

## 基本信息

姓名: 王少博  
电话: 15901220820  
邮箱: wshaobo1998@163.com  
籍贯: 河北保定

年龄: 27 岁  
毕业院校: 北京科技大学 (QS 前 500)  
学历: 本科



## 个人优势

- 以终为始, 价值驱动, 拥有强烈的主人翁意识与产品思维, 坚持从业务的根本痛点出发, 以终为始地设计技术方案, 追求创造可被清晰量化的商业价值。
- 极速学习, 拥抱前沿, 对 AI 领域的前沿技术抱有极大的热情与好奇心, 具备在短时间内将前沿理论转化为高可用工程实践的能力。
- 深度钻研, 系统性解决问题, 擅长运用第一性原理进行思考, 系统性地调试与根因分析。

## 技术栈

- 大模型: 熟练掌握 langchain/langgraph 框架, 熟练使用 Dify、coze、n8n 等低代码平台快速搭建产品 demo, 掌握 RAG、agent、提示词工程、模型微调方法, 熟练掌握 Milvus、Chroma 等向量数据库, 熟悉 MCP 协议
- 深度学习: 熟练使用 Pytorch 深度学习框架, 了解 TensorFlow, 熟练掌握多头 self-attention 机制、Transformer 架构、BERT 模型原理, 熟悉 RNN、LSTM、GRU 模型原理, 熟悉 CNN 相关图像模型原理
- 工程能力: 精通 python 语言, 面向对象编程思想, 熟练使用数据分析及可视化库 Numpy、Pandas、Matplotlib, 熟练操作 Linux 开发环境、Conda 虚拟环境搭建及 WinSCP、Putty 软件, 熟悉 Git 代码协作管理
- 模型部署: 熟练掌握 vllm、Ollama 推理框架、FastAPI 框架、Docker 容器化部署

## 项目经历

## 基于 LangGraph 的动态法律分析智能体

## 大模型应用工程师

2025.04 至 2025.07

项目描述: 为解决公众对法律知识获取门槛高、信息时效性差的痛点, 开发了一款先进的法律智能问答助手。项目核心是从传统线性 RAG 范式升级为基于 LangGraph 的动态智能体架构。

- 架构设计: 主导技术选型 (LangGraph, Milvus), 设计了包含意图路由、多工具协同 (本地检索/网络搜索)、信息综合的循环图工作流, 使系统能根据问题复杂度动态规划执行路径。
- 高精度检索: 部署并优化了 Milvus 向量数据库, 设计了结合 BM25 与稠密向量的混合检索策略, 并引入二阶段重排 (Reranker), 提升了检索的精准度。
- 数据工程: 针对法律文本特性, 设计并实现了结构化分块策略, 确保了知识库内容的完整性与高保真度。
- 模型优化: 通过 LoRA 对大模型进行指令微调, 专注于提升模型在法律领域的表达专业性和逻辑一致性。
- 业务效果: 与基线 RAG 模型相比, 答案准确率提升了约 30%, 检索召回率提高至 90%。端到端平均响应时间比暴力检索所有数据源的方式减少约 40%。

## AI 驱动故障诊断问答系统

## 项目负责人

2023.10 至 2024.03

项目描述: 面对地铁车辆运维中故障排查效率低下的核心痛点, 成功交付了一套创新的智能问答系统。该系统基于 RAG 技术, 为一线技术人员提供源自复杂技术手册的即时、精准的故障处理方案。

- 架构设计: 负责整体技术框架的选型与搭建, 战略性地选择 Dify 低代码平台, 以加速开发周期。
- 数据策略与执行: 负责项目中关键的数据处理流程, 整合了 3000 多页维修手册及故障案例库。
- 业务效果: 成功实现了运维团队工单处理效率提升 40% 的里程碑式目标, 将维修知识查询耗时从平均 20 分钟缩短至 3 分钟, 节省技术人员手册查阅时间约 200 小时/月, 年度节省人力成本约 80 万元。

## 工作经历

上海科大讯飞信息科技有限公司

大模型应用工程师 (实习)

2025.04-2025.07

北京京港地铁有限公司

技术部工程师

2020.09-2024.03