Juan Diego Londoño Chavarría

AI Engineer | Sistemas agénticos, RAG y LLMOps en AWS

Medellín, Colombia · Disponible remoto | juandiego0329@gmail.com | linkedin.com/in/juandlch | github.com/JuanDLCH
Portafolio: juandlch.dev

4+ años en software | 2+ años liderando IA | 10+ proyectos con clientes | 5 industrias | Certificado AWS y Claude

PERFIL PROFESIONAL

AI Engineer con 4+ años en ingeniería de software y 2+ años llevando IA generativa a producción en AWS. Diseño y entrego sistemas agénticos, RAG híbrido, IA documental y agentes de voz en tiempo real, desde el descubrimiento con el cliente hasta su infraestructura. Mi base full-stack (Python, TypeScript, serverless) hace que cada solución incluya APIs, evaluación, observabilidad y estimación de costos, no solo un prompt. Casos de estudio y un Voice Lab en vivo en **mi portafolio**.

HABILIDADES PRINCIPALES

IA / LLM: agentes de IA, flujos agénticos, LangGraph, LangChain, LlamaIndex, MCP, RAG (búsqueda híbrida, reranking), embeddings, búsqueda vectorial, prompt engineering, tool/function calling, salidas estructuradas, Claude, OpenAI, Gemini, Amazon Nova

LLMOps: evaluación automatizada de LLMs (agentes evaluadores), trazas con LangSmith, registro y costeo por llamada, observabilidad con CloudWatch, CI/CD para IA (GitHub Actions, Terraform, AWS SAM)

Voz y documentos: Deepgram, ElevenLabs, Amazon Transcribe, Google Cloud STT/TTS, WebSockets, WebRTC, Textract, OCR

AWS: Bedrock, Bedrock AgentCore, Lambda, Step Functions, EventBridge, API Gateway, DynamoDB, Aurora, S3, CloudWatch, ECS/Fargate y más | **Cloudflare:** Workers, Durable Objects, D1, Cron Triggers, Zero Trust Access, Turnstile

Backend y datos: Python, TypeScript, Node.js, FastAPI, REST, PostgreSQL, pgvector, Redis, Redshift, Glue, Docker

Frontend: React, Next.js, Tailwind, Angular, Vue | **Integraciones:** WhatsApp API, Google Maps (GCP), HubSpot, Twilio, n8n

EXPERIENCIA PROFESIONAL

Lead AI Engineer | Trade EC | Quito, Ecuador (Remoto) **Sep 2024 – Oct 2026**

- Lideré un equipo de 3 personas que entregó soluciones de IA a la medida (chatbots, RAG/OCR, automatización de voz y texto) en AWS para restaurantes, retail, cine, finanzas y educación, coordinando con los equipos de front-end, back-end y QA.
- Diseñé sistemas agénticos y RAG (LangGraph, LangChain, LlamaIndex) con reescritura de consultas, recuperación híbrida (palabras clave + vectores), reranking y un agente validador que revisaba cada respuesta para mantenerla fundamentada en los datos del cliente.
- **LLMOps:** Construí agentes evaluadores que probaban cada cambio en ciclo y registré y costeé cada llamada a LLM (LangSmith, CloudWatch, Aurora): las decisiones de modelo y prompt se basaban en latencia, costo y calidad.
- Fui el referente técnico del equipo en descubrimiento con clientes, arquitectura y estimación de costos; aporté arquitectura, documentación y estimaciones al proceso de AWS Generative AI Competency de la empresa.
- Usé desarrollo asistido por IA (Claude Code, Cursor) con revisión humana estricta: cada PR se revisaba antes de llegar a producción.

Full Stack Engineer | Trade EC | Quito, Ecuador (Remoto) **Sep 2023 – Ago 2024**

- Prototipé los chatbots y herramientas LLM que dieron origen a la práctica de IA de la empresa y a sus primeras entregas a clientes.
- Entregué software de campañas y herramientas de datos para Juan Valdez: backend, integraciones, despliegue en la nube y reportes.

Full Stack Engineer | Stellion Tech SAS | Medellín, Colombia (Híbrido) **Jun 2022 – Jul 2023**

- Entregué 2 ERP de punta a punta para clientes como SPCI: backend, módulos de interfaz, pruebas y despliegue (Python, JavaScript, AWS).

PROYECTOS DESTACADOS CON CLIENTES

- **Muyu Education:** plataforma agéntica que transcribe clases grabadas (Amazon Transcribe, Step Functions) y genera evaluaciones en JSON.
- **ANEFI Ecuador:** pipeline agéntico de OCR que extrae datos estructurados de documentos financieros.
- **KFC Ecuador:** agentes de voz en tiempo real para toma de pedidos: STT/TTS en streaming, interrupciones y tool calls al sistema de pedidos.
- **Multicines, InsureHero, El Hornero, Sushicorp, JXBS:** chatbots LLM y asistentes de WhatsApp (Meta Cloud API) integrados con sistemas de negocio.

EDUCACIÓN E IDIOMAS

Ingeniería de Sistemas, Universidad de Medellín **Graduado 2022**

Idiomas: Español – Nativo | Inglés – Profesional (nivel de trabajo)

CERTIFICACIONES

- **Claude Certified Architect – Foundations (CCA-F)**, Anthropic – 2026
- **AWS Certified AI Practitioner**, Amazon Web Services – 2024 (también insignia Early Adopter)
- Bootcamp: Generative AI and LLM Applications, Udemy – Jun 2024