吴军鹏

工作年限： 5 年
语言： IELTS 6.5
联系方式： 19315108009
邮箱： 398550293wjp@gmail.com

教育背景

2018.07 - 2020.08	塔斯马尼亚大学	信息技术与系统	硕士
2012.09 - 2016.06	南京理工大学	化学工程与工艺	本科

工作经历

2024.03 – 至今	京亿联控（北京）科技有限公司	NLP 算法工程师
2022.10 – 2024.02	北京启辰云景信息科技有限公司	NLP 算法工程师
2020.09 – 2022.08	上海众调信息科技有限公司	大数据开发工程师

专业技能

- **Agent 开发框架：** 基于 LangGraph 框架构建 Agent 系统，熟悉 MCP 协议、A2A 通信协议 及 Function Call 架构，支持多 Agent 协同工作流设计与外部工具集成落地，同时熟悉 Dify、CrewAI、Coze 等工作流框架。
- **RAG 应用开发：** 基于 LangChain、BM25、Milvus、RagAS 等技术构建 RAG 系统，支持多模态文档处理（OCR 等），实现企业级智能问答与知识检索。
- **模型高效微调：** 熟练应用 LoRA、QLoRA 等微调技术，支持小样本场景下的模型参数优化，高效提升大模型表现。
- **提示词工程：** 熟练掌握 CoT、ToT、Reflexion 等高级提示词技术，优化模型输出与任务执行。
- **深度学习与 NLP：** 掌握 PyTorch、Transformers、FastText 等框架，熟悉 Bert 系列模型，具备神经网络模型设计与开发能力。
- **机器学习算法：** 熟悉 XGBoost、随机森林、K-Means 等算法，支持分类、回归与聚类任务。
- **模型优化和压缩：** 熟悉 模型量化、知识蒸馏、剪枝 等技术，显著降低模型推理延迟与资源消耗，保持高性能。
- **大数据技术栈：** 熟悉 Hadoop、Hive、Spark 生态及 Flink 实时计算框架，熟练使用 Redis、HBase、Kafka 等非关系型数据库与消息队列。
- **云计算平台：** 熟悉 阿里云 Dataworks、Maxcompute 等大数据平台，具备云环境下的开发与部署经验。
- **编程语言：** 具备扎实的 Python 编程能力，熟悉面向对象设计与开发规范。
- **其他编程语言：** 熟悉 Java 开发，掌握 SQL 语句（支持 MySQL 等关系型数据库），具备 Linux 环境开发与部署能力。

项目经验

项目名称：

Fat Uncle Rental 智能租车助手 2025.02-至今

技术架构：

LangChain+LangGraph+OCR+FastMCP+FastAPI+MinHash+LSH+Docker+Milvus+GLM-4-9B-Chat-Tool-32k（LoRA 微调）+MySQL+Redis

项目介绍：

针对 Fat Uncle Rental 公司的核心业务（常规租车、房车租赁、机场接送、景点包车及配套保险/附加服务）中，传统人工咨询单次耗时高（10-15 分钟）、订单转化率低、人工成本高的这些核心痛点，同时结合了 2025 年多 Agent 协同+组件化架构技术的趋势，由此主导构建了覆盖“全场景车型匹配-订单操作-合规校验（含房车专项资质）-导航服务（常规/房车适配）-路线规划（景点/机场定制）-附加服务”的全流程智能助手，以实现业务效率与转化效果的双提升。

核心职责：

- **负责全场景租车业务工具模块 + MCP 组件封装**：
  - 将业务拆分为了常规租车、房车租赁、机场接送、景点包车、附加服务 5 大类，每类拆解为 “只读工具”（车型查询、房车功能说明、机场接送站点、景点路线库、附加服务清单）与 “敏感操作工具”（订单创建、房车资质审核、接送调度确认、包车订单修改、保险购买），同时按业务功能分类封装 MCP 微服务，通过 MCP 实现与 Agent 的解耦调用，优化后初期组件复用率达 60%。
- **负责双层多 Agent 架构设计**：
  - 设计实现了 “A2A 宏观调度 + LangGraph 微观执行” 的混合架构。在宏观层面，利用 A2A 协议 和 Redis Streams 构建分布式消息总线，实现 Supervisor Agent 对 5 个子业务单元（房车/接送机等）的异步任务分发与状态管理。
  - 在微观层面，为每个子 Agent 设计了基于 LangGraph 的 ReAct 状态机，定义了从“思考-工具调用-结果反思”的完整逻辑闭环，并用幂等键 + DB 状态单调推进确保复杂业务流（如资质审核+库存检查）的原子性执行。
- **负责 LLM 与 MCP 的协同优化**：
  - 对 GLM-4-9B-Chat-Tool 进行全场景租车 LoRA 微调，构建了可扩展的知识更新链路：可以拉取外部信息（北美房车资质要求、机场航班动态、景点开放信息等）写入知识库（RAG/MySQL），同时结合 4000 + 华人租车全场景术语，使得中文指令意图识别准确率从 82% 提升至 97%（实测）。
- **负责 MCP 组件业务落地**：
  - 支付模块集成支付宝 MCP 和银联 MCP；导航模块接入高德地图 MCP 海外版（适配房车尺寸）；景点包车模块集成景点路线 MCP（覆盖华人热门景点）；新增机场接送调度 MCP（对接航班查询 API）、保险 MCP（接入北美租车保险服务商）等。
- **负责分层记忆架构落地**：
  - 通过上下文窗口+ LangGraph 的 checkpointer+RAG 向量库来实现记忆的分层管理。
- **负责数据链路优化：**
  - 使用 MinHash+LSH 替代传统去重方案，结合数据清洗流程完成了常规租车 / 房车 / 包车订单和景点文档的去重，实现常规交规 + 房车特殊交规、景点开放时间与接送调度的关联，提升了检索精准度。

