Braulio Otavalo

AI/ML Engineer

De Las Uvas, N52-378, Quito, 170124, Ecuador
braulio.otavalo@gmail.com • LinkedIn • Github
+593 984 349 747

SUMMARY

AI/ML Engineer with 7+ years of experience delivering end-to-end AI systems, from business problem framing to production deployment and monitoring. Expertise in machine learning, deep learning, LLMs, and agentic AI systems. Skilled in designing scalable, secure, cloud-native architectures, including data pipelines, model development, evaluation, and MLOps for production-grade AI solutions.

PROFESSIONAL EXPERIENCE

AI/ML Engineer Jun 2025 - Present
BairesDev, USA (Remote)

- Designed and implemented of a production-grade multi-agent AI copilot for solar system design, leveraging TypeScript, LangGraph, and Model Context Protocol (MCP) to implement advanced agentic patterns (planning, tool use, reflection, and human-in-the-loop workflows).
- Architected end-to-end agentic AI systems, from business problem framing to production, enabling early detection of design errors and improving solar project feasibility (financial and site constraints), reducing design iteration cycles and operational risk.
- Built scalable Retrieval-Augmented Generation (RAG) pipelines with vector databases to ground LLMs (OpenAI, Vertex AI, Claude) in proprietary engineering data, applying hybrid retrieval, context engineering, and agentic RAG strategies for high-precision outputs.
- Designed and implemented a comprehensive evaluation suite for Evaluation-Driven Development across RAG and agentic AI systems, covering dataset creation, benchmarking, and continuous validation. Developed trajectory-based (glass-box), step-level (white-box), and final-output (black-box) evaluations; integrated RAG-specific metrics (faithfulness, context precision/recall) and agentic metrics (task completion rate, cost per task, token efficiency). Leveraged MLflow, DeepEval, PyRIT, LLM-as-a-judge, and Langfuse with OpenTelemetry to enforce guardrails, safety, observability, and production reliability.
- Developed agent observability and reliability pipelines using OpenTelemetry and Langfuse, enabling hierarchical tracing of multi-step agent workflows, real-time monitoring of latency, error rates, and cost, and systematic debugging of reasoning failures in production.
- Optimized agent performance and cost-efficiency by analyzing task completion per million tokens, balancing accuracy vs. cost trade-offs, and reducing unnecessary tool calls and LLM invocations across multi-agent workflows.
- Designed event-driven, cloud-native architectures using NATS and GCP (Vertex AI, Cloud Functions, Pub/Sub), supporting scalable, resilient, and observable AI systems in production environments.

ML Engineer Aug 2024 - Jul 2025
Awana, USA, Remote

- Led the development of AI strategies for Awana's recruiting process, automating applicant-job description matching, and improving success rates by 20%.
- Designed and executed the AI Skills Accelerator, a 12-week training program for junior developers, enabling them to deliver clean ML projects 30% faster.

- Designed and implemented Generative AI-based solutions for ChildNEXUS, including semantic recommenders and chatbots, leveraging vector embeddings, vector DB, graph DB, LLMs, and RAG pipelines, which improved user engagement by 15% through automated, real-time responses.
- Built and maintained scalable, cost-efficient ML workflows on GCP, AWS, and Azure, reducing operational costs by 25% through automated MLOps pipelines using Docker, Kubernetes, Terraform, and GitHub Actions.

Data Scientist Consultant - STC Feb 2024 - Jun 2024
World Bank Group, USA, Remote

- Developed a web application for infrastructure project prioritization based on the World Bank's Infrastructure Prioritization Framework, automating data analysis, geospatial processing, and non-linear optimization.
- Conducted buffer and zonal statistic analysis on geospatial data, replicating complex calculations to ensure accurate project prioritization.
- Created interactive data visualization dashboards using Plotly, Leaflet, Leafmap, and Chart.js, providing stakeholders with real-time insights into project prioritization outcomes.

Head of International Connectivity Nov 2021 - Feb 2023
National Telecommunications Corporation, Ecuador, Remote

- Forecasting future international traffic patterns to Network Access Points (NAPs), enabling accurate estimation of new capacity requirements, and supporting strategic procurement decisions based on applied ML for time series data forecasting.
- Built data visualization dashboards using internet traffic data, providing actionable insights for decision makers to optimize network performance and reduce costs.
- Oversaw occupancy reports, 95th percentile calculations, and international network connectivity availability using the Cacti server dataset, ensuring optimal network performance and resource allocation.

