张嘉玮

AI 数据平台与 Agent 工程负责人 · 研发负责人

2017年开始软件研发 | 硕士 | 离职，随时到岗
邮箱：zchaw127@gmail.com | 个人网站：hire.jiawis.com
常驻杭州，全国机会均可沟通

个人优势

- 从 0 到 1 主导 SolarSense AI 数据平台，9 个月上线生产；同时负责 SolarSense 与 QApex 两条产品线，带领 20+ 人研发、运维、测试团队。Java 后端出身，Python / Go / Vue 全栈。
- Agent 工程化：主导业务 Agent 平台 Solar Harness，MCP 网关统一接入 150+ 平台工具、25 个 Skill；重构上下文工程，推理成本降低近一半，任务成功率提升 40%。AI 原生研发，多智能体协作开发。
- 海量数据处理：全模态（图像、视频、点云、音频、文本、时序）数据流水线，TB 级数据集上传与版本管理，万级任务秒级调度；生产单周 300 万+ 请求，可用性 99.99%。

核心技能

Agent 与大模型：MCP、Skill、确定性工作流与委派、上下文与记忆管理、Prompt 与缓存优化、LLM 网关、评测与变异测试、LangChain。

数据处理：多模态入库与清洗转换、MySQL 任务队列（SKIP LOCKED / 幂等 / 租约）、DAG Worker、LakeFS 版本与血缘、分片直传。

后端与架构：Java、Spring Boot / Cloud、Go；网关与多租户鉴权、事务一致性、Kafka / RabbitMQ、Redis、MySQL / MongoDB / Oracle。

模型与交付：统一推理契约与按需扩缩、CPU 推理优化；Docker、k3s、Argo CD / GitOps、CI 门禁、可观测、ARM64 离线交付。

全栈：Vue 3、TypeScript、React、Node.js、Electron、WebGL。

工作经历

杭州景联文科技有限公司 | 研发负责人　2025.09 - 2026.09

- 主导 SolarSense 与 QApex 两条产品线，从 0 到 1 搭建技术体系，带领 20+ 人团队 9 个月上线生产；一年 5,800+ 次提交、评审 400+ 个 MR，2～3 周稳定迭代一个版本。
- 主导业务 Agent 平台与海量数据处理链路：MCP 接入 150+ 工具，推理成本降低近一半；万级任务提交从分钟级降到 0.89 秒，TB 级数据集分片直传零中断。
- 建设 GitOps 发布、CI 质量门禁与全链路可观测，告警覆盖 100%、可用性 99.99%；整套平台 27 个镜像离线交付至 ARM64 国产化环境。

JiawisTech Solutions LLC | 技术合伙人　2024.02 - 2025.05

- 负责 AI 客服技术方案、模型选型、提示词与 Java 核心接口；Python / LangChain 文本分析，组织小团队完成联调与部署。

东方证券 | 全栈工程师　2020.03 - 2022.01

- 负责委托与回报消息模块的Java化开发及跨系统联调，使用Kafka / Netty完成消息通信与实时推送，结合Redis锁与状态机处理并发一致性。
- 使用Python与Vue开发交易延迟监控工具，结合ELK日志检索、可视化和告警支持性能排查及异常定位。

Linkage Group | 全栈工程师　2017.09 - 2019.07

- 参与政务房产交易与智慧城市系统，带领 5 人团队；负责 MySQL 优化、RabbitMQ 异步与 Redis 缓存，独立开发汽车服务小程序。

教育经历

马凯特大学 | 计算机与信息科学 | 硕士　2022.05 - 2024.01

南京大学金陵学院 | 新闻学 | 本科 | 全日制　2012 - 2016

项目经历

张嘉玮 | Agent 工程 · 数据平台 · 架构 · 交付

Solar Harness 业务 Agent 运行时

2026.03 - 2026.09

架构负责人

平台内置的业务 Agent：把自然语言需求拆成工作流，调用平台已有的数据集、预处理、模型与复核能力逐步执行，写操作经人确认，每一步留痕可回放。

- **工具与编排**：MCP 网关统一接入 150+ 平台工具与 25 个 Skill；补齐 8 个确定性 YAML 工作流、Agent 间委派与 MCP 标准客户端 / 服务端三项 P0。
- **上下文工程**：先量再改，982 次主链路调用逐段校准 token 构成（工具定义占 40.5%、历史仅 8.8%）；稳定段前置使缓存前缀从 888 增至 2,777 字符，单个 prompt 12,726 → 9,631 token。
- **可靠性**：工具目录 28 → 20，加入循环检测与预算上限；修复反思调用孤儿消息，失败率 34.7% → 0；用户消息永不丢弃，压缩记录可追溯。
- **安全与评测**：写操作确认门绑定参数、一次性消费；46 个方法权限码隔离；528 个用例 0 失败，6 个变异全部被测出，24 例回归失败 53 → 20。

技术栈：Java、Spring、MCP、LLM 网关、Milvus、YAML 工作流、JUnit

SolarSense AI 数据平台

2025.09 - 2026.09

架构与研发负责人

面向数据采集、处理、标注、质检、审核与导出的 AI 数据平台；从第一个 commit 起建设，9 个月上线云上生产，同时支持客户离线私有化部署。

- **平台架构**：11 个 Java 服务 + Python Worker + 模型服务 + 标注工具矩阵 + 桌面端；网关、OAuth2 与 HMAC 内部调用鉴权，三层多租户隔离。
- **数据流水线**：Manifest → MySQL 任务队列 → DAG Worker → 交付状态追踪；56 个算子 × 预览 / 批量 112 格全部真跑并校验 SHA-256；疏通 4 万+ 任务积压（优先级倒挂）。
- **入库与版本**：按 TB 级单数据集设计上传与版本链路，presigned 分片直传、完成时列目录核对；数据版本与血缘基于 LakeFS，生产回填 18,370 条血缘事件全部对账。
- **多模态与性能**：21 个标注工具覆盖 10 类模态，单帧点云百万点级；CPU 分割 6,000 → 640 ms，4K 图内存峰值 4,078 → 1,907 MB；100 并发 p95 50 ms。

技术栈：Java、Spring Cloud、Python、Go、Vue 3、MySQL、Redis、MongoDB、LakeFS、对象存储、k3s、Argo CD

多模型AI客服与文本分析

2024.03 - 2025.03

核心开发

面向电商客服与脱敏文本分析场景，集成模型API，交付业务问答和分析流程，并构建用户行为推荐原型。

- **核心开发**：负责Java应用服务、模型选型、提示词设计和核心接口，完成对话上下文与业务接口集成。
- **Python与编排**：使用Python / LangChain组织文本分析和模型调用，支持脱敏反馈文本的情绪倾向聚合分析与提示词迭代。

技术栈：Java、Python、Spring Boot / Spring Cloud、Redis、MongoDB、Docker、Nginx、模型API、LangChain

个人 AI 项目（业余）

2025.09 - 2026.10

独立开发

日常自用的 AI 应用，原则是模型只做擅长的部分，能确定性计算的不交给模型。

- **小票厨房**：视觉大模型识别小票进食材库，按到期排菜谱，健康分按膳食指南公式确定性计算；识别结果与小计自动对账，中英双语。
- **阅读库 Agent 与 Skill 工作台**：贴链接即读正文、OCR 配图、写四段总结并入 Obsidian；发版、回滚、环境体检、排障等运维经验沉淀为带护栏的 Skill。

技术栈：Vision LLM、Claude Code、Codex、MCP、Obsidian、React / TypeScript