成果：

- 中文指令识别准确率 97%（全场景 1000 + 样本实测），驾照 + 房车资质合规校验误判率 0.3%（1000 + 样本实测），支付宝 MCP 跨境支付成功率 99.1%（500 + 笔实测）。
- 项目完成预计覆盖 85% 常见咨询意图（按意图库覆盖率统计），平均处理时长从 10–15 分钟降到 4 分钟（基于脚本回放/压测）。
- MCP 组件复用率达 60%，多 Agent+MCP 的解耦设计已完成北美 3 个核心州业务适配（实测）。

项目名称：

中医药知识图谱智能问答系统 2024.03-2025.01

技术架构：

Python+LangChain+DeepSeek+Neo4j+Faiss+LangGraph+Streamlit+Requests+BeautifulSoup+Git

项目介绍：

面向中医药连锁门店业务场景，为了解决店员专业知识查询繁琐、知识与门店商品脱节、服务效率偏低等痛点，而构建了融合知识图谱、大模型与业务系统的智能问答模块。它搭建了“专业知识问答+精准商品推荐+业务数据联动”一体化服务，助力门店提升服务质量与运营效率。

核心职责：

- **负责项目系统架构设计和搭建：**
  - 参与设计并搭建贴合中医药连锁门店的全流程系统架构。包含多源数据采集、知识图谱构建、多智能体交互和业务联动四大模块，实现了从用户自然语言输入到知识反馈和商品推荐的全闭环。
- **负责多源数据采集与处理：**
  - 通过 Requests+BeautifulSoup 爬取公开中医药专业知识和行业资讯，同时对接连锁门店内部商品库、销售记录等业务数据，并且完成数据清洗与结构化解析，为知识构建提供支撑。
- **负责数据中实体和实体以及关系的抽取：**
  - 通过 LangChain+DeepSeek 大模型，来抽取中医药领域核心实体（中药、病症、方剂等）及关联关系（功效、配伍、禁忌等）。并且通过 LoRA 微调优化专业术语识别的精度。然后，依托 Neo4j 构建结构化知识图谱，搭配 Faiss 搭建向量索引，实现语义检索与模糊匹配。
- **负责 LangGraph 多智能体工作流的搭建：**
  - 利用 LangGraph 搭建多智能体工作流。完成了意图识别、实体匹配、Cypher 查询语句的生成与其合法性的校验，并且精准区分用户中医相关/非相关问题，提取用户问题中的核心实体并匹配知识图谱节点，同时与 MySQL 业务库关联来实现知识解答与门店商品推荐的无缝衔接。

成果：

- 构建包含 3 万+中药、症状、功效、疾病实体及 5 万+知识关系的 Neo4j 知识图谱；语义检索召回率较关键词检索提升约 18%，实现 95%常见中医问答场景覆盖。
- 用户测试整体答案满意度提升至 85%+，可高效支撑中医药连锁门店一线服务需求，减少人工查询依赖，提升店员服务效率与顾客体验，助力门店沉淀核心数字资产、优化运营决策。

项目名称：

舍弗勒工业级文档检索系统（二期） 2023.05-2024.02

技术架构：

LangChain+BERT+PEFT(LoRA)微调+FastAPI+ReRanker+BM25+汉明算法+Milvus+Mysql+Redis+SGlang

项目介绍：

依托项目一期的工作，结合 2023 年大模型与 RAG 技术爆发趋势，升级原系统为 RAG 系统，并解决一期人工标注不准、权限控制粗放、无智能总结的痛点，打造自然语言交互的智能检索系统。

