Rafi Riyaz

MSc AI | Stanford University Online ML Certified
+44 7774874773 | rafa.works313@gmail.com | LinkedIn | GitHub | Portfolio | Hugging Face

EXPERIENCE

Machine Learning Engineer
Oct. 2025 – Present
Mercor, USA — London, UK (Remote)

- **Rubrics Academy Fellow:** Focused on complex prompt reasoning and technical problem-solving through first-principles architectural design with strict constraints on using AI tools.
  Engaged in rigorous technical challenges emphasizing first-principles reasoning and human-led architectural design under strict constraints on using AI tools.
- **Project Launchpad:** Contributed to Meta AI Researchs expansion of OpenAIs MLE-bench by analyzing buggy code blocks and implementing fixes in LLM-generated solutions for complex ML tasks, utilizing the AIRA-dojo development environment and Jupiter mega-cluster powered by **6 NVIDIA H100 GPUs on AWS** for rapid large-scale experimentation.
  Developed high-quality plan-and-code pairs for post-training data and supported collection of debugging traces from fixing LLM-generated Python code, **merged 30+ pull requests** covering task conversions, dataset integrations, bug fixes, and evaluation improvements.
- **Project Vulcan:** Collaborated with Meta to extend OpenAIs MLE-bench into a comprehensive benchmark by transforming Kaggle competitions into reproducible, Docker-based evaluation tasks.
  Extended MLE-bench with recent NeurIPS, ICML, and ICLR datasets across computer vision, NLP, time-series, and tabular domains to reflect modern ML challenges.

Co-Founder and AI Engineer
Sep. 2025 – Present
FeedHire, UK — London, UK (Remote)

- Orchestrated a **microservices architecture** decoupling high-frequency data collection from AI extraction logic using **Python (FastAPI)** and **Docker** for independent scalability.
- Developed a multi-stage AI extraction pipeline with a **fallback chain (Groq, Bedrock, Ollama)**, ensuring high-accuracy job property extraction even during provider rate limits.
- Implemented a production-grade HTTP client featuring **circuit breakers**, **connection pooling**, and **exponential backoff** to maintain service reliability across external API dependencies.
- Engineered an automated **"Gold Standard" dataset generator** that captures raw/structured data pairs to facilitate future model distillation and performance benchmarking.
- Automated social media engagement by integrating **LinkedIn and Discord APIs**, using AI to generate post summaries with custom visuals for real-time promotion.
- Built scalable scrapers for **Telegram, Reddit, and GitHub** with delta-detection logic, reducing redundant processing and optimizing resource utilization.

Founding ML Research Engineer
Mar. 2025 – Nov 2025
Curify-Ai, China — London, UK (Remote)

- Implementing end-to-end video translation pipelines with transcription, translation, voice cloning, and lip-sync synchronization, optimizing multi-stage workflows to reduce processing time
- Developing temporal alignment algorithms for voice-sync generation using GPT-based post-editing to resolve audio overlapping and improve visual-audio synchronization
- Built and deployed **FastAPI microservices** (ChatterBox multilingual TTS, WhisperX transcription, PaddleOCR) with Docker on Azure cloud
- Integrated state-of-the-art models including ElevenLabs and XTTS for voice cloning
- Worked on scalable backend with PostgreSQL, queue-based job orchestration, credit/subscription systems, and RESTful APIs

AI/ML Research Intern
Jul. 2024 – Oct. 2024
City, University of London, UK — London, UK

- Integrated clinical and phylogenetic data to enhance ML models for predicting lung cancer patient survival using a novel dataset with no prior research
- Experimented with survival model techniques and feature engineering to improve patient survival time predictions
- Collaborated with Dr. Robert Noble (Oxford alumni) and Dr. Tillman Weyde, integrating mathematical and CS expertise to interpret results and refine models (publication in progress)

AI Engineer **Sep. 2022 – Sep. 2023**
Webomates, USA Mumbai, India (remote)

- Deployed and monitored Machine Learning models on AWS EC2, with strong emphasis on utilizing advanced functionalities and reliability of Linux systems
- Designed and implemented a novel approach using **AWS SQS** to replace Flask-based request handling, leading to team-wide adoption and **30-50% increase in efficiency**
- Developed a multi-modal ML pipeline for detecting feature changes on web pages using HTML, user interaction logs, and visual data
- Implemented a hybrid model combining XGBoost for HTML data, CNN for image data, and Random Forest for final feature detection, **improving accuracy by 25%**
- Designed Flask application leveraging OpenAI API to generate and enhance test cases for TestOps

Machine Learning Engineer Intern **Oct. 2021 – Jan. 2022**
ResoluteAI, India Mumbai, India (remote)

- Worked on a U-NET Neural Network model architecture to detect defects in fabric videos, handling video-to-frame conversion, image augmentation, and model training/testing
- Led a team of 4 interns on image annotation tasks with OpenCV
- Extracted regions of interest (ROI) and labeled objects using the Canny edge detection algorithm

PROJECTS

- End-to-end video translation pipeline using Whisper large-v3 and multilingual voice cloning (Chatterbox), with automated audio-video synchronization, **11+ languages** support, and deployment on Hugging Face Spaces (demo).
- Fine-tuned Meta-Llama-3-8B using LoRA on the Mac M3 dataset; released on Hugging Face with **200+ downloads**.
- Implemented Super Resolution Residual Network (SRResNet) and Super-Resolution Generative Adversarial Network (SRGAN) to enhance image resolution, proposing new improvements to the models, achieving better performance on benchmarks. Also presented first baselines for the Fréchet Inception Distance (FID) following their work.
- Completed a thesis-based internship research project on predicting cancer patient survival times using clinical and genetic data, applying linear and non-linear regression methods, with a publication in progress.

EDUCATION

MSc Artificial Intelligence (Grade: Merit) **Sep. 2023 – Oct. 2024**
City, University of London London, UK

- **Achievements**: Selected for ML in Lung Cancer Research Internship

BSc Information Technology (Grade: Distinction) **Jul. 2019 – Mar. 2021**
University of Mumbai Mumbai, India

AWARDS & ACHIEVEMENTS

- **AIUK 2025 Specialist:** Invited to present at UK national event for mental health AI contributions. **(Mar 2025)**
- **Winner of Harmony AI Challenge in collaboration with UCL and 3 other universities. (Jan 2025)**
- **Co-Founder, Feedhire:** Scaled AI job platform to 1500+ global users traffic in 6 months. **(Sep 2025)**
- **Fast-Track Promotion:** Promoted within 3 months of a 6 month internship at Webomates, USA. **(Jan 2023)**

SKILLS

Machine Learning & Deep Learning: Python, model development, evaluation, production ML systems
NLP & LLMs: Transformers, LoRA fine-tuning, Whisper, text similarity, prompt-based reasoning
Computer Vision: CNNs, GANs (SRGAN), U-Net, image and video processing
Cloud & MLOps: AWS, Azure, Docker, FastAPI, Flask, Linux, Git