MOHSEN MOSTAFA

Independent Researcher — Physical & Geometric AI
Cairo, Egypt | +20 102 556 5499 | mohsen.mostafa.ai@outlook.com
GitHub | ORCID 0009-0004-4478-0317 | Kaggle | LinkedIn | Personal site

RESEARCH PROFILE

Independent researcher in Physical & Geometric AI (2023–present) and AI team lead with 8+ years of applied AI, computer vision and GenAI experience in industry. My research addresses world models and control for embodied systems along three lines:

- **Principled latent world models** — Dynamic LeJEPA, Decoupled Entropic-Geometric Planning (DEGP) and The Vacuity Barrier, including pre-registered studies whose negative results are reported openly.
- **Control-theoretic foundations of learned models** — "How Deep Is the Control?" quantifies Lie-bracket depth–width trade-offs and shows that a fitted control-affine model generically reports the counting bound of its own function class rather than the structure of the true system.
- **Uncertainty-aware and physics-informed architectures** with provable robustness bounds.

My recent work is released with code and was run on free-tier cloud GPUs. I am seeking a doctoral position to continue this program under formal supervision.

EDUCATION

B.Sc., Management Information Systems — High Institute for Specialized Technological Studies, Cairo *2000–2005*
Research training: Research Methodology (University of London); Total Data Quality (University of Michigan). Professional certificates are listed below.

RESEARCH OUTPUT (PREPRINTS)

All entries are single-author preprints and have not yet been peer reviewed. An up-to-date list is on ORCID.

World models, planning and control

1. How Deep Is the Control? Lie-Bracket Depth–Width Trade-offs, the Generic Bias of Learned Models, and a Negative Result on Structural Priors. 2026. Zenodo; code: GitHub. (paper underreview in Journal of Dynamical and Control Systems Springer)
2. The Vacuity Barrier: Conditional Lower Bounds and Measured Premises for Latent World Model Planning. 2026. Preprints.org 202610.1780.
3. Decoupled Entropic-Geometric Planning (DEGP): Verified Conditioning and a Measured Contact-Prediction Boundary in Latent World Models. 2026. Research Square. (paper underreview in Journal of Intelligent & Robotic Systems Springer)
4. Dynamic LeJEPA: Maximum-Entropy Representations for Sequential Prediction and Latent Planning with Theoretical Guarantees. 2026. Research Square; interactive demo. (paper underreview in Journal of Machine Learning Springer)
5. Gaussian-Constrained LeJEPA Representations for Unsupervised Scene Discovery. 2026. arXiv:2602.07016.

Uncertainty-aware, physics-informed and optimization methods

6. Geometric Foundations of Racing Dynamics: How Gradient Descent Adapts Network Capacity on Data Manifolds. 2026. Preprints.org 202604.0964.
7. Bayesian PASA: Provably Stable Adaptive Activation with Uncertainty Quantification. 2026. Research Square.
8. HPST: A Hybrid Physics-Spectral-Threshold Framework for Fluid Flow Analysis. 2026. Preprints.org 202603.1281.
9. R-LayerNorm: Robust Layer Normalization with Adaptive Noise Suppression. 2026. Preprints.org 202603.0060.
10. Bayesian R-LayerNorm: Uncertainty-Aware Adaptive Normalization with Provable Robustness Bounds. 2026. Preprints.org 202603.0450.
11. Uncertainty-Aware Classifier with Physics-Based Rejection (UA-PBR). 2026. Preprints.org 202603.0748.
12. EPANG-Gen: A Curvature-Aware Optimizer with Uncertainty Quantification for Scientific ML. 2026. Preprints.org 202603.0692.

PROFESSIONAL EXPERIENCE

AI Scientist & Lead — GraviLog (Remote, US/UAE) *Sep 2025 – Present*

- Designed an end-to-end multimodal (CV + LLM) system architecture with TensorRT/ONNX optimization, cutting latency by 40%.
- Defined the technical roadmap for 3 products and presented research strategy to C-level stakeholders.
- Led the RAG retrieval pipeline (Pinecone + embeddings); mentored 4 engineers in MLOps and system design.

Senior AI Developer (Research Focus) — OMC Technologies (Remote, Limerick, Ireland) *May 2023 – Aug 2025*

- Built LLM agents (Falcon, DeepSeek, Qwen, Llama) automating 3 workflows (−40% manual effort); deployed a production RAG system (LangChain + Pinecone) that reduced hallucinations by 35%.

