

王增焕



📍 深圳 ✉ w905840774@gmail.com ☎ 17665252185
🌐 wzhnbsixsixsix.github.io 🗯 wzHandsomeboy92 📄 wzhnbsixsixsix

教育背景

2022.09-2024.07 广东工业大学|本科 | 计算机科学与技术|广州, 中国

主修课程: 数据结构、计算机网络、操作系统、计算机组成原理、Java

2024.09-2026.06 University of Exeter | Bachelor | Computer Science | Exeter,UK

主修课程: Data Science、Deep Learning、Web Development、Database Theory

实习经历

深圳华大基因科技有限公司 | 前沿技术研发部 (X) | LLM(Agent) 后端开发实习生 2025.7-2025.11

- **核心系统与高性能 AI Agent 构建:** 基于 Dify 主导实时检索 workflows、问诊助手和血型 Agent 开发, 深度优化 Pubmed API 检索流程 (10 分钟缩至 1 分钟), 并完成 Milvus 向量数据库部署, 实现 RAG 系统落地。
- **前沿血型文献助手 (PaSa 项目) 研发:** 部署并且复现论文实验; 按论文标准开发多个数据生成脚本, 构建高质量数据集 (10w+)。
- **系统评估、多个论文实验复现部署以及轮询接口开发:** 搭建 LLM 评估 workflows 和并发测试脚本, 解决知识图谱并发及超时瓶颈; 独立完成开源 AI 科研发现项目 SciAgents 部署以及公司自研知识图谱的迁移、开发轮询接口并且迁移至 Dify。

杭州胜事影视文化有限公司 | AIGC 内容部 | AI Agent 开发 2025.5-2025.7

- **AI 小说转剧本工作流:** 基于 Coze 平台开发百万字小说改编的 AI Agent 流程, 涵盖导入、篇幅规划、剧本生成与原文匹配评估, 显著提升动态漫/短剧制作效率
- **LumiLove 智能陪聊系统:** 设计并实现服务架构, 前端使用 Next.js + React + Tailwind, 后端采用 Spring Boot + MySQL, RAG 对话由独立 FastAPI 服务处理。
- **RAG 记忆系统:** 基于 LangChain 实现向量化对话历史检索 (ChromaDB + SentenceTransformers), 支持上下文注入、角色记忆与 token 成本监控。
- **系统架构与部署:** 采用前后端分离 + 微服务架构, 部署于 AWS EC2/Amplify, 使用 JWT 实现权限控制, 结构化数据由 RDS (PostgreSQL) 托管。

项目经历

InsightFlow — 智能投研/竞品分析多智能体系统 (RAG + MCP + Memory) [InsightFlow](#)

2026.02 – Now

项目简介: 面向“行业情报/竞品分析”场景的自动化分析团队: 输入领域或公司/框架名, 系统自动完成联网检索+内部知识库检索、对比分析与可视化, 输出结构化深度报告。

技术栈: Python, FastAPI, AgentScope (Message Hub), Hybrid RAG (BM25+Vector), Milvus/ChromaDB, Reranker (BGE Reranker), MCP Server/Client, Matplotlib

- **多智能体协作:** 以 Manager/Researcher/Analyst/Operator 角色分工, 基于 Message Hub 设计任务拆解、结果汇总与质量审查的闭环流程;
- **Hybrid RAG 与重排:** 实现关键词 (BM25) + 向量检索的混合召回, 并引入 Rerank (BGE-Reranker) 提升证据命中率与引用可追溯性;
- **结构化记忆:** 短期使用 DictDialogMemory, 长期将“实体/结论/证据链接”写入向量库 (Long-term Memory), 支持跨会话对比与复用;
- **Skills 工具化:** 封装 Search/Fetch/Clean (HTML 清洗) 等工具调用接口, 支持动态网页处理 (可选 Play-

wright/Selenium) 以增强数据覆盖;

- MCP 解耦执行层: 搭建本地 MCP Server (文件/SQLite/内部数据), Operator 通过 MCP 写入分析结果表并落盘生成 Markdown 报告与对比图表。

LifeMirror (人生镜像) — PKM + Agentic Workflow 的全息人生伴侣系统 + 可以外包给 MCP (多智能体 + 记忆/RAG + MCP) 2026.01 – Now

[LifeMirror](#)

项目简介: 以“第二大脑/人生伴侣”为目标, 通过日记与目标 (KPI) 持续学习用户偏好与情绪模式; 结合工具调用与计划执行, 实现从“被动记录”到“主动规划”的闭环。

技术栈: Python, AgentScope (Workflow/MsgHub), FastAPI, Hybrid RAG (BM25+Vector), Milvus/ChromaDB, SQLite/PostgreSQL, MCP Protocol, Scheduler/Push

- 分层多智能体架构: 设计 Observer/Planner/Coach/Reviewer 角色分工, 基于 AgentScope Workflow 完成任务拆解、协作调度与质量复核;
- 深度记忆与画像: 短期采用 Window Memory; 长期将日记切片写入向量库用于 Episodic RAG, 并通过 Reflector 夜间任务更新 Persona Profile (兴趣/价值观/情绪标签权重);
- 混合检索策略: 实现 BM25 + 向量相似度的 Hybrid Search, 兼顾“那一天”的精确回忆与“某种感觉”的语义召回, 提升可解释性与命中率;
- MCP 工具链集成: 封装 mcp-zhihu / mcp-xiaohongshu 工具 (search_topics / get_career_advice / search_notes / get_travel_guide), Planner 动态调用外部信息生成行动指南;
- 执行与复盘闭环: Scheduler 定时推送每日 To-Do、周报/月报; Reviewer 对比计划与实际偏差输出改进建议, 持续写回长期记忆。

