Siddhant Shah

siddhantshah.me — mail[at]siddhantshah[dot]me — linkedin.com/in/thesidshah — github.com/thesidshah

Summary —

Applied ML + Software Engineer focused on **AI evaluation**, simulation systems, and ML-driven workflows. Strong **Python**, **FastAPI/Flask/Django**, and **Java/Spring Boot** background; ship production platforms, multimodal LLM pipelines, and tooling for structured insights, agent workflows, and data-centric evaluation. Curious about **LLMs**, **agentic AI**, and **model behavior analysis**; driven to make AI systems more measurable, testable, and reliable at scale.

Skills

Languages Python, Java, TypeScript, JavaScript, SQL, Bash

AI/ML PyTorch, HuggingFace, OpenAI (GPT, Whisper), Anthropic Claude, LangChain, GANs, NLP, OCR

Evaluation Model evaluation, structured/`json_schema` decoding, precision/recall/F1 harnesses, synthetic data

Simulation Agent workflows, interactive system modeling, MCP protocol

Backend FastAPI, Flask, Django, Spring Boot, Node.js, SQLAlchemy, REST, microservices

Frontend React, Next.js, Capacitor, Tailwind CSS, Storybook

Data Pandas, NumPy, PySpark, Redis, RQ, SSE, ffmpeg, ETL, MLflow

Databases PostgreSQL, MongoDB, DynamoDB, MSSQL, SQLite, Redis

Cloud Azure (Functions, Blob, Queue), AWS (EC2, S3, SQS, DynamoDB), Docker, Terraform, GitHub Actions

Observability Prometheus, Grafana, instrumentation, logging

Testing Pytest, Jest, React Testing Library, Playwright, JUnit, CI/CD

Experience

Cambridge Mobile Telematics **Apr 2026–Present**
Software Engineer II *Cambridge, MA*

- Deliver end-to-end features on the customer Portal — a Django, React, and Terraform-provisioned AWS microservice platform serving **10,000+ visits/hour** — spanning Postgres, DynamoDB, and SQS-backed services.
- Shipped the 100 Hz sensor data alignment effort, carrying it from statistical analysis through backend services to the React UI, wiring the change across multiple microservices and validating it with end-to-end testing.
- Integrated crash severity classification into the Portal, and currently landing Agero partner data ingestion across the same service boundaries.
- Authored design documents and reconstructed missing documentation for legacy subsystems, making previously opaque services safe for the team to extend.

TERVA Corporation (0→1 Startup) **Aug 2023–Apr 2026**
Software Engineer II *Boston, MA*

- Architected and shipped 3 full-stack platforms (Daily Reports, Site Observations, Admin Console) from database design through production deployment, serving **500+ users** across distributed teams.
- Engineered a microservices backend with **FastAPI** handling **10K+ daily API calls**, implementing Microsoft OAuth authentication, Redis caching, and MongoDB with optimized indexing strategies.
- Built a granular RBAC system using the Composite design pattern, enabling fine-grained role and permission composition across **50+ project teams** with zero security incidents.
- Integrated **LLM-based summarization** (OpenAI GPT + Whisper) to generate executive summaries and KPIs from raw reporting data.
- Automated PDF generation and email distribution using Azure Functions and Queue Storage, reducing manual reporting effort by **70%** and delivery latency by **60%**.
- Developed an event-driven notification service delivering real-time Microsoft Teams alerts with fault-tolerant retries and **99.5%** delivery success.
- Designed a configuration diff engine supporting nested objects, unordered arrays, ignore rules, and duplicate detection, tracking report changes across **1000+ daily submissions**.
- Migrated the admin panel to SSR with code splitting, improving initial load time by **40%** and reducing client bundle size by **35%**.
- Integrated Azure Blob Storage with signed URLs for secure image uploads, processing **5000+ images monthly** across draft and published containers.
- Implemented **observability and instrumentation** (Prometheus + Grafana) to derive structured insights and detect anomalous system behaviors.

- Improved backend throughput **15%** through query tuning, lean data-access patterns, and selective field projection.
- Expanded reporting UX with annotation support, issued-date metadata, issue-reporting endpoints, an administrator statistics dashboard, and multi-tenant team synchronization.
- Built Jest / React Testing Library suites reaching **85% coverage** on critical business logic, deep-diff behavior, and form validation; enforced pass gates before production deploys.
- Established CI/CD with self-hosted GitHub Actions runners, Dockerized services, automated test gates, and vulnerability scanning, cutting deployment time by **25%**.

KeelWorks Foundation **Jun 2023–Jun 2024**
ML Engineer & Full Stack Developer *Remote*

- Built NLP-driven content evaluation components and internal tooling supporting dataset creation and labeling workflows.

Northeastern University — IT Services **Sep 2022–Jul 2023**
Digital Experience Assistant *Boston, MA*

- Instrumented frontend interactions and improved accessibility, contributing user behavior insights to design iterations.

SS&C Intralinks **Oct 2022–Dec 2022**
Software Engineer Co-op *Waltham, MA*

- Reduced JavaScript bundle size by **30%** through strategic code splitting, improving page load performance for an enterprise document management platform.
- Developed and tested **30+ React components** using Storybook with Component Story Format, maintaining a **90% test pass rate** in CI/CD.
- Built accessible UI components with React, SASS, and Backbone.js, resolving multiple critical UI bugs in a large enterprise codebase through structured debugging and behavior tracing.

