Juan Lara

LLM & ML Engineer | Generative AI & RAG Systems

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

Professional Summary

LLM/ML Engineer with 3+ years of experience developing production-ready generative AI solutions and scalable machine learning systems. Combines strong theoretical foundations in Computer Science and Mathematics with hands-on expertise in large language models, RAG architectures, and end-to-end NLP pipelines. Experienced in the complete AI engineering stack including model fine-tuning, vector databases, prompt engineering, and cloud deployment. Successfully led development of multilingual NLP systems processing large-scale datasets, with recognition for operational excellence across international markets. Skilled at bridging research and production environments to deliver scalable solutions that transform complex data into actionable business insights.

Professional Experience

LLM/ML Specialist

July 2025 - Present

GenomAI

Danville, USA (Remote)

- **Architect AI-powered clinical decision support systems** integrating RAG pipelines with vector databases and LLM fine-tuning techniques to deliver real-time, evidence-based treatment recommendations for healthcare professionals.
- **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.
- **Optimize inference pipelines and model deployment** on cloud infrastructure, implementing efficient scaling strategies and performance monitoring to support production-level healthcare applications with sub-second response requirements.

Supervisor: Noemi Pérez

Research Associate | ML Specialist

Sep 2022 - July 2025

Harvard University

Boston, USA (Remote)

- **Built end-to-end ML pipelines** integrating clustering algorithms, XGBoost models, and NLP techniques to analyze large-scale organizational datasets, revealing key insights on firm learning strategies and technology adoption patterns.
- **Designed mathematical frameworks** for modeling organizational hierarchies and technology shocks, providing formal validation through simulation-based approaches that underpin upcoming academic publications.
- **Automated research workflows** that accelerated data-driven insight generation for firm behavior analysis, translating quantitative research into actionable recommendations for reskilling strategies.

Supervisor: Jorge Tamayo

Data Scientist

Ipsos

Feb 2024 - Jan 2025

Bogotá, D.C., Colombia (Hybrid)

- **Engineered production-ready applications** for geospatial analysis and segmentation using ML models and robust data-processing pipelines on Google Cloud Platform, enhancing operational efficiency across multiple data sources.
- **Developed TextInsight**, a Python-based NLP library leveraging generative AI and NetworkX graph analysis, reducing text processing time from hours to under one hour and earning Total Ops Star Employee recognition across LATAM.
- **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

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 Skills

Core Competencies

Areas of technical expertise where I have the most depth and experience

LLM & RAG Systems

Advanced expertise in designing and deploying production-ready RAG architectures with vector databases, embedding optimization, and multi-stage retrieval systems, specializing in LLM fine-tuning using PEFT methods (LoRA, QLoRA) and prompt engineering for domain-specific applications.

ML Engineering & MLOps

Expertise in building scalable ML applications using Python frameworks (LangChain, Streamlit, FastAPI) with efficient inference pipelines. Strong background in model deployment, CI/CD for ML systems, and production optimization of language model solutions.

AI System Architecture

Expertise in designing scalable AI system architectures using cloud infrastructure and microservices patterns. Skilled in distributed computing, container orchestration, and building fault-tolerant systems that handle high-volume AI workloads with optimal performance and reliability.

AI Safety & Evaluation

Specialized in AI model evaluation, bias detection, and safety monitoring for production systems. Experienced in compliance frameworks (HIPAA, GDPR), implementing audit trails, and developing comprehensive testing strategies to ensure responsible AI deployment.

Technical Proficiency

Languages, frameworks, and tools with proven experience and impact

LLM & RAG Systems

Large Language Models	*LLaMA, OpenAI GPT, Gemma, Fine-tuning*
RAG & Vector Databases	*Chroma, FAISS, Embedding Strategies*
PEFT & Optimization	*LoRA, QLoRA, Quantization, Inference*

ML Engineering & NLP

ML Frameworks	*PyTorch, Hugging Face, TensorFlow*
NLP Libraries	*Transformers, NLTK, spaCy, NetworkX*
Traditional ML	*Scikit-learn, XGBoost, Statistical Modeling*

Cloud & MLOps

Cloud Platforms	*GCP-Vertex, AWS-SageMaker, Cloud Storage*
MLOps & Deployment	*Docker, MLflow, Kubernetes*
CI/CD & Automation	*GitHub Actions, Git, Automated Testing*

Development & Applications

Programming Languages	*Python, SQL, Bash*
Web & API Development	*FastAPI, Streamlit, React*
Data Processing	*PySpark, Pandas, Distributed Systems*

AI Agents & Automation

Agent Frameworks	*LangChain, AI Agents, Multi-Agent Systems*
Workflow Automation	*n8n, Prompt Engineering, Evaluation*
Visualization	*Matplotlib, Plotly, Interactive Dashboards*

Additional Training

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

Credential ID: AIEDA0019827293059

Curso de LangChain — July 2025
Platzi — *Certificate*

Credential ID: dd0e8538-8e8f-4ed9-acae-5192ba8faf18

Curso de NLP con Python — July 2025
Platzi — *Certificate*

Credential ID: 520eb925-05d2-4298-ae08-187d5a2bae0a

Fundamentals of MCP May 2025
Hugging Face *Certificate*

Credential ID: juanlara

Bases de datos SQL April 2025
Platzi *Certificate*

Credential ID: 539844d2-3b5e-43e9-ae00-d68331327f26

AI Agents Fundamentals February 2025
Hugging Face *Certificate*

Credential ID: juanlara

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.

Research Interests

🔍 **Continual Learning & Lifelong AI**
Studying methods that allow models to learn continuously from streaming data without catastrophic forgetting, with applications in dynamic environments and evolving tasks.

🔍 Explainable AI (XAI)

Developing interpretable machine learning models and post-hoc explanation techniques to improve transparency, accountability, and trust in AI systems.

🔍 AI for Systems Optimization

Applying ML and optimization techniques to improve system-level performance in scheduling, resource allocation, and operations research.

Languages

Spanish	Native
English	Advanced