个人信息

姓名：潘立勇　　性别：男　　工作地点：北京、天津

电话：173-0124-0429　　邮箱：polanpan1963@163.com　　求职意向：大模型Agent应用研发

核心优势

- 了解深度学习、神经网络基础、熟悉Transformer架构,熟悉大语言模型LLM应用研发相关技术栈
- 熟悉大模型RAG和Agent应用研发技术，拥有深入理解和实践经验
- 熟练微服务架构设计,高并发系统优化,熟悉Flink、Hadoop、Kafka、Elasticsearch等大数据技术
- 拥有4年智能客服领域经验，成功落地多个运营商客户项目，累计创造千万级业务价值
- 拥有4年AIOPS智能运维行业经验，主导头部银行智能运维项目技术方案实施，具备丰富的跨部门协作与客户沟通能力

工作经历

北京必示科技/上海鼎茂信息技术 | 资深研发 /主管

2021年3月 - 2023年9月 | 2023年9月 - 2025年5月

- 管理8人技术团队，参与智能运维平台架构设计与开发，负责头部银行项目对接落地，完成合同需求并协助项目完成验收回款

北京舜和通达数字网络科技有限公司 | 智能客服项目研发负责人

2017年1月 - 2021年3月

- 管理20人技术团队，负责全景智能客服中心设计与开发，覆盖浙江移动、山东移动等客户，日均处理呼叫量超10万次

北京寰宇通桥国际教育 | 高级Java开发

2015年6月 - 2016年12月

- 参与公司留学申请办公系统，海外住家管理等系统研发

北京索思教育科技 | C语言&Java讲师

2012年4月 - 2015年6月

- 作为C语言二级考试讲师培训1000+学员,作为Java讲师培训学员近百名

专业技能

- **编程语言**：Java、Python
- **大语言模型（LLM）**：了解深度学习、神经网络基础、熟悉Transformer架构，了解模型蒸馏、模型LoRa微调训练
- **大模型应用架构设计**：熟悉常用大模型应用架构设计模式，构建高性能、低成本、安全合规的大模型应用
- **流行智能体平台工具**：熟悉Coze、Dify、摩搭社区等主流智能体工具平台

- **MCP**：熟悉MCP（Model Context Protocol）协议，熟悉使用FastMCP实现python版本的MCP server、client功能
- **RAG（Retrieval-Augmented Generation）**：
  - 熟悉Langchain、LlamaIndex生态实现RAG应用研发
  - 熟悉使用普通文本、结构化文档、PDF文档，CSV文档、表格文档数据处理
  - 熟悉父子索引、句子滑动窗口等索引优化
  - 熟悉图片等多模态数据嵌入检索
  - 熟悉FAISS、Milvus、Chroma等向量数据库使用
  - 熟悉检索前查询问题重写、分解、扩展等优化，检索后重排、压缩、校正等优化，提供高效准确的检索结果
  - 熟悉使用RAGAS、DeepEval、TruLens等工具评估RAG系统
- **Agent应用开发**
  - **LangChain**：熟悉使用LangChain生态构建Agent应用
  - **LangGraph**：熟练使用LangGraph进行复杂Agent流程编排，支持多步骤任务分解与状态管理
  - **CrewAI**：使用CrewAI来构建智能体应用
- **微服务**：熟练使用spring生态构建JAVA分布式服务,熟练使用FastApi、Uvicorn、Gunicorn 部署Python后台服务
- **大数据**：Kafka、Elasticsearch、Flink、Hadoop
- **数据库**：MySQL、Redis、MongoDB 、Milvus
- **工具链**：Cursor、TongyiLingma、Linux、Nginx、Git、Jenkins、Docker、Gitlab等

项目经验

AutoPPT | 资深研发\项目负责人

2025年3月 - 2025年6月

- **项目介绍**: AutoPPT 是一个基于多模态 AI 技术的智能助手，旨在提升企业办公自动化流程的效率。它能够处理语音、图像和文本等多种输入形式，通过精确的提示工程和强大的自然语言处理能力，为用户生成高质量的 PowerPoint 演示文稿。ChatPPT 不仅简化了信息收集和内容创作过程，还通过自动化的报告生成和分析功能，帮助企业快速、准确地完成各类汇报和展示任务，从而显著提升工作效率和业务价值。
- **技术栈**： Gradio、python-pptx、python-docx、Langchain、Ollama、deepseek-r1、Qwen2.5、Stable Diffusio、Docker

AIOPS智能运维平台 | 资深研发\项目负责人

2021年3月 - 2025年5月

- **项目背景**：智能运维平台，帮助大型企业数据中心在海量监控数据中提前发现异常、定位 故障、预测风险，降低人工干预，提高企业IT系统可用性和运营管理效率，助力企业数字化转型。
- **技术栈**： React、SpringCloud、Nacos、Redis、Mysql、Kafka、Elasticsearch、Flink、DolphinScheduler、Hadoop、langchain、Qwen2.5、Milvus
- **个人贡献**：
  - 使用Langchain生态构建智能根因推荐Chat-Ops智能体：对接头部银行指标数据，使用RAG构建本地指标数据知识库支持，根据用户输入进行语义理解使用FunctionCall能力结合智能排障系统，给出推荐根因并对接事件处置实现问题修复
  - 参与公司Arcana Maas平台知识库模块研发，实现Langchain|LlamaIndex实现RAG嵌入、检索，使用Milvus向量数据库，使用 Ragas评估检索结果

    - 管理8人技术团队，参与智能运维平台架构设计与开发，使用springcloud生态构建分布式服务，实现分钟级别故障识别定位，给出故障根因对接处置系统完成故障修复,主导头部银行智能运维项目对接落地，完成项目合同要求需求并协助项目完成验收回款
- **项目成果**：服务头部银行客户，总合同金额超1000万元

全景智能客服中心 |研发负责人

2017年1月 - 2021年3月

- **项目背景**：为运营商提供云端智能客服系统，并支持本地私有化部署，支持机器人外呼、人机协同等智能场景
- **技术栈**：Vue、SpringCloud、Nacos、MongoDB、MySQL、socket-io、Redis、Kafka、FreeSwitch
- **个人贡献**：
    - 管理20人技术团队，完成分布式高并发外呼系统设计、研发，集成中科院基于规则的语义NLP引擎实现自动外呼机器人
    - 负责浙江移动智能外呼项目落地，实现日均处理通话量10万次
- **项目成果**：累计对接运营商客户近10家，年营收超近千万

行业短信发送平台 | 研发负责人

2019年10月 - 2021年3月

- **项目背景**：为企业提供短信验证码、营销通知等服务，支持日均800万条发送量，峰值达2000万条
- **技术栈**：Vue、SpringCloud、Netty、MySQL、Redis、Kafka、Elasticsearch
- **个人贡献**：
    - 主导系统架构重构，采用异步消息队列+分布式重构，提升系统处理能力,完成项目营收近千万

教育背景

河北联合(华北理工)大学 | 自动化专业 | 本科
2010年9月 - 2014年6月