



Kit Kyo

AI · Agent · LLM Infra | Full-Stack

me@sagamiyun.me SagamiYun

Summary

AI Agent / LLM infrastructure engineer with a full-stack background. Focused on sovereign / on-prem AI agent infrastructure and edge LLM / MoE inference (deployability × data residency × close-to-metal). Experienced in building and operating multi-node k3s clusters from scratch, A2A message buses, unified LLM inference gateways, and observability stacks. 5 years of Java/Go full-stack foundation, now deep in LLM agent engineering and GPU inference optimization.

Experience

- TLZS Corporation

2026.3 - 2026.6

Platform / Infrastructure Engineer (SRE)
- Z-SOFT Corporation

2025.2 - 2025.6

Systems Development Engineer
- Huanzhiyu Technology (Inner Mongolia)

2023.2 - 2024.6

Java Developer
- Xunya Shikong Technology (Tianjin)

2022.6 - 2022.9

Java Developer (Intern)

Projects

- Sovereign AI Cluster / Fleet Platform Engineering (SRE)

2026.3 - 2026.6

k3s Prometheus/Grafana litellm KEDA Tailscale/Liqo vast.ai A100

 - Built and operated a production-grade 8-node bare-metal k3s cluster from scratch (multi control-plane + etcd, 58 namespaces / 197 deployments); LVM storage, nftables/PPPoE networking, Tailscale + Liqo cross-cluster federation; 70 days of stable uptime, handling 60+ node/disk/network/GPU operations incidents.
 - Built a self-hosted observability + AI infrastructure stack: Prometheus/Grafana/Loki/Jaeger + Tetragon eBPF; litellm unified LLM gateway, Langfuse observability, KEDA autoscaling, Kaniko in-cluster CI; vast.ai A100 GPU orchestration (vLLM/Ollama, QLoRA fine-tuning of Gemma4); delivered business systems including jp-schedule-hub and jp-esign-forge.
- Mamelog - Social Content Platform for Residents in Japan

2025.3 - 2025.6

Golang GoZero Flutter Kubernetes

 - RED/Xiaohongshu-style content-sharing app: distributed video upload/processing, marketplace and payments, AI content recommendation and Q&A powered by Ollama + Gemini;

Profile

- Kit Kyo / M / 28
- Tokyo, Japan

Education

Inner Mongolia Agricultural University
Computer Network Technology
2020.9 - 2023.7

AI / Agent / LLM

- Claude Agent SDK MCP
- LangGraph / A2A vLLM Ollama
- llama.cpp MoE edge inference
- Gemma / Qwen RAG

Infrastructure

- Kubernetes / k3s
- NATS / JetStream
- Docker / Kaniko
- Tailscale / Liqo
- nftables / PPPoE Infisical
- systemd / launchd
- Linux (Debian/macOS)
- vast.ai GPU

Backend & Full-Stack

- Go (Gin/GoZero)
- Java (Spring Boot/Cloud) Python
- TypeScript / Node React Vue
- Three.js / WebGL Flutter
- MySQL / Redis RabbitMQ

K8s circuit breaking/load balancing, supporting 100 concurrent users.

Integrated AI Tools Platform

2023.10 - 2024.4

Golang NextJS Stable Diffusion ChatGPT

- Unified platform integrating multiple AI services such as image generation and conversational models; user/token balance management, remote multi-instance Stable Diffusion monitoring, ChatGPT API token distribution, supporting 100 concurrent users.

© 2026 Kirit Kysil Updated: June 2026

Government Grid-Management System

2023.5 - 2023.12

Java Spring Boot React Cesium

- Community housing/population/grid-officer workflow management system, integrating GIS maps and a data dashboard; the front end uses Cesium for 3D building visualization that reconstructs real-world cityscapes.