Juan Lara

AI & ML Engineer | Generative AI

larajuand@outlook.com | Web Page | Bogotá, Colombia
+57 315 512 8464 | julara | JuanLara18

Professional Summary

ML Engineer with 3+ years architecting and deploying production machine learning systems across healthcare and enterprise domains. Combines rigorous Computer Science and Mathematics foundation with hands-on expertise designing end-to-end ML infrastructure—from model fine-tuning and optimization to scalable deployment pipelines. Builds production-ready solutions leveraging LLMs, RAG architectures, vector databases, and MLOps frameworks on cloud platforms. Led development of systems that streamlined large-scale data processing and automated analytical workflows, earning recognition for operational excellence across international markets. Skilled at bridging research innovation with production requirements to deliver systems that drive measurable business impact.

Professional Experience

LLM/ML Specialist — July 2025 - Present
GenomAI — *Danville, USA (Remote)*

- **Architect and deploy AI-powered clinical decision support systems** with RAG pipelines, vector databases, and fine-tuned LLMs, serving real-time recommendations through FastAPI microservices with comprehensive health checks and logging infrastructure.
- **Build and optimize ML deployment pipelines** with Docker containerization and CI/CD automation, implementing scalable cloud infrastructure that supports production healthcare applications with sub-second response requirements.
- **Develop HIPAA-compliant generative AI solutions** using advanced prompt engineering and retrieval-augmented generation to process multimodal medical data while ensuring regulatory compliance and patient privacy protection.

Supervisor: Noemi Pérez

Research Associate | ML Specialist — Sep 2022 - July 2025
Harvard University — *Boston, USA (Remote)*

- **Built end-to-end ML pipelines** processing large-scale organizational datasets (clustering, XGBoost, NLP), revealing insights on firm learning strategies that informed reskilling recommendations.
- **Designed scalable data processing frameworks** for modeling complex organizational systems, implementing validation pipelines and automated testing that ensured reproducibility across large-scale simulations.
- **Automated research workflows** accelerating data-driven insight generation, translating quantitative analysis into actionable business recommendations for technology adoption strategies.

Supervisor: Jorge Tamayo

Data Scientist Feb 2024 - Jan 2025
Ipsos *Bogotá, D.C., Colombia (Hybrid)*

- **Engineered production-ready** geospatial analysis and segmentation applications using ML models and automated data pipelines on Google Cloud Platform, enhancing operational efficiency across multiple data sources.
- **Developed a modular Python NLP library** deployed via PyPI and integrated into production tools, automating analytical workflows that reduced creative team processing time from hours to minutes and earning LATAM-wide operational excellence recognition.
- **Streamlined analytical workflows** through automated Python pipelines, significantly reducing manual processing while enabling dynamic real-time reporting and cross-functional analytics.

Supervisor: Sandra Pastrán

Software Engineer (Freelance) Jan 2023 - Present
Independent Consultant *Bogotá, Colombia (Remote)*

- Design and deliver full-stack web applications with custom database architectures and REST APIs, implementing version control workflows and deployment automation for clients across education and business sectors.
- Manage complete project lifecycle from requirements gathering to deployment, translating client needs into technical solutions while providing ongoing technical mentoring in programming, mathematics, and data science.

Supervisor: Juan Lara (Me)

Education

B.S. in Computer Science Feb 2019 - Nov 2023
Universidad Nacional de Colombia, Bogotá D.C.
Emphasis on Machine Learning GPA: 4.7/5.0
Director: Omar Duque Gomez Detailed List of Exams

B.S. in Mathematics Feb 2018 - Jun 2022
Universidad Nacional de Colombia, Bogotá D.C.
Emphasis on Applied Mathematics GPA: 4.7/5.0
Director: Omar Duque Gomez Detailed List of Exams

Technical Baccalaureate in Business Administration Feb 2015 - Nov 2017
Centro Educativo los Andes, Bogotá D.C.
Emphasis on Business Administration GPA: 4.5
Director: Cristian Santiesteban Detailed List of Exams

Technician in Maintenance of Computer Equipment Nov 2015 - Dec 2016
Servicio Nacional de Aprendizaje - SENA, Bogotá D.C.
Emphasis on Corrective Software GPA: 4.6/5.0
Director: Carlos Wilches Detailed List of Exams

Technical Skills

Core Competencies

Key technical areas combining hands-on experience with production systems

Machine Learning Engineering

Build and deploy LLM fine-tuning solutions (LoRA, QLoRA, PEFT), RAG architectures with vector databases (Chroma, FAISS), and embedding optimization. Work with PyTorch, TensorFlow, Hugging Face, and LangChain to develop domain-specific AI applications using advanced prompt engineering.

MLOps & Cloud Infrastructure

Deploy ML systems on GCP and AWS with Docker containerization and CI/CD pipelines. Implement model monitoring, inference optimization, and scalable cloud architectures. Familiar with Kubernetes orchestration and cloud-native deployment patterns.

Data Engineering & Scale

Build and optimize large-scale data processing systems with Spark and parallel computing. Design efficient ETL pipelines, optimize databases (SQL/NoSQL), and profile performance to handle high-volume workloads efficiently.

Research & Problem Solving

Combine analytical rigor with practical engineering to solve complex problems. Design experiments, apply mathematical modeling, and validate solutions systematically. Translate research insights into production systems through data-driven decision making and iterative testing.

Technical Proficiency

Core technologies and tools with proven production experience

AI & ML Stack

ML Frameworks	*PyTorch, TensorFlow, Hugging Face, Scikit-learn*
LLMs & RAG Systems	*LLaMA, GPT, Fine-tuning, Vector DBs*

Cloud & Deployment

Cloud Platforms	*GCP (Vertex AI), AWS (SageMaker)*
MLOps & Containers	*Docker, Kubernetes, CI/CD, MLflow*

Development & Data

Languages & APIs	*Python, SQL, FastAPI, REST APIs*
Data Processing	*PySpark, Pandas, Distributed Systems*

Additional Training

AI Engineer for Developers Associate — July 2025
DataCamp — *Certificate*

Credential ID: AIEDA0019827293059

Artificial Intelligence Expert Certificate (CAIEC) — November 2024
Certiprof — *Certificate*

Credential ID: TLZVDQTVTGG-XWHHHQPTQ-RDJFLDLRK

Artificial Intelligence Bootcamp May-October 2024
Talento Tech Cymetria *159 hours*

Credential ID: 2518458921

DevOps Certification October 2024
Platzi *Comprehensive Program*

Credential ID: cc4cfe8a-d78a-4883-8a75-ca90931151f6

Algorithmic Toolbox 2023
Coursera *Advanced Course*

Credential ID: 8GR62BCT499V

Distinctions & Awards

Total Ops Star Employee - LATAM April 2024
Ipsos

Recognized for developing TextInsight, demonstrating exceptional initiative, technical expertise, and commitment to operational excellence across Latin America operations.

Best Averages Scholarship 2018-2023
Universidad Nacional de Colombia

Awarded for 10 consecutive semesters to the top 15 students with highest academic performance in the program, maintaining excellence throughout entire academic career.

Areas of Interest

ML System Reliability & Monitoring
Improving production ML systems through robust monitoring, testing strategies, and continuous model validation to ensure consistent performance.

LLM Optimization & Fine-tuning
Advancing techniques for efficient model adaptation and deployment, including PEFT methods and inference optimization for resource-constrained environments.

AI Compliance & Safety
Implementing frameworks for responsible AI deployment in regulated industries, focusing on privacy protection and audit trail systems.

Languages

Spanish	Native
English	Advanced