

Haiyun Hu

AI Applications & Agent Engineering | FDE | Cloud Native & Full Stack

Hangzhou, China | +86 135 1618 6232 | huhenry.bj@hotmail.com | github.com/huhenry

18 years of software engineering experience | Bachelor's degree

Professional Summary

Software engineer and architect with 18 years of experience, focused on AI applications, agents, harness engineering, and data pipelines. Build systems from data cleaning through retrieval and production delivery. Combine requirements discovery, integration, and full-stack development for Forward-Deployed Engineering roles, backed by Kubernetes and Istio product development at Alauda.

Technical Skills

- **AI and Agents:** LLM / VLM, RAG, OCR, LangGraph, MCP, ASR / TTS, context and memory; harnesses, Skills, execution evidence, regression gates; Claude Code / Codex.
- **Data Engineering:** ETL, normalization, deduplication, invalid-row review, checkpoints, taxonomy governance, indexing; PostgreSQL / pgvector, OpenSearch, Redis, RabbitMQ.
- **Full Stack:** Go, Python, C++20, TypeScript; Gin, FastAPI, Next.js / React, WebAssembly; Rust project work.
- **Cloud Native and Delivery:** Kubernetes, CRDs / controllers, Istio, Docker / GPU containers, Harbor, Jenkins, GitHub Actions, OpenTelemetry, Prometheus / Grafana, Playwright.

Recent Experience

Hangzhou Jiepei Technology | Software Engineer

Nov 2025 - Sep 2026 | Hangzhou

- Led AI assistant, knowledge retrieval, and EDA platform development for hardware engineers and manufacturing workflows, covering business integration, full-stack implementation, and production delivery.
- Built deployment workflows for multiple languages and GPU inference across 60+ production containers, with image digest verification, health checks, browser acceptance, and rollback references.

Selected Projects

aibot AI Assistant | Document Intelligence and Business Integration

Mar 2026 - Sep 2026

- Extend LibreChat with MinerU parsing and RAG chunks that retain page numbers and bounding boxes; clicking a retrieved citation highlights the corresponding passage in the original PDF.
- Integrate vision models for datasheet parameter extraction; deliver file spaces, tool artifacts, and previews, integrating 21 permission categories and quotas with business authentication.

EDA Data Governance and hwkb | Large-Scale ETL and RAG

Nov 2025 - Sep 2026

- Build multi-source component and hardware knowledge pipelines. Project-stage scale: 18.3M SKUs, 251M search documents, 68K hardware projects, 1.42M BOM rows, and 1.06M vector chunks.
- Stream GB-scale CSV data in Go; normalize source fields, part numbers, and currencies. Use SHA-256 fingerprints for deduplication, retain invalid rows and review reasons, and resume from file-offset checkpoints.
- Implement taxonomy, parameter, and substitute-relation governance with PostgreSQL job scheduling, versioned indices, and alias swaps for recoverable ETL and zero-downtime index rebuilding.
- Connect VLM, OCR, and embedding workers with leases and heartbeats. Move bge-large-zh embeddings to GPU TEI: identical single-batch processing falls from 90.5s to 0.31s, approximately 292x faster.

Harness Engineering | Quality Infrastructure for Coding Agents

Aug 2026 - Present | github.com/huhenry/harness-engineering

- Develop a zero-dependency Node.js CLI assessing Instructions, Tools, Environment, State, Feedback, and Loop, with reusable instructions, state records, and verification scaffolds.
- Execute declared checks and record exit codes, duration, and evidence freshness; implement report diffs, GitHub Action regression gates, Skills distribution, handoffs, command checks, stop conditions, and budget constraints.

Selected Projects Continued

Soulmate Voice Agent | Real-Time Interaction and Long-Term Memory

Apr 2026 - Present

- Build a Python voice service, FastAPI management API, and Next.js UI. Connect VAD, ASR, LangGraph tool execution, and TTS over WebSocket / Opus for ESP32, browser, and desktop clients.
- Implement user-level memory and RAG with pgvector, character profiles, context budgets, trusted-turn writes, and speaker identity gates; resolve concurrent memory updates and guest-context contamination.

Navix Container Platform and AI Diagnosis Prototype | Cloud Native and AIOps

May 2026 - Present | github.com/huhenry/navix

- Develop a Go control plane and Envoy xDS data plane with Docker discovery, Hub / site Agents, mTLS, versioned desired state, last-good recovery, and failure-isolation acceptance workflows.
- Integrate a Python / LangGraph diagnosis sidecar with triage, reasoning, reporting, and an eight-round limit; develop control-plane topology / metrics and SSH collectors as diagnostic data interfaces.

kicad-cli-mcp | Professional EDA Tools for Agents

Jun 2026 - Sep 2026 | github.com/huhenry/kicad-cli-mcp

- Expose KiCad ERC / DRC, BOM, Gerber / drill, and STEP / GLB exports through HTTP APIs and seven typed MCP tools sharing a durable job engine.
- Implement synchronous / asynchronous jobs, idempotency, lease recovery, and HMAC callbacks using Go and PostgreSQL; add process isolation, outbound constraints, container delivery, CI, and security scans.

Screenshot Data Structuring | OCR Cleaning and Traceable Delivery

Oct 2026

- Use Swift Vision OCR and Python to structure payment and order screenshots: deduplicate orders, reconcile payments / refunds, classify records, link and redact evidence, and deliver a web ledger with synchronized Excel exports, human review, and audit trails.

Previous Experience

UD TRUCKS | Software Engineer

Nov 2024 - Oct 2025 | Tianjin / Collaboration with Japan

- Delivered API observability with Next.js, TypeScript, MUI, Prisma, and Azure KQL; built WebSocket device control and IBM MQ / ACE integrations with idempotency, dead-letter queues, and compensating replay.

Anhui Tianhongji | IT Department Manager

Mar 2022 - Oct 2024

- Led ERP, OA, BI, and business-system integration; built Kubernetes, Harbor, and Jenkins pipelines for image builds, deployment, staged releases, and monitoring.

Alauda | Architect

Jun 2018 - Mar 2022 | Beijing

- **Container Platform Development:** Extended Kubernetes with CRDs and controllers for declarative application deployment, staged releases, scaling, and retirement, with visual management interfaces.
- **Service Governance Products:** Designed Istio / ASM traffic routing, circuit breakers, rate limits, mTLS, topology, and tracing; built a Spring Cloud service framework and container delivery toolchain.

Earlier Software Engineering and Technical Leadership

Jul 2008 - Jun 2018

- Developed Java / .NET full-stack, WebGIS, and distributed integrations in automotive, healthcare, and energy. At DeepGlint (Oct 2017 - Jun 2018), implemented DeepEye Cloud containerization, Kubernetes migration, and Go deployment automation for vision models.

Education and Languages

China University of Petroleum (East China) | Bachelor's degree, Geographic Information Systems | Sep 2004 - Jul 2008

English: CET-6; technical documentation and cross-border collaboration | Chinese: Native