Assem Alhomsi

Dearborn Heights, Michigan | P: (313)-247-0367| assem.h2001@gmail.com

AI Engineer specializing in LLM deployment, RAG systems, and scalable backend services. Experienced in building production-ready AI pipelines, vector search systems, and agentic workflows using GPU infrastructure and modern orchestration tools.

WORK EXPERIENCE

Machine Learning & Generative AI Engineer — boxMind.ai

July 2023 – July 2025

- Configured and deployed on-prem LLMs on NVIDIA A100 infrastructure, optimizing environment setup and model serving to improve inference throughput by ~20%.
- Built scalable Flask inference services for internal and external AI access, enabling consistent low-latency responses for LLM-powered applications.
- Developed and maintained RAG pipelines using Milvus, LangChain, and LangGraph, with RAGAS and LangSmith evaluation improving retrieval accuracy by 15–20%.
- Engineered agentic LLM systems for automated content generation and workflow automation, reducing manual work by ~30%.
- Created data ingestion pipelines, embedding workflows, and text preprocessing modules supporting multi-document retrieval and enterprise knowledge integration.
- Automated deployment, testing, and monitoring with Docker, GitHub Actions, and Airflow to support continuous delivery of AI microservices.
- Designed secure, maintainable backend components using Flask, SQL, and containerized services powering production AI features.

AI Development Intern — ISS (Software Hive)

August – September 2022

- Built OCR pipelines using EasyOCR, PaddleOCR, and Keras, achieving 100% license plate recognition accuracy in testing scenarios.
- Performed model performance tuning, image preprocessing, and integration with NVR systems.

EDUCATION

Beirut Arab University Beirut, Lebanon

Bachelor’s of Computer Engineering 2019-2023

Additional Experience (Part-time/Side Roles)

Handshake AI Remote

Fellow November 2025 – Present

- Performing data annotation and quality review tasks to support model training and evaluation.

W Institute Livonia, MI

AI, Python, & Robotics Tutor December 2025 – Present

- Teaching foundational coding, robotics, and introductory AI concepts to students.

Technical Skills

LLMs & AI: LLM deployment (A100, vLLM, Hugging Face models), LangChain, LangGraph, embeddings, agentic workflows, RAG pipelines, OpenAI API
Backend & APIs: Flask, REST APIs, Docker, Linux server administration, CI/CD, SQL (MySQL/PostgreSQL)
RAG & Vector DBs: Milvus, Chroma, Pinecone, FAISS, RAG evaluation (RAGAS, LangSmith)
Automation & Data Engineering: Apache Airflow, GitHub Actions, n8n, Spark, data ingestion pipelines
ML/CV: PyTorch, TensorFlow, scikit-learn, OpenCV, OCR models
Cloud: AWS (Lambda, API Gateway, S3), containerized microservices