Juan Diego Londoño Chavarría

AI Engineer | Agentic Systems, RAG & LLMOps on AWS

Medellín, Colombia · Open to remote | juandiego0329@gmail.com | linkedin.com/in/juandlch | github.com/JuanDLCH
Portfolio: juandlch.dev

4+ yrs software engineering | 2+ yrs leading AI delivery | 10+ client projects | 5 industries | AWS & Claude Certified

PROFESSIONAL SUMMARY

AI Engineer with 4+ years of software engineering and 2+ years taking generative AI to production on AWS. Design and ship agentic systems, hybrid RAG, document AI and realtime voice agents, from the client's first discovery call to the infrastructure they run on. Full-stack roots (Python, TypeScript, serverless) mean every solution ships with APIs, evaluation, observability and a cost estimate, not just a prompt. Case studies and a live Voice Lab on **my portfolio**.

CORE SKILLS

AI / LLM: AI agents, agentic workflows, LangGraph, LangChain, LlamaIndex, MCP, RAG (hybrid search, reranking), embeddings, vector search, prompt engineering, tool/function calling, structured outputs, Claude, OpenAI, Gemini, Amazon Nova

LLMOps: automated LLM evaluation (evaluation agents), LangSmith tracing, per-call logging and cost tracking, CloudWatch observability, CI/CD for AI (GitHub Actions, Terraform, AWS SAM)

Voice & Documents: 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 and more | **Cloudflare:** Workers, Durable Objects, D1, Cron Triggers, Zero Trust Access, Turnstile

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

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

PROFESSIONAL EXPERIENCE

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

- Led a 3-person team delivering custom AI solutions (chatbots, RAG/OCR, voice and text automation) on AWS for clients in food service, retail, cinema, finance and education, coordinating cross-team development with the front-end, back-end and QA teams.
- Designed agentic and RAG systems (LangGraph, LangChain, LlamaIndex) with query rewriting, hybrid keyword + vector retrieval, reranking and a validator agent that checked answers before they were returned, keeping responses grounded in client data.
- **LLMOps:** Built evaluation agents that tested every change in a loop and logged and costed every LLM call (LangSmith, CloudWatch, Aurora), so model and prompt decisions were driven by latency, cost and quality data.
- Owned client discovery, architecture and cost estimation as the team's technical reference; contributed architecture, documentation and estimates to the company's AWS Generative AI Competency process.
- Used AI-assisted development (Claude Code, Cursor) under strict human review: every PR was reviewed and nothing reached production unchecked.

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

- Prototyped the chatbots and LLM tooling that grew into the company's AI practice and its first AI client deliveries.
- Delivered campaign software and data tooling for Juan Valdez: backend services, integrations, cloud deployment and reporting.

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

- Delivered 2 ERP systems end to end for clients including SPCI: backend logic, UI modules, testing and deployment (Python, JavaScript, AWS).

SELECTED CLIENT WORK

- **Muyu Education:** agentic platform that transcribes recorded classes (Amazon Transcribe, Step Functions) and returns structured JSON evaluations.
- **ANEFI Ecuador:** agentic OCR pipeline extracting structured data from financial documents for downstream workflows.
- **KFC Ecuador:** realtime voice agents for order taking: streaming STT/TTS, interruption handling and tool calls into ordering systems.
- **juandlch.dev (personal):** live AI lab (chat + realtime voice agents) on Cloudflare Workers, Durable Objects and D1, with per-IP and per-call cost caps, Turnstile, and an admin guarded by Zero Trust Access.
- **Multicines, InsureHero, El Hornero, Sushicorp, JXBS:** LLM chatbots and WhatsApp (Meta Cloud API) assistants integrated with business systems.

EDUCATION & LANGUAGES

B.Sc. Systems Engineering, Universidad de Medellín **Graduated 2022**

Languages: Spanish – Native | English – Professional working proficiency

CERTIFICATIONS

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