Ali Asgher Mohammed

Karachi, Pakistan +92-3328041453
amohammed.bee20seecs@seecs.edu.pk Linkedin Github **Portfolio**

Education

National University of Sciences & Technology October 2020 – May 2024
Bachelors in Electrical Engineering (GPA: 3.80/4.00) *Islamabad, Pakistan*

- **Awards:**
  - NUST High Achievers Award 2022-2024
  - NUST Merit Based GPA Scholarship 3rd - 6th Semester

Experience

Sketric Solutions Oct 2023 - Present
Machine Learning Engineer *Remote*

- Designed and deployed a production-grade end-to-end AI pipeline integrating OpenAI Agents API, custom workflow layers, S3 asset management with presigned URLs, SQS-based credit processing, and robust retry/error-handling for SaaS chat applications
- Reduced Time-to-First-Token (TTFT) by 70–90% by implementing real-time SSE streaming for LLM responses, achieving sub-100ms first-token delivery with support for stream cancellation and heartbeats
- Cut LLM API costs by 60–75% through intelligent context management that automatically summarizes conversations at 80% token capacity, compressing history to 500 words while preserving continuity and avoiding context overflow
- Improved database throughput by 3–5× using connection pooling, async DynamoDB operations with adaptive retries, and transactional batch writes, reducing per-operation latency from 10–20ms to 2–5ms
- Implemented production MLOps infrastructure including real-time LLM response streaming with cancellation support, automatic context summarization at 80% token capacity, distributed tracing with span collection, and async background persistence reducing API response latency by decoupling database writes from request handling
- Led development of a multi-platform AI agent integration system (Instagram, WhatsApp, Telegram, Zapier) using AWS Lambda, API Gateway, and DynamoDB, with secure webhooks (HMAC, OAuth 2.0), async event processing, and lifecycle management—achieving 99.9% reliability and 300% reach expansion.

Information Processing and Transmissions Lab June 2023 – December 2023
Researcher *NUST*

- Worked on maximizing data rates of energy harvesting devices in CR-NOMA networks using deep reinforcement learning (DRL)
- Solved a continous action spaced optimization problem using DRL and convex optimization
- Contributed in writing a research paper as a first author to publish our findings in a journal publication

TruID June 2022 – September 2022
Machine Learning Intern *NUST*

- Completed hands-on labs on various Python libraries for data visualisation and handling
- Handled data extraction from XML annotation files
- Worked in a team to devise strategies for passive liveness detection using deep learning

Projects

AquaGuard (Final Year Project) | *Drone Avionics, Embedded Systems, Python Programming*

- Developing a complete end-to-end product of a search and rescue UAV to autonomously detect and rescue a drowning person by dropping a float (Nominee of Rector's Gold Medal, Top 7 FYPs)

SketricGen | *AI Workflows, Agentic Orchestration, AWS*

- Full-stack SaaS platform for building and deploying AI chatbots across websites and messaging platforms (WhatsApp, Instagram, Telegram, Slack, Zapier), enabling agents to integrate with 2800+ apps. Built with AWS Amplify Gen 2, OpenAI GPT, and serverless architecture, featuring no-code agent creation, knowledge base training, visual workflow editor, and real-time chat infrastructure using AWS Lambda (Python), DynamoDB, and S3)

Resumes Ranker | *GenAI, AWS amplify, AWS DynamoDB*

- Designed and implemented an end-to-end solution tailored for the recruiting industry, enabling efficient scoring and detailed analysis of bulk resumes. Leveraged cutting-edge technologies, including OpenAI and cloud computing, while employing advanced prompt engineering techniques to optimize the outputs of large language models (LLMs).

Contract Sense | *Python, OpenAI, AWS*

- Developed an innovative solution for a contracting company to proactively identify potential government tenders by leveraging web scraping and Generative AI (GenAI). The solution enabled the company to detect tenders faster, significantly improving their bid success rate by 100 %.

Visualizer AI | *Python, OpenCV, Meta SAM, Ultralytics YOLO, Computer Vision*

- Developed innovative algorithms for computing perspective transforms of walls, floors, and countertops in 2D indoor images using semantic scene understanding. These algorithms matched the performance of industry leaders who had been refining their solutions for 5+ years, while delivering results 3x faster

Drown AI | *Python, OpenCV, Ultralytics YOLO, MLOPs, SQLite*

- Developed an end-to-end analytics solution for Dubai beaches, utilizing AI to detect drowning and swimming individuals. Led the backend development, including crafting database operations with SQLite and implementing ML model training pipelines using MLflow

SniperPlate | *OpenCV, Ultralytics, Parseq, RPI*

- Deployed an Automatic License Plate Recognition (ALPR) system on a Raspberry Pi 5 by training custom object detection and OCR models. Enhanced the model's processing speed from 600 ms to 120 ms by implementing NCNN model quantization techniques. Initially achieving an overall frame rate of 2.5 FPS, we optimized performance to 8.0 FPS through advanced techniques such as threading and queue management in Python.

Publications

- A. A. Mohammed, M. W. Baig, M. A. Sohail, S. A. Ullah, H. Jung and S. A. Hassan, "Navigating Boundaries in Quantifying Robustness: A DRL Expedition for Non-Linear Energy Harvesting IoT Networks," in *IEEE Communications Letters*, doi: 10.1109/LCOMM.2024.3451702. Link to paper

Technical Skills

Languages: Python, C/C++
Technologies: OpenCV, Pytorch, TensorFlow, Flask, SQLalchemy, TensorRT, Numpy, Timm, Albumentations, Hugging Face, Jetson Nano, Raspberry PI, Gazebo, ROS, Git
Concepts: Data Structures & Algorithms, Artificial Intelligence, Computer Vision, Robotics, Object Oriented Programming
Certifications:

- Neural Networks and Deep Learning - DeepLearning.ai
- Improving Deep Neural Networks: Hyperparameter Tuning, regularization, and optimization - DeepLearning.ai
- Structuring your Machine Learning Projects - DeepLearning.ai
- Convolutional Neural Networks - DeepLearning.ai
- AI Capstone Project with Deep Learning - IBM
- Deep Learning with PyTorch : Object Localization - Coursera
- What is Data Science? - IBM
- Tools for Data Science - IBM
- 5-Days boot camp on Deep Learning for Object Detection and Semantic Segmentation In Visual Data