

Ali Ahmed Samir

Senior AI Engineer & Delivery Owner · Computer Vision · Agentic AI · On-Premise Delivery

Abu Dhabi, UAE · +971 50 608 2836 · ali.samir.1744@gmail.com · linkedin.com/in/ali-ahmed-samir · github.com/Supernova1744 · ali-samir.com

SUMMARY

AI engineer and delivery owner at Tahaluf AI Emarat Technical Solutions, building production computer vision and agentic AI for UAE government and enterprise clients. Architected an agentic AI framework, orchestrated with LangGraph and combined with computer vision models in one pipeline, that monitors most of Abu Dhabi's public parks for the Department of Municipalities and Transport (DMT). Removed a request-serialization bottleneck in a federal ID-verification API by migrating inference to ONNX Runtime, raising sustained per-node concurrent throughput from about 1 to roughly 100 requests on CPU alone, no GPU involved. Manages deployment inside security-restricted government entities (ICP) under strict on-premise and enterprise-security constraints, mentors AI engineers on implementation quality, and leads an active engagement porting the team's AI solution onto a global automotive OEM's onboard compute unit. Native Arabic speaker based in Abu Dhabi, with direct GCC public-sector delivery experience.

EXPERIENCE

Tahaluf AI Emarat Technical Solutions · Senior AI Engineer & Delivery Owner

Sep 2023 – Present

Government technology firm: AI and computer vision systems for UAE federal entities and international public-sector clients

- Own end-to-end delivery across Tahaluf's UAE government AI portfolio, setting technical and quality standards, mentoring AI engineers on implementation, and managing deployment inside security-restricted entities (ICP) under strict on-premise constraints.
- Built an agentic AI framework, LangGraph-orchestrated agents combined with computer vision models in a single pipeline, for Abu Dhabi's Department of Municipalities and Transport (DMT) that monitors violations across most of the emirate's public parks.
- Diagnosed a request-serialization bottleneck inside a MediaPipe/TensorFlow Lite computer vision component on a federal ID-verification API (ICP); migrated the inference path to ONNX Runtime and validated it through unit, integration and load testing, raising sustained concurrent throughput per CPU-only node from about 1 to roughly 100 in-flight requests.
- Accelerated multi-model inference throughput by 127% (36 to 82 FPS) by re-architecting the inference engine with optimized matrix operations and hardware-accelerated GPU execution.
- Re-architected the AI pipeline of a mobile Smart City surveillance platform processing 180 simultaneous live video feeds across Abu Dhabi, cutting system failures by 20% and making the platform fault-tolerant under continuous RTSP stream load.
- Led MLOps standardization and hands-on training for 7 engineers, establishing data-versioning workflows that cut model training cycle time across the department.

Tahaluf AI Emarat Technical Solutions · Artificial Intelligence Engineer

Jan 2021 – Sep 2023

- Reduced OCR inference time by 65% at 99.84% text accuracy by architecting a custom single-shot text recognition model with a convolutional backbone optimized for speed and accuracy.
- Recovered a failing proof of concept from 28% to 89% accuracy by diagnosing hidden integration constraints and redesigning the architecture to synchronize the AI and embedded subsystems.

SELECTED PROJECTS

- **Smart Parks, agentic AI park compliance (DMT, Abu Dhabi).** LangGraph, agentic AI, computer vision.
- **ID-verification API performance engineering (ICP).** ONNX Runtime, MediaPipe, TensorFlow Lite, concurrency engineering.
- **Generative AI Fan Dashboard (Presidential Cup, UAE).** Personalized social posts for match viewers based on team affiliation and sentiment about the match; LLM content generation.

- **Edge AI deployment (global automotive OEM, ongoing).** Leading a collaboration porting the team's AI solution onto a vehicle's onboard compute unit as part of a vehicle-integration program.

OPEN SOURCE & WRITING

- **Vision.rs** (Rust, 38 stars): object detection with a gRPC API for remote inference. **YOLOv8-nms** (Python): NMS incorporated into the YOLOv8 graph. **open-multi-agent-rs** (Rust): multi-agent LLM orchestration. **Football_Vision_AI** (Python, built from a public tutorial).
- Substack, *Deep Diaries*: ddbyali.substack.com

EDUCATION, RESEARCH & CERTIFICATIONS

- BSc Computer Science, Faculty of Computers and Information, Kafr El Sheikh University, 2016–2020.
- Evolutionary Algorithm-Based CNN for Predicting Heart Diseases, *Computers & Industrial Engineering* 161 (Elsevier, 2021).
- AWS Certified Machine Learning, Specialty (earned Nov 2021). Udacity: Agentic AI Nanodegree; Generative AI Nanodegree; AI Agents with LangChain. Udemy: Product Management for AI and Data Science.

SKILLS

Languages: Python, Rust, SQL, Bash/Shell

Agentic AI & LLM: LangGraph, LangChain, LlamaIndex, RAG, HuggingFace Transformers, PEFT/LoRA fine-tuning

Computer vision & inference: OpenCV, object detection, ONNX, TensorRT, NVIDIA Triton, vLLM, model optimization, edge AI

MLOps & delivery: Docker, CI/CD, MLflow, AWS, on-premise and security-restricted deployment, data versioning, Linux

APIs & tools: Flask (REST APIs), Git, vector databases (Qdrant, pgvector, Weaviate)

Languages spoken: Arabic (native), English (professional)

Client systems are not public.