Jiawei Zhang

AI Data Platform & Agent Engineering Lead | Head of Engineering

Software engineering since 2017 | M.S. Computer and Information Science, Marquette University (USA) | Available now
zchaw127@gmail.com | hire.jiawis.com
Hangzhou, China · Open to Europe, Australia & New Zealand, incl. PhD / research · Will relocate

PROFILE

- Led the SolarSense AI data platform from zero to production in 9 months; owned two product lines (SolarSense, QApex) and a 20+ person engineering, operations and QA team. Java backend at the core, full-stack across Python, Go and Vue.
- **Agent engineering:** led the Solar Harness business-agent platform, 150+ platform tools and 25 skills behind one MCP gateway; rebuilt context engineering to cut inference cost by nearly half and raise task success by 40%. AI-native, multi-agent development.
- **Data at scale:** an all-modality pipeline (image, video, point cloud, audio, text, time series), TB-scale upload and versioning, sub-second scheduling for 10k+ tasks; 3M+ production requests a week at 99.99% availability.

CORE SKILLS

Agents & LLMs: MCP, skills, deterministic workflows and delegation, context and memory, prompt and cache optimization, LLM gateways, evaluation and mutation testing, LangChain.

Data: Multimodal ingest and transformation; MySQL task queues (SKIP LOCKED, idempotency, leases); DAG workers; LakeFS versioning and lineage; multipart upload.

Backend: Java, Spring Boot / Cloud, Go; gateways and multi-tenant auth, transactional consistency, Kafka / RabbitMQ, Redis, MySQL / MongoDB / Oracle.

Models & Delivery: Unified inference with scale-to-zero, CPU inference optimization; Docker, k3s, Argo CD / GitOps, CI gates, observability, ARM64 air-gapped delivery.

Full-Stack: Vue 3, TypeScript, React, Node.js, Electron, WebGL.

EXPERIENCE

Hangzhou Jinglianwen Technology Co., Ltd. | Head of Engineering
2025.09 - 2026.09

- Led two product lines, SolarSense and QApex, building the stack from zero and shipping to production in 9 months with a 20+ person team; 5,800+ commits and 400+ MRs reviewed in a year, a stable release every 2-3 weeks.
- **Led the business-agent platform and the large-scale data path:** 150+ tools via MCP, inference cost cut by nearly half; 10k-task submission from minutes to 0.89 s; TB-scale datasets uploaded directly with zero interruptions.
- Built GitOps releases, CI quality gates and end-to-end observability with 100% alert coverage and 99.99% availability; delivered the full 27-image platform offline to ARM64 domestic infrastructure.

JiawisTech Solutions LLC | Technical Partner
2024.02 - 2025.05

- Owned technical design, model selection, prompts and core Java APIs for AI customer service; Python / LangChain text analysis; led small-team integration and deployment.

Orient Securities | Full-Stack Engineer
2020.03 - 2022.01

- Owned Java migration and integration of order and execution-report messaging, using Kafka / Netty for messaging and real-time push, with Redis locks and state control.
- Developed Python / Vue latency monitoring with ELK log search, visualization and alerts for performance investigation and incident diagnosis.

Linkage Group | Full-Stack Engineer
2017.09 - 2019.07

- Government property-transaction and smart-city systems, leading a five-person team; MySQL tuning, RabbitMQ and Redis; independently built an automotive WeChat Mini Program.

EDUCATION

Marquette University
2022.05 - 2024.01

Master's degree | Computer and Information Science

Jinling College, Nanjing University
2012 - 2016

Bachelor's degree | Journalism | Full-time

Selected Projects

Jiawei Zhang | Agent Engineering · Data Platforms · Architecture · Delivery

Solar Harness · Business-Agent Runtime

2026.03 - 2026.09

Architecture Lead

The platform's built-in agent: it turns a plain-language request into a workflow and runs it step by step on the platform's dataset, preprocessing, model and review services, with human approval for writes and a replayable log.

- **Tools and orchestration:** 150+ platform tools and 25 skills behind an MCP gateway; added three P0s, deterministic YAML workflows (8), agent delegation and a standard MCP client / server.
- **Context engineering:** measured before changing, calibrating the token mix of 982 main-chat calls (tool definitions 40.5%, history only 8.8%); stable-first ordering grew the cache prefix from 888 to 2,777 characters and cut a prompt from 12,726 to 9,631 tokens.
- **Reliability:** tool catalog 28 → 20 with loop detection and budget caps; fixed orphaned reflection messages (failure rate 34.7% → 0); user messages are never dropped and compaction stays traceable.
- **Safety and evaluation:** approvals bound to exact parameters and consumed once; permission isolation for 46 methods; 528 tests, zero failures, all 6 mutations caught, 24-case regression failures 53 → 20.

Stack: Java, Spring, MCP, LLM gateway, Milvus, YAML workflows, JUnit

SolarSense · AI Data Platform

2025.09 - 2026.09

Architecture & Engineering Lead

An AI data platform for collection, processing, annotation, QA, review and export; built from the first commit to cloud production in 9 months, with air-gapped on-prem delivery for customers.

- **Architecture:** 11 Java services, Python workers, model serving, an annotation-tool suite and a desktop client; gateway with OAuth2 and HMAC-signed internal calls, three-layer multi-tenant isolation.
- **Pipelines:** Manifest → MySQL task queue → DAG workers → tracked delivery states; 56 operators × preview / batch, all 112 cells run with SHA-256 checks; cleared a 40k+ job backlog caused by priority inversion.
- **Ingest and versioning:** upload and versioning designed for TB-scale datasets, presigned multipart upload verified by listing at completion; LakeFS versioning and lineage, with 18,370 lineage events backfilled and reconciled in production.
- **Modalities and performance:** 21 annotation tools across 10 modalities with million-point LiDAR frames; CPU segmentation 6,000 → 640 ms, 4K-image peak memory 4,078 → 1,907 MB; p95 50 ms at 100 concurrent users.

Stack: Java, Spring Cloud, Python, Go, Vue 3, MySQL, Redis, MongoDB, LakeFS, Object Storage, k3s, Argo CD

Multi-Model AI Customer Service & Text Analysis

2024.03 - 2025.03

Core Developer

Integrated model APIs for e-commerce customer service and de-identified text analysis, delivering question-answering and analysis workflows and building a user-behavior recommendation prototype.

- **Core development:** Owned Java services, model selection, prompt design and core APIs, integrating conversation context with business interfaces.
- **Python and orchestration:** Used Python / LangChain for model calls and aggregate sentiment analysis of de-identified feedback, with prompt iteration.

Stack: Java, Python, Spring Boot / Spring Cloud, Redis, MongoDB, Docker, Nginx, Model APIs, LangChain

Personal AI Projects

2025.09 - 2026.10

Independent

AI tools I use every day, built on one rule: models do only what they are good at, and anything deterministic stays deterministic.

- **Receipt Kitchen:** a vision model reads receipts into a pantry, recipes are planned by expiry, and the health score is computed from dietary-guideline formulas; recognized items reconcile against the subtotal. Bilingual.
- **Reading-library agent and skill workbench:** paste a link to get the text, image OCR and a four-part summary filed in Obsidian; release, rollback, environment checks and triage captured as skills with guardrails.

Stack: Vision LLM, Claude Code, Codex, MCP, Obsidian, React / TypeScript