Northeastern University — Research **May 2022–Jul 2022**
Graduate Research Assistant *Boston, MA*

- Built a Flask-based analytics platform on an AWS S3 data warehouse, streamlining neuroimaging data collection from **20+ MRI experiments**.
- Automated a data preprocessing pipeline using Python and MRIQC, improving processing efficiency by **25%** across **500GB+** of imaging data.

Northeastern University — Khoury College **Aug 2021–May 2022**
Teaching Assistant (AI & Systems) *Boston, MA*

- Supported Foundations of AI and Systems coursework; assisted students with ML fundamentals and algorithmic behavior analysis.

Globant **Dec 2020–Aug 2021**
Associate Software & Data Engineer *Pune, India*

- Developed a unified ServiceNow platform for a Fortune 100 client, eliminating **70%** redundant reporting and saving **500+ man-hours annually**.
- Built **50+ Power BI dashboards** integrated with Microsoft SQL Server, enabling global operational visibility across **10+ business units**.
- Developed Python + SQL ETL powering analytics dashboards, and optimized PySpark jobs on Azure Databricks (**45% reduction** in compute time).
- Automated workflows and alerts using Agile delivery, standardizing reporting pipelines for enterprise-scale operations.

Visava Labs (0→1 Startup) **Jul 2018–Jun 2020**
Machine Learning Engineer (R&D) *Pune, India*

- Led ML research on generative design systems, developing GANs that produced **1000+** architectural floor plan variations from input constraints.
- Built a Java-based expert system generating **100+ valid configurations** per input for architectural planning workflows.
- Developed low-latency ML features deployed in production; contributed technical demos and product pitches that secured **80%** of first-round startup funding.

IoTIoT.in **Feb 2018–Jun 2018**
Machine Learning Engineer Intern *Pune, India*

- Built a CNN-based facial recognition pipeline optimized for ARM devices, reducing compute requirements through minimized facial landmark representations.

Projects

Reel — Multimodal Video-to-Structured-Data Platform **reel.siddhantshah.me**

- Designed and shipped a **32K-LOC** production platform (57 REST endpoints across 6 Flask blueprints, 24 tables, 51 React components) turning Instagram reels into schema-validated recipes, itineraries, and workouts.
- Built a multimodal extraction pipeline fusing Whisper transcripts, caption metadata, and on-screen text (OCR plus vision LLM) over ffmpeg-sampled frames, using perceptual dHash dedup to drop redundant frames before inference.
- Engineered an async RQ/Redis job queue with real-time SSE progress streaming over Redis pub/sub, subscribe-before-emit race protection, heartbeats, and dialect-aware duplicate-request locking on PostgreSQL and SQLite.
- Capped inference at **$0.50/reel** using tiered provider chains (DeepSeek/Qwen-VL → OpenAI → Claude), adaptive per-reel frame and vision budgets, and P50/P90/P99 cost telemetry, while suppressing hallucination with strict `json_schema` decoding at temperature 0, per-field evidence citations, and an **offline precision/recall/F1 eval harness**.
- Ported the same Next.js codebase to iOS and Android with Capacitor, driving both targets from one env var and adding a custom Swift plugin plus iOS Share Extension for sharing reels directly from Instagram.

Developer Tooling, Agents & Evaluation

- **AI Agent Simulation (Pokemon MCP)** — MCP-based simulation system for modeling interactive agent behaviors and evaluating decision patterns. GitHub
- **FastAPI Backend Template** — production-ready scaffold with async SQLAlchemy, Alembic migrations, environment-aware config, observability, auth, and CI/CD suitable for agent/evaluation tooling. GitHub
- **MCP Developer Tool** — TypeScript MCP tooling with automated CI/CD; contributed architecture docs to the Microsoft Amplifier ecosystem.
- **DeepSeek OCR Analysis** — technical breakdown of compression strategies, model architecture, and implications for evaluation and token reduction.
- **Spring Boot Fintech Assessment** — structured API design, tests, and performance instrumentation. GitHub
- **Churn Prediction LSTM App** — end-to-end ML system at **85% accuracy**; includes data pipelines and evaluation workflows.

Earlier Work — Distributed ML, NLP & Computer Vision

- **Estimating Muscle Force for Amputees** (Spark, Horovod, PyTorch, MLlib) — parallelized a 2D CNN to a **4x speedup** at 3.7% RMSE, and showed PCA gains plateau beyond 8 partitions. [Report]
- **Topic Modeling & Extractive Summarization** (Gensim, NLTK, Scikit-learn) — applied NMF, truncated SVD, and LDA over **1000+** TF-IDF-normalized documents, reaching ROUGE ∼0.3. [Code]
- **Image Convolutions from Scratch** (Python, NumPy, OpenCV) — built a filtering and Harris-corner framework in raw NumPy with automatic border and outlier removal; written up on Medium.

Education

Northeastern University **Boston, MA**
M.S., Computer Science (AI & ML Specialization)
Pune University **Pune, India**
B.E., Computer Engineering (Networks & ML)

Certifications

- **Microsoft Certified: Azure Fundamentals (AZ-900)**
- **Microsoft Certified: Azure AI Fundamentals (AI-900)**
- **ServiceNow Certified System Administrator**

Publications

- Shah et al., *An Intuitive Study: Intrusion Detection Systems and Anomalies — How AI can be used as a tool to enable the majority, in the 5G era*, IEEE

Open Source & Community

- Active contributor across 50+ GitHub projects; engaged in agent, evaluation, and backend OSS communities.
- Volunteer at ODSC.