Swaminathan Chellappa

Software Engineer | (213) 829-4803 | swaminathanchellappa5@gmail.com | LinkedIn | Github | Portfolio

EDUCATION

University of Southern California — August 2024 - May 2026
Master of Science in Computer Science — *Los Angeles, CA*

Dayananda Sagar College of Engineering — August 2018 - May 2022
Bachelor of Engineering in Computer Science — *Bangalore, India*

TECHNICAL SKILLS

Languages: JavaScript, TypeScript, Python, Kotlin, C#, SQL
Frontend: React, Next.js, Vite, Redux, Mantine UI, MUI, HTML/CSS
Backend & Databases: Node.js, Express, FastAPI, .NET, MongoDB, MySQL, ChromaDB
Testing & DevOps: Jest, Vercel, JUnit, Espresso, Mockito, GCP, AWS (S3), Azure, Docker, SLURM, Linux, Git, Splunk, JIRA
AI/ML: LangChain, LangGraph, HuggingFace, Ollama, Prompt Engineering, RAG, FAISS, PyTorch, LLMOps

PROFESSIONAL EXPERIENCE

USC DILL LAB — December 2025 – Present
Graduate Research Assistant

- Improved annotation F1 score by **~30%** over zero-shot baseline by designing a "Live Codebook" system that synthesizes annotator corrections into structured LLM prompt rules, enabling iterative model improvement without retraining
- Cut annotation time from **3 hours to 12 minutes** by building a React frontend orchestrating batched sample review, inline span/label correction, and a live codebook feedback loop across Express and FastAPI backends
- Eliminated dependency on commercial LLM APIs to comply with HIPAA and data retention policies by implementing a local data anonymization pipeline achieving **87% accuracy** on sensitive entity detection
- Enabled scalable **multi-GPU inference** on USC CARC HPC by deploying a full-stack Vite app via SLURM, orchestrating MongoDB, Ollama across 4 GPUs, and async task queuing to manage resource contention

KODERAI (AI Code Generation Startup) — June 2025 – August 2025
AI Software Engineering Intern

- Secured internal AI workflows against prompt injection attacks by developing **5+ production API endpoints** integrating LLM guardrails including content filtering, model-based filtering, and cosine similarity thresholds
- Reduced API retries by **~35%** and cut new provider integration effort by 50% by building a provider-agnostic inference platform with telemetry and token-level cost tracking
- Reduced Expo (React Native) compile times by **~40%** by automating bundling and deployment to cloud VMs and optimizing Gradle build scripts
- Improved production accuracy by **~26%** by designing a .NET LLM benchmarking and routing framework evaluating 32 commercial/open-source models for task-specific agents, optimizing across latency, cost, and output quality

WELLS FARGO – Digital Technology & Innovation (DTI) — August 2022 – June 2024
Software Engineer

- Contributed to app rating improvement from **4.6 to 4.8 stars** by building a Python dashboard analyzing 10M+ app reviews using sentiment classification and clustering to surface feature-level pain points and release-wise rating trends
- Served **23M+ users** by shipping 10+ reusable MVVM UI components for BillPay and Alerts modules on the mobile banking platform
- Delivered **100% accessibility compliance** across BillPay and Alerts by integrating TalkBack, Screen Magnification, and Accessibility Focus, meeting WCAG standards
- Reduced post-release regression defects by **38%** by maintaining 90%+ unit test coverage across UI and business logic layers via TDD using JUnit, Espresso, Mockk, and Mockito

PROJECTS

F1 Pit Strategy Reasoner

- Improved pit strategy recommendation accuracy by **30%** over base model by fine-tuning Mistral-7B with QLoRA (4-bit quantization, LoRA rank 16) on curated historical F1 race data spanning 4 seasons
- Reduced hallucinated strategy calls by building a RAG pipeline over **4 seasons** of race telemetry, weather, and tire degradation data using FAISS and LangChain, grounding outputs in real race conditions
- Achieved **sub-4s response latency** for real-time strategy queries by containerizing the full inference stack with Docker and designing a FastAPI serving layer

PII Redactor for Gemini (Chrome Webstore | Github)

- Achieved **98% PII detection accuracy** with 2s inference by building a privacy-first Chrome extension running Llama-3-8B locally in-browser to detect and redact sensitive information from prompts before sending to Gemini
- Eliminated manual re-insertion of redacted terms by implementing a **response rehydration system** that automatically restores redacted information in Gemini's responses using cached term mapping

Breast Cancer Classification in Mammograms (Contributing Research | Paper)

- Improved classification accuracy from **60% to ~90%** (F1: 89.5%) by designing a novel Sobel-Canny-Gabor hybrid preprocessing pipeline for mammogram and ultrasound images, contributing to a published study in iJOE
- Achieved **23% improvement** in Peak Signal-to-Noise Ratio over traditional edge detection methods by addressing discontinuous edge detection and noise reduction in digital pathology images