CV (AI Application Development Focus)

Personal Information

- **Name**: Pan Haitao
- **Location**: Shanghai, China
- **Current Status**: Open to work
- **Phone**: +86-19286470192
- **Email**: manbuzhe2009@qq.com
- **LinkedIn**: www.linkedin.com/in/haitaopan
- **Languages**: Chinese (Native), English (Conversational)
- **Personal Website**: https://www.svc.plus (Live **ASKAI Assistant** Demo)

Education

Period	Degree	School	Major
2006.9 –2010.6	Bachelor	Changchun Institute of Technology	Electrical Engineering & Automation

---

Desired Roles

- AI Application Engineer / LLM Engineer / AI Agent Developer
- AIOps / MLOps Engineer (Application-focused)
- Platform Engineer (LLM Application + DevOps Integration)

Profile Summary

- 14 years of experience in infrastructure and platform engineering across manufacturing, finance, telecom, and internet sectors; recently focusing on **AI + Operations/Business applications**.
- Hands-on expertise in **end-to-end AI application engineering**: data ingestion → retrieval/indexing → RAG & Agent design → API/service deployment → observability & canary release → GitOps/CI/CD.
- Strong in **Kubernetes, GitHub Actions, GitOps, Terraform, Pulumi**, bridging **LLM applications** with **cloud-native platform engineering**.

- Practical experience in **AIOps and observability**: leveraging eBPF / OpenTelemetry / pgvector + LLM for **root cause analysis, Runbook Q&A, alert deduplication, and automated incident response**.

Core Skills

- **LLM & Agent**
  - Frameworks: LangChain / LlamaIndex / OpenAI Assistants API / function calling
  - Agent Design: tool orchestration, retrieval chains, memory management, task planning & execution
  - Models & Services: OpenAI / Claude / Qwen, Llama, Ollama, vLLM, Text & Embeddings APIs
- **RAG & Retrieval**
  - Embeddings & Rerank: bge-m3 / e5 / common rerankers
  - Vector Databases: PostgreSQL + pgvector, Elasticsearch (kNN + BM25), Qdrant, Milvus
  - Document Engineering: segmentation, metadata indexing, hybrid retrieval strategies
- **MLOps / AIOps**
  - Data pipelines: ETL, nearline ingestion of logs/metrics/traces into knowledge bases
  - Evaluation: retrieval recall, answer accuracy, latency & cost monitoring, regression testing for prompts
  - Governance: prompt/versioning, safety guardrails, rollout/rollback strategies
- **Platform & Automation**
  - Cloud-Native: Kubernetes, Helm, ArgoCD, FluxCD
  - CI/CD: GitHub Actions, GitLab CI, Jenkins (AI app build, evaluation, release pipelines)
  - IaC: Terraform, Pulumi (AWS/GCP multi-cloud)
- **Programming & Backend**
  - Python, Shell (familiar with Go, Rust, JavaScript)
  - FastAPI / Flask / gRPC / WebSocket, task queues (Celery/Redis)
  - Observability: Prometheus, Grafana, OpenTelemetry, DeepFlow

---

Selected AI Projects & Achievements

1) **AIOps Runbook Q&A (RAG + Agent)**

- **Scenario**: Unified Runbook/alerts/change logs/monitoring docs into a retrieval-based Q&A system for SREs.

- **Tech**: LangChain, pgvector, bge-m3, BM25 hybrid retrieval, ArgoCD read-only API, Grafana API.
- **Pipeline**: GitHub Actions for daily incremental indexing; PR-triggered regression tests.
- **Impact**: Reduced mean time to resolution by 30%+; average query response < 1.5s (cache hit).

2) **Infra Docs-to-Code DevRel Copilot**

- **Scenario**: Query Terraform modules, Helm values, FAQs and auto-generate MR drafts.
- **Tech**: LlamaIndex + pgvector, schema-constrained chunks, GitHub Actions PR Bot.
- **Impact**: Reduced MR preparation time from hours to minutes.