Undersecretary of Non-contributory Assurance May 2021 - Sep 2021
Ministry of Economic and Social Inclusion, Quito - Ecuador

- Supervised designing and implementing a proxy means test leveraging machine learning to classify more than 9 million individuals from national social registry databases into extreme poverty, poverty, and non-poverty segments.
- Oversaw data preparation, feature engineering, model training, testing, and evaluation, ensuring methodological rigor and alignment with program objectives.
- Guided the team in assessing the model's sensitivity, fairness, and explainability to support the targeting of transparent and equitable beneficiaries for social programs.
- Ensured delivery of robust, interpretable models to inform policymakers and enhance resource allocation.

AI Consultant Sep 2018 - Present
OA7, Ecuador, Remote

- Designed architecture solutions and implemented projects using AWS, Azure, Google Cloud, Docker, and Kubernetes.
- Developed computer vision, machine learning, and deep learning applications with Python and Web frameworks.
- Led end-to-end machine learning projects, including data engineering, model training, and maintenance using MLOps tools for experiment tracking, model logging, deployment, and monitoring.

EDUCATION

UNIVERSITY OF SOUTHAMPTON, UNITED KINGDOM Nov 2020
MSc in Artificial Intelligence, Merit (Upper-second class honour)

NATIONAL POLYTECHNIC SCHOOL, ECUADOR Jun 2014
BSc in Electronic and Control Engineering, Cum Laude (9/10)

HONORS AND AWARDS

SANTANDER SCHOLARSHIP - MIT PROFESSIONAL EDUCATION 2020
Leading digital transformation program.
CHEVENING SCHOLARSHIP (FCDO) 2019-2020
Full scholarship for international students to study for a master's degree in the United Kingdom.

CERTIFICATES & COURSES

Claude Code in Action - Anthropic 2025

AI-Assisted Development - KodeKloud 2025

AWS Certified Cloud Practitioner - AWS Training and Certification 2025

AI For Business Specialization, University of Pennsylvania - Coursera 2025

MLOps Zoomcamp, DataTalksClub 2024

Climate Change AI Summer School 2024, Climate Change AI 2024

Oxford Machine Learning Summer School - OxML 2023 , AI for Global Goals 2023

Nonlinear Dynamics: Mathematical and Computational Approaches, Santa Fe Institute 2021

Leading Digital Transformation, MIT Professional Education 2020

Big Data for Businesses, Fundación Telefónica Movistar 2019

SKILLS

Programming & Systems: Python, TypeScript, C++, SQL, HTML/CSS
AI/ML & Deep Learning: PyTorch, TensorFlow, JAX, Scikit-learn; model development, fine-tuning (LLMs, CV), evaluation, and optimization
LLM & Generative AI: RAG pipelines, prompt engineering, fine-tuning (LoRA/PEFT), inference optimization (vLLM, TensorRT-LLM, llama.cpp, Ollama); vector DBs (Pinecone, Qdrant, Elasticsearch); LangChain, RAGAS
Agentic AI Systems: Multi-agent architectures, planning/reasoning patterns, tool use, memory systems (Mem0, Zep); LangGraph, Google ADK; evaluation (DeepEval), guardrails and safety (Guardrails AI)
MLOps & Cloud: AWS, GCP, Azure; Docker, Kubernetes, Terraform; MLflow, Kubeflow, CI/CD (GitHub Actions); scalable, secure, cloud-native deployments
Data Engineering: Pandas, PySpark, Airflow, DVC, Great Expectations; data pipelines, validation, and governance
Observability & Evaluation: OpenTelemetry, Langfuse, evaluation-driven development, LLM-as-a-judge, benchmarking (RAG + agentic systems), cost and performance optimization
Software & AI Lifecycle: Business problem framing, system design, experimentation, deployment, monitoring, and continuous improvement of production AI systems
Leadership & Business: Stakeholder management, technical leadership, product thinking, cross-functional communication, delivering AI solutions aligned with business impact

---

LANGUAGES

- Spanish (native)
- English (fluent)