Samad Rehan

Lucknow, India samadrehan123456@gmail.com +919369476890 in/samadrehan samadrehan02.github.io

SUMMARY

ML Engineer with production experience deploying LLM-powered systems and optimizing GPU inference pipelines. Specialized in RAG architectures, structured generation, and constraint-aware decision systems with proven ability to reduce latency by 25-40% and cut compute costs by 30% in real-world deployments. Strong foundation in end-to-end ML system design from model selection through cloud deployment and monitoring.

SKILLS

Programming Languages: Python, C++

ML Systems: LLM deployment, RAG, instruction tuning, structured generation, GPU inference, latency optimization

MLOps & Cloud: Docker, AWS (EC2, S3, IAM, ECR), CI/CD, model monitoring, versioning, Linux

Frameworks: PyTorch, Scikit-learn, Hugging Face Transformers, FastAPI, XGBoost

EXPERIENCE

Artificial Intelligence and Machine Learning Engineer

Augurs Technologies **December 2025 - Present, Lucknow, India**

- **LLM Inference Optimization:** Designed and deployed GPU-accelerated inference pipelines across local and cloud environments, reducing end-to-end latency by 25-40% and lowering per-request compute cost by ~30% through batching strategies and model quantization.
- **Production API Development:** Built high-throughput FastAPI services supporting RAG pipelines with sub-second P95 latency at thousands of daily requests, implementing caching layers and async processing for improved reliability.
- **Full-Stack ML Ownership:** Independently architected, deployed, and operated multiple production ML systems end-to-end, including model selection, backend services, cloud deployment (AWS EC2, S3), and operational monitoring.
- **System Reliability:** Established model versioning, error logging, and performance monitoring practices ensuring >99% uptime across deployed services.

PROJECT

Constraint-Aware Dynamic Pricing Platform

October 2025 - February 2026

- Built end-to-end pricing optimization system with constraint solver respecting margin floors, daily change bounds, and inventory levels, deployed with FastAPI backend and real-time WebSocket updates.
- Designed synthetic demand modeling addressing cold-start SKUs, enabling controlled strategy evaluation across 1,000+ product scenarios.
- Created operator dashboard with SKU-level price inspection, alert management, and manual override capabilities.

Automated Doctor's Report from Patient verbal Transcript

December 2025 - January 2026

- Deployed real-time pipeline converting Hindi/Hinglish medical conversations into structured clinical notes with <2s end-to-end latency using custom entity extraction and schema-constrained LLM generation.
- Reduced structurally invalid outputs by 40-55% through grammar-guided decoding compared to free-form generation.
- Implemented session-aware speaker diarization, improving transcription continuity across interruptions by 30% in production testing.

Schema-Aware NL2SQL Decision Assistant

December 2025 - January 2026

- Developed schema-aware natural language to SQL translator achieving >95% syntactic validity across PostgreSQL and MySQL environments through dynamic schema injection and query validation layers.
- Implemented conversation context persistence enabling multi-turn refinement, improving follow-up query success by 35-45%.
- Built query safety system preventing unauthorized operations, reducing malformed query attempts by 50% in testing.

EDUCATION

Bachelor of Technology - Computer Science and Engineering with specialization in Artificial Intelligence and Machine Learning

Vellore Institute of Technology Chennai Campus • 2025 • 8.00

CERTIFICATIONS

CSSOAI Introduction to Artificial Intelligence with Python

Harvard University (edx) • Oct 2025

Amazon AWS Certified Cloud Solutions Architect - Associate

Amazon Web Services • Mar 2025

Amazon AWS Certified Cloud Practitioner

Amazon Web Services • Nov 2024