Shivani Singh

San Francisco, CA | +91 9818155639 | shsingh844@gmail.com | LinkedIn | GitHub

SUMMARY

Full-stack software engineer with 3+ years building production AI applications and scalable infrastructure. Expertise in 0→1 product development spanning frontend to distributed backends. Proven track record reducing infrastructure costs through Kubernetes optimisation and accelerating development cycles through scalable architecture. Pursuing doctoral research in predictive analytics while maintaining technical edge through independent work.

TECHNICAL SKILLS

AI & ML: LLM Integration & Deployment, Production ML Systems, Agent Systems, Model Serving & Orchestration, Prompt Engineering, Claude API, OpenAI, LangChain, RAG, Statistical Analysis, Predictive Analytics, TensorFlow, PyTorch, scikit-learn

Development: Python, Ruby, React, TypeScript, JavaScript, Flask, Node.js, RESTful APIs, WebRTC, Full-Stack Development, Microservices, Event-Driven Architecture

Data & Infrastructure: R, SQL, ETL/ELT Pipelines, PostgreSQL, Airflow, Databricks, Prometheus, Grafana, Docker, Kubernetes, CI/CD, AWS, Azure, Terraform, Tableau, Power BI, BigQuery

EXPERIENCE

Independent Software Consultant | Remote — Nov 2022 - Present

- Built two production AI applications demonstrating end-to-end product ownership from architecture to deployment, with active user engagement and high system reliability
- Architected intelligent prompt engineering system processing multimodal data (telemetry, weather, user history) for actionable insights with strong user satisfaction
- Developed PostgreSQL-backed analytics platform with geospatial algorithms, improving query response times through intelligent caching strategies
- Integrated Claude API and OpenAI GPT-4 for context-aware recommendations, creating real-world applications showcasing advanced LLM capabilities

Mixhalo | San Francisco, CA — Apr 2022 - Nov 2022

Full Stack Engineer

- Led 0→1 development of real-time translation prototype that became company's flagship monetisation product, directly influencing future investor presentations and product roadmap pivot toward translation services
- Designed scalable WebRTC architecture through systematic evaluation of multiple frameworks, reducing future product R&D cycles and establishing reusable patterns for all streaming features
- Reduced infrastructure costs by 67%, architecting Kubernetes-based deployment system that improved deployment frequency from weekly to daily, eliminating DevOps bottlenecks for engineering team
- Built end-to-end observability platform capturing user interactions and system metrics, dramatically reducing incident response time and enabling data-driven product decisions that shaped company roadmap

Extreme Networks | San Jose, CA — Mar 2020 - Apr 2022

Software Engineer

- Architected company-wide data platform consolidating multiple siloed systems into unified warehouse processing millions of events monthly, dramatically reducing executive reporting cycles across business units

- Delivered self-service analytics platform with multiple Tableau/Power BI dashboards heavily used across organisation, dramatically reducing ad-hoc reporting requests and freeing capacity for strategic initiatives
- Delivered multiple production features for ExtremeIQ cloud networking platform, managing thousands of enterprise devices, gaining deep expertise in TCP/IP, DNS, VPNs that accelerated customer issue resolution
- Established comprehensive testing framework integrated into CI/CD pipeline, reducing production incidents by 45% and deployment rollbacks significantly, enabling team to accelerate feature velocity

Frontend Engineering Intern

- Led UI redesign that significantly increased customer satisfaction through user research and iterative testing, becoming key differentiator in sales demos and customer retention
- Created reusable component library using Dojo framework, significantly reducing code duplication and cutting feature development time across multiple product teams

RESEARCH & EDUCATION

Westcliff University | Irvine, California, USA | Remote (Weekends) Oct 2022 - Present

Doctor of Business Administration (Business Intelligence & Data Analytics)

- Developing ML-based financial contagion model analysing historical bank failures with ensemble methods, achieving strong predictive accuracy; research progressing toward peer-reviewed publication

Boston University | Boston, MA, USA Sep 2018 - Jan 2020

M.S., Computer Information Systems (Database Management & Business Intelligence)

- *GPA: 3.60*

Indira Gandhi Delhi Technical University | New Delhi, India Aug 2014 – Jun 2018

B.Tech, Electronics & Communication Engineering

- *CPI: 64.37*

NOTABLE PROJECTS

Windborne Constellation Tracker | Live Demo

- Built a production AI analytics platform for weather balloon operations, processing real-time telemetry data with Claude API integration that dramatically reduced analysis time while maintaining high system reliability

Sustainable Fashion Analytics Platform | Live Demo

- Launched AI-powered sustainable fashion platform with strong user engagement through GPT-4-driven personalised recommendations and geospatial brand matching, demonstrating practical consumer-facing AI applications