



Siria Sadeddin

Machine Learning Engineer

Nürnberg, Germany

+49 15510 359475 | siriasadeddin@gmail.com

Available: Immediately / Work authorization: EU / Germany

Profile

Machine Learning Engineer with 7+ years of experience building, deploying, and operating AI systems in production environments. Experienced in LLM deployment, inference pipelines, model serving, GPU-based workloads, and integration of state-of-the-art AI models into enterprise platforms.

Strong background in bridging machine learning and infrastructure, including model conversion pipelines, production inference systems, API integration, containerized deployments, and scalable AI services. Experience working in highly technical and research-intensive environments including Wolfram Research and CERN.

Particularly interested in AI platform engineering, LLM infrastructure, RAG systems, and large-scale deployment of foundation models.

Core Skills

AI/ML: Deep Learning, Transformers, LLMs, Vision Models, Scientific ML

Systems: Model deployment, inference pipelines, distributed AI systems, Linux environments

Frameworks: PyTorch, TensorFlow, Hugging Face, ONNX, Wolfram Language

Computer Vision: YOLO, Vision Transformers, multimodal models

MLOps: Docker, REST APIs, production pipelines, ETL systems

Programming: Python, SQL, Bash

Professional Experience

Machine Learning Engineer

Wolfram Research Inc.

02/2021 – Present

Remote

- Designed and maintained production inference pipelines for large AI models integrated into the Wolfram ecosystem
- Built model conversion workflows (PyTorch, TensorFlow, ONNX) enabling deployment of state-of-the-art AI models across heterogeneous execution environments
- Re-implemented transformer architectures when direct export paths were unavailable, ensuring compatibility with production infrastructure
- Worked extensively with ONNX Runtime and optimized inference workflows for large neural network models
- Developed and deployed computer vision systems exported to ONNX and CoreML for production applications
- Reconstructed and integrated ESMFold protein-folding models into Wolfram's scientific computing stack

Machine Learning Engineer – LLM Systems

CERN OpenLab – BEAMS Department

07/2024 – 09/2024

Geneva, Switzerland

- Developed LLM-based assistant systems for scientific documentation and code understanding within CERN's BEAMS infrastructure
- Benchmarked and analyzed LLM models for code summarization and generation
- Fine-tuned and evaluated Llama 3 models under hardware GPU constraints
- Designed evaluation pipelines for LLM performance in scientific software environments (C++, Python, Bash)
- Created structured Question-Answer datasets from technical documentation using LLaMA-Index
- Collaborated closely with physicists and engineers to translate real operational needs into AI system requirements

Machine Learning Engineer

Global Hitss Colombia (CLARO Datacenter)

10/2018 – 02/2021

Bogotá, Colombia

- Developed and deployed a production-grade ticket classification system achieving 96% accuracy for automated IT service routing
- Designed ML-driven automation pipelines integrated into Power BI dashboards used for real-time operational decision-making
- Built ETL pipelines for large-scale structured and unstructured data processing
- Designed and evaluated anomaly detection systems (autoencoders) for predictive infrastructure monitoring (research-to-prototype phase)
- Worked closely with infrastructure teams to identify opportunities for automation and process optimization

Education

M.Sc. Artificial Intelligence

Johannes Kepler University Linz, Austria

2025 – Present

B.Sc. Physics

Universidad Simón Bolívar, Venezuela

2019

Languages

Spanish (Native) | English (C1 – professional proficiency) | German (B1 – actively learning)

Publications

- Estimating Water Levels through Smartphone-Imaged Gauges: Comparative Analysis of ANN, DL, and CNN Models (Springer, 2024)
- Facial Emotion Recognition and Handling Data Imbalance in Machine Learning (ICTACT Journal, 2023)

Certifications

- Generative AI with Large Language Models (DeepLearning.AI & AWS)
- Convolutional Neural Networks (DeepLearning.AI)
- Machine Learning with Python (IBM)
- Docker for Machine Learning Systems (Udemy)

Nürnberg, June 14, 2026