ATMANE AYOUB

AI/ML Engineer | Full-Stack Software Architect

Email: atmaneayoub10@gmail.com | **Mobile**: +971566937208 | **LinkedIn**: linkedin.com/in/atmaneayoub/ | **Address**: Abu Dhabi, UAE.

SUMMARY

AI/ML Engineer and Full Stack Software Engineer with experience delivering end-to-end, production-grade solutions. I specialized in developing robust AI systems using PyTorch, TensorFlow, and large language models (LLMs), with deep hands-on expertise in retrieval-augmented generation (RAGs), autonomous agents, computer vision, NLP, and advanced deep learning. I combine strong machine learning foundations with a full-stack engineering background to build scalable web and mobile applications using FastAPI, Django, React, Next.js, and modern cloud technologies. I follow best practices in MLOps, deploying reliable, high-performance systems with Docker, Kubernetes, and CI/CD pipelines.

EDUCATION

Master of Science (MSc) in Artificial Intelligence | University of Science and technology Oran | **Algeria** | 2020 – 2022

- Focus: Machine learning, deep learning, computer vision, NLP
- Thesis: AI-based weather forecasting system for Algeria

Bachelor of Science (BSc) in Computer Science | University of Science and technology Oran | **Algeria** | 2017 – 2020

- Focus: Software engineering, algorithms, databases, web/mobile development
- Final project: Recycling platform with website and mobile app to connect households with waste collectors.

WORK EXPERIENCE

AI/ML Engineer | Shory Insurance | June 2025 – present | Onsite | Abu Dhabi, UAE

- Designed and implemented scalable RAG (Retrieval-Augmented Generation) pipelines using vector databases, embedding models, and prompt orchestration to power knowledge-aware chatbots and virtual assistants for internal insurance workflows.
- Architected and deployed scalable RAG and agentic systems using vector databases, hybrid embeddings, semantic search, and dynamic prompt orchestration—powering knowledge-aware chatbots and automated insurance workflows.
- Designed and implemented enterprise ingestion & ETL pipelines (Airflow, cloud-native orchestration) for document chunking, metadata enrichment, embedding generation, and continuous knowledge-base updates.
- Built Shory's standardized RAG/agentic architecture, defining retrieval schemas, memory strategies, hybrid search routing, evaluation frameworks, and company-wide best practices for production AI systems.
- Engineered advanced OCR pipelines for RAG ingestion, converting structured/unstructured documents (policies, claims, forms, emails) into high-fidelity, chunkable representations optimized for retrieval.
- Developed high-accuracy OCR and document-understanding systems for IDs, passports, and fraud detection using PaddleOCR, DeepSeekOCR fine-tuned LLMs/VLMs, and hybrid computer vision techniques.
- Created multimodal AI pipelines (NLP + CV + tabular ML) for claim triage, document intelligence, and anomaly detection—significantly reducing manual review cycles and enhancing fraud-detection accuracy.
- Fine-tuned and deployed LLMs and vision transformers (LoRA, QLoRA, PEFT) and built scalable model-serving architectures (FastAPI, Docker, cloud-native) supporting low-latency, high-throughput inference.
- Built and trained deep learning models in PyTorch and TensorFlow for document classification, image quality enhancement, fraud pattern detection, and domain-specific feature extraction—integrating these models into Shory's RAG, OCR, and underwriting automation pipelines to boost accuracy and end-to-end system performance.

