

SAMARTH CHAUHAN

Dwarka, New Delhi, India

📞 +91-8586948175 ✉️ samarth18410@iiitd.ac.in [in](#) [LinkedIn](#) [Github](#) [Portfolio](#)

Software engineer with 4 years of experience building and running ML-powered products at scale: speech-scoring assessment platforms used by 100,000+ students, LLM features in production, on-premise and offline model deployments, and high-traffic search. I work across serving infrastructure for compute-heavy inference, training-data pipelines, and the product built around the model. Python, PyTorch, Django, FastAPI, React, AWS, GCP.

Education

Indraprastha Institute of Information Technology, Delhi

August 2018 - June 2022

Bachelor of Technology in Computer Science and Bioscience

CGPA: 7.99

Experience

Wadhvani AI

Jan 2024 – Present

Software Development Engineer 2

New Delhi, India

- Built a standalone spoken English assessment platform where ML models score reading, speaking, vocabulary and listening tasks, with adaptive practice pathways. 100,000 students onboarded and 2.5M assessments served, scaling to 500,000 students.
- Ran the serving setup for compute-heavy inference windows. Processed 100,000+ assessments at 25 concurrent ML inferences per minute (40s average latency) with no downtime, after reworking queue management and resource allocation.
- Shipped LLM features in production, including Llama-based paragraph generation and a “Lecture Segmentation” feature built in two weeks.
- Designed the search engine for a MOOC platform on BigQuery with custom ranking. 500,000+ searches served and 180,000+ course clicks driven.
- Built a configurable field data collection platform with offline and online inference pipelines, audio, camera and location capture, and a portal for checking data quality before ingestion. 5,000 collectors and 200,000 entries in four months.
- Deployed a model on an on-premise server inside a hospital PACS workflow, and built offline-first radiology and dermatology apps handling bidirectional sync, image caching and device memory limits. Migrated sensitive scan data from cloud to on-premise storage with verification checks.
- Wrote a cloud-agnostic object storage wrapper now used by 5 projects in the organisation, so teams keep one codebase across cloud providers.

Wadhvani AI

Sept 2022 – Dec 2023

Software Development Engineer 1

New Delhi, India

- Owned the radiology data and annotation platform: React frontend handling 100,000+ DICOM uploads, APIs over object storage and a relational DB feeding training pipelines, and annotation job management, distribution and analysis.
- Built the TB differentiated care Django backend (schema, REST APIs, serializers, tests) and automated hospital mapping ingestion with custom management commands.
- Managed AWS infrastructure and deployments (EC2, RDS, S3, Pinpoint), dockerized services behind multi-worker Nginx, and reduced recurring cloud cost by merging similar services and shutting down idle instances.
- Added OCR to a healthcare imaging app with custom container images, worked with the ML team to cut model response times, and standardised backend templates (tests, logging, structure) reused in later projects.

Technical Skills

Languages: Python, JavaScript, SQL

ML: PyTorch, TensorFlow

Backend: Django, FastAPI, REST APIs, queues, offline-first sync

Data: BigQuery, PostgreSQL, MySQL, MongoDB

Cloud/DevOps: AWS (EC2, RDS, S3), GCP, Docker, Nginx, Linux, Git

Frontend: React, PWAs