HAMZA BOUGHANIM

AI & Backend Engineer | Production LLM & MLOps

Based in Morocco | Open to Remote & International Opportunities

hamzaboughanim06@gmail.com | hamzaboughanim.com | +212677060232

PROFESSIONAL SUMMARY

AI Engineer & Backend Engineer with 3+ years of experience building and scaling production-ready LLM and ML systems. Specialized in model training, fine-tuning, RAG pipelines, and multi-agent orchestration, with a strong backend foundation in FastAPI, vector databases, and MLOps. Experienced in delivering secure, high-performance AI services from architecture to production.

PROFESSIONAL EXPERIENCE

Freelance AI Engineer – Satellite Computer Vision

GeoAP *Dec 2025 – May 2026 | Remote, Morocco*

- **End-to-End Satellite CV Pipeline:** Architected and delivered a production-grade MLOps pipeline for detection and semantic segmentation of agricultural and urban objects from satellite imagery, processing **180 source images** into **12,000+ training tiles** via overlap-aware tiling with zero data leakage across train/val/test splits.
- **Multi-Model Architecture:** Trained three complementary models — `YOLO11m` object detection (mAP50 **0.836**), `YOLO11m` instance segmentation (mAP50-mask **0.737**), and fine-tuned **SAM1 ViT-H** semantic segmentation (mIoU **0.804**) — covering **10 object classes** including buildings, trees, solar panels, swimming pools, cultivated parcels, rivers, and roads.
- **SAM1 Fine-Tuning:** Fine-tuned SAM1 ViT-H mask decoder (4M trainable params, 300M-param encoder frozen) with pre-computed embedding caching across **9,958 tiles**, reducing epoch time from **88 min to 5 min** on a T4 GPU. Implemented custom Dice + Focal loss and `torch.amp` mixed-precision training.
- **GIS-Ready Output:** Inference results exported as **GeoJSON FeatureCollections** with real-world WGS84 coordinates, enabling detected objects and segmentation masks to be visualised directly on real maps in **Google Earth**, **QGIS**, **Leaflet**, and any GIS-compatible tool.
- **SAHI Tiled Inference:** Integrated Slicing Aided Hyper Inference (SAHI) for detection on high-resolution satellite imagery, enabling robust small-object detection across arbitrary image resolutions.
- **FastAPI Inference Service:** Built a `FastAPI` REST API with four inference endpoints, lazy singleton model loading, and an interactive dark-theme frontend with live legend and GeoJSON download.
- **MLOps & CI/CD:** Configured a full `DVC` pipeline DAG (12 reproducible stages) with Roboflow dataset versioning. Implemented `GitHub Actions` CI/CD with linting, 25 unit tests, API smoke tests, and Docker build validation on every push.
- **Tech Stack:** `PyTorch` · `YOLO11m` · `SAM1 ViT-H` · `SAHI` · `FastAPI` · `rasterio` · `Shapely` · `OpenCV` · `DVC` · `Roboflow` · `Lightning AI` (T4 GPU) · `Docker` · `GitHub Actions`

AI Engineer / Software Engineer

Assure Solutions – Client: MAMDA-MCMA (Major Insurance Group)
Sep 2023 – Present | Rabat & Kenitra, Morocco

- **AI-Powered Document Automation Platform:** Architected and deployed an intelligent document processing system integrating OCR, NLP, and vector search for insurance workflows.
- **Hybrid OCR Engine:** Developed a Tesseract + EasyOCR pipeline achieving **91.3% accuracy** on Arabic/French documents with intelligent text extraction and reconstruction.
- **Semantic Search Engine:** Built Pgvector vector database with HNSW indexing enabling sub-second semantic retrieval across **760+ insurance documents**, improving information access by **>40%**.
- **LLM Integration:** Implemented Phi-3 LLM via Ollama for local Named Entity Recognition and document classification with **>90% precision**, automating manual categorization.
- **Secure Infrastructure:** Deployed S3 storage with AES-256 encryption for GDPR compliance and full audit trails, implementing Zero-Trust security principles.
- **Full-Stack Interface:** Delivered web application for document upload, semantic search, and validation, integrating ChatBot Assistant for contextual queries and summarization.
- **Insurance System Automation:** Engineered automation for RENP system (Recovery of Unpaid Installments) and formal notice generation using JasperReports, streamlining client communication.
- **Performance Optimization:** Optimized SQL Server stored procedures and resolved critical production bugs, reduc-

ing system latency by ~**30%**.

Software Engineer

Assure Solutions *Aug 2023 – Present | Kenitra, Morocco*

- Developed and maintained core insurance modules (Client, Victim, Sinister Management) using PHP Phalcon and SQL Server.
- Designed and implemented a secure RBAC/PBAC system for 11 distinct user profiles, controlling access to sensitive data and features.
- Architected a Document Management System (GED) using MinIO for secure storage and retrieval of insurance documents.
- Integrated OTP authentication to strengthen security for user authorization flows.

