

Sukhman Virk

Full-Stack Software Engineer | Cloud Infrastructure | Backend Systems | Machine Learning

(858) 308-6022 · sukhmanvirk26@gmail.com · [LinkedIn](#) · [GitHub](#)

EXPERIENCE

Minfy Technologies (AWS Premier Partner)
Associate Software Engineer

McLean, VA
August 2024 – Present

- Architected a full-stack multi-tenant weather-data marketplace: a **FastAPI** backend with **Cognito** JWT auth, tenant-scoped access, cursor-paginated APIs a **TypeScript/React** dashboard for discovery, subscriptions, and API-key management.
- Built an internal **Claude Code** delivery platform adopted org-wide across **900+** engineers, running spec, backlog, TDD build, and handover as gated stages, with telemetry attributing token cost per engineer, project, and phase.
- Engineered a production RAG system on **AWS Bedrock Knowledge Base** using **Cohere** embeddings, semantic chunking, and metadata-filtered **OpenSearch Serverless** vector search, cutting query resolution to **under 10 seconds**.
- Designed an **MCP** server via **FastMCP** exposing role-gated tools to **Bedrock** agents, enabling hybrid retrieval across **PostgreSQL** and vector stores with permissions enforced at the tool layer.
- Built a multi-tenant **RBAC** admin portal for fleet management in **React.js** and **GoLang**, isolating data per tenant.
- Shipped a secure file-management service in **Flask** with **KMS** encryption to meet the client's data-protection requirements, sustaining **10,000+** daily requests at **99.99%** uptime.
- Led greenfield builds for enterprise clients from architecture to deployment, and migrated on-premises workloads to **AWS** with automated failover and zero-downtime cutovers.

Reliance Industries Limited (JIO)
Software Engineering Intern

Mumbai, India
July 2023 – September 2023

- Built a **C++ WebRTC** P2P application for a Fortune 500 company enabling high-throughput transfer of images, video, and messages at up to **4K** resolution.
- Automated ICE negotiation via Google STUN, cutting connection setup time by **90%** and reducing end-to-end media latency by **30%**.

HealthCare.com

Software Engineering Intern

Miami, FL
June 2022 – August 2022

- Implemented a full-stack **AES-256** cryptographic applet in **Next.js** and **Java** protecting **50,000+** user records with zero post-launch vulnerabilities.
- Optimized decryption-key retrieval via **AWS KMS** and integrated Okta 2FA, eliminating manual key rotation across the entire platform.

PROJECTS

Algorithmic Market Making System | Kalshi
Co-Founder & Developer

January 2025 – Present

- Deployed a production-grade market maker quoting in **sub-50ms** across multiple prediction markets with **\$10K+** in traded volume, featuring dynamic spread optimization and inventory-based skewing.
- Modeled fill probability and markout using **gradient-boosted ensembles** with **Bayesian smoothing** on **500K+** trade events, incorporating queue-position and order-book imbalance signals to reduce adverse selection.
- Optimized quoting via **EV maximization** over spread, size, and cancel timing with inventory variance as a penalty term and toxicity-aware throttles during volatile moves.

Pandera | Quantitative Research Platform
Co-Founder

March 2025 – June 2025

- Built a full-stack quant research platform (**Next.js**, **Flask**, **PostgreSQL**) with a daily ETL pipeline ingesting equities, options chains, and Treasury yields.
- Implemented **Black-Scholes** IV inversion and options Greeks with spline interpolation for smooth implied volatility surfaces, and applied LRU caching to cut per-query runtime by **4×**.
- Published **cuda-kernels** to PyPI: CUDA-accelerated autocorrelation and sum reduction running **91%** faster than NumPy.

Airline Sentiment NLP Pipeline

Researcher | COGS 118B

January 2024 – March 2024

- Developed an end-to-end NLP pipeline on **14,600** tweets: **TF-IDF** features, **SVM** and Logistic Regression for 3-class sentiment, **K-means** with **PCA** clustering validated by Silhouette scoring, and **LDA** topic modeling to surface latent complaint themes.

SocialCompass | Real-Time Location App
Developer | CSE 110 – Software Engineering

January 2023 – March 2023

- Engineered a real-time **Android** app in Java rendering friends' live locations via persistent location streams and a SQL-backed friend graph, with a full test suite (**Robolectric**, **JUnit**, **Espresso**) and **GitHub Actions** CI/CD.

EDUCATION

University of California, San Diego

San Diego, CA

B.S. Mathematics – Computer Science | Minor: Cognitive Science, Economics

Major GPA: **3.95** | Cumulative GPA: **3.83** | **Provost Honors**

Relevant Coursework: Data Structures & Algorithms, Software Engineering, Systems Programming, Supervised Machine Learning, Statistical Analysis, Numerical Analysis, Econometrics, Game Theory

SKILLS

Languages: Python, SQL, C++, Java, C, GoLang, TypeScript, JavaScript, PHP

Frameworks & Libraries: React.js, Next.js, Node.js, Flask, Django, FastAPI, PyTorch, TensorFlow, NumPy, Pandas, OpenCV, CUDA, Redis
Cloud & DevOps: AWS (S3, ECS, ECR, DynamoDB, Cognito, Lambda, Bedrock, KMS, CloudWatch, QuickSight), Docker, GitHub Actions CI/CD, Cloud-Native Architecture

Domains: Machine Learning, NLP, RAG Systems, Agentic Systems, Market Microstructure, Cryptography

AI-Assisted Dev: Claude Code, Cursor, GitHub Copilot

LEADERSHIP

SangamSD – Tech Lead, Board Member

San Diego, CA | September 2022 – June 2024

- Built a QR code check-in system cutting queue wait times by **50%** and eliminating unauthorized entry across **10+** events, and maintained a website serving **7,000+** monthly visitors.