



# ACHRAF HABIB

AI Engineering Student

## EDUCATION

 [linkedin](#) |  [github](#)

### • National Institute of Applied Sciences and Technology (INSAT)

Sep 2020 - Sep 2025

Software and AI Engineering degree

## EXPERIENCE

### •Part-Time AI Engineer

Lightray Technologies

Sep 2024 - Present

- Developed financial market prediction models using LLaMA3.1 405B to mimic professional investor workflows
- Established a fine-tuning pipeline with Snowflake, reducing training time by 30%, leveraging a multi-GPU environment with peer-to-peer communication between GPUs for enhanced efficiency
- Managed inference with TensorRT-LLM and Triton, optimizing model deployment
- Built a custom dataset through web scraping techniques, resulting in a 25% increase in training data volume for new LLMs
- **Technical environment:** LLaMA3.1 405B, LLaMA3.1 70B, LoRA, Snowflake framework, multi-GPU with p2p communication, TensorRT-LLM, Triton inference server

### •MLOps Intern

Appicare.AI

Jun 2024 - Sept 2024

- Fine-tuned LSTM, GRU, and Transformer models on the MIMIC-III dataset, improving prediction accuracy by 25%
- Built a web app for real-time patients data prediction and Containerized the app with Docker, deployed on Azure, and established an MLOps pipeline with MLflow for tracking and versioning, increasing model reproducibility by 40%
- **Technical environment:** MLflow, Docker, Streamlit, GitHub Actions, Azure Container Registry, Azure Container Instance, Azure Monitor, MIMIC Dataset

### •Generative AI Intern

VERMEG

Jul 2023 - Aug 2023

- Engineered a chatbot to simplify access to Vermeg's internal Palmera framework, using Langchain's RAG framework, improving response accuracy by 25%
- Optimized document retrieval and response time by 30% using vector embeddings and efficient similarity searches
- Architected a user-friendly Streamlit app for real-time chatbot interactions
- **Skills:** Large Language Models, LangChain, ChromaDB, StreamLit

## ACADEMIC PROJECTS

### • Deployment of a Containerized Application with Kubernetes on AKS via GitLab CI/CD Apr 2024 - May 2024

Insat

- Designed and implemented a containerized deployment on AKS using GitLab CI/CD and Helm charts, aimed at comparing different deployment strategies in terms of downtime. This project resulted in a 40% improvement in deployment speed, and a 20% reduction in deployment failures after assessing blue/green and recreate strategies

### •Full Stack Application

Dec 2024 - Feb 2024

Insat

- Implemented DEVHUB, a comprehensive full-stack web application aimed for efficient management of projects, customers and consultants
- Employed Angular for the frontend and NestJS for the backend, implementing robust guards, middleware, and JWT authentication to ensure a secure and scalable application architecture

### •Image Anomaly Detection

Jan 2024 - Jun 2024

Insat

- Completed an end-of-year school project focused on Image Anomaly Detection using deep learning, implementing an autoencoder-based algorithm that enhanced accuracy and reliability in detecting anomalies during image reconstruction

## TECHNICAL SKILLS

**Programming Languages:** Python, Java, TypeScript, JavaScript, C, SQL

**Libraries & Frameworks:** Scikit-learn, TensorFlow, PyTorch, Seaborn, Flask, Angular, NestJS, Streamlit, HuggingFace, Matplotlib, LLama Factory, Selenium, Snowflake, Groq, Langchain, openAI

**Tools:** MLflow, Azure, GitHub Actions, CI/CD, Kubernetes, helm, Terraform, Linux, Vector Databases, Docker, TensorRT-LLM, Micro-services, Grafana, Prometheus

## EXTRA CURRICULAR ACTIVITIES

### •Google developer student club: Active member

Sep 2022 - Jan 2023

### •Securinets Club: Active Member

Sept 2022 - Jun 2023

## LANGUAGES

**Arabic:** Native   **English:** Fluent   **French:** Fluent