SELECTED AI & BACKEND PROJECTS

Premium Secure AI Backend (Zero-Trust LLM Infrastructure)

Personal Project / Production Blueprint

- Architected a **production-grade, GPU-aware AI inference backend** unifying OpenAI and local `Ollama` models (`Phi` , `Llama-2` , `Mistral`) behind a single hardened `FastAPI` service.
- Implemented **enterprise-grade security**: `Argon2-id` password hashing (GPU-resistant), short-lived scoped JWTs, LLM firewall for pre-inference prompt injection/PII blocking, and `Redis` based rate limiting (requests, tokens, GPU-seconds).
- Designed **full audit trails** with structured JSON logging for compliance (SOC 2, GDPR) and built a `Docker Compose` stack for one-command deployment (API, Redis, Ollama).
- **Demonstrated Impact**: Secured AI service pattern where stolen credentials/tokens yield minimal, auditable access directly applicable to regulated healthtech/fintech AI agent roles.

M3AE – Multi-Modal Medical Vision & Language Model

Research Project | `PyTorch` , `Transformers` , `Multi-Modal Language`

- Developed a **Transformer-based multi-modal architecture** using Masked Autoencoders for joint understanding of medical images and clinical text.
- Pre-trained on the **ROCO dataset (81k radiology image-text pairs)** and fine-tuned on **VQA-RAD**, achieving **state-of-the-art results** on medical visual question answering.
- Applied deep learning techniques: CNN feature extraction, Transformer attention mechanisms, and contrastive learning for vision-language alignment.

AI-Powered Chatbot (RAG)

Professional Project | `FastAPI` , `ChromaDB` , `LangChain` , `React`

- Implemented a Retrieval Augmented Generation **RAG** chatbot using **HuggingFace LLMs (Flan-T5, BART)** and **ChromaDB**.
- Utilized **LangChain** for orchestration, enabling context-aware conversations and document-based Q&A.
- Built an interactive **React** frontend for real-time chat visualization and seamless user interaction.

Cryptocurrency Forecasting & Automated Trading System

Personal Project | `Python` , `Scikit-learn` , `XGBoost` , `LSTM` , `Binance API`

- Created a **ML-powered forecasting pipeline** using ARIMA, XGBoost, Random Forest, and LSTM models for ETH/USDT prediction.
- Built a **custom backtesting framework** with dynamic risk controls, achieving **5.41% ROI** and **MAE of 0.0065** in volatile market conditions.
- Integrated classical ML and deep learning approaches with real-time market data and sentiment analysis (via LLMs) into an `n8n` automated workflow.

CORE TECHNICAL EXPERTISE

AI & LLM Engineering

- **Production LLM Systems:** RAG pipelines, agent orchestration (`LangChain`), evaluation frameworks, prompt engineering, fine-tuning/ `LoRA`

- **Model Integration:** OpenAI API, `Ollama` (`Phi`, `Llama`, `Mistral`), AWS Bedrock, Google Gemini
- **Vector Databases:** `ChromaDB, Pinecone, pgvector`
- **Agent Systems:** Multi-agent workflows, tool calling, MCP, A2A architectures, autonomous reasoning
- **Trust & Safety:** LLM firewalls, hallucination control, audit logging, Zero-Trust security

Machine Learning & Deep Learning

- **Architectures:** Transformers, CNNs, RNNs/LSTMs, Masked Autoencoders, BERT-based models
- **Frameworks:** `PyTorch`, `TensorFlow`, `Scikit-learn`, `Pandas`, `NumPy`, `Keras`
- **Classical ML:** SVM, Decision Trees, K-Means, PCA, XGBoost, ensemble methods
- **Lifecycle:** Research to production training, validation, optimization, deployment
- **Computer Vision:** `OpenCV`, `YOLO11`, `SAM`, `SAHI`, OCR (`Tesseract`, `PaddleOCR`, `EasyOCR`), `VLM`, satellite imagery analysis

Backend & Cloud Engineering

- **Languages:** `Python` (Advanced), `TypeScript` / `JavaScript`, `PHP`, `Java`
- **Frameworks:** `FastAPI`, `Flask`, `NestJS`, `Spring`, `Laravel` / `Phalcon`
- **APIs:** REST, gRPC, WebSockets, OAuth2/JWT, RBAC/PBAC
- **Databases:** `PostgreSQL`, `SQL Server`, `MySQL`, `Redis`, `MongoDB`
- **DevOps & MLOps:** `Docker`, `Kubernetes`, `DVC`, CI/CD (`GitLab CI`, `GitHub Actions`), `Lightning AI`
- **Observability:** logging, monitoring, performance profiling
- **Workflow Automation:** `n8n`, workflow orchestration, ERP/CRM integration

CERTIFICATIONS

- **Artificial Intelligence on Microsoft Azure** — Microsoft (Coursera), March 2026
- **Indexing, Performance Optimization & Functions in SQL Server** — Microsoft (Coursera), March 2026
- **AI for Business Professionals** — HP LIFE (HP Foundation), April 2026

EDUCATION

- **Master's Degree in Artificial Intelligence & Virtual Reality:** Ibn Tofail University, Morocco (2025)
- **Bachelor's Degree in Computer Science & Mathematics:** Ibn Tofail University, Morocco (2023)

LANGUAGES

- **Arabic (Native), English (Professional), French (Professional)**