ALI HAMAMA

AI Engineer | Production AI & Applied Machine Learning

+971-54-250-7300 | ali.y.hamama@gmail.com | Dubai, UAE

PROFESSIONAL SUMMARY

AI Engineer with hands-on experience building, integrating, optimizing, and deploying production AI systems in Python. Experienced in computer vision, PyTorch-based model workflows, vision-language models, and GPU inference optimization using TensorRT and NVIDIA Triton Inference Server. Skilled in LLM API integration, retrieval-augmented generation (RAG), embeddings, and vector search for knowledge retrieval and question-answering applications. Background includes model evaluation, modular system integration, robotics, and enterprise workflow automation, translating technical and business requirements into maintainable software solutions.

TECHNICAL SKILLS

Programming: Python, C++
AI / ML: PyTorch, Computer Vision, Segmentation, OCR, Vision-Language Models, Hugging Face Transformers
LLM Applications & Retrieval: LLM APIs, Retrieval-Augmented Generation (RAG), Embeddings, Vector Search, Vector Databases
AI Deployment: NVIDIA TensorRT, NVIDIA Triton Inference Server, GPU Inference, Dynamic Batching, Model Optimization
Robotics: ROS2, NVIDIA Isaac Sim, LiDAR / Camera Perception
Enterprise Systems: Bitrix24 CRM, Workflow Automation, CRM Data Modeling, Dashboards
Current Learning Focus: Structured Outputs, Tool Calling, FastAPI

PROFESSIONAL EXPERIENCE

Micropolis Robotics and AI | 2024 - Present

AI Engineer

- Build and maintain computer-vision pipelines for video and image processing, improving robustness across real-world operating conditions.
- Develop zero-shot vision workflows using SigLIP vision-text embeddings and prompt-based scoring to enable flexible scene understanding without task-specific retraining.
- Lead optimization and production integration of AI model stacks by profiling bottlenecks, enabling batching, and tuning GPU inference with TensorRT and NVIDIA Triton Inference Server for lower latency and higher throughput.

Robotics Software Engineer

- Worked with NVIDIA Isaac Sim and collaborated on internal simulation extensions supporting autonomous-system workflows.
- Developed ROS2 tools and software components to streamline robotics pipelines and improve developer productivity.
- Researched and prototyped solutions for autonomous vision and perception challenges, translating findings into practical system improvements.

Meraki Real Estate Developer | 2023 - 2024

CRM Data Analyst

- Translated business requirements and operational workflows into Bitrix24 CRM configurations and automated processes.
- Designed CRM data models supporting reliable reporting, customer tracking, and business operations.
- Created dashboards and workflow automations to improve visibility and reduce repetitive operational work.

SELECTED PROJECTS

AI-Powered Dubai Real Estate Assistant | Ongoing Project

Developing an intelligent agent to simplify access to Dubai real estate data, service information, and property listings.

- Enable natural-language queries about historical transactions in Dubai’s real estate market.
- Provide answers grounded in the Dubai Land Department (DLD) service guide and FAQs.
- Aggregate property listings through web scraping across multiple real estate websites, matching results to users’ preferences, budgets, and desired locations.
- Build toward an **agentic RAG** architecture, with **LangChain** for workflow and tool orchestration, **BGE-M3** for multilingual text embeddings, and **pgvector** for vector storage and similarity search.

Technology stack (in use or planned): Agentic RAG, LangChain, BGE-M3, PostgreSQL/pgvector, web scraping.

Production GCC ANPR System - End-to-End AI Deployment

- Collected and curated a real-world dataset reflecting GCC plate formats, driving environments, and camera conditions.
- Benchmarked multiple detection and OCR approaches on a custom evaluation set and selected models using measurable accuracy and latency criteria.
- Designed a modular end-to-end ANPR pipeline covering stages, interfaces, post-processing, and outputs.
- Fine-tuned key models and optimized GPU inference with TensorRT FP16, dynamic shapes, and batching where applicable.
- Packaged the pipeline as NVIDIA Triton Inference Server services and integrated it with the central platform for reliable streaming and communication across modular components.

Drive-able Area Extraction for Robot - Research Project

- Fused LiDAR and camera data to identify drivable regions and filter corresponding areas from LiDAR point clouds.
- Used AI segmentation models to detect drivable areas in image space across varying scene conditions.
- Converted segmentation masks to point-cloud representations, applied calibration and coordinate transformations, and produced a refined non-drivable point cloud.

EDUCATION

Tishreen University | 2016 - 2022

Bachelor's Degree in Information Technology Engineering

Major: Artificial Intelligence

LANGUAGES

Arabic - Native | English - B2