3) **Observability Data Q&A (Logs/Traces/Metrics)**

- **Scenario**: Natural language to observability query (metrics, traces, logs).
- **Tech**: Toolformer-style router, functions like query_metrics(), search_traces(), get_logs().
- **Impact**: Enabled first-line engineers to self-serve context queries, reducing SRE load.

Projects & Open Source Work

- **ASKAI (Live Demo)**: https://www.svc.plus —conversational AI assistant with tool orchestration.
- **XScopeHub (AIOps Suite)**: Vector/OpenTelemetry/Postgres integration for nearline ETL & retrieval, with evaluation scripts.
- **XCloudFlow / XConfig**: CI/CD + IaC modular framework, supporting canary/rollback for LLM apps.
- **Navi (Desktop AI Assistant)**: Cross-platform agent with guided tasks, knowledge Q&A, and memory-enabled tool orchestration.

Professional Experience

Beijing Yunshan Century Network Technology Co., Ltd. —Senior Support Engineer (2024.11–2025.09)

- **Stack**: Linux, Kubernetes, DeepFlow, NPM & APM, eBPF, Hybrid Cloud
- **Responsibilities**: On-site deployment, tuning, and ops of observability systems; AI Agent PoC delivery.

- **Key Achievements**:
  - Optimized DeepFlow collector performance for heavy traffic (SynTao Finance).
  - Delivered **AI Agent PoC** for COSCO Shipping (containerized deployment & QA testing).
  - Supported observability deployments for energy & telecom clients (Shanghai Power, Zhejiang Mobile, Jiangsu Telecom).

Tesla (Shanghai) Co., Ltd. —Site Reliability Engineer (2024.01–2024.05)

- **Stack**: Linux, Kubernetes, GitHub Actions, GitOps (ArgoCD), Ansible, Helm, Jenkins
- **Achievements**:
  - Developed Jenkins pipelines for industrial control (Ignition) automated deployment/upgrade.
  - Built GitHub Actions + ArgoCD pipelines for ITSM & DMP systems, enabling full automation.
  - Automated Grafana dashboards & alerts with Ansible Playbooks (Infra as Code).

Huaxun Network Systems Co., Ltd. —Senior Cloud Solution Architect (2022.03–2023.10)

- **Stack**: AWS/AliCloud, Terraform, Pulumi, GitLab CI, GitOps (FluxCD)
- **Highlights**:
  - Re-architected AWS EKS Terraform modules (migration, upgrade, addon mgmt).
  - Delivered Syngenta DevOps platform (EKS + GitLab) with IaC & app pipeline templates.
  - Built multi-cloud monitoring & IAM integration with Keycloak.

UCloud Technology Co., Ltd. —Senior Solution Architect (2020.07–2021.11)

- **Focus**: Hybrid cloud design, large-scale Kubernetes migration, CI/CD modernization.
- **Projects**:
  - GrowingIO: Designed LB migration with automated load testing via Ansible.
  - Blockchain client: Automated creation of 1000+ VMs via Python + API for elastic compute.

Alauda (灵雀云) —Delivery Engineer (2018.05–2020.06)

- **Focus**: PaaS upgrades, DevOps pipelines, containerized platform delivery for finance clients.
- **Achievement**: Led 6 major upgrades of China Everbright Bank’ s CPaaS container platform.

Deepin Technology —Software Engineer (2015.05–2018.04)

- **Focus**: Deepin Server OS V15 development, automated packaging & CI/CD pipeline.
- **Achievement**: Reduced release cycles by 30% via automation scripts.

KnownSec —Operations Engineer (2013.11–2015.04)

- **Focus**: IT infra ops, CDN optimization, automated monitoring with Saltstack/Nagios.

Inspur —System Software Engineer (2013.05–2013.10)

- **Focus**: Linux server administration, monitoring & automation scripting.

CS2C (China Standard Software) —Software Engineer (2011.05–2013.04)

- **Focus**: Porting Linux to Loongson CPU, Koji-based RPM automation.