

Andrés Quintero Bedoya

Website: <https://dresdevsite.web.app/> | Email: hi@andresqb.dev |

LinkedIn: <https://www.linkedin.com/in/andresqb198/>

Senior AI Engineer | AI Infrastructure & GenAI Systems

Professional Summary

Senior AI Engineer with 5+ years of experience architecting and scaling production-grade GenAI systems and infrastructure. Expert in bridging the gap between research and production by building resilient AI infrastructure, automated MLOps pipelines, and observability stacks. Proven success in reducing cloud costs by 30% and managing 15+ AI products. Committed to building reliable, distributed systems that drive business impact.

Professional Experience

Senior AI Engineer | Itti (Paraguay)

03/2025 – Present

- Spearheaded the design and operational strategy for 15+ production-grade GenAI systems for Grupo Vásquez, ensuring system reliability and performance at scale.
- Led the AI infrastructure roadmap, implementing best practices for distributed AI systems on AWS to ensure high availability and low latency.
- Acted as the technical bridge between research and operations, defining standards for model deployment, monitoring, and production-grade reliability.

Site Reliability & GenAI Engineer | Guane Emerging Technologies

10/2023 – 03/2025

- Architected a 30% reduction in cloud infrastructure costs by optimizing Kubernetes resource allocation, container orchestration, and auto-scaling policies on GCP.
- Engineered a high-throughput, agentic RAG platform capable of automating the analysis of 10,000+ judicial decisions, significantly improving legal document processing efficiency.
- Built and maintained comprehensive observability stacks using Prometheus, Grafana, and Slack-integrated alerting to minimize MTTR (Mean Time to Resolution).
- Automated CI/CD pipelines, streamlining the deployment of ML services and reducing time-to-production for AI models.

Applied AI Research Engineer | Universidad de Antioquia

02/2023 – 02/2025

- Engineered a high-performance Colombian Sign Language (LSC) translation system, leveraging deep learning architectures (GRU/LSTM) to achieve 91.17% F1 score.
- Designed the data engineering architecture to process and manage a high-volume corpus of 1,500+ videos.
- Optimized model training and inference pipelines to ensure feasibility for real-time application.

Data Analyst | Experimentality Labs

12/2021 – 08/2022

- Developed AWS-based ETL pipelines utilizing Terraform and Python, facilitating seamless data integration for business intelligence products.

Selected Projects

Matroide — AI Observability

- Developed a specialized observability prototype for LLM-based applications, focusing on tracing, cost monitoring, and risk assessment for distributed systems.

LSC Translator

- Architected a scalable, production-ready deployment on Kubernetes, enabling real-time AI translation with 91.17% F1 accuracy.

Skills

- **AI & GenAI Systems:** LLMs, Agentic AI, RAG, LangChain, LlamaIndex, Transformers, Prompt Engineering.
- **Infrastructure & Cloud:** AWS, GCP, Azure, Kubernetes (K8s), Docker, Terraform, Distributed Systems.
- **MLOps & Observability:** Prometheus, Grafana, Jaeger, Tempo, CI/CD, Cost Optimization, LLM Tracing.
- **Backend Engineering:** Python, FastAPI, Microservices, Async Processing, REST APIs.
- **Data & Storage:** PostgreSQL, MongoDB, SQL, ETL/ELT, NumPy, Pandas.

Education

- **B.S. in Systems Engineering** | Universidad de Antioquia