Rafi Riyaz
MSc AI | Stanford University Online ML Certified
● +44 7774874773 ● rafa.works313@gmail.com ● LinkedIn ● GitHub ● Portfolio ● HuggingFace

Work Experience

Mercor, London (remote) **ML Engineer (Independent Contractor)** ***Oct 2025 - Present***

- Contributing to **Meta AI Research's expansion of OpenAI's MLE-bench** through **Project Vulcan and Project Launchpad**, converting Kaggle and ML competitions into standardized, Docker-containerized benchmark tasks with reproducible setups and evaluation metrics.
- Extended MLE-bench with recent NeurIPS, ICML and ICLR datasets across CV, NLP, time-series and tabular domains, ensuring the benchmark reflects modern ML challenges.
- Actively worked with a **Jupiter mega-cluster** environment, powered by **6× NVIDIA H100 instances on AWS**, enabling rapid experimentation
- **Merged 30+ PRs** covering task conversions, dataset integrations, bug fixes, evaluation improvements and documentation updates.
- **Developed high-quality *plan-and-code* pairs for post-training data** and helped collect debugging traces from experts fixing LLM-generated Python code.

Curify-Ai, London (remote) **Founding ML Research Engineer** ***Mar 2025 -Present***

- Implementing on **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
- Worked on implementing comprehensive test suites for authentication, video processing, credits management, and admin functionality

City, University of London, London, UK **AI/ML Researcher** ***Jul-Oct 2024***

- 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** integrated mathematical and CS expertise to interpret results and refine models. *(publication in progress)*

Webomates, Stamford, USA (remote) **AI Engineer** ***Sep 2022-Sep 2023***

- Deployed and monitored Machine Learning models on **AWS Elastic Compute Cloud (EC2),** with a strong emphasis on utilising the advanced functionalities and reliability of Linux systems.
- Designed and implemented a novel approach using **AWS SQS service** to replace Flask-based request handling for API applications, leading team-wide adoption and production redeployment, **resulting in around 30-50% increase in efficiency.**
- Developed a **multi-modal machine learning** 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 overall accuracy by 25%.**
- Designed and built a Flask application leveraging **Openai API** to generate and enhance test cases for TestOps.
- Deployed several **NLP and computer vision models**, now actively running in production.
- Authored **articles for AI in the software testing domain**, featured on the company website.
- Developed multiple Python scripts to automate excel reports utilising **AWS Lambda**.
- Gained extensive experience with Amazon Web Services (AWS), including **AWS Lambda, ECS, EC2, IAM, Cloudwatch, S3, SQS, Elastic search.**

- Primarily worked with Linux, OpenAI API, Flask, AWS, SQL, Jenkins, Kibana, NLP, Git, Bitbucket**,** building APIs, data collection, and exploratory data analysis (EDA).
- Recognised for exceptional contributions and promoted by the 3rd month of a 6-month internship

ResoluteAI**, Mumbai, India** **Machine Learning Engineer** ***Oct 2021-Jan 2022***

- 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 with visualisations.
- **Led a team of 4 interns** on image annotation tasks with Open CV.
- Extracted regions of interest (ROI) and labeled objects using the **Canny edge detection algorithm.**
- An intern position to understand and work with production grade **computer vision techniques and projects**.

Education

City, University of London **MSc Artificial Intelligence** ***2023-2024***

- **Grade achieved: Merit**

University of Mumbai **BSc Information Technology** ***2019-2021***

- **Grade achieved: 8.5/10 CGPA**

Research & Projects

- **Co-founded** Feedhire, a **unique AI global job discovery platform,** leveraged **NLP and open-source LLMs** for automated job post extraction, **dockerized services for cloud deployment on Oracle, implemented logging, monitoring, and iteratively improved features**. **Grew to 1000+ users within the 3 months,** with traffic from multiple countries including the **US, UK, and India.** ***(September 2025)*** *(Live)*
- Developed **VideoTrans (Hugging Face Space)** – an end-to-end video translation and voice cloning pipeline that automatically transcribes videos using Whisper large-v3, translates content via Google Translate, and generates voice-cloned audio in the target language using Chatterbox multilingual TTS, with temporal alignment and automated audio-video synchronization. Supports 11+ languages including English, Hindi, Spanish, French, German, Japanese, Korean, and Chinese. *(**October 2025**) (Live)*
- **Fine-tuned Meta-Llama-3-8B model with Low-Rank Adaptation (LoRA)** on M3, released as **F1llama** on Hugging Face, achieving domain-specific adaptability with **200+ downloads**.
- 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.
- 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.

Awards & Achievements

- Invited as **AI specialist to AIUK 2025 (Alan Turing Institute)** by Harmony in **recognition of contributions to their mental health production tool**; compensated **£1,250** for two-day event ***(March 2025).***
- **Winner of Harmony NLP Challenge (DOXA AI Platform)** - Fine-tuned Transformer model reducing MAE from 24 to 20.544, improving psychology survey question alignment with human assessments; £250 prize. Challenge organized by **Harmony with UCL and European universities *(January 2025).***
- Selected for **ML in Lung Cancer Research Internship** (City, University of London CS & Mathematics) using **UK Cancer Research TracerX data**

Certifications

- **AI Agents Course,** *HuggingFace* ***May 2025-Present***
- **Langchain for LLM Application Development***, DeepLearning.ai* ***Aug 2023***
- **Neural Networks and Deep Learning,** *DeepLearning.ai, Coursera* ***Oct 2022***
- **Machine Learning by Andrew Ng,** *Stanford Online, Coursera* ***Jun 2022***

- **Python Master Certification Course,** *Perfect E-Learning* ***Sep 2022***
- **Python and ML for Data Science,** *Kaggle* ***Jan 2021***