MOHSEN MOSTAFA

Applied AI Research Scientist / AI&ML Lead For Data Infrastructure
Cairo, Egypt | +20 102 556 5499 | mohsen.mostafa.ai@outlook.com
https://mohsenmostafa1.github.io/Mohsen.github.io/

PROFESSIONAL SUMMARY

ML Lead & Research Engineer with 8+ years in AI/ML/CV/GenAI and 10+ years in IT infrastructure.
Proven track record of production deployments with measurable business impact.
Built end-to-end MLOps pipelines (Kafka, Spark, Hadoop, MLflow, Kubeflow) for real-time Robotics & IoT anomaly detection.

CORE SKILLS

- **Research & Theory:** Physics-Informed Neural Networks (PINNs), Geometric Deep Learning, Bayesian Uncertainty, Curvature-Aware Optimization.
- **Languages & Data:** Python, C++, Pandas, SQL, PySpark
- **ML / DL Frameworks:** PyTorch, TensorFlow, XGBoost, scikit-learn, Isolation Forest, LSTM Autoencoders
- **GenAI / LLM:** Hugging Face, Falcon, DeepSeek, Llama, LangChain, RAG, ColBERT
- **Computer Vision:** YOLOv8, OpenCV, Mask2Former
- **Streaming & Big Data:** Apache Kafka, Spark Structured Streaming, Hadoop HDFS, Parquet
- **MLOps & DevOps:** MLflow, Kubeflow, Airflow, Docker, Kubernetes, KServe, FastAPI, GitLab CI
- **Databases:** PostgreSQL, MongoDB, Redis, Pinecone
- **Cloud & Infra:** AWS SageMaker/EC2, GCP Vertex AI, Prometheus, Grafana, ONNX, TensorRT

PROFESSIONAL EXPERIENCE

AI Lead — *GraviLog (Remote)* Sep 2025 – Sep 2026
- End-to-end AI system architecture for multimodal CV+LLM (TensorRT/ONNX) → 40% latency reduction.
- Defined technical roadmap for 3 products; delivered C-level presentations.
- Led RAG retrieval (Pinecone + embedding); mentored 4 engineers in MLOps & system design.

Senior AI Developer — *OMC Technologies (Remote)* May 2023 – Aug 2025
- Built AI agents (Falcon, DeepSeek, Qwen, Llama) → automated 3 workflows, 40% manual effort reduction.
- Production RAG (LangChain + Pinecone + Falcon/DeepSeek) → 35% hallucination reduction.
- Multimodal healthcare AI → F1 ↑19%, latency 210ms→145ms (30% via streaming + FP16).
- Fraud detection AUC ↑17%, false positives ↓22% (FastAPI, p99 ≤ 95ms @ 500 RPS).
- Architected an InsurTech PoC for automated roof damage assessment, fine-tuning SAM 3 via Roboflow to segment drone imagery; designed a custom metric pipeline translating mIoU masks into real-world area (squares) and MAPE-driven cost estimations.
- Managed AI infrastructure with 24 GPUs (12 nodes) + 15 CPU nodes for real-time IoT anomaly detection; optimized distributed training and inference using TensorRT, Kubeflow, and KServe.
- Mentored 5 juniors → 2 promotions.

AI Developer & Researcher — *Nexgro (South Africa)* Apr 2020 – Apr 2023
- Grain quality CV pipeline → 98% accuracy across 6 types.
- Drone/satellite anomaly detection (NDVI + YOLOv8) → manual inspection efficiency ↑160%.
- LLM RAG assistant on technical manuals; document automation saved 15 person-hours/week.
- RL logistics prototype → travel distance ↓12%.

Data Scientist — *Nexgro* Mar 2018 – Mar 2020
- Forecast models → logistics costs ↓18%.
- Predictive maintenance prevented 3 breakdowns, saving ~$45K.
- Inventory optimization → spoilage ↓18%.

Systems Administrator — *Egyptian Co.* Jun 2013 – Mar 2018
- IT infrastructure management for 200+ users → 99.8% uptime; multi-site network & servers.
- Automation scripts reduced manual ticket handling by 25%.
- Self-directed pivot to data science (certifications 2016–2018); managed LAN/WAN, firewalls, backup for 150+ nodes.

IT Network Administrator / Specialist — *International Co. for Trade* Aug 2005 – Jun 2013
- Administered LAN/WAN, firewalls, and backup solutions for 150+ nodes (Administrator, 2009–2013).
- First-line network support, hardware procurement, and vendor coordination (Specialist, 2005–2009).

Independent Researcher | Physical & Geometric AI (Remote) | *2023 – Present*

- Founded and lead an independent research program bridging geometric deep learning theory and production ML constraints.
- Proved 4 theorems on gradient descent on data manifolds; developed EPANG-Gen (curvature-aware optimizer) and Bayesian normalization layers (R-LayerNorm).
- **Production Translation:** Currently extending research into SafePredict (Kaggle) — real-time shadow monitor for humanoid robotics using Triton/Kafka — and LePhyJEPA (RGB-D predictive architectures).

SELECTED PROJECTS (PERSONAL, OPEN SOURCE)

Multi-Tenant LLM Serving Platform on Kubernetes — GitHub

- Self-service platform serving DeepSeek and Llama via vLLM behind one unified FastAPI gateway, with JWT auth, per-tenant usage tracking and Redis rate limiting (100 req/min per tenant).
- Kubernetes deployment with HPA autoscaling, GitLab CI/CD (build → test → deploy), and Prometheus/Grafana observability.

Supply-Chain-Risk-Intelligence — GitHub

- Real-time IoT anomaly detection: Kafka → Spark Streaming → MongoDB, with raw events archived to HDFS as Parquet and a nightly Airflow ETL into a PostgreSQL warehouse.
- Kubeflow-orchestrated retraining tracked in MLflow; best model promoted to KServe for real-time inference and monitored in Grafana.

PUBLICATIONS & RESEARCH PREPRINTS

- Achieved 8 first-author preprints using only free-tier cloud GPUs (T4/P100), demonstrating extreme resource efficiency.
- **Geometric & Physics-Aware AI:** Published 3 papers on fluid flow (HPST), curvature-aware optimization (EPANG-Gen), and geometric racing dynamics (Preprints.org/arXiv).
- **Uncertainty-Aware Architectures:** Published 4 papers introducing Bayesian R-LayerNorm, Bayesian PASA, and robust classifiers with provable bounds.
- **Maximum Entropy Representations for Sequential Prediction and Latent Planning with Theoretical Guarantees:** Dynamic LeJEPA as a Principled World Model
- Geometric Deep Learning for Point Cloud Segmentation — under revision for WACV 2027.
- Physics-Informed Neural Networks for Fluid Flow — arXiv 2025.

Full list on ORCID: 0009-0004-4478-0317

Link for papers : https://mohsenmostafa1.github.io/Mohsen.github.io/publications.html

EDUCATION

B.Sc. Information Systems
High Inst. for Specialized Technological Studies, Egypt — 2000 – 2005

LANGUAGES

Arabic (Native) | English (Fluent – C1/C2)

CERTIFICATIONS

- Deep Learning Specialization – DeepLearning.AI
- TensorFlow Advanced Techniques – Google
- Google Cloud MLOps
- Total Data Quality – University of Michigan

LEADERSHIP & MENTORING

- Mentored 5+ engineers (2 promotions); led cross-functional AI initiatives.
- Conducted internal LLM/MLOps workshops.
- Designed full MLOps showcase: Kafka → Spark Streaming → HDFS → MongoDB → PostgreSQL → MLflow → Kubeflow → FastAPI for real-time predictive maintenance.