Samad Rehan

Lucknow, India
+91 9369476890 | samadrehan123456@gmail.com | LinkedIn | Github

Machine Learning Engineer with hands-on experience building, deploying, and scaling LLM-powered systems, production NLP pipelines, and GPU-accelerated inference backends. Strong background in LLM fine-tuning, evaluation, RAG systems, ASR, and MLOps, with proven impact on latency, cost, and reliability in production environments.

SKILLS

Programming Languages: Python, C++, C, Java

Machine Learning Libraries: PyTorch, Hugging Face Transformers, Scikit-learn, XGBoost, Model Training & Evaluation, Feature Engineering, Offline & Online Evaluation

LLMs and NLP: LLMs, Instruction Tuning, Prompt Engineering, Retrieval-Augmented Generation (RAG), Text-to-SQL (NL2SQL), Information Extraction, Named Entity Recognition (NER), Speech Recognition (ASR), Hindi & Hinglish NLP

Computer Vision: Image Classification, Object Detection, Face Recognition, OpenCV MLOps, Cloud & Deployment

CloudOps and Deployment: AWS (EC2, S3, IAM, ECR), Docker, GPU-Accelerated Inference, FastAPI, Model Versioning & Monitoring, CI/CD for ML Pipelines, Linux, REST APIs

WORK EXPERIENCE

Machine Learning Engineer — Lucknow, India

Augurs Technologies — Dec 2025 - Present

- Applied MLOps best practices (model/version tracking, monitoring, reproducible deployments) to GPU-accelerated LLM inference pipelines on cloud infrastructure, reducing deployment regressions by ~60% and improving system reliability and rollback time by ~2×.
- Designed and deployed LLM-powered systems using both local GPU-hosted and managed cloud models, optimizing prompts and inference paths to achieve ~25–40% lower latency and ~30% lower compute cost per request.
- Built and maintained FastAPI-based inference backends serving LLMs, RAG pipelines, and downstream services with authentication and structured logging, sustaining sub-second P95 latency at thousands of requests/day.
- Performed LLM fine-tuning and evaluation (instruction tuning, response calibration, accuracy testing), improving task compliance and output accuracy by ~35–45%, and reducing invalid or off-policy responses by ~2–3× across production workloads.

PROJECTS

Automated Doctor's Report from Patient verbal Transcript — Dec 2025 - Jan 2026

Built a real-time Hindi medical transcription system using FastAPI, Vosk ASR, and a custom NLP pipeline to extract clinical entities (symptoms, medications, diagnoses) from doctor–patient conversations.Integrated an instruction-tuned LLM (llama3.1 via Ollama) with strict prompt controls to generate safe, nonhallucinated OPD notes from the extracted structured data.Implemented session-scoped speaker diarization, Hindi/Hinglish normalization, and a modern web UI for live transcription, structured JSON inspection, and one-click clinical note export.

Conversational Chatbot with specialized NL to SQL functionality — Nov 2025 - Dec 2025

Built an NLSQL chatbot capable of translating natural language queries into safe, validated SQL across multiple schemas, while also functioning as a general-purpose conversational assistant.Implemented context-aware query handling with session memory, follow-up resolution (e.g., pronouns, references), and schema-aware routing to maintain accurate multi-turn interactions.Deployed the system using FastAPI with integrated LLM inference, query execution, and response formatting, supporting both structured database answers and free-form conversational replies.

Stock Market Prediction using Transformer Models Dec 2024 - Apr 2024

Built a full-stack ML application using TCN Transformer achieving 98% R^2 score in stock forecasting on historical datasets.Deployed real-time prediction pipelines handling thousands of data points per second.Optimized backend performance, reducing query latency by 40% through caching and async APIs.

EDUCATION

Vellore Institute of Technology

Bachelor of Technology - Computer Science and Engineering with specialization in Artificial Intelligence and Machine Learning - GPA: 8.00 Sep 2021 - Sep 2025

Lal Bahadur Shastri Smarak JAVM Senior Secondary School

Intermediate Mar 2019 - May 2020

CERTIFICATIONS

Amazon AWS Certified Cloud Practitioner | Amazon Web Services Nov 2024

Amazon AWS Certified Cloud Solutions Architect - Associate | Amazon Web Services Mar 2025

CS50AI Introduction to Artificial Intelligence with Python | Harvard University (edx) Oct 2025

Microsoft Azure Fundamentals AZ-900 | Microsoft Dec 2023

Python Institute Certified PCEP Python Programmer | The Python Institute Aug 2022