Shreyas Nikam

shreyas.s.nikam@gmail.com | +1 (857) 204-4757 | linkedin.com/in/nikam-shreyas
github.com/shreyas-nikam | nikam-shreyas.com | leetcode.com/u/nikamshreyas

PROFESSIONAL EXPERIENCE

AI Software Engineer → Tech Lead, **QuantUniversity**, Boston, MA — Jan 2024 - Present

- Architected large-scale backend services for LLM-powered platforms using FastAPI, Redis, PostgreSQL, and Docker, supporting multi-tenant internal users across production environments reducing end-to-end production time by **85%**
- Designed and implemented a modular LLM integration layer with request routing, prompt versioning, response validation, and provider fallback, **cutting inference costs by 65%** based on billing telemetry
- Owned CI/CD and deployment strategy using Docker, GitHub Actions, and AWS EC2, **enabling zero-downtime blue-green releases** across 4+ production servers
- Scaled and maintained distributed workloads by integrating background workers, task queues, and real-time WebSocket updates for long-running jobs
- Led technical design reviews and mentored engineers on system architecture, deployment strategy, and operational best practices

Data Scientist, **Changing the Present**, Boston, MA — Sep 2023 - Jan 2024

- Scaled data collection pipelines to **400k+ profiles across 100+ universities**, reducing manual effort by **70%**
- Built **real-time dashboards** in Streamlit with drill-down filters, enabling stakeholders to monitor live data
- Designed **WebSocket-based real-time update systems** using Redis pub/sub, powering instant task updates and alerts

Data Researcher Intern, **Boompanda**, Pune, India — Jan 2021 - Jun 2021

- Built data ingestion pipelines scraping **150k+ event listings**, converting unstructured HTML into relational datasets
- Orchestrated **Airflow DAGs** for automated cleaning, transformation, and feature generation
- Developed a **content-based recommender system**, improving CTR by **22%**, validated via A/B testing on **5k users**

Software Developer Intern, **Tech Square**, Pune, India — Jun 2019 - Jan 2020

- Designed real-time dashboards ingesting **MQTT/Modbus streams** from 20+ industrial sensors
- Implemented rule-based anomaly detection, identifying failures within **2 seconds**, reducing downtime by **18%**
- Operationalized alerting pipelines with SMTP notifications, improving incident response times

Software Development Intern, **Riskpro**, Pune, India — Mar 2019 - May 2019

- Curated **1.2M credit records** via Python scrapers targeting CIBIL-suite portals, eliminating manual data entry errors
- Pipelined **2 Airflow ETLs** converting raw HTML/CSV into SQL tables, trimming daily ingestion time to 25 min
- Innovated a Tkinter desktop GUI for lookup; honored with the **Best Performance Award 2019** at college hackathon

EDUCATION

Northeastern University, Boston, MA — Sep 2021 - May 2023
Master of Science, Artificial Intelligence (AI) — GPA: 4.0 / 4.0
Student Success Guide (Leadership) - Fostered inclusivity while helping new students navigate cultural transitions
Teaching Assistant - Supported 150+ graduate students through design reviews and system critiques

Pune University, Pune, India — Sep 2017 - May 2021
Bachelor of Engineering in Computer Engineering — GPA: 9.60 / 10.00

RELEVANT PROJECTS

AI-Assisted SAR Generation [Live Application] [Github] [Documentation] [User Guide]

- Developed an LLM-powered tool to generate first-draft **Suspicious Activity Reports**, reducing narrative drafting time by **~60%** while ensuring compliance through human-in-the-loop review

Github Package Health [Github] [Demo]

- Engineered a **full-stack package-health platform using FastAPI and React**, integrating AI-driven dependency analysis and actionable insights to improve codebase hygiene for users across multiple repositories

WarcParser [Github]

- Built an efficient **WARC file processing pipeline** in Python with async IO and multithreading, scaling dataset generation from Common Crawl for downstream training of large language models

TECHNICAL SKILLS

LLMs & Agents: OpenAI, LangChain, LlamaIndex, RAG
Backend: Python, FastAPI, Node.js, C++, Java, JavaScript, REST APIs, WebSockets
Data & Infra: Airflow, PostgreSQL, Docker, AWS (EC2, Docker-based deployments), GCP
CI/CD: GitHub Actions, Blue-Green Deployments