徐俊豪

电话：13566540167　邮箱：xujie_search_sth@163.com　性别：男



职业技能 PROFESSIONAL SKILL

- Agent 架构：熟练使用 LangChain 框架及多智能体协作，设计意图识别、任务拆解与通信逻辑。
- MCP 协议：通过 MCP Server 实现 LLM 与业务系统标准化连接。
- RAG 优化：掌握 RagAS 模型评估、文档切片、向量存储（Milvus）及检索优化。
- Prompt 工程：运用 CoT、Few-Shot、ReAct 提升模型指令遵循度。
- NLP 任务：精通语义相似度计算、关键词提取、NER 等。
- 深度学习：使用 PyTorch/TensorFlow 开发文本分类、序列标注模型。
- 传统机器学习：应用决策树、SVM、XGBoost/LightGBM 算法。
- 编程功底：Python 数据结构算法，异步编程（asyncio），Linux/Shell 脚本。
- 存储与缓存：MySQL 关系型数据建模，Redis 缓存加速、分布式锁。
- 特征工程：数据清洗、转换、特征提取，结果可视化（Matplotlib/Seaborn）。

工作经历 WORK EXPERIENCE

浙江沣华医药科技有限公司　NLP算法工程师　**2023.09-至今**

- 探索与应用前沿NLP技术：深入研究包括大型语言模型（LLM）在内的先进自然语言处理技术，并成功将其应用于医疗健康领域，提升了信息处理效率及准确性达30%。
- 模型架构设计与优化：主导开发复杂深度学习模型，通过引入创新方法有效解决了梯度消失和过拟合问题，显著提高了模型性能，准确率提升至95%以上。
- 跨部门协作与项目落地：与产品、业务团队紧密合作，推动NLP技术在搜索优化、内容自动生成等多个实际场景中的应用，加速了产品迭代周期，缩短了新功能上线时间约20%。
- 数据挖掘与分析：对大量文本及用户行为数据进行深度分析，实现热点话题自动识别、关键事件抽取等功能，为决策提供了强有力支持，助力公司做出更精准的市场策略。
- NLP基础组件建设：根据具体需求定制开发了一系列核心NLP工具，如自动摘要生成器、高效分词器等，极大丰富了公司技术栈，支撑了多个重要项目的顺利实施。

项目经历 PROJECT EXPERIENCE

基于A2A协议的智能医药代表助手　**2025.08-2026.01**

****项目背景：****

医药代表外勤时需要频繁切换不同系统和工具，功能分散、数据割裂导致工作效率低下。构建多智能体协作助手，通过A2A协议实现Agent间协作，通过MCP协议实现工具与模型解耦。

****技术栈：****

Python Asyncio、LangChain、Python A2A、OpenAI API、MCP协议、FastMCP、Pydantic、Streamlit、MySQL、SQLite、ChromaDB、BERT、Requests、BeautifulSoup、Schedule

工作内容：
- **多智能体架构设计**：基于LangChain框架搭建Tool-Calling Agent与链式调用逻辑，集成langchain_openai组件对接OpenAI API，结合Python A2A协议构建多智能体协作网络，实现Agent间非阻塞异步通信与任务分发，支持医学知识Agent、客户查询Agent、拜访记录Agent、竞品情报Agent等专业Agent，客户答疑响应效率提升40%
- **MCP协议实现**：采用mcp库开发FastMCP服务，实现工具（RAG检索、nl2sql、CRM写入、数据采集）与大模型解耦，依托Pydantic完成底层数据校验，降低模块依赖耦合度，系统维护成本减少30%，支持工具灵活扩展
- **前端开发**：基于Streamlit开发对话式前端界面，通过st.write_stream实现流式输出，st.session_state管理会话状态，嵌入自定义CSS/Markdown优化界面样式，支持移动端适配，支撑日均200+医药代表会话需求，操作满意度达4.7/5.0
- **混合存储设计**：搭建MySQL+SQLite+ChromaDB混合存储体系，MySQL存储天气及客户关联数据，SQLite负责本地订单、用户信息轻量存储（支持离线使用），规划ChromaDB用于RAG向量存储，根据数据特点选择存储方式，检索响应速度提升30%
- **数据采集流水线**：构建数据采集流水线，通过Requests调用外部API，结合Schedule实现天气数据定时爬取，集成BeautifulSoup/Lxml储备网页解析能力，保障多源数据（内部数据库、第三方API、互联网）实时更新，数据时效性提升60%
- **性能优化**：基于Asyncio实现并发任务处理，优化多Agent协作与数据爬取效率，搭配日志体系定位性能瓶颈，使单任务处理时长缩短50%，满足高并发业务场景需求（200+QPS）

项目成果：
覆盖90%+医药代表日常工作场景；通过MCP协议与A2A协作实现业务模块灵活扩展，政策检索准确率达93%，订单处理效率提升35%，客户答疑响应效率提升40%，大幅降低人工操作成本。

