MIR IMAAD ALI

AI / MACHINE LEARNING ENGINEER · APPLIED LLM & AGENTIC SYSTEMS

Dubai, UAE | mirimaadali1@gmail.com | +971 56-3347201 | GitHub | LinkedIn | Golden Visa Holder

PROFESSIONAL SUMMARY

AI/ML engineer who turns state-of-the-art models into production-grade systems deployed for enterprise clients. Owns problems end-to-end — architecture, data, model design, deployment, and iteration — for LLM applications, agentic reasoning workflows, and scalable inference infrastructure serving long-context production traffic. Published researcher in Transformer-based applications. Strong in Python, PyTorch, and TensorFlow, with deep experience operating LLM pipelines (RAG, multi-agent systems, evaluation) and shipping real systems for regulated, real-world environments.

PROFESSIONAL EXPERIENCE

Sentiment AI **July 2025 – Present**

AI Engineer / Technical Lead

- Own the full technical stack for an enterprise AI behavioral-analytics product in production: LLM application architecture, agentic pipelines, inference infrastructure, and deployment — embedded in a small team shipping directly to an enterprise client.
- Designed and deployed a multi-tenant LLM pipeline for long-context production inference (~55K tokens/call), building a node-based DAG executor on Temporal with a Postgres semantic layer; evaluated and benchmarked open-weight models on dedicated GPU endpoints (2×H100) against hosted APIs for cost, latency, and quality.
- Built the applied-LLM tooling stack — PydanticAI, DSPy, LiteLLM routing, RAGAS/Langfuse evaluation — and a framework that dynamically spins up to 40 customizable voice agents that autonomously converse and are scored by an external LLM judge.
- Engineered a context-aware conversational system that dynamically switches behavior on detected user emotion, and architected real-time STT and nudge/analytics integrations (Genesys AudioHook, D365) for enterprise contact-center deployments.
- Drove data-residency and compliance architecture (regional inference via cloud-native endpoints, SOC 2) so AI systems could be deployed for regulated enterprise and government-adjacent clients.

Chalmer St. Consulting **July 2023 – January 2024**

Software Engineer Intern

- Engineered a scalable auditing platform with React.js, Node.js, and Firebase, applying design patterns aligned with production ML system deployment.
- Refactored data models and real-time pipelines for low-latency performance, improving operational productivity by 30% and enabling production-grade usage across multiple teams.

SELECTED PROJECTS

PaperBot — AI-Powered Research Assistant **June 2025**

- Architected an end-to-end document QA (RAG) system with LangChain, FAISS, and GPT for semantic understanding of research papers.
- Built a modular Django + REST backend integrating PDF parsing, chunking, vector embedding, and LLM retrieval — cutting query response time by 50% with improved search accuracy.

Continuous Sign Language Recognition (CSLR) for ASL **May 2025**

- Built a CSLR system on the How2Sign dataset using a ResNet18 + BiLSTM architecture for sign recognition from video, optimizing the preprocessing and feature-extraction pipelines.

PUBLICATIONS

Evaluating Deep Learning & Transformer Models for Suicidal Ideation Detection **Dec 2025**

- Accepted at the 5th Intl. Conference on Computers and AI Technology (CAIT 2024), Hangzhou, China (ISBN: 979-8-3315-3088-4). Achieved 98.6% accuracy with BERT, outperforming LSTM-CNN and Bi-LSTM-CNN baselines.

PromptRefine — 14th EAI ArtsIT Conference **Nov 2025**

- Designed a framework to refine vague, ambiguous, or biased prompts before they reach LLMs; evaluated across proprietary models (OpenAI, Anthropic) for clarity, intent alignment, and bias mitigation.

EDUCATION

Heriot-Watt University **September 2021 – July 2025**

BSc. Computer Science (Honours)

- First Class Honours, 4.0 GPA — UAE Golden Visa recipient.

SKILLS

LLM / GenAI: Agentic Workflows, RAG, Reasoning Models, Multi-Agent Systems, Prompt Engineering, LLM Evaluation (RAGAS, Langfuse), PydanticAI, DSPy, LiteLLM, LangChain, FAISS

ML / DL: PyTorch, TensorFlow, Transformers, JAX (familiar), Traditional ML (SVM, Regression, Tree-based, Clustering), Model Evaluation & Optimization

Backend & Infra: Python, Go, FastAPI, Django, Flask, Node.js, Temporal, Postgres, SQL, Docker, Kubernetes, CI/CD, Git

Production AI: GPU Inference (H100), Long-Context Pipelines, Data Residency / Compliance (SOC 2), Real-Time Systems, Agile (Scrum)

CERTIFICATIONS

AI-900: Microsoft Azure AI Fundamentals · Deep Learning: Getting Started · NLP with TensorFlow · AI Foundations: Thinking Machines · ML Foundations: Linear Algebra