Saba Firdaus Ansaria

Machine Learning Engineer

Chertsey, England, UK | sabafirdaus.cs@gmail.com | +44 7435 846168 | LinkedIn | GitHub | Portfolio

Professional Summary

Machine Learning Engineer with an **M.Sc. Robotics (Distinction)** and an **M.Sc. Applied Mathematics**, experienced in taking machine learning models from research through to production. Built and administered a self-hosted evaluation platform end-to-end, including data pipelines, object storage integration and a live scoring/monitoring service. Took a fault-detection model into live commercial deployment and built a production-style Retrieval-Augmented Generation service with a FastAPI/Streamlit front end. Strong Python foundation across PyTorch, TensorFlow and Scikit-learn, comfortable across the full model lifecycle from development through deployment and monitoring, and used to collaborating across engineering, research and stakeholder teams. **Right to work in the UK; no sponsorship required.**

Key Skills

Languages: Python, SQL, Bash

ML & Deep Learning: PyTorch, TensorFlow, Scikit-learn, NumPy, Pandas, CNNs, RNNs, NLP, ASR, Computer Vision, Model Evaluation & Validation

Generative AI & LLMs: Retrieval-Augmented Generation (RAG), Ollama, Qdrant, BM25 & Hybrid Retrieval, Cross-Encoder Reranking, Sentence-Transformer Embeddings, Prompt Engineering, LangChain, FAISS, Quantised LLMs (GGUF)

MLOps & Deployment: Model Deployment & Monitoring, Data/Model Pipeline Design, FastAPI, Streamlit, Docker & Docker Compose, CI/CD (GitHub Actions), Fly.io, Render, MinIO / Object Storage, SMTP & Service Integration

Tools & Practices: Git/GitHub, GitHub Codespaces, Codabench, Linux, Technical Documentation, Cross-functional Collaboration

Mathematics & Statistics: Linear Algebra, Probability & Statistics, Differential Calculus, Statistical Modelling, Experimental Design

Professional: Stakeholder Communication, Clear Written & Verbal Communication (6+ yrs teaching), Problem-Solving

Experience

Research Intern — Speech & Language AI **Dec 2025 – Present**
School of Computer Science, The University of Sheffield *Sheffield, UK (Remote)*

- Built and maintained a self-hosted Codabench ASR Ranking Challenge platform end-to-end, configuring SMTP for service notifications and integrating MinIO object storage.
- Developed a Kendall Tau scoring pipeline to rank and monitor submitted models, maintaining it as production infrastructure.
- Produced formal technical documentation of the system's architecture and operation for the supervising professor.

Freelance Machine Learning Engineer **Feb 2025 – Nov 2025**
Independent (freelancer.com) *Surrey, UK (Remote)*

- Built and deployed a document-based question-answering (RAG) system using quantised Llama 2, LangChain, FAISS, Sentence Transformers and Streamlit, achieving sub-5-second CPU response times for on-device search across medical PDFs.
- Scoped, built and shipped the project end-to-end with full autonomy; released open-source with reproducible documentation.

Software Consultant **Apr 2024 – Jan 2025**
Spagiria Consulting (Contract) *Maastricht, NL (Remote)*

- Developed a now-commercialised fault-detection system for a chemical production pipeline using multisensory data and image processing with Python, OpenCV and Raspberry Pi.
- **Reduced manual inspection time by 20% and increased fault-detection accuracy by 30%**, taking the system from R&D through to live commercial deployment.
- Worked directly with client stakeholders to scope requirements, validate data quality, and translate model output into actionable reporting for a non-technical audience.

Secondary School Teacher — Mathematics **Dec 2014 – Aug 2021**
Maria's Day School, Maria's Group of Institutes *Howrah, India*

- Designed and delivered an ICSE-aligned mathematics curriculum for 1,000+ students over six years, developing strong skills in simplifying complex quantitative material for varied audiences.

Key Projects

Aunt Flo Assistant — Hybrid RAG On-Device Chatbot github.com/sfansaria

- Engineered a hybrid retrieval pipeline (dense vector search + BM25 + cross-encoder reranking) over a locally-served LLM (Ollama/Qwen2.5), then productionized it as three containerized microservices on Fly.io. Debugging cross-service networking, memory-constrained inference, and persistent vector storage to ship a live, publicly accessible RAG system.

Multisensory Fusion & Low-Latency Sleep Apnoea Detection github.com/sfansaria

University of Sheffield — M.Sc. Thesis

- Built a real-time paediatric sleep apnoea detection system using multisensory NHS clinical data, comparing SVM, logistic regression and neural network models for accuracy and latency.
- **Achieved 98.1% accuracy at 9.46 ms CPU latency** with a multichannel CNN fusion model, deployable on low-power wearable devices.

Automatic Speech Recognition (ASR) github.com/sfansaria

University of Sheffield — M.Sc. Project

- Developed multiple CNN architectures (baseline, regularised, residual) for audio-based speech recognition in TensorFlow, extracting spectrograms, filterbanks and MFCCs as features.
- **Improved accuracy from 69.68% to 74.64%** across model variants, with full evaluation using confusion matrices and performance plots.

Industrial Vision System — Commercialised Spagiria Consulting

- Designed and deployed a real-time monitoring fault-detection system using Python, OpenCV and Raspberry Pi integrated with IoT sensors; reduced manual inspection time by 20% and increased accuracy by 30%, transitioning from R&D to full production.

Education

M.Sc. Robotics — First Class with Distinction **Sep 2021 – Nov 2022**
The University of Sheffield *Sheffield, UK*
Specialisation: Data Modelling & Machine Learning, Health AI.
Thesis: *Multisensory Fusion and Low-Latency Sleep Apnoea Detection* — Supervisor: Prof. Lyudmila Mihaylova.

M.Sc. Applied Mathematics — First Class **Jul 2013 – Jul 2015**
The University of Burdwan *India*
Specialisation: Linear Algebra, Probability & Statistics, Calculus.

B.Sc. Mathematics **Jul 2010 – Jul 2013**
The University of Calcutta *India*

Certifications

- **Generative AI with LLMs** — Coursera / DeepLearning.AI Jun 2024
- **Machine Learning Specialisation** (3 courses) — Coursera / Andrew Ng Jul–Aug 2024
- **McKinsey Forward Programme** — McKinsey & Company Jul 2025