医药市场情报RAG分析系统 2025.05-2025.12

项目背景：
医药代表外勤时查询产品信息需查阅纸质资料耗时5分钟以上，内部资料分散在各个部门，第三方数据（米内网、药智网）与内部数据割裂。构建RAG系统整合多源数据，支持快速查询。

技术栈：
RAG、LangChain、Milvus、BGE-M3、FastAPI、uvicorn、PyTorch、Transformers、BERT、Redis、MySQL、RapidOCR、DashScope API、RagAS、定时任务调度

工作内容：
- **多源数据融合**：整合内部员工上传文档（PDF/DOCX/PPT/图片）和第三方API数据，使用RapidOCR处理扫描文档，ChineseRecursiveTextSplitter进行父块（1200字符）+子块（300字符）分层切分，BGE-M3向量化，支持跨数据源检索，查询准确率达92%
- **混合检索优化**：设计BM25快速匹配（Redis缓存，命中率85%以上）+向量语义检索（稠密+稀疏向量）的混合架构，通过RRF算法融合结果，BGE-Reranker重排序，复杂问题召回率提升35%，平均响应时间从5秒降到2秒内
- **多策略检索路由**：基于BERT实现查询分类（产品咨询/竞品分析/客户沟通/市场情报），根据分类结果自动选择直接检索、子查询分解、HyDE、回溯检索等策略，关键问题准确回答率提升22%
- **数据管理流程**：实现文档上传审核模块，自动提取元数据、OCR识别、内容质量评估，审核周期从2-3天缩短到1天；构建第三方数据同步模块，定时任务调用API/爬取数据，清洗去重标准化后批量入库，支持增量更新
- **领域优化**：构建医药领域专业术语词典（药品名、疾病名、化学式），在文本预处理和向量化阶段进行实体识别和标准化，专业术语识别准确率从70%提升到95%，检索准确率提升25%

项目成果：
覆盖90%+医药代表工作场景，90%高频问题毫秒级响应，查询效率提升10倍，数据覆盖率达85%以上，查询满意度88%，成为公司核心销售工具。

企业员工管理智能问答系统 2023.10-2024.03

项目背景：
面向公司管理层和HR部门需求，构建整合员工信息、人事政策、历史决策记录的企业员工管理智能问答系统。解决员工信息查询困难（需登录多个系统耗时10分钟以上）、人事政策理解偏差、历史决策难追溯、数

据分析效率低等痛点，提升管理效率和决策质量。
技术栈：
RAG、LangChain、Milvus、BGE-M3、FastAPI、PyTorch、Transformers、MySQL、Redis、BERT、数据分析、权限控制、RagAS评估
工作内容：
- 构建多源数据融合架构，整合HR系统结构化数据（员工信息、绩效、考勤）和非结构化文档（简历、政策、会议纪要），使用BGE-M3进行文档向量化，支持跨数据源检索，信息查询准确率达92%
- 实现混合检索机制，结合SQL精确查询（员工ID、部门等）和向量语义检索（模糊查询、政策解读），通过智能路由和结果融合，复杂管理问题的召回率提升38%，平均响应时间控制在2秒内
- 设计分级权限控制体系，基于角色（老板/HR/部门经理）实现数据访问权限控制，敏感信息自动脱敏，在检索前进行权限预过滤，保障数据安全，获得管理层信任
- 实现员工数据分析功能，整合RAG检索与数据分析模型，支持员工留存分析、绩效趋势分析、培训效果评估等，自动生成管理洞察报告，决策支持准确率达85%
- 构建历史决策记录库，整合会议纪要、决策文档、案例库等，支持时间维度检索和相似案例匹配，历史决策查询效率提升50%
- 通过Redis缓存高频查询（员工基本信息、常见政策），响应时间从10分钟优化到1分钟以内，日均查询量2000+，用户满意度达4.7/5.0
项目成果：
成功交付员工管理系统，覆盖公司90%+人事管理查询需求；管理层查询员工信息和政策的时间从平均10分钟降低到1分钟，效率提升90%；员工信息查询准确率达92%，数据分析自动化后效率提升240倍；实现分级权限控制，敏感信息访问零泄露，获得老板和HR部门高度认可，成为公司核心管理工具。

教育经历 EDUCATION

绍兴文理学院元培学院 本科 **2018.09-2023.06**

自我评价 SELF ASSESSMENT

- 背景经历：拥有近半年的NLP算法工程师经验，在浙江沣华医药科技有限公司期间，专注于将前沿NLP技术应用于医疗健康领域，成功提升了信息处理效率及准确性达30%，并主导了多个复杂深度学习模型的设计与优化工作。
- 专业技能：精通大型语言模型（LLM）等先进自然语言处理技术的应用，擅长解决梯度消失和过拟合问题，显著提高模型准确率至95%以上；同时具备数据挖掘与分析能力，能有效识别热点话题和关键事件，为决策提供支持。
- 个人优势：展现出较好的技术创新能力与跨部门协作能力，具备清晰的语言表达能力，能精准对接跨部门需求、高效传递技术思路，推动新技术快速落地，缩短产品迭代周期约 20%；具有较强的逻辑思维能力和项目管理经验，善于统筹策划，确保项目高效推进。