- Shipped a multimodal healthcare AI system (F1 +19%; latency 210 ms → 145 ms) and a fraud-detection service (AUC +17%, false positives −22%; p99 ≤ 95 ms at 500 RPS).
- Architected an InsurTech proof of concept for roof-damage assessment from drone imagery (fine-tuned SAM 3; mIoU-to-area and MAPE-driven cost estimation).
- Managed AI infrastructure (24 GPUs across 12 nodes, 15 CPU nodes) for real-time IoT anomaly detection with TensorRT, Kubeflow and KServe; mentored 5 junior engineers, leading to 2 promotions.

AI Developer & Researcher — Nexgro (South Africa) *Apr 2020 – Apr 2023*

- Built a grain-quality computer-vision pipeline (98% accuracy over 6 grain types) and drone/satellite anomaly detection (NDVI + YOLOv8; +160% inspection efficiency).
- Optimized 3D point-cloud segmentation for LiDAR/DSM data to sub-300 ms inference on edge devices; deployed an LLM RAG assistant over technical manuals (15 person-hours saved per week).

Data Scientist — Nexgro (South Africa) *Mar 2018 – Mar 2020*

- Forecasting models cut logistics costs by 18%; predictive maintenance prevented 3 breakdowns (~$45K saved); inventory optimization reduced spoilage by 18%.

Systems Administrator — Egyptian Co. *Jun 2013 – Mar 2018*

- Managed IT infrastructure for 200+ users at 99.8% uptime; automation cut manual ticket handling by 25%. Moved into data science through self-directed study (certifications 2016–2018).

IT Network Administrator / Specialist — International Co. for Trade *Aug 2005 – Jun 2013*

- Administered LAN/WAN, firewalls and backup for 150+ nodes; vendor coordination and first-line support.

SELECTED ENGINEERING PROJECTS

SafePredict — Real-Time Shadow Safety Monitor for Humanoid Robot Policies *Jul 2026*

- Built a secondary safety monitor over a LeRobot v0.6.0 vision-language-action policy: zero-disk streaming ingestion, temporal windows from delta timestamps, and programmatic safety labels derived from finite-difference joint acceleration.
- Trained a 96×96-input CNN with 20-pass MC-Dropout uncertainty, exported it to ONNX, and ran a 30 Hz streaming loop in which synthetic collision shocks cross the E-STOP threshold within one inference cycle (~15 ms time-to-alert on a T4); wrote the Triton repository and config, then validated the same I/O schema through the ONNX Runtime backend. Kaggle notebook

Operationalizing Fetal Ultrasound — ONNX Segmentation and Structured Risk Reports *Nov 2026*

- Four-class abdominal segmentation (artery, liver, stomach, vein) over ~1,600 ultrasound images: SAM-encoder backbone, ultrasound-specific augmentation, pseudo-labelling, 5-fold ensemble and EMA, exported to a single ONNX graph.
- Served the ensemble in a Gradio app that returns annotated structures and a guideline-based report in under 10 s; the risk-stratification layer is rule-based and marked for clinical validation. Hugging Face demo

TECHNICAL SKILLS

Deep learning & probabilistic ML: PyTorch, TensorFlow, Hugging Face, JEPA, CLIP, YOLO, Mask2Former; Bayesian deep learning, uncertainty quantification, natural-gradient and curvature-aware optimization

Computer vision & robotics software: OpenCV, 3D point-cloud segmentation (LiDAR/DSM), multi-view geometry, ROS2

GenAI / LLM: RAG, ColBERT, LangChain, DeepSeek, Qwen, Llama, streaming inference

MLOps & serving: MLflow, Kubeflow, Airflow, Docker, Kubernetes, KServe, FastAPI, TensorRT, ONNX, quantization, GitLab CI, GCP Vertex AI, AWS SageMaker

Data engineering: Kafka, Spark Structured Streaming, Hadoop/HDFS, Pandas, Dask, Polars, PySpark, PostgreSQL, MongoDB, Redis, Pinecone

CERTIFICATIONS AND LANGUAGES

Research Methodology (University of London); Total Data Quality (University of Michigan); Deep Learning Specialization (DeepLearning.AI); TensorFlow Developer and Advanced TensorFlow (Google); Google Cloud MLOps; IBM Data Science Professional Certificate; Cisco CCNA 200-301 and ENCOR 350-401.

Languages: Arabic (native); English (fluent, C1/C2).