及实现部分服务器返回信息的封装。

使用诸如装饰器模式，单例模式等多种设计模式对框架进行开发及优化。

大量使用装饰器对 SDK 进行功能集中化和便捷化

商砼ERP　前后台开发　2022.06-2023.09

基于 SpringCloud 框架的分布式微服务后端管理平台，业务包括基础、销售、生产、财务模块，旨在实现水泥工厂常规运营的线上化支持。项目不仅实现了与厂中 LED、中控生产等硬件设备的顺畅对接，也成功推动了工厂的全自动化生产销售。后端代码量达到 34 余万行，其中个人负责模块核心代码超过 4 万行。

技术栈：Spring Cloud、Java、Vue、MySQL、GIT、硬件设备接口（LED、AUDIO）、Python、Feign

项目职责：负责全部生产模块及部分其他模块的业务技术设计及后端开发，参与文档的编写工作。

个人成果：

基于注解及反射等原理，成功实现全项目通用的大部分工具类，提高了代码的灵活性和可维护性。

独自负责开发了前端复杂页面，确保了用户友好的界面和良好的交互体验。

利用 Python 开发了基于项目框架的全自动代码生成工具，极大地提高了开发效率，能够生成全套后端代码。

项目地址：公司内部项目

教育经历

大连海事大学　本科　通信工程　2018-2022