

胡海云

AI 应用与 Agent 工程化 | FDE | 云原生与全栈架构

杭州 | 13516186232 | huhenry.bj@hotmail.com | github.com/huhenry

18 年软件研发经验 | 本科 | 求职方向: AI 应用架构、Agent 平台研发、AI FDE

职业概述

18 年软件研发与架构经验, 近期聚焦 AI 应用、Agent、Harness 与数据工程。建设文档助手、语音 Agent、MCP 工具和大规模数据管线, 贯通多源采集、清洗治理、检索与业务交付。具备 FDE 所需的需求拆解、系统集成、全栈开发和生产验收能力; 曾在灵雀云研发 Kubernetes 扩展与 Istio 治理产品。

核心技术

- AI 与 Agent:** LLM / VLM、RAG、OCR、LangGraph、MCP / Tool Calling、ASR / TTS、上下文与记忆; Harness、Skills、验证证据与回归门禁; Claude Code / Codex。
- 数据工程:** 多源 ETL、字段标准化、指纹去重、异常数据复核、断点续跑、分类治理、索引重建; PostgreSQL / pgvector、OpenSearch、Redis、RabbitMQ。
- 全栈开发:** Go、Python、C++20、TypeScript; Gin、FastAPI、Next.js / React、WebAssembly; Rust 项目实践。
- 云原生与交付:** Kubernetes、CRD / Controller、Istio、Docker / GPU 容器、Harbor、Jenkins、GitHub Actions、OpenTelemetry、Prometheus / Grafana、Playwright。

最近工作经历

杭州捷配科技有限公司 | 开发工程师

2025.11 - 2026.09 | 杭州

- 面向硬件工程师和制造业务, 主导 AI 助手、知识检索与 EDA 平台建设, 负责业务集成、全栈开发和生产交付。
- 建设多语言与 GPU 推理发布体系, 覆盖 60+ 生产容器; 以镜像 digest、健康检查、浏览器验收和回滚锚点保障交付。

重点项目经验

企业 AI 助手 aibot | 业务集成与全栈交付

2026.03 - 2026.09

- 基于 LibreChat 二次开发, 扩展 MinerU 文档解析与 RAG 分片链路, 保留页码和 bbox; 前端点击检索引用可直接定位并高亮 PDF 原文。
- 接入视觉模型抽取数据手册参数; 交付文件空间、工具产物沉淀和在线预览, 将 21 类权限矩阵与配额接入业务鉴权; 完成容器化发布。

EDA 数据治理与 hwkb 知识库 | 大规模 ETL 与 RAG

2025.11 - 2026.09

- 建设多源元器件与硬件知识数据管线; 项目阶段规模达 1,830 万 SKU、2.51 亿搜索文档、6.8 万硬件项目、142 万 BOM 行及 106 万向量分片。
- 用 Go 流式处理 GB 级 CSV, 统一多源字段、型号与币种; 以 SHA-256 指纹去重, 保留异常行、原因及复核记录, 通过文件偏移 checkpoint 实现断点续跑。
- 打通标准分类、参数与替代关系治理, 采用 PostgreSQL 任务队列并发调度、版本化索引与别名切换, 实现可恢复 ETL 和零停机索引重建。
- 串联 VLM、OCR 与嵌入 Worker, 以租约 / 心跳治理异步任务; 将 bge-large-zh 迁至 GPU TEI, 同一单批耗时 90.5s → 0.31s, 约 292 倍提速。

Harness Engineering | AI 编程 Agent 的工程质量工具

2026.08 - 至今 | github.com/huhenry/harness-engineering

- 开发零依赖 Node.js CLI, 围绕 Instructions、Tools、Environment、State、Feedback、Loop 六类子系统评估仓库, 并生成可复用的指令、状态与验证脚手架。
- 真实执行验证命令并记录退出码、耗时与时效证据; 实现报告 diff、GitHub Action 回归门禁和 Skills 分发, 配套状态交接、命令拦截、停止条件与预算约束。

重点项目经验续

Soulmate 语音 Agent | 实时交互与长期记忆

2026.04 - 至今

- 建设 Python 语音服务、FastAPI 管理 API 和 Next.js 界面；通过 WebSocket / Opus 串联 VAD、ASR、LangGraph 工具调用与 TTS，支持 ESP32、网页和桌面客户端。
- 以 pgvector 实现用户级长期记忆与 RAG；设计角色画像、上下文预算、可信回合写入和声纹身份门禁，解决并发记忆更新与访客上下文串入问题。

Navix 容器平台与 AI 诊断原型 | 云原生与 AIOps

2026.05 - 至今 | github.com/huhenry/navix

- 开发 Go 控制面与 Envoy xDS 数据面，支持 Docker 发现、Hub / 站点 Agent、mTLS、版本化期望状态和 last-good 恢复，并建设故障隔离验收流程。
- 集成 Python / LangGraph 诊断 sidecar，实现分诊、推理和报告流程及最多 8 轮的停止条件；开发控制面拓扑 / 指标与 SSH 采集工具，为 AI 诊断提供数据接口。

kicad-cli-mcp | 把专业 EDA 能力开放给 Agent

2026.06 - 2026.09 | github.com/huhenry/kicad-cli-mcp

- 将 KiCad ERC / DRC、BOM、Gerber / 钻孔、STEP / GLB 导出封装为 HTTP API 与 7 个类型化 MCP 工具，共用持久化任务引擎。
- 基于 Go + PostgreSQL 实现同步 / 异步任务、幂等、租约恢复与 HMAC 回调，加入进程隔离和出站约束；通过容器、CI 和安全扫描交付可追踪产物。

多源截图数据结构化 | OCR 清洗与可追溯交付

2026.10

- 通过 Swift Vision OCR 与 Python 整理付款和订单截图，完成订单去重、实付 / 退款口径核对、分类汇总、逐项证据关联和脱敏；交付网页台账及同步 Excel，支持人工确认与审计。

以往工作经历

UD TRUCKS | 开发工程师

2024.11 - 2025.10 | 天津 / 对接日本团队

- 使用 Next.js、TypeScript、MUI、Prisma 与 Azure KQL 交付 API 可观测平台；建设 WebSocket 设备控制和 IBM MQ / ACE 消息集成，完善幂等、DLQ 与补偿重放。

安徽天宏基 | 信息化部经理

2022.03 - 2024.10

- 带领团队推进 ERP、OA、BI 与业务集成；建设 Kubernetes、Harbor 和 Jenkins Pipeline，覆盖镜像构建、部署、灰度上线及监控告警。

北京灵雀云 | 架构师

2018.06 - 2022.03 | 北京

- 容器平台研发**：以 CRD + Controller 扩展 Kubernetes，将应用发布、灰度、扩缩容和下线纳入声明式生命周期管理，并提供可视化管理能力。
- 服务治理产品**：设计 Istio / ASM 治理平台，产品化流量路由、熔断限流、mTLS、服务拓扑和调用链追踪；建设 Spring Cloud 服务框架与容器化交付工具链。

早期研发与技术管理经验

2008.07 - 2018.06

- 覆盖汽车、医疗与能源领域的 Java / .NET 全栈、WebGIS 和分布式集成；2017.10 - 2018.06 在格林深瞳推进 DeepEye Cloud 容器化、Kubernetes 迁移与 Go 自动发布，支撑视觉模型部署。

教育与语言

中国石油大学（华东） | 地理信息系统本科 | 2004.09 - 2008.07

英语 CET-6 | 英文技术文档阅读与跨国团队协作