



乔磊

AI · Agent · LLM Infra | Full-Stack

✉ me@sagamiyun.me

🌐 SagamiYun

👤 个人简介

AI Agent / LLM 基础设施工程师，全栈背景。专注主权 · on-prem AI agent 基础设施与边缘 LLM / MoE 推理（可部署性 × 数据驻留 × close-to-metal）。有从零搭建并运维多节点 k3s 集群、A2A 消息总线、统一 LLM 推理网关与可观测体系的经验；5 年 Java/Go 全栈底子，现深耕 LLM Agent 工程与 GPU 推理优化。

🏢 工作经历

TLZS 株式会社	2026.3 - 2026.6
平台 / 基础设施工程师 (SRE)	
z-SOFT株式会社	2025.2 - 2025.6
系统开发工程师	
内蒙古环之宇科技有限责任公司	2023.2 - 2024.6
Java 开发工程师	
天津讯雅时空科技有限责任公司	2022.6 - 2022.9
Java 开发工程师 (实习)	

🔧 项目经验

主权 AI 集群 / 舰队平台工程 (SRE)	2026.3 - 2026.6
k3s Prometheus/Grafana litellm KEDA Tailscale/Liqo vast.ai A100	
• From-scratch 搭建并运维生产级 8 节点裸金属 k3s 集群（多控制面 + etcd），自建 Prometheus/Grafana/Loki/Jaeger + Tetragon eBPF 可观测体系；Tailscale + Liqo 跨集群联邦，稳定运行 70 天。	
• litellm 统一 LLM 网关、Langfuse 可观测、KEDA 自动扩缩、Kaniko 集群内 CI；vast.ai A100 GPU 编排 (vLLM/Ollama, QLoRA 微调 Gemma)，交付 jp-schedule-hub、jp-esign-forge 等业务系统。	
まめログ - 在日社交内容平台	2025.3-2025.6
Golang GoZero Flutter Kubernetes	
• 类小红书内容分享 App：分布式视频上传/处理、商城与支付、基于 Ollama+Gemini 的 AI 内容推荐与问答；K8s 熔断降级/负载均衡，支持 100 并发。	
综合 AI 工具整合平台	2023.10-2024.4
Golang NextJS Stable Diffusion ChatGPT	
• 集成图像生成、对话模型等多种 AI 服务的统一平台；用户/Token 余额管理、远程 SD 多实例监控、ChatGPT API 令牌分发，支持 100 并发。	
政府网格化管理系统	2023.5-2023.12
Java Spring Boot React Cesium	
• 社区房屋/人口/网格员流程管理系统，集成 GIS 地图与数据大屏；前端用 Cesium 实现建筑物三维可视化、还原城市实景。	

👤 基本信息

• 乔磊 / 男 / 28岁

🎓 教育经历

内蒙古农业大学

计算机网络技术

2020.9-2023.7

🛠️ AI / Agent / LLM

Claude Agent SDK	MCP
LangGraph / A2A	vLLM Ollama
llama.cpp	MoE 边缘推理
Gemma / Qwen	RAG

🏗️ 基础设施 / Infra

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

🔗 后端 & 全栈

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

