张新生

电话：18394331838 邮箱：2113561932@qq.com 年龄：22
职位名称：大模型应用开发工程师

教育经历

兰州文理学院 软件工程 本科 2020-09 ~ 2024-06
GPA：3.55/4.0
主修课程：Python程序设计、Java企业级应用开发、数据结构、操作系统、Linux操作系统及应用等

荣誉奖项

第十五届全国大学生数学竞赛 省级一等奖
团体程序设计天梯赛 省级一等奖
2022-2023年度三好学生
2022-2023年度国家奖学金

实习经历

携程集团 2025-02 ~ 2025-06
大模型应用算法工程师
主要工作：

- 负责对Agent生成的数百条Bad Case进行归类分析，定位出“知识过时”与“意图遗忘”两大核心问题，为模型迭代提供了关键数据洞察。
- 完成RAG“答案忠实度”评估模块的开发，将该环节的人工评测效率提升超10倍，并量化验证了新策略能将准确率提升8%。
- 针对特定业务场景设计并验证多种Prompt方案，其中“角色扮演”指令经A/B测试可将用户满意度评分有效提升0.6分（5分制）

项目经历

JobPilot-基于多 Agent 协作的AI自动化求职系统 2025-03 ~ 2025-06
Agent开发工程师 https://github.com/lucky-carpZ/find_job.git
项目描述：为解决求职流程繁琐、效率低下等痛点，**独立设计并从零到一构建了AI自动化求职系统。**该系统通过多个智能 Agent 协作，可自主完成职位搜索、简历-岗位精准匹配、个性化求职信生成及自动投递的全链路流程，旨在大幅提升求职效率与成功率。
技术栈：Python | LangChain | Selenium | DeepSeek API
核心贡献：

- **架构设计：**主导设计并实现了**分层协作式多 Agent 架构。**通过中心协调器（Orchestrator）调度专用 Agent，确保了系统高稳定性、高可控性与高扩展性。
- **智能决策：**利用 LangChain 和 LLM 赋予了 Agent 的**任务规划、分析推理与文本生成**能力，使其能够理解复杂指令并产出高质量内容。
- **工程实践：**实现 LLM-规则引擎双轨容错机制以支持无缝回退，并通过持久化 Cookie 与状态记录解决自动化登录及任务幂等性难题。

Mind-具备推理链的本地RAG Agent 2024-08 ~ 2024-12
RAG应用开发工程师 https://github.com/lucky-carpZ/RAG.git
项目描述：为探索在**数据隐私和低成本**场景下部署智能对话系统的可行性，设计并构建了一套本地化部署的、具备高级推理能力的 RAG 应用。该系统不仅能基于本地知识库进行精准问答，还能**调用外部工具**获取天气实时信息，并通过**可解释的思考链**展示其决策过程。
技术栈：Python | Streamlit | LangChain | Ollama | FAISS | Agno
核心贡献：

- 设计并实现 ReAct 推理范式，通过显式思考链 <think> 提升 Agent 行为的**可解释性与可控性。**
- 将外部 API 封装为**可调用工具 (Tools)**，使 Agent 能突破自身知识局限，获取实时信息。
- 基于 Ollama 与 FAISS 构建了完整的本地化 RAG 管道，实现私有数据环境下的精准问答。
- 为解决传统 RAG 的语义割裂问题，设计并实现了**基于摘要增强**的两阶段检索算法，显著提升了复杂问题的回答质量。

个人技能

- 具备基于 Ollama 框架进行大模型（DeepSeek, Qwen）本地化部署与集成的实践经验。
- 熟练使用 LangChain 构建复杂的 Agent、实现 RAG 管道以及封装工具链。
- 深入理解并实践过多种 Agent 架构，包括ReAct 推理范式、多 Agent 协作系统、工具使用等。
- 掌握从文档处理、向量化 (FAISS)、到检索与上下文注入的全流程 RAG 技术。
- 能够设计和优化高质量的 Prompt，以引导模型完成特定任务并提升输出质量。
- 语言能力：CET-4，有良好的读写能力。