

Bryan Javier Cáceres Duarte

Medellin - Santiago | CO +57 302 1226914 - CL +56 9 4032 2987 | javier.duarte.cf@gmail.com
[LinkedIn](#) | [GitHub](#) | [Medium](#)

Professional Summary

AI Engineer with 5+ years of software engineering experience (5+ YOE), including 3+ years as a Tech Lead driving technical roadmaps and 2+ years delivering production-grade AI solutions. Dual AWS certified (AI Practitioner and Cloud Practitioner) with strong full-stack foundations. Specialized in architecting and operating production-grade AI systems, focusing on Agentic AI, orchestration with LangChain/LangGraph, and deep observability.

I build resilient software by integrating AI capabilities through context management, *tool use*, advanced RAG, and comprehensive monitoring. I combine a product-oriented mindset with deep technical rigor to deliver *end-to-end* AI solutions: from initial prototyping and development to production deployment and post-launch monitoring.

Technical Skills

- * **AI Engineering & LLM Ops:** Agentic AI Systems, LangChain, LangGraph, LangChain Deep Agents, RAG (Retrieval-Augmented Generation), Prompt Engineering, Hallucination Monitoring (LangSmith/Langfuse), Embeddings, Vector Search
- * **Fullstack Development:** Python (Advanced), TypeScript, Node.js, Next.js, React, API Integration, Microservices, RESTful Services
- * **Cloud & Infrastructure:** AWS (Bedrock, Serverless, Lambda, EC2, S3, ECS, ECR, CloudWatch, SAM, RDS, App Runner, Amplify), Terraform (IaC), Docker, PostgreSQL, PgVector, Redis, CI/CD (GitHub Actions), CodeArtifact
- * **AI Development Tools:** MCPs, Cursor | Advanced Claude Code (proficient in all primitives and the Claude Agent SDK to embed into CI/CD pipelines and automations; authored custom Claude Code Plugin)
- * **Other:** Self-hosted n8n, Sentry, Advanced Excel

Featured Project — Production-Grade AI

Panóptico AI — panopticoai.com

Production-grade autonomous news intelligence platform that clusters and analyzes event feeds using agents, embeddings, and AI workflows. [Read Technical Blog Post](#)

Professional Experience

Vera (Fintech) | United States (Remote)

AI Engineer | April 2026 – Present | 6 months

- * Transitioned the agent's core from a centralized single-router architecture to an orchestrated network of subagents using **LangChain Deep Agents**, specializing domain handling (financial and operational) and reducing user intent drift.
- * Designed and implemented the user long-term memory architecture (personal and financial context), managing asynchronous ingestion and prioritization via a dedicated attention-queue pipeline.
- * Re-engineered the voice agent over WebSockets and WebRTC streaming, integrating an in-flight memory layer and low-latency RAG for real-time conversational responses.
- * Engineered an internal, CLI-driven LLM evaluation framework integrated into CI/CD pipelines, drastically reducing hallucinations and standardizing production observability with Langfuse.

FirstJob | Santiago, Chile

Tech Lead (Promotion) | July 2022 – March 2026 | 3 years 8 months

- * Architected and deployed an autonomous agentic system using LangGraph and LangChain, featuring multi-step reasoning (*autonomous tool-calling*), long-term memory management, and cyclical clustering/refinement pipelines.
- * Built a hybrid RAG engine using PostgreSQL and PgVector (semantic vector search + full-text search), optimizing context window packing to reduce inference latency and compute costs.
- * Deployed continuous agent monitoring and tracing using LangSmith (later migrated to Langfuse), ensuring output auditing and proactive hallucination mitigation.
- * Significantly reduced AI operational expenditure (OpEx) and inference latency by implementing *prompt caching*, batch generation APIs, and dynamic multi-provider routing.
- * Developed content safety architectures based on the Generator-Validator pattern (*Guardrails*), enforcing strict structured outputs (*JSON Schema / Pydantic*) and distributed multi-provider rate limiting.
- * Defined team-wide AI engineering standards, leading cross-functional adoption of Claude Code and integrating Claude Agent SDK into CI/CD for automated PR code reviews and continuous synchronization of `CLAUDE.md` and `AGENTS.md`.
- * Automated 100% of the job post auditing workflow using an autonomous multi-agent system, eliminating human error and enforcing strict platform editorial compliance.
- * Architected and scaled both frontend and backend for an internal greenfield CRM, centralizing disparate product and operational databases to establish a single source of truth for Data and Product teams.
- * Led the architectural migration of mission-critical operational components to an AWS Serverless environment, optimizing infrastructure costs and accelerating agile deployments through CI/CD pipelines.
- * **Platform Project Leader (Promotion)** — *April 2022 – July 2022*
- * **Project Coordinator (Promotion)** — *July 2021 – April 2022*
- * **People Data & Ops Trainee** — *January 2021 – June 2021*

Education & Certifications

- * **AWS Certified AI Practitioner** | Amazon Web Services
- * **AWS Certified Cloud Practitioner** | Amazon Web Services
- * **IBM AI Developer Specialization** | IBM
- * **AI Engineering & Full Stack Development** | Continuous Learning — Ongoing study of scalable architectures, LLM systems, and modern software design patterns
- * **B.S. in Certified Public Accounting & Auditing (Contador Auditor)** | INACAP Technological University of Chile

Languages

- * **Spanish:** Native
- * **English:** Professional Working Proficiency (B2 — Fluent technical communication)