Sanskar Thapa

US Citizen | Los Angeles, CA | sanskar1016pro@outlook.com | sanskarfolio.com | github.com/sskarz

Education

California State University of Los Angeles GPA: 3.8/4.0 (Dean's List)
Master of Science in Computer Science Expected May 2026
Bachelor of Science in Computer Science Completed May 2025

Skills

Languages: Python, TypeScript, JavaScript, Go, SQL, Swift, Java, C#
Frameworks and Libraries: React Native, React, Vue.js, Express.js, Executorch, PyTorch, Tensorflow, LangChain
Cloud and Technologies: AWS (EC2, S3, Lambda), Docker, PostgreSQL, Elasticsearch, Linux, Git

Experience

Software Engineer Intern, Northrop Grumman – Rome, NY June 2024 – August 2024

- Reduced user onboarding time by 67% (from 15 to 5 minutes) by generating AI responses from a corpus of over 4000 pages of military manuals stored in AWS S3 using a RAG system with Phi-3 Medium (LLM) and Elasticsearch
- Enhanced LLM resource efficiency by 50% by applying precision reduction on weights and activations (quantization), yielding significant computational savings while maintaining model accuracy
- Deployed and scaled containerized applications using Docker on AWS EC2 instances (Red Hat Enterprise Linux 8), ensuring production readiness and simplifying the development workflow

Project Lead, The Aerospace Corporation – Los Angeles, CA August 2024 – May 2025

- Led a team of 10 students to develop a Vue.js web app, visualizing global satellite coverage on a world map using Leaflet to identify active and dead satellite zones globally
- Engineered a 4-endpoint Express.js REST API to provide real-time satellite visualizations, fetching and ingesting thousands of longitude, latitude, and PDOP data points into PostgreSQL every 30 minutes

Founder, Aery – Los Angeles, CA Jan 2025 – Present

- Created Aery, a cross-platform React Native email client that surfaces messages in a swipe-card UI and runs on-device LLMs for email summarization, categorization, and smart-reply; shipped to 100+ users on Apple Testflight
- Boosted AI inference performance by 3x by embedding the Executorch runtime, strategically managing trade-offs between model size and on-device memory constraints to enable real-time summarization and smart-reply features
- Enhanced app security by implementing OAuth 2.0 with JWT-based authentication and encrypting session tokens in on-device key stores, ensuring secure user login and session management

Projects

Sight: Vision Impairment Assistance Web App github.com/jasonly027/Sight

- Developed a React web app for the visually impaired centered on a YOLOV5 object detection model capable of identifying 80 distinct object classes from the user's surroundings
- Built a cross-platform haptic feedback system that translates object detection events into distinct vibration patterns, providing nuanced environmental cues to visually impaired users and enhancing interactive accessibility

Executorch Library (Meta, PyTorch) github.com/pytorch/executorch/pull/8522

- Enhanced project portability by converting absolute CMake paths to relative ones, simplifying build setup across environments, allowing for new contributors to run the React Native iOS demo previously broken

GPT-2 AI User Imitation github.com/sskarz/GPT-2-AI-User-Imitation

- Fine-tuned a 124M-parameter GPT-2 model on a 5MB personal text message dataset utilizing Tensorflow and PyTorch, generating responses that achieved an 80% success rate in blind tests for user imitation