Kitana Toft

San Francisco Bay Area - US Citizen

+1-808-747-5517 kitanatoft@gmail.com github.com/kctoft kitanatoft.com linkedin.com/in/kitana

Experience

Super Micro Computers July 2023 – Present
Technology Enablement Engineer *San Jose, California*

- Deployed production-ready AI applications on next-gen hardware including NVIDIA B200, H100, A100, RTX Pro 6000, and Intel Gaudi 2/3 accelerators for enterprise data center environments
- Coordinated cross-functional technical demos and product launches across engineering, marketing, and sales teams for two consecutive quarters
- Fine-tuned and optimized state-of-the-art LLMs (Llama 3.x, GPT-4, Mistral) and multi-modal models (GPT-4V, LLaVA, CLIP) for real-time inference and benchmarking
- Created technical video tutorials for YouTube that demonstrate AI deployment workflows, handling full production pipeline from scripting to post-production
- Built scalable deployment pipelines using PyTorch, TensorFlow, LangChain, and Docker, enabling rapid model iteration and testing

SproutLabs Smart Irrigation September – December 2022
Software Engineer — Apprenticeship *Santa Cruz, California*

- Developed full-stack smart irrigation platform with admin APIs, microservices architecture, and automated CI/CD deployment

Projects

Multi-Modal AI Applications 2023 – 2024
Vision-Language Models on Supermicro Infrastructure *Python, PyTorch, Transformers, Gradio*

AI Image & Video Generation Suite 2023
Stable Diffusion XL & Video on Intel Gaudi 2 *Python, PyTorch, Gradio*

LLM Chatbot with RAG 2023
Llama 3 70B on Supermicro Gaudi 2 AI Server *Python, PyTorch, LangChain, Docker*

CruzCal - Academic Scheduling Tool 2022
Senior Capstone Project *Next.js, React, TypeScript, Jotai, Jest*

Education

University of California, Santa Cruz Graduated December 2022
BSE in Computer Engineering with Honors, Computer Science Minor *Santa Cruz, California*

Foothill College Graduated June 2020
AS in Computer Science, Engineering, & Mathematics *Los Altos, California*

Technical Skills

AI/ML Frameworks: PyTorch, TensorFlow, Transformers, LangChain, LlamaIndex, vLLM, Gradio, CUDA, oneAPI
LLMs & Models: Llama 3/3.1/3.2, GPT-4/3.5, Mistral, Stable Diffusion XL/Video
Multi-Modal Models: GPT-4V, LLaVA, CLIP, BLIP
Hardware: NVIDIA B200, H100, A100, RTX Pro 6000; Intel Gaudi 2/3, Max Series, Flex Series
Languages: Python, Golang, TypeScript, JavaScript, C++, C, Java, SQL, Conversational Japanese
Developer Tools: Docker, Kubernetes, Git, VS Code, Jupyter, GCP, Firebase, Linux
Web Technologies: React, Next.js, Svelte, FastAPI, Node.js
Content Creation: Video Production, Cinematography, Adobe Premiere Pro, Technical Writing
Project Management: Agile, SCRUM, Cross-functional Leadership, Stakeholder Management

Media Engagement & Technical Communication

YouTube Technical Tutorials: Produced video series on AI deployment and optimization

- Stable Diffusion XL on Gaudi 2 AI Server
- Llama 3 70B Chatbot Implementation
- Stable Video Diffusion Tutorial

Technical Presentations: Delivered live demonstrations of AI capabilities for diverse audiences
Certifications: Project Management Practitioner (De Anza College)