罗欢

算法工程师 · AI Agent & NLP

985+211 背景，NLP 方向硕士，中共党员，专注 AI Agent 系统工程化落地。独立从 0→1 设计并交付企业级生产 AI Agent 系统（LangGraph + Milvus + vLLM + FastAPI），面向 500+ 员工的工时管理系统集成上线；曾在小米（小爱同学 NLP 组）实习参与 IOT Agent 优化；SCI 论文在投。具备从算法选型、架构设计到生产部署与监控运维的完整交付能力。

lh229470989@gmail.com

15974340206

yriri.site (Personal Site)

mailto:lh229470989@gmail.com (Email)

Work

算法工程师

Jul 2025 – Invalid Date

独立从 0→1 设计并交付公司工时管理平台 AI 智能助手系统，覆盖需求分析、算法选型、系统架构、服务开发到 Docker 生产部署全流程，面向 500+ 员工的工时管理系统集成上线。

- Function Calling 架构改造：将「意图分类→参数提取」两步级联重构为单次 FC 调用，消除分类边界误差传播；本地 vLLM 因 schema prefill 延迟略增，托管 API（DashScope）场景因省去一次网络往返预期缩短 20-40%；LangGraph StateGraph 编排 RAG + Tool Agent 双层 Agent 架构；SpringBoot WebFlux 代理 SSE 全链路流式输出
- RAG 混合检索 + 重排序：LangChain EnsembleRetriever（Milvus 60% + BM25 40%）+ MultiQuery 改写 + CrossEncoder Reranker 精排 Top-5；自建 50 条 Q-A 评测集（覆盖工时填报规则、请假流程、福利政策、考勤管理、岗位津贴等多类业务场景），Recall@5 = Milvus 100% / Hybrid 98%（当前知识库规模下 BM25 噪音主导，文档扩展后预期混合反超）；意图分类精度 87%（2000 条测试集）
- SQL Agent 三层安全：硬规则（语句白名单 + 关键字黑名单 + 列黑名单）拦截 25% 恶意请求，LLM 语义改写辅助层处理 75%，权限校验注入用户级 WHERE 条件；6 级权限体系（employee → superAdmin）
- 全栈工程化 + 端到端可靠性：Docker Compose 7 服务编排（ai-service / Milvus + etcd + minio / Redis / Prometheus / Grafana），反向 SSH 隧道

穿透公网到内网 GPU 服务器；自研 Prometheus 5 类业务指标 + Grafana 8 面板可观测体系；端到端发现并修复 11 项跨集成边界缺陷

- 工具参数智能解析：(user, project) 二维历史众数 + LLM 候选白名单兜底 + cachetools.TTLCache，解决 LLM 输出参数不稳定与用户口语化 → 业务 ID 映射问题

算法工程师（实习）

Jul 2024 – Oct 2024

- 大模型生成数据流程：构建大模型生成标注数据 pipeline 替代 jsgf 文法 + 人工标注，系统评估 10 个通用大模型并选型落地；基于该 pipeline 对 icsf 小模型做全量微调部署上线，新品类数据集建设时间缩短约 50%
- 难负样本优化：引入 LRM Loss 让模型聚焦难负样本，子设备 sf 识别准召率提升 3.2%
- 小样本召回优化：正负样本模板构造捕捉知识库信息，其他意图 f1 提升 2.3%；歧义词过滤（衣架品类）f1 提升 1.6%
- GSB 上线评估：IOT Agent 与线上模型对比评估，修复 200+ bad cases

Education

中国地质大学（武汉）

硕士 in **测绘工程**

Sep 2022 – Jul 2025

Courses

- 数据挖掘与机器学习
- 高级程序设计
- 现代空间信息系统软件工程
- 矩阵理论

中山大学

本科 in **地理信息科学**

Sep 2017 – Jul 2022

Courses

- 遥感图像处理
- 机器学习程序设计基础
- 软件工程
- 空间分析与应用
- 数据结构

Projects

基于多模态社交媒体数据的城市内涝数据挖掘与制图

Mar 2024 – Mar 2025

research

硕士研究课题。设计多模态洪涝信息检索与严重程度识别框架，应用于洪水范围精确制图

- 融合 BERT 全量微调文本分支 + CLIP 基于 PEFT/LoRA 微调图像分支（兼顾显存效率与跨模态对齐），zero-shot 建设数据集后 finetune，洪涝分类 f1 96%、程度分类 f1 79%，较基线分别提升 5% 和 8%
- Bi-LSTM-CRF NER 提取 POI 位置信息用于制图，f1 87%
- 论文在投 Mining and Mapping Urban Flooding Information Using Social Media Data and Multimodal Machine Learning，《International Journal of Applied Earth Observation and Geoinformation》（SCI, JCR Q1）

BERT CLIP PEFT LoRA zero-shot Bi-LSTM-CRF NER 多模态

Skills

AI Agent & 大模型

LangGraph LangChain Function Calling RAG Prompt Engineering Qwen LLaMA GPT

机器学习与微调

PyTorch Transformer PEFT LoRA vLLM 推理加速

检索与 NLP

Milvus Faiss BGE M3E BM25 jieba CrossEncoder Reranker BERT Bi-LSTM-CRF

工程基础设施

FastAPI Docker Redis Prometheus Grafana MySQL SSE

编程语言

Python Java

Languages

中文

母语

英文

六级 480，可流畅阅读技术文档