

Daniel Esteban Cárdenas Jerez

Medellín, Colombia (Remote) · +57 318 884 8757 · daniestebanc@hotmail.com linkedin.com/in/dancarjes · github.com/carjes232 · dev.danielcardenas.dev

AI Automation / RAG Engineer

AI engineer building production-oriented GenAI systems: RAG pipelines, voice/document automation, computer vision/OCR, Python APIs, vector search, and edge ML telemetry. Strong in Python/FastAPI/Flask, PostgreSQL/pgvector, FAISS, Gemini/Azure OpenAI, Docker, MQTT, STM32/FreeRTOS, and reliability-minded evaluation.

Experience

Lean Tech — Artificial Intelligence Engineer · Medellín · Feb 2026 – Present - Building Python-based AI automation, LLM integration workflows, and maintainable AI features for production-oriented business use cases.

Solenium — ML & Firmware Engineer · Medellín · Jan 2025 – Feb 2026 - Led ML development for a smart-meter NILM system with real-time energy disaggregation at **10 Hz** and **sub-2-second latency**. - Built retrainable ML workflows plus a FaceID field-attendance PWA with on-device recognition, liveness, GPS/project checks, offline queueing, and photo-backed publishing. - Contributed to STM32 acquisition, OTA updates, watchdogs, OpenSSL-backed data handling, JWT Flask/Django APIs, and **six field-device** deployments.

Magnus — Machine Learning Engineer (LLMs & RAG) · Remote · May 2023 – Sep 2024 - Built Python services for Azure-hosted LLM workflows supporting clinical simulations and survey feedback analysis. - Implemented RAG workflows with **FAISS** and **PostgreSQL/pgvector**, plus Flask APIs, async inference/summarization services, embedding comparisons, prompt evaluation, and quality monitoring.

XM S.A. E.S.P — Market Development Analyst · Medellín · Jul 2024 – Jan 2025 - Shipped a **PyQt6 + Pandas** SQL/Excel/API reconciliation workbench with tolerance rules and audit exports, cutting reconciliation from **4-6 hours** to **under 30 minutes**.

Universidad Nacional de Colombia — Teaching Assistant · 2023 – 2024 - Guided STM32F4/FreeRTOS robotics labs: sensor fusion, odometry, PWM/PID, onboard A*, debugging, and experiment documentation.

Selected Projects

Multilingual Multi-Hop RAG — FastAPI · pgvector HNSW · BGE-M3 · Ollama: local-first RAG backend with translate/retrieve/cite/synthesize prompting, recall@k evaluation, and p95 latency target under 5 seconds. **Solenium FaceID Attendance PWA** — Flask · PWA · IndexedDB · Geolocation: field attendance system with local face recognition/liveness, offline queueing, GPS/project checks, and photo-backed event publishing. **Telegram Inventory Bot with Gemini Voice Parsing** — Python · Gemini · SQLite: voice messages become confirmed sales, reservations, payments, inventory changes, and customer debt records.

Skills

LLMs / AI: RAG, embeddings, prompt engineering, Azure OpenAI, Gemini, OCR, computer vision, pgvector, FAISS, recall@k, citation checks **Backend / Data:** Python, FastAPI, Flask, Django, Pandas, PostgreSQL, SQLite, Docker/Compose, REST APIs, async workers **Edge / Embedded:** STM32, FreeRTOS, C/C++, MQTT, OTA, watchdogs, telemetry, Raspberry Pi, OpenCV, ROS2

Education & Certifications

B.Sc. Ingeniería Física — Universidad Nacional de Colombia · 2019 – 2025 IBM Deep Learning with PyTorch, Keras and TensorFlow · AWS Academy Cloud Foundations · Cloud Computing Foundations · SQL · Applied Plotting, Charting & Data Representation in Python

Languages

Spanish (Native) · English (C1) · Portuguese (A2)