Samit Shah

San Jose, CA | Portfolio | (602) 574-3772 | linkedin.com/in/samitnshah/ | sshah229@asu.edu | github.com/sshah229

EDUCATION

- **Master of Science in Software Engineering, Arizona State University** Aug 2023 - May 2025
- **Bachelor of Technology in Computer Engineering, NMIMS University** Aug 2019 - June 2023

EXPERIENCE

Full-Stack AI infra engineer: Unsloth AI, San Jose, CA, USA Dec 2025 - Current

- Standardized **LoRA** configs and training hyperparameters across **60+ LLM** models with auto-mapped defaults; shipped **6 production PRs**, cutting onboarding time by **40%** and configuration failures by **30%**.
- Built dynamic fine-tuning orchestration (epoch and step-based training), enabling **3 times faster iteration** and **25% better GPU utilization**.
- Developed the LLM training and evaluation AI platform using **React, Gradio, FastAPI, and Python**, powering end-to-end training and monitoring workflows and handling **100+ training runs per week** with consistent performance.

Software engineer Intern: RoundTechSquare, Tempe, AZ, USA Jan 2025 - May 2025

- Deployed Dockerized **Node.js and React** microservices behind **Nginx** with autoscaling and connection pooling, cutting median latency from **130ms to 90ms** and sustaining **250+ concurrent sessions**.
- Built a GPT-4 intent mapper across **180 intents** with batched inference and streaming, **processing 3.5K+ messages per day** and improving availability and throughput.
- Implemented metrics and structured logging, sped incident triage and enabled end-to-end observability across services.

Software engineer Intern: NOMURA, Mumbai, India Jun 2022 - Nov 2022

- Optimized 35 Informatica workflows, cutting ETL runtime from **120 to 22 min (81%)** and unblocking downstream analytics.
- Orchestrated migration pipelines, achieving **95% data integrity** and enabling accurate forecasting for business teams.
- Automated workflow performance tests on multi-TB datasets, hardening data operations.

Data Engineer Intern: CBIA, Mumbai, India Nov 2020 - Jun 2021

- Built a centralized **SQL Server** data warehouse on **Azure** using Data Factory and Azure Pipelines, integrating 6+ sources to standardize schemas and enable near real time reporting.
- Designed a star schema with partitioning and clustered indexes, tuning queries and adding materialized views to improve **response times by 72%** and cut redundant lookups**.**
- Implemented an ETL automation pipeline to integrate Planning & Resource Management data, reducing manual effort by **20 hours/month per project** while optimizing **cost to $3 per million-row run**, improving operational efficiency.
- Shipped Power BI dashboards, **eliminating 60%** of manual reporting with self-serve insights**.**

Artificial Intelligence Developer intern: CBIA, Mumbai, India Apr 2019 - Jul 2019

- Built **ML analytics** for inventory decisions, **improving customer insights by 20%** and strengthening stock planning.
- Developed **NLP** pipelines for **NER** and text classification in **Python** with **spaCy** and **scikit-learn**, scheduled via Airflow, **reducing manual reporting 60%** and standardizing insights.
- Shipped **BERT**-based NER with **ONNX** Runtime behind **FastAPI**, cutting inference latency **40%** and lifting tagging **F1 8%.**

RELEVANT PROJECTS

Soul-Support: Agentic LLM powered AI-mental health companion Mar 2025

- Led end-to-end React/Node build streaming real-time 3D avatar chats by embedding Blender WebGL scenes with the **Gemini** API, sustaining **120+ msgs/min** with **p95 less than 200ms**.
- Designed a multi-agent pipeline (context keeper, crisis detector, emotion logger) hitting **97%** risk-detection precision, auto-dialing contacts within **5s**, and logging **15K+** journal entries per day.
- **DevHacks 2025 AI for Innovation winners** from **100+ teams** for Soul-Support's multi-agent crisis-support design, **97%** risk detection, and **5s** emergency escalation.

Panacea: Enhanced Electronic Health Records System Jan 2022

- Built **HIPAA-compliant ETL** on Azure Data Factory and Functions, normalizing **HL7/FHIR** across hospital DBs to cut record latency **60%** and ensure fault-tolerant access.
- Served NER and disease-prediction models on **AKS** behind **API Management** with App Insights, delivering observable, versioned **REST endpoints** for real-time clinical support.
- Enforced privacy and governance with **PII masking**, **RBAC**, **Key Vault** secrets, and row-level security in clinician dashboards.

TECHNICAL SKILLS

Relevant technologies- TypeScript, React, Next.js, Node.js, JavaScript, REST APIs, API Development, API Integration, Data Flows, Debugging, Distributed Systems, Web Development, Frontend Development, Backend Development, Full Stack Development, AI, Machine Learning, Conversational AI, Voice Technologies, Git, Linux, Docker, AWS.

ACHIEVEMENTS

Won 3 Hackathons | 3 Microsoft Certifications- Power Platform, Azure, SQL | Tableau Certified | Published AI Research paper