AI/ML Engineer | eData Information | June 2023 – June 2025 | Onsite | Dubai, UAE

- Designed and deployed real-time AI valuation system for vehicles using ensemble regression models and deep learning pipelines, significantly improving market-aligned pricing and quote generation.
- Built an intelligent VIN decoding system using sequence-to-sequence models and ensemble learning techniques, boosting decoding accuracy and robustness for downstream vehicle analytics.
- Developed and integrated multi-agent AI systems and domain-specific chatbots to automate insurance assistance and underwriting, leveraging RAG architectures, context-aware agents, and fine-tuned LLMs.
- Led the creation of an AI-powered scraping and data validation platform using agentic workflows, active monitoring, and dynamic data mapping to ensure high-quality, real-time vehicle data ingestion.
- Fine-tuned open-source LLMs using LoRA, QLoRA, and mixed-precision techniques to optimize domain adaptation and inference performance for automotive and insurance use cases.
- Implemented retrieval-augmented generation (RAG) systems using custom embedding pipelines and Hugging Face Transformers, powering knowledge-grounded assistants and decision support tools.
- Managed full MLOps pipelines using MLflow, Kubernetes, and Docker, with production observability enabled through Prometheus and integrated model monitoring dashboards.
- Developed and containerized scalable backend services with FastAPI, orchestrating training and inference workflows using PyTorch Lightning, TensorFlow, and scikit-learn to support rapid deployment and reproducibility.
- Delivered cloud-native AI solutions across GCP Vertex AI, Azure ML Studio, and Amazon SageMaker, ensuring secure, scalable deployments.

Software Engineer | Talal Tech | Feb 2022 – June 2023 | Remote | Medina, Saudi Arabia

- Engineered scalable backend systems and APIs using Python, Django, and Django Rest Framework, addressing complex business requirements with high efficiency and precision, ensuring performance and reliability in production environments.
- Excelled in Python backend development, leveraging modern libraries and tools for system integration, data processing, and API optimization, contributing to the seamless execution of multiple high-impact software projects.
- Played a key role in full-stack development, proficiently integrating React.js and Next.js to deliver rich, dynamic user interfaces and enhancing frontend development, ensuring responsive, user-centric web applications.
- Implemented CI/CD pipelines with Docker, Kubernetes, and modern DevOps practices to streamline the development, testing, and deployment processes, enabling fast and reliable releases of scalable applications.
- Designed and executed unit and integration testing frameworks in Python, ensuring robust backend functionality, system scalability, and high-quality software delivery by preventing regression and minimizing defects.

Mobile Applicaitons Developer | HoskaDev | Mar 2020 – Feb 2022 | Remote | Oran, Algeria

- Led the development of multiple cross-platform mobile applications using the Flutter framework, demonstrating expertise in hybrid app development for Android and iOS platforms.
- Proficiently translated UI designs from XD and Figma into intuitive user interfaces, collaborating seamlessly with backend and design teams to ensure cohesive integration and exceptional user experiences.
- Demonstrated versatility in delivering high-quality mobile applications both independently and as part of a collaborative team, including successful deployment to Google Play Store and Apple App Store.
- Built full-stack mobile applications, integrating backend services (using FastAPI, Django, etc.) and cloud solutions to deliver seamless, end-to-end user experiences.
- Implemented robust state management solutions (e.g., Provider, Riverpod, BLoC) to enhance app performance, scalability, and maintainability across multiple production projects.
- Integrated third-party APIs and SDKs (e.g., Firebase, Google Maps, Stripe) to enable real-time features such as authentication, payments, and location-based services.

LANGUAGE, SKILLS AND CERTIFICATES

Languages: Arabic (native), French (fluent), English (fluent)

Core Skills:

- **AI/ML:** PyTorch, TensorFlow, scikit-learn, Hugging Face Transformers, PyTorch Lightning
- **LLMs & GenAI:** Fine-tuning (LoRA/QLoRA), RAGs, multi-agent systems, open-source LLMs, prompt engineering
- **MLOps & Deployment:** MLflow, Airflow**,** Prometheus, Docker, Kubernetes, CI/CD pipelines
- **Software Engineering:** Python (advanced), C#, C++, Dart, Java Script, FastAPI, Django, React.js, Next.js, .NET, Flutter
- **Cloud Platforms:** GCP Vertex AI, Azure ML Studio, Amazon SageMaker

Certifications (Selected):

- IBM AI Developer Specialization
- DeepLearning.AI Machine Learning Specialization (with Stanford University)
- NVIDIA Fundamentals of Deep Learning
- Certified Data Scientist & Machine Learning Professional – London International Research Centre & Dubai Knowledge