

肖小平

15064225750 | xiaoxiaoping2023@163.com | xxping.xyz | github.com/XP-Xiao

男 · 硕士应届 (2027 届) · 目标岗位: AI 应用开发工程师



教育经历

青岛理工大学 · 电子信息 · 硕士

2024.09 - 2027.06

Disentangled VIB with Hyper-Cluster for Few-Shot Class-Incremental Audio Classification (SCI Q1 · 一作 · 2026.06 录用)

青岛理工大学 · 计算机科学与技术 · 本科

2020.09 - 2024.06

实习经历

路觅教育南京分公司 (Lumist) · AI 应用研发实习生

2026.04 - 2026.06

AI 教学服务平台 | 师资智能匹配、简历 AI 解析与课堂质量风控

- 背景与目标: 公司 AI 团队围绕导师全生命周期构建 LLM 能力, 解决师资匹配效率、招聘筛选与教学质量监控问题; Java 主后端 + Python AI 服务 (LangGraph 编排, Milvus 混合检索 + MySQL, Qwen 系模型)。
- 我的职责: 主导师资智能匹配、简历 AI 解析与课堂分析三条 AI 功能的全链路实现。
 - 师资智能匹配-检索链路: 构建"资料录入 → RAG 检索 → LLM 推荐理由 → 人工确认"workflow; Milvus 混合检索 (expr + dense/BM25 + RRF + Rerank) 与偏好映射对齐语义; 三态状态机失败自动降级人工。
 - 简历 AI 解析落地: 多格式简历统一转图 → 千问 3.6-Plus 多模态一步完成识别与归纳, 直出结构化字段与研判报告 (匹配度/风险预警/课程建议); 异步状态机 + 失败重试, 人工审查定稿后才向量化入 Milvus。
 - 课堂分析与风控: 构建"云录制 → 中英 ASR 转写 → 分段聚合"异步管线, 课堂总结与环节回顾可视化, 分析结论关联转写原文可回放核实, 违禁检测分级归类输出学生/品控双视图。
 - AI 质量评测与迭代: 以 200 条历史确认为正例离线消融, Recall@5 由 63% 提升至 81% (MRR 同步提升); 解析线即时护栏 + ai_raw 全量 diff 校验; 课堂线 quote 反查验引用, 以教盲评分歧率校准品控口径。

项目经历

自进化 Agent 系统 · github.com/XP-Xiao/SproutAgent

2026.06 - 2026.08

- 项目简介: 基于 Python 实现极简自进化 Agent, 以 9 个原子工具与约 100 行 Agent Loop 为核心, 通过 LLM 工具调用获得终端、文件、浏览器等系统级控制, 任务执行路径自动固化为可复用 Skill 实现自我进化。
- 个人贡献: 独立实现 Agent Loop、多工具调度与 Skill 沉淀机制的框架核心, 解决长任务稳定性问题。
 - Agent Loop 工程化: 实现 LLM 循环调用与工具结果回注, 通过每 10 轮重置工具描述控制上下文膨胀; 设计 EXITED / 任务完成 / 超轮上限三级退出与 next_prompt 链式驱动, 保证长任务稳定推进。
 - 工具与记忆体系: 9 个原子工具 (代码执行/文件补丁/浏览器控制/人工确认) 并行调度; L0-L4 分层记忆 (元规则/极简索引/全局事实/任务 Skill/会话归档), 记忆只源自成功的工具调用结果, 索引层控制规模防膨胀。
 - Skill 自进化: 实现任务执行路径自动固化为可复用 Skill 的机制, 遵循"不预载技能、靠执行沉淀"的设计哲学, 使用越久能力越强。
 - 评测体系: 固定任务集 9 题 (开源基准子集 7 + 自建 2), 隔离工作区运行, grader 自动判分并汇总报告; 成功率 89% (8/9), 平均 3.1 万 Token、6 次调用; 沉淀 Skill 后同集第二遍 Token -69%、耗时 -43%。

Agentic RAG 知识库问答 · github.com/XP-Xiao/LuminaRAG

2025.12 - 2026.03

- 项目简介: 基于 LangGraph 与 FastAPI 构建多轮 RAG 问答系统, 支持 MinerU 结构化文档解析入库、混合检索、查询改写、工具调用与 HITL, 覆盖文档上传到答案生成全链路, Milvus + PostgreSQL/Redis + Docker 部署。
- 个人贡献: 独立设计检索链路、问答路由与评测体系, 以证据分级路由与离线评测降低幻觉。
 - 问答路由: 复杂度快路径分流 (简单直答省 token 延迟 / 复杂 Send 扇出子 Agent 并行合成); 证据评判五路路由 (直接作答/查询改写/澄清追问/范围选择/无知识), 证据不足自动改写或澄清; LLM 给建议、代码做裁决。
 - 三级分块与混合检索: L1/L2 存 PostgreSQL、L3 入 Milvus (bge-m3 + 服务端 BM25 双索引); 双路召回 + RRF 融合 + Rerank, 同父块命中 ≥ 2 自动合并; 失败自动降级 (纯稠密兜底)。
 - HITL 人机协同: 复杂/歧义问题中断追问, 恢复点存图外, 图外直线恢复不重跑全图; 恢复后针对性二次检索补新信息, 预算继承; 支持多轮追问自然收敛。
 - 评测体系: 自建 20 问 smoke 评测集 (开源数据集改编 + 10 篇自建语料), ingest → run → score → clean 全流水线, 独立 collection 跑全链路、主库零污染; 用例通过率 90%、路由准确率 95%、Recall@5 达 94%。

技能与其他

编程语言: Python, TypeScript, JavaScript, SQL

AI 工程: RAG, Agent, 工作流编排, Prompt 工程, Function Calling, LangChain / LangGraph, MCP, 评测与回归

工程基础: FastAPI, Vue / React, Docker, Git, Linux, Jenkins

数据与工具: MySQL, PostgreSQL, Redis, Milvus, Apifox, Cursor / Claude Code 等 AI 编程助手

语言: 英语 CET-4; SCI Q1 期刊论文一作 (独立完成英文写作), 可流畅阅读英文技术文档