Shaurya Negi

+91 9694953568 | shauryanegi17@gmail.com | linkedin.com/in/shauryanegi | github.com/shauryanegi

Education

Georgia Institute of Technology — Atlanta, GA, USA
Master of Science in Computer Science, Specialization in Machine Learning — *Aug. 2022 – May 2025*
Selected Coursework: Deep Learning (CS 7643), Reinforcement Learning (CS 7642), NLP (CS 7650), Graduate Algorithms (CS 6515)

Manipal University Jaipur — Jaipur, India
Bachelor of Technology in Computer Science and Engineering — *Aug. 2013 – Aug. 2017*

Experience

Arya.ai - An Aurionpro Company — Mumbai, India
Senior Research Scientist — *Apr. 2025 – Present*

- Architected a ReAct-based Agentic RAG pipeline for credit memos using hybrid retrieval and Cross-Encoder re-ranking; integrated web-search tools to slash end-to-end reporting lifecycle by 85% (3 weeks → 3 days).
- Steered technical strategy for 3 enterprise GenAI products; implemented Model Context Protocol (MCP) to standardize agent-to-data interfaces, decoupling inference logic from backend silos and reducing custom integration boilerplate by 60%.
- Benchmarked 8 Dense and MoE models (Llama-3, Mistral) for BFSI compliance; engineered a Human-in-the-Loop eval framework reducing entity extraction hallucinations from 23% to 4% (F1 score 0.72 → 0.89).

Senior Data Scientist — *May 2024 – Mar. 2025*

- Architected a segmented cash-flow forecasting system; implemented a routing layer to decouple deterministic (recurring) payments from stochastic (XGBoost) transactions, improving accuracy by 18%.
- Led monthly business reviews with C-suite to drive AI strategy, demonstrating 35% decision-making efficiency gains from GenAI integration.
- Fine-tuned Llama-3.1 adapters on curated synthetic data (distilled from larger models + expert validation); optimized serving via vLLM/PagedAttention to slash P99 latency by 22% on production GPUs.

Data Scientist — *Jan. 2022 – Apr. 2024*

- Scaled Bank-Statement-Analyzer API to 18 countries; processed 1M+ documents/month with 98% extraction accuracy, displacing 40% of manual review load.
- Developed Text-to-SQL engine for complex BFSI schemas using few-shot prompting and schema linking; improved query success rate to 85%, driving 30% user engagement growth.

SankalpTaru Foundation — New Delhi, India
Data Analyst — *Mar. 2020 – Aug. 2021*

- Fine-tuned EfficientNet-B7 on custom field imagery to diagnose sapling health; implemented confidence-based routing to automate 85% of audits, flagging only low-confidence samples (<0.75) for expert review.
- Developed a semantic search engine using BERT embeddings to align Corporate CSR mandates with afforestation projects, increasing partner conversion rates by 15%.

IBM (Client: Etihad Airways) — Gurugram, India
Technical Analyst — *Dec. 2018 – Jan. 2020*

- Engineered Python-based ETL scripts to parse unstructured server logs, automating anomaly detection and reducing Mean Time to Detection (MTTD) by 40% for global airline infrastructure.

SankalpTaru Foundation — New Delhi, India
Program Executive (Operations & Analytics) — *Aug. 2017 – Nov. 2018*

- Spearheaded the migration of analog field records to structured SQL databases, establishing the organization's first centralized dataset for tracking 50,000+ ecological assets.

Skills

Programming: Python, C++, SQL, R, Bash
AI / ML Frameworks: PyTorch, TensorFlow, LangChain, LlamaIndex, vLLM, Model Context Protocol (MCP)
AI / ML Techniques: NLP, Agentic RAG, QLoRA (PEFT), Cross-Encoders, Knowledge Distillation, MoE
Infrastructure: Kubernetes, Docker, OpenShift, Podman, FastAPI, AWS, GCP, HA/DR
Data Engineering: Apache Spark, Kafka, NiFi, FAISS, Milvus, PostgreSQL, MongoDB, Redis