Jury LLM — 多模型 + 人类加权评审系统 (架构与全栈) 2025.10 – 2025.12

[Jury LLM](#)

项目简介: 构建“多模型评审团 + 人类参与”的评估系统: 多个大模型对主题打分并在分歧较大时触发辩论; 随后对“匿名理由”进行盲投 (人类+模型均参与), 最终由首席法官模型综合加权结果生成 Markdown 终稿报告。

技术栈: Python, AgentScope (条件节点流/中断点), LiteLLM (OpenRouter 路由), Jupyter Notebook (交互与人工盲投), YAML (配置中心), python-dotenv (密钥管理)

- 人类参与与赋权: 在 human_authority 节点根据用户履历评估人类权重 (0.5–1.5), 人类同时参与打分与匿名盲投, 提升可信度与可解释性;
- 分歧驱动的可控辩论: 通过标准差阈值触发 0–N 轮 debate (超轮次直接投票), 对偏离均值者发起“再考虑”提示并记录辩论日志;
- 匿名盲投机制: 在 prepare_vote 将“作者 + 理由 + 分数”随机匿名化为 Option 1/2/... 并保存真实映射; 在“模型投票前”设置中断点, Notebook 展示选项供人类先投票;
- 多模型统一调用与健壮解析: 基于 LiteLLM 统一消息格式与错误兜底, 支持不同供应商模型路由; 对模型输出进行 JSON 清洗/解析, 降低非结构化返回导致的失败率;
- 加权终审与可配置: 盲投获胜者额外加权 (模型或人类 +0.5), 评审团模型列表、辩论阈值/轮次与法官模型均可在 jury_config.yaml 配置。

探店笔记 - Java 后端开发

2024.12–2025.4

[Dianping](#)

项目简介: 基于 SpringBoot 的分布式点评类 APP 后端服务, 支持短信登录、探店点评、优惠券秒杀、每日签到、好友关注、粉丝推送等功能, 为 C 端用户提供流畅的交互体验。

技术栈: SpringBoot, MyBatis-Plus, Redis, Nginx, MySQL, Redisson

- 多级缓存: 采用 Redis+本地缓存实现用户/商品毫秒级响应, 解决集群 Session 共享问题, 降低数据库压力;
- 结合布隆过滤器拦截无效查询, 动态 TTL 机制防止缓存雪崩;
- 全局唯一 ID: 使用 Redis 实现全局唯一 ID 生成, 并通过乐观锁 (CAS) 进行库存控制, 解决超卖问题;
- 秒杀活动优化: Redisson 分布式锁保障资源互斥, Lua 脚本实现原子库存扣减及流水记录, 基于 Redis Stream

构建有序异步队列；

- 用户行为管理：ZSet 实现点赞排行榜，结合 Set 管理用户关注关系，BitMap 构建签到选项；

脑肿瘤自动识别模型开发

2025.1-2025.3

- 基于 Keras 框架自主搭建卷积神经网络 (CNN) 模型，开发用于 MRI 影像中脑肿瘤的自动检测与分类系统。项目涵盖数据预处理、模型设计、训练调优与性能评估等完整流程，最终模型在测试集上达到 98.6% 的分类准确率，表现出较强的泛化能力。项目成果包括系统实现代码与详细评估报告，全面展示了在医学图像分析方向的工程能力与科研总结能力，为深度学习在医疗影像诊断中的应用提供了可行性参考。

个人技能

编程语言：Python, Java, JavaScript/TypeScript, SQL

LLM / Agent / RAG：Diffy, Coze, LangChain, LangGraph, LlamaIndex, AgentScope (Workflow/Message Hub), LiteLLM (OpenRouter 路由), Hybrid RAG (BM25+Vector), Reranker (BGE Reranker), SentenceTransformers, Prompt Engineering, LLM 评估工作流 (AI Evaluate AI)

后端开发：FastAPI, Spring Boot, Django, RESTful API, JWT, MyBatis-Plus, Redis Stream, Lua 脚本, 分布式锁 (Redisson)

数据库 / 向量库 / 中间件：MySQL, PostgreSQL, Redis, Milvus, ChromaDB, RDS

工程化 / 云：Docker, CI/CD, AWS EC2, AWS Amplify, Nginx, Linux

数据与科研：PubMed API 检索与优化, 数据生成脚本与数据集构建, PyTorch, Keras (CNN), Matplotlib

AI 辅助开发：熟练使用 Claude Code (Open Code CLI)、Antigravity(Cursor) 等 AI 编程工具，具备 Prompt 工程能力，建立 AI 辅助开发规范体系，开发自定义 Skill 指令和 MCP，实现 AI 深度参与开发全流程。

语言能力：英语 (IELTS 6.5)，具备良好的英文沟通与阅读能力。