

# Juan Lara

AI & ML Engineer | Generative AI

 larajuand@outlook.com  
 +57 315 512 8464

 Web Page  
 julara

 Bogotá, Colombia  
 JuanLara18

## Professional Summary

ML Engineer with 3+ years architecting and deploying production machine learning systems across healthcare, retail, and research domains. Combines a 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

**AI Engineer**  
*Falabella*

November 2025 - Present  
Bogotá D.C, Colombia (Hybrid)

- **Develop computer vision and deep learning solutions** for real-time operational monitoring, managing the full ML lifecycle from data collection and model training to production deployment preparation on GCP infrastructure.
- **Build data processing pipelines** integrating OCR engines, anomaly detection models, and cloud data warehouses, enabling automated extraction and quality validation across large-scale operational datasets.
- **Design ML system integrations** connecting modern AI capabilities with legacy enterprise platforms, implementing scalable data flows that bridge existing infrastructure with production-ready machine learning services.

Supervisor: José Rodriguez

**LLM/ML Specialist**  
*GenomAI*

June 2025 - December 2025  
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**  
*Harvard University*

**Sep 2022 - July 2025**  
*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**  
*Ipsos*

**Feb 2024 - Jan 2025**  
*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

## Education

**M.S. in Artificial Intelligence (In Progress)**

**Jan 2026 - Nov 2027**

*Universidad de los Andes, Bogotá D.C.*

GPA: ...

Emphasis on Reinforcement Learning

[Detailed List of Exams](#)

Director: -

**B.S. in Computer Science**

**Feb 2019 - Nov 2023**

*Universidad Nacional de Colombia, Bogotá D.C.*

GPA: 4.7/5.0

Emphasis on Machine Learning

[Detailed List of Exams](#)

Director: [Omar Duque Gomez](#)

**B.S. in Mathematics**

**Feb 2018 - Jun 2022**

*Universidad Nacional de Colombia, Bogotá D.C.*

GPA: 4.7/5.0

Emphasis on Applied Mathematics

[Detailed List of Exams](#)

Director: [Omar Duque Gomez](#)

**Technical Baccalaureate in Business Administration**

**Feb 2015 - Nov 2017**

*Centro Educativo los Andes, Bogotá D.C.*

GPA: 4.5

Emphasis on Business Administration

[Detailed List of Exams](#)

Director: [Cristian Santiesteban](#)

## 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.*

#### 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.*



#### 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.*



#### 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

*DataCamp*

*July 2025*

*Certificate*

**Credential ID:** AIEDA0019827293059

### Artificial Intelligence Expert Certificate (CAIEC)

*Certiprof*

*November 2024*

*Certificate*

**Credential ID:** TLZVDQTVTGG-XWHHHQPTQ-RDJFLDLRK

**Artificial Intelligence Bootcamp**  
*Talento Tech Cymetria*  
Credential ID: [2518458921](#)

May-October 2024  
159 hours

**DevOps Certification**  
*Platzi*  
Credential ID: [cc4cf8a-d78a-4883-8a75-ca90931151f6](#)

October 2024  
Comprehensive Program

**Algorithmic Toolbox**  
*Coursera*  
Credential ID: [8GR62BCT499V](#)

2023  
Advanced Course

## Distinctions & Awards

**Total Ops Star Employee - LATAM**  
*Ipsos*

April 2024

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

**Best Averages Scholarship**  
*Universidad Nacional de Colombia*

2018-2023

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

### Q ML System Reliability & Monitoring

*Improving production ML systems through robust monitoring, testing strategies, and continuous model validation to ensure consistent performance.*

### Q LLM Optimization & Fine-tuning

*Advancing techniques for efficient model adaptation and deployment, including PEFT methods and inference optimization for resource-constrained environments.*

### Q 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