Advait Shinde

+1 (408) 210 1461 | advaitss11@gmail.com | advaitshinde.com | linkedin.com/in/advait-shinde | github.com/adv-11

Education

San Jose State University — San Jose, CA
M.S. in Artificial Intelligence; GPA: 3.92 — *Expected May 2027*

Savitribai Phule Pune University — Pune, India
B.E. in Computer Engineering; Honors in AI/ML — *May 2025*

Experience

Sandisk — Machine Learning Intern | *Milpitas, CA* — May 2026 – Aug. 2026

- Fine-tuned CLIP for failure-case retrieval, achieving **92% Recall@1** and cutting turnaround by **2–3 hours**.
- Built a CLIP → Q-Former → small language model pipeline, fine-tuned with LoRA, to generate failure descriptions from images; achieved **70% exact match** and **84% F1**.
- Replaced the Matplotlib image-generation pipeline with CUDA rasterization, increasing throughput **10×** (20 to 200 images/s) and improving retrieval score by 2%.
- Deployed models on on-premises GPUs with thread-based parallelism and query latency **below 50 ms**.

Synechron — Artificial Intelligence Intern | *Pune, India* — Aug. 2024 – July 2025

- Benchmarked small language models (Llama, Qwen, Gemma, Mistral, Phi) on fintech NLP tasks to design an LLM API gateway for model routing on private GPU clusters.
- Optimized LLM inference with PyTorch, KV caching, and speculative decoding, **reducing inference costs 20%**.
- Built an organization-wide, OWASP-aligned LLM guardrails framework for bias, toxicity, PII, and prompt injection, adopted by **10+ internal applications**.
- Delivered a client proof of concept for multimodal retrieval-augmented generation (RAG), using vision-language models, CLIP, and ColPali embeddings on on-premises infrastructure.
- Co-designed a client's agentic RAG workflow to automate contract generation, **cutting generation time 50%**.

Projects & Open Source

Spacecraft Pose Estimation — Demo

- Built a 6-DoF pose pipeline (HRNet heatmaps to EPnP/LM) on FRESH event-RGB data; trained only on synthetic data, with 24 real sequences held out for final testing.
- Designed 8 RGB-event fusion strategies, including a physics-based saturation gate; added glare augmentation, reducing rotation error under synthetic glare from **13.7° to 5.8°** on validation data.
- Ran 129+ ablations with pre-registered comparisons and bootstrap statistics; corrected a 59° ground-truth error and a synthetic-to-real event encoding mismatch from the reference paper.

Snitch-Stitch — CLI Security Auditor | 1st Place, 500.co Hackathon — PyPI | GitHub

- Built a multi-agent CLI to detect vulnerabilities in AI-generated code, score severity, and generate patches; secured 500.co sponsorship after the win.

Agent2Agent (A2A) — Open Source Contribution — Pull Request

- Contributed an AutoGen currency-conversion agent to the A2A samples repository (3 merged pull requests) with session management, multi-turn streaming, and JWK-authenticated webhook notifications.

Technical Skills

Languages: Python, C++, SQL, R
ML Libraries: PyTorch, Hugging Face Transformers, TensorFlow, OpenCV, NumPy, pandas
AI Systems: RAG, LoRA, LangChain, LangGraph, AutoGen, FastAPI, LiteLLM, MCP, A2A
Infrastructure: CUDA, Git, Docker, Redis, Google Cloud Platform, AWS Bedrock, Azure AI
Coursework: Machine Learning, Deep Learning, Reinforcement Learning, Recommender Systems

Recognition: 4× hackathon winner; Amazon ML Summer School (2024); LA Hacks mentor (2026); Best Outgoing Student (B.E.)