Arjun Kantamsetty

972-971-9108 | arjunkan2003@gmail.com | linkedin.com/in/arjun-kantamsetty-820/ | github.com/arjunk820

Education

Tufts University Sep. 2021 - May 2025
Bachelor of Arts in Computer Science, Minor in Mathematics *Medford, MA*

Technical Skills + Interests

Tools: AWS, Azure, Terraform, Docker, Kubernetes, Helm, GitHub Actions, Grafana, PyTorch, FastAPI, LangChain, Hugging Face
Interests: Chess, Poker, Bhangra, Golf, NBA, Soccer, DJ

Experience

AI Infrastructure Engineer May 2025 – Present
WEX *Portland, ME*

- Architected a distributed training platform using JupyterHub, Ray, Kubernetes, and SageMaker to orchestrate **CPU/GPU workloads** for WEX's modeling teams, enabling scalable high-compute AI services across AKS and EKS environments.
- Implemented an end-to-end **observability stack** with Helm, Kubernetes, and ArgoCD, delivering visibility across all environments for services supporting **millions of customers** — executed 3× faster than comparable organization initiatives.
- Led several **company-wide** trainings on building AI applications with RAG and multi-agent workflows, equipping both technical and non-technical teams with the skills to securely integrate AI into internal workflows; trained **over 200 engineers** across multiple departments.

AI Engineering Intern Feb. 2025 - Jul. 2025
BPRHub *San Francisco, CA*

- Built Octo, a production-grade agentic AI system for a manufacturing compliance platform, accelerating manufacturing policy audits and boosting compliance through integrated evaluation and optimization pipelines by **45%**
- Achieved a **25x speedup** in RAG inference by semantic chunking, optimized batching, and embedding model evaluation, and benchmarking vector database performance for low-latency retrieval.

Software Engineering Intern Aug. 2024 – Jan. 2025
Levo.ai *Austin, TX*

- Leveraged LLMs to enhance the readability of API documentation by **30%**, improving client comprehension and streamlining developer onboarding.
- Integrated AI-driven documentation enhancements into existing engineering workflows, enabling **faster product adoption** and reducing support overhead for technical teams.

Projects

Upright | *React, TypeScript, Go, AWS* Oct. 2025 – Present

- Built a posture-monitoring desktop app that collects anonymized session data to deliver **posture insights and fine-tune pose estimation models**, enabling personalized ergonomics feedback.
- Deployed full-stack infrastructure on **AWS** (Lambda, API Gateway, RDS) with privacy-first data handling and analytics via PostHog, supporting **100+** users.

Awards

WEX Hackathon Finalist | *Python, Golang, Gradio, Docker, Kubernetes, PostgreSQL* Aug. 2025

- Built OpsCopilot, a **multi-agent AI system** that integrates with incident management and observability tools to help on-call engineers detect, diagnose, and **resolve production issues faster**.

Jumbohack Best Overall Hack & Education Track Winner | *Hugging Face, Next.js, AWS* Feb. 2025

- Presented to an audience of **500+**; built an end-to-end platform that **fine-tuned an ASR model** (60% → 95% accuracy) to transcribe STEM lectures with domain-specific terminology, addressing hidden accessibility barriers in education.