# ISMAIL ASHRAF

AI/ML Engineer

Cairo, Egypt | +20 109 796 7127 | ismailashraf32@gmail.com

github.com/ismail-ashraf | linkedin.com/in/iismmail | kaggle.com/iismmail

## PROFESSIONAL SUMMARY

AI/ML Engineer focused on building intelligent, scalable, and production-oriented AI solutions. Experienced in developing Machine Learning and Deep Learning models, NLP and Computer Vision applications, and LLM-powered systems using Retrieval-Augmented Generation (RAG) and Agentic AI architectures. Skilled in Python, PyTorch, LangChain, LangGraph, and FastAPI, with a strong interest in model optimization, AI deployment, and designing reliable AI-driven backend systems. Driven by enthusiasm for exploring emerging AI technologies and turning them into practical, real-world applications.

## TECHNICAL SKILLS

**Programming:** Python, SQL

**Machine Learning & Deep Learning:** Machine Learning, Deep Learning, PyTorch, Scikit-learn, CNNs, LSTMs, NumPy, Pandas, Data Preprocessing, Model Training & Evaluation

**NLP & Computer Vision:** NLP, Computer Vision, OpenCV, Hugging Face Transformers, Sentence Transformers

**Generative AI & Agentic Systems:** Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Agentic AI, LangChain, LangGraph, Prompt Engineering, Embedding Models, Semantic Search, Tool Calling, LLM API Integration

**ML Engineering & Deployment:** MLflow, ONNX, Model Experiment Tracking, Model Deployment

**Backend Development:** FastAPI, RESTful APIs, Pydantic, Asynchronous Programming (async/await), Dependency Injection, Clean Architecture

**Databases & Vector Search:** PostgreSQL, SQLite, ChromaDB, FAISS, Vector Databases, Similarity Search

**Developer Tools:** Git, GitHub, Linux, Docker

**Software Engineering:** Object-Oriented Programming (OOP), Data Structures & Algorithms, Logging, Error Handling

## TRAINING & PROFESSIONAL EXPERIENCE

**Agentic AI Intern** Present

*Echo Inc.*

- Currently learning and implementing Retrieval-Augmented Generation (RAG) systems through a structured curriculum, covering RAG versus fine-tuning, prompt engineering, document chunking strategies, and a TF-IDF retrieval baseline.
- Studying the retrieval algorithms behind RAG and applying them in hands-on pipelines: dense embedding-based semantic search with FAISS compared against BM25 keyword search, cross-encoder reranking, hybrid retrieval, query transformation and routing, and modular RAG design.
- Upcoming curriculum stages: agent architecture and the perception-action loop, LangChain chains and memory, the ReAct pattern, tool calling, single- and multi-agent orchestration, and LLMOps topics such as async concurrency and deployment.
- Final stage planned: knowledge graphs (Neo4j and Cypher), GraphRAG, and agentic RAG architecture.

**Machine Learning Intern** July 2026 – August 2026

*FlyRank — Marketing Department*

- Applied AI and Machine Learning techniques to real business problems within the Marketing department, working with business and marketing data.
- Processed and prepared business and marketing datasets for analysis and modeling, with attention to data quality.
- Trained Machine Learning models to identify issues and anomalies in datasets, so that marketing analysis and optimization decisions rest on reliable data.
- Used data-driven insights to support marketing analysis and Generative Engine Optimization (GEO) efforts, applying AI techniques within business workflows.

**Machine Learning & Deep Learning Training** May 2026 – June 2026
*National Telecommunication Institute (NTI) — Intensive Technical Training Program*

- Machine Learning: studied supervised and unsupervised learning, including regression, classification, and clustering algorithms, and built practical projects covering model training, evaluation, and performance assessment with Scikit-learn, NumPy, and Pandas.
- Deep Learning for Computer Vision (DeepLearning.AI coursework): neural network fundamentals, Convolutional Neural Networks (CNNs), image classification, object detection and localization, image segmentation, and transfer learning for practical vision applications.
- Deep Learning for NLP and sequence models (DeepLearning.AI coursework): text preprocessing and sequence representation, Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and sequence modeling for language-related applications.

**Generative AI Training**
*Information Technology Institute (ITI) / NVIDIA*

- Studied core Generative AI concepts and prompt engineering techniques for working with generative models.
- Applied these concepts through practical AI application development.

**InnovEgypt Entrepreneurship Program**
*ITI & TIEC*

- Led a team project, conducted market research, and delivered a structured business proposition; awarded Grade A.

## SELECTED PROJECTS

**LexiCore — Enterprise NLP & RAG Backend** GitHub Repository
**Technologies:** *Python, FastAPI, PyTorch, Transformers, GLiNER, ONNX, PostgreSQL, ChromaDB, Celery*

- Engineered a modular NLP backend for automated legal document processing, integrating document ingestion, classification, Named Entity Recognition (NER), and intelligent retrieval through RESTful APIs.
- Implemented an asynchronous document processing pipeline using Celery, enabling background PDF ingestion, text extraction, chunking, and embedding generation.
- Developed a Retrieval-Augmented Generation (RAG) pipeline using BGE embeddings and ChromaDB, supporting semantic search and context retrieval for legal question answering.
- Integrated transformer-based NLP models and ONNX model workflows, with PostgreSQL for structured document metadata and persistent storage management.

**VisionCore — Real-Time Multi-Camera Computer Vision System** GitHub Repository
**Technologies:** *Python, PyTorch, YOLO11, ONNX, OpenCV, NumPy, FastAPI, React, SQLite, Threading*

- Engineered a real-time computer vision system for multi-camera video processing and pothole detection, using YOLO11 with an optimized ONNX inference workflow.
- Designed a concurrent streaming architecture with producer-consumer components and frame buffering to prevent processing backlogs and maintain real-time stream responsiveness.
- Implemented dynamic batching to combine frames from multiple streams into shared inference batches, improving inference throughput.
- Developed object tracking using Kalman Filters to maintain object identities across consecutive video frames.
- Built a monitoring dashboard with analytics and system logs, providing visibility into detection activity and system performance.

**SafeSphere — AI-Powered Content Moderation System** GitHub Repository
**Technologies:** *Python, FastAPI, PyTorch, TensorFlow, Transformers, NLP, Computer Vision, React, SHAP*

- Developed an integrated content moderation platform combining NLP and Computer Vision to identify toxic text and violent visual content.
- Implemented transformer-based text classification for toxicity detection and deep learning-based video analysis for violence detection.
- Built unified moderation workflows supporting text, image, and combined content analysis through RESTful APIs.
- Integrated SHAP-based explainability to provide insights into model predictions, alongside configurable moderation thresholds.
- Developed an interactive dashboard for content analysis, moderation results, and system monitoring.

## EDUCATION

**Bachelor of Science in Business Information Systems (B.Sc. BIS)** Expected Graduation: 2028
*New Cairo Academy – Faculty of Commerce* | Third-Year Student