

教育背景

东北大学 - 库里计算机科学学院	理学硕士	计算机科学	2024.09 - 2027.05
硅谷圣塔克拉拉大学 - 文理学院	理学学士	计算机科学	2020.09 - 2024.06

实习经历

赛索德智能 (具身智能) Agent 算法实习生	中国, 上海	2026.09-2026.12
----------------------------	--------	-----------------

- 负责机器人端 Agent 架构搭建, 以及核心算法的设计与实现。
- 与运控、视觉团队协作, 负责 Agent 与运动控制、视觉感知模块的联调。

博世中国 Agent 开发实习生	中国, 无锡	2026.05-2026.08
--------------------	--------	-----------------

邮件助手 — 主导 Dify 后端工作流(日活 100+)

- 主导邮件助手后端工作流设计与落地, 支持未读邮件批量摘要与多轮回复草稿; 设计多工厂租户路由与成本隔离、轻量模型前置分流及异常自动降级, 在控制模型调用成本的同时支撑多工厂稳定使用。
- 基于 30 条真实邮件样本建立架构与模型对比评测, 以端到端延迟与输出质量驱动技术选型; Agent → Workflow 使 P95 52s → 18s, 进一步实施模型分层路由, 使 70%样本耗时再降 50%+, 质量无下降。
- 在平台由他方统管、无数据库权限的约束下, 通过并行分支埋点并复用平台去重能力, 将用量监控从应用级细化至用户级, 支持用户维度的调用量与成本追踪。

智能排班系统 POC

- 将物料、运力、人车时段等约束建模为 5 张 MySQL 表, 将取数能力封装为 MCP 服务并编写排班 Skill, 使 deepagents Agent 可基于实时约束生成排班方案并通过多轮对话动态调整, 将传统固定规则排班转化为自然语言驱动的交互式决策 POC。

Next Play Games 软件开发实习生	美国, 圣何塞	2025.09-2025.12
---------------------------	---------	-----------------

- 主导 B2B SaaS 教练平台全栈开发; 建立数据库变更工程化流程: 版本化 migration 经 CI 在 PR 与合并后逐级校验, 并同步至 dev/prod, 配套回滚策略, 解决多人并行开发下的 schema 不一致风险。
- 重构数据链路: 采用 Supabase 前端直连 (RLS 隔离), 业务逻辑下沉为 Postgres RPC, 端到端 P95 ~420ms → ~170ms; 移动端 1 秒轮询改为 Realtime WebSocket 订阅, 单端日均请求 ≈2.9 万 → 数十次。
- 主导团队 AI 辅助开发: 以 AGENT.md 统一 10+ 人团队的 Coding Agent 开发规范, 3 名新成员据此完成 onboarding; 通过 GitHub Actions 接入 AI PR review, 让大体量 AI 生成代码可审阅, CI 失败时自动建 issue 跟踪。
- 基于 OpenAI API 开发 AI 教练功能, 搭建自动化 fine-tuning 流水线: 累计 200+ 条样本, 完成 5 轮全自动迭代, 零人工介入。

北京格上富信 后端开发 实习生	中国, 北京	2025.06-2025.08
-------------------	--------	-----------------

- 在 7 年历史的 Flask/MongoDB 代码库中维护与新增 5 个核心 RESTful API, 向后兼容扩展, 生产零中断、无回归。
- 优化列表接口数据返回策略, 引入分页替代全量查询, 接口响应时间从 300 - 500ms 降至 200ms。

项目经历

KnowHive — 本地优先的 RAG + Agent 知识库桌面应用	[Github Repo]
--------------------------------------	-------------------------------

TypeScript/Bun · Vercel AI SDK v7 · Ollama + DeepSeek · Langfuse

- 检索链路: 设计统一文档中间层, 按文档类型选择分块策略, 采用父子块两级结构; 向量 + FTS5 关键词双路召回, RRF 融合、交叉编码器精排后展开父块补全上下文。经 RAGAS A/B 验证 context_precision 0.788 → 0.862 后才设为默认; 另修复 FTS5 不切分中文导致关键词检索静默失效的问题 (recall +0.054)。
- Agent 工具层: 在 AI SDK 工具循环上自建 7 个知识库与会话历史工具, 加入调用上限与重复检索拦截, 保证必定产出文本回答; 写操作 fail-closed, 需人工审批 (HITL)。
- 长期记忆: 将历史对话提炼为语义、偏好、情景三类记忆, 分别走向量召回或直接注入; “答案仅存在于历史对话中”的问题, 正确率 0 → 100%
- 拒答与注入防御: 基于精排分数设置拒答阈值, 无解问题正确拒答率 32% → 91% (误拒 4%); 以 spotlighting 隔离外部内容, 注入攻击成功率 0.40 → 0.27。后续一次重构使成功率回升至 0.47, 被自建红队测试在合并前捕获并修复。

Go 聊天室 — 实时通信服务端	[Github Repo]
------------------	-------------------------------

Go · Gin · Redis · WebSocket / WebRTC

- 构建支持私聊、群聊与 P2P 视频通话的实时聊天后端: goroutine-per-connection 加读写协程分离, 有缓冲 channel 实现非阻塞消息传递, 聊天与信令链路采用差异化背压; 复盘定位出 Hub 事件循环内同步持久化导致的吞吐瓶颈, 并给出异步化与分片方案。
- 实现无状态 JWT 鉴权 (业务接口零 DB 查询) 与 Redis 会话管理; 自建 WebRTC 信令服务, 支持 ICE 候选排队和信令、媒体两级断线恢复。

其他技能

- 编程语言: Python, JavaScript/TypeScript, Go
- 框架: Next.js, FastAPI, Flask, Gin, Vercel AI SDK, LangChain, LangGraph, deepagents
- 数据库: PostgreSQL, MySQL, MongoDB, Redis, SQLite
- 工具与 DevOps: Git, Supabase, Docker, Dify, Github Actions, Langfuse

个人介绍

技术方向: AI Agent/LLM 应用开发, 具备全栈工程与生产环境后端经验, 习惯以量化评测驱动技术选型。中美两地实习经历, 可中英文工作。业余爱好: 游戏、羽毛球、旅行 (已深度游历 10+ 国家)。