Vedant Agarwal

(510) 984-8759 | vedantagarwal2008@berkeley.edu | linkedin.com/in/vedant2008 | github.com/vedantagarwal-web | vedantagarwal.xyz

Education

University of California, Berkeley — Berkeley, CA
Bachelor of Arts in Applied Mathematics & Physics — *Aug 2021 – Dec 2025*

Technical Skills

Languages: Python, C++, Java, TypeScript/JavaScript, Go, Rust, SQL, Swift
AI & ML: PyTorch, TensorFlow, Hugging Face, LoRA, RAG, VLMs, FAISS
Agentic Systems: LangGraph, LangChain, MCP, Tool Calling, Playwright, Computer Use
Backend & Infra: FastAPI, Node.js, Temporal, Redis, PostgreSQL, Docker, Kubernetes, AWS (Fargate, EC2, Lambda)
Frontend & Mobile: React, Next.js, React Native, Tailwind CSS, WebGPU

Experience

Software Engineer — Feb 2026 – Present
Hive Labs (Titan Holdings) — *San Francisco, CA*

- Architected an end-to-end multi-agent AI platform that autonomously processes insurance claims by coordinating voice calls, browser automation, and LLM reasoning.
- Built an LLM-driven orchestration layer that loads claim state, plans next actions, and dispatches work to specialized agents (voice, browser, sheets). Implemented durable execution on Temporal with Redis-based distributed locking and retry logic for multi-call campaigns.
- Engineered a real-time voice pipeline on ElevenLabs Conversational AI and Telnyx SIP that places live calls to insurance payers, dynamically injects claim context, and parses structured outcomes from transcripts.
- Built vision-based browser agents using Playwright and VLMs that autonomously navigate insurance portals (Availity, Kern), handle multi-step forms, and programmatically solve MFA via Microsoft Graph API.
- Deployed as forward-deployed engineer on-site at a major toy manufacturer, integrating AI automation pipelines into legacy order processing infrastructure and owning the client relationship.

Software Engineer Intern — Dec 2025 – Feb 2026
RunAnywhere (YC W26) — *San Francisco, CA*

- Extended core inference SDK to support multi-provider model routing with optimized tokenization and batching. Applied LoRA fine-tuning to compress models for efficient edge deployment.
- Implemented WebGPU inference backend, enabling fully local browser-based agents with GPU-accelerated execution and zero server round-trips.

Software Engineer Intern — Oct 2025 – Dec 2025
Tegore-AI (YC X25) — *San Francisco, CA*

- Designed an LLM tool-calling architecture that dynamically renders interactive React components mid-conversation, enabling an AI tutor to draw diagrams and quiz students across voice and text modalities.
- Built backend services (FastAPI, TypeScript, PostgreSQL) with prompt-chaining, contextual retrieval, and concurrent workers, improving tutor response accuracy by 27% and reducing latency by 35%.

Machine Learning Engineer Intern — Sep 2024 – Dec 2024
The Mind Company — *San Jose, CA*

- Built a real-time brain-computer interface: CNN-based EEG classifier with Common Spatial Patterns achieving 92% accuracy on 4-class motor imagery. Quantized to INT8 and deployed on Raspberry Pi with sub-50ms inference.
- Rebuilt signal processing pipeline with ICA artifact rejection and adaptive bandpass filtering (+12 dB SNR). Added GPU-accelerated PyTorch training with mixed precision, cutting iteration time 3×.

Leadership & Activities

AI Entrepreneurs @ Berkeley: Builder-in-Residence. Led applied ML workshops, mentored 50+ technical founders, and organized showcase events with 200+ attendees.