KITANA TOFT

AI Infrastructure Engineer · Developer Advocate · Solutions Engineer
+1-808-747-5517 · kitanatoft@gmail.com · kitanatoft.com · github.com/kctoft · linkedin.com/in/kitana

SUMMARY

AI Infrastructure Engineer and Developer Advocate with end-to-end ownership across deployment, content, and go-to-market — building and shipping production AI systems on NVIDIA B200, Intel Gaudi 3, and AMD MI350X while driving developer awareness through 14 technical videos, 26.2K+ views, and live demos at NVIDIA GTC, SC, and Cloud Fest. Sole producer of all video content from scripting and engineering to filming and post-production. Proven ability to translate complex GPU infrastructure into compelling developer-ready solutions for enterprise and government audiences.

EXPERIENCE

Technology Enablement Engineer · **Super Micro Computers** *San Jose, CA | July 2023 – Present*

- Sole producer of 14 technical videos end-to-end — scripting, engineering, filming, and post-production — generating 26.2K+ views and 524 likes — spanning product launches, developer education, and sales enablement for Supermicro's enterprise GPU line (top video: 4.2K views, HGX B200 Air & Liquid Cooled).
- Delivered live AI capability demos at NVIDIA GTC, SC, and Cloud Fest — supporting sales pipeline for enterprise and government accounts by translating next-gen GPU capabilities into compelling, customer-ready demonstrations.
- Built and deployed production inference, RAG, and multi-agent pipelines on NVIDIA B200/H200/H100, Intel Gaudi 3/2, and AMD MI350X — covering full stack from hardware bringup to model serving with PyTorch, LangChain, LlamaIndex, and vLLM.
- Deployed and validated AI workloads across Supermicro's full next-gen hardware portfolio — NVIDIA B200/B300, RTX Pro 6000 BSE, H200/H100, A100, Intel Gaudi 3/2 — from initial hardware bringup through production model inference.
- Built multi-agent orchestration pipelines using CrewAI and MCP server architecture; deployed NVIDIA AI Enterprise Blueprints including Llama 3 70B, Stable Diffusion XL, and multi-modal models (LLaVA, CLIP) for enterprise demonstrations.

Software Engineer · **SproutLabs Smart Irrigation** *Santa Cruz, CA | Sept – Dec 2022*

- Delivered full-stack smart irrigation platform with admin APIs, microservices architecture, and automated CI/CD pipeline; served dual role as engineer and Project Manager leading a 3-person team across 3 Agile sprints.

PROJECTS

Career Copilot — RAG Chatbot Portfolio Assistant | *2025–2026* | Claude API · SvelteKit · Supabase · pgvector · TypeScript

- Production RAG chatbot embedded in kitanatoft.com — answers any question about experience, projects, and skills using vector search over a live Supabase knowledge base powered by the Claude API.

LLM Chatbot with RAG on Gaudi 2 | *2023* | Python · PyTorch · LangChain · Docker

- Deployed Llama 3 70B with full RAG pipeline on Supermicro Gaudi 2 AI Server; documented end-to-end setup as a top-performing tutorial (1.1K+ views).

Multi-Modal AI Applications | *2023–2024* | Python · PyTorch · Transformers · Gradio

- Deployed Vision-Language Models (LLaVA, CLIP, BLIP) and generative AI pipelines on Supermicro enterprise infrastructure; produced demo content reaching 3.4K+ views.

EDUCATION

University of California, Santa Cruz *Graduated Dec 2022*
B.S. Computer Engineering with Honors · Computer Science Minor · Santa Cruz, CA

Foothill College *Graduated June 2021*
A.S. Computer Science, Engineering & Mathematics · Los Altos, CA

Certifications: Project Management Practitioner (De Anza College)

TECHNICAL SKILLS

AI/ML & LLMs: PyTorch, LangChain, LlamaIndex, vLLM, Transformers, Gradio, CUDA | Llama 3, GPT-4, Mistral, Stable Diffusion XL, LLaVA, CLIP
Hardware: NVIDIA B200/B300, RTX Pro 6000 BSE, H200/H100, A100 | Intel Gaudi 3/2, Max, Flex | AMD MI350X
Languages & Tools: Python, TypeScript, JavaScript, C++, SQL | Kubernetes, Docker, GitHub Actions, GCP, Linux, Supabase, pgvector
Web & Content: React, Next.js, Svelte, FastAPI, Node.js | Video Production, Adobe Premiere Pro, Technical Writing, Live Demos