核心职责：

- **负责 Llama 2-7B Base 模型 LoRA 微调：**
  - 基于 PEFT 完成意图识别任务训练，结合岗位信息与查询内容判定：图纸 / 合同需求，还是选择 RAG 流程，两项任务准确率分别达 83%、88%。
- **主导 系统入口的 Bert 模型微调，实现二分类任务：**
  - 精准区分用户问题是否需大模型回答，搭建专属回复语料库，大幅减少无效模型调用。
- **开发实现三层记忆机制：**
  - 构建瞬时 / 短期 / 长期记忆体系，通过异步检索 + 动态合并精准控制上下文长度，适配大模型 Token 上限约束，完成大模型上下文信息高效构建。
- **优化查询性能：**
  - 对向量检索逻辑和缓存策略进行优化，基于 Redis 缓存高频查询结果，实现系统 P99 延迟≤2s，

稳定支撑日均 2000 + 次并发请求。

- **负责提示词体系开发：**
  - 定制化开发采购、电力工程师、销售等 8 大部门、30 + 岗位专属提示词体系，针对性提升大模型输出内容与业务场景的相关度。
- **实现权限标量校验机制：**
  - 基于文档公开 / 内部 / 机密标签关联岗位权限矩阵，确保文档使用等级匹配用户权限，私密文档仅做向量匹配、不送入大模型
- **优化标注准确性：**
  - 针对一期人工标注偏差问题，通过 Chinese-CLIP 实现图纸图文跨模态特征提取，筛选 1 万份模糊样本开展二次校验，将标注准确率提升 20%，F1 值达 84%。

成果：

- 核心性能显著提升，整体查询响应时间从 5 秒压缩至 3 秒内，意图识别与文档检索综合准确率超 82%，业务问题回答满意度达 80%，高效支撑业务决策闭环。
- 合同/图纸核心信息提取准确率达 86%；三层记忆机制有效保障对话连贯性，大幅降低人工查询与核验成本，提升业务流转效率。
- 岗位定制提示词使大模型输出相关度提升 12%，权限校验零误判。

项目名称:

舍弗勒工业级文档检索系统（一期） 2022.10-2023.04

技术架构：

Transformers+PyTorch+OCR+BM25+Milvus+MySQL+FastAPI+SimHash64 +汉明算法+TextTiling

项目介绍：

国内业务线的 ERP 系统中沉淀了海量合同、工业图纸、合规文件等非结构化文档，传统人工翻阅的检索模式效率低下，单次查询耗时达 5-10 分钟。项目一期以文档数字化基建为核心目标，完成 10 万份文档的电子化处理与向量化存储，搭建基础检索功能体系，为核心业务决策提供高效、精准的文档数据支撑。

核心职责：

- **主导 10 万份非结构化文档电子化与标注全流程：**
  - 采用 PP-OCRv3 实现扫描件 / 图片类文档文字提取，解决 PDF/PNG/CAD 等 8 类格式兼容性问题；优化 OCR 后数据清洗规则，修正字符错误、格式错乱问题，将文档有效率从 60% 提升至 83%；牵头人工标注文档类型、核心关键词、概要及存储路径，通过 SimHash + 汉明算法完成文档去重，剔除重复文档 8000 + 份。
- **负责文档预处理与语义向量化：**
  - 基于 TextTiling 语义切分方案，完成长文档的语义连贯片段拆分；基于 Sentence-BERT 模型结合施耐德行业语料开展增量微调，生成 768 维向量实现文档语义编码，让模型深度适配行业专属语料特征。
- **搭建混合检索存储与检索体系：**
  - 构建 Milvus+MySQL 分布式存储架构，开发关键词标量 + Dense+Sparse 向量混合检索核心代码，实现查询语句与文档的精准语义匹配检索，落地基础检索功能。

成果：

- 完成 10 万份文档全量数字化处理，构建包含 23 万条语义片段的基础知识库，系统可稳定支撑日均 300 + 次检索查询。
- 大幅优化检索性能，响应时间缩短至 5 秒内，文档语义匹配准确率达 78%，核心业务文档召回率超 80%。
- 实现 DFS 文档系统路径精准回显，核心业务人员文档查询效率提升 90%，为二期 RAG 技术落地夯实

高质量数据基础。

自我评价

- 5 年 NLP 与大数据开发经验，具备完整的项目生命周期管理能力；
- 精通大模型和大数据技术栈，善于解决复杂业务问题；
- 具备良好的编程习惯和代码质量意识，擅长封装常用工具类；
- 团队协作能力强，乐于分享技术经验，融入团队文化；
- 持续学习前沿技术，对新技术保持强烈求知欲；
- 具备较强的问题解决能力，能够快速定位和解决生产环境问题；