SEBASTIAN RAMOS REGALADO

AI Engineer — Agentic AI · LLM Systems · Cloud AI

Quito, Ecuador (Open to Remote) | +593 99-701-2463 | sramosre@gmail.com | LinkedIn | GitHub

PROFESSIONAL SUMMARY

AI Engineer with 6 years of experience building and deploying production-ready AI and LLM-powered systems. Specialized in agentic AI architectures, retrieval-augmented generation (RAG), and intelligent automation that integrate with APIs, databases, and enterprise workflows. Proficient in Python, PyTorch, and TensorFlow and the modern generative-AI stack — LangChain, LangGraph, CrewAI, and vector databases (Pinecone, Weaviate, FAISS) — with hands-on deployment across AWS, Azure, and Google Cloud Vertex AI. Technical leader who has led teams of up to 5 and mentored 6, taking AI initiatives end-to-end from prototype to production.

TECHNICAL SKILLS

Generative AI & LLMs: OpenAI API, Anthropic Claude, LangChain, LangGraph, CrewAI, AutoGen, Retrieval-Augmented Generation (RAG), prompt engineering, embeddings, semantic search

Agentic AI & Harness: MCP (Model Context Protocol) servers, custom tools & sub-agents, tool-using agents, multi-agent architectures, memory & context management, autonomy boundaries & guardrails, orchestration

Evals & Observability: Offline eval sets, LLM-as-judge with human calibration, CI regression gates, tracing & tool-call logs, LangSmith / Langfuse, cost & latency budgets

Machine Learning & Deep Learning: PyTorch, TensorFlow, scikit-learn, deep learning (MLPs, CNNs), NLP, feature engineering, model optimization

MLOps & Cloud: Docker, Kubernetes, CI/CD, MLflow, model deployment & monitoring, AWS, Azure, Google Cloud (Vertex AI), Microsoft Fabric, Databricks

Databases & Vector Stores: Pinecone, Weaviate, FAISS, PostgreSQL, MySQL, SQL Server, MongoDB

Full-Stack & Automation: TypeScript / JavaScript, internal tools & admin UIs, REST APIs, webhooks, n8n, Zapier, Make, event-driven systems

Programming: Python (advanced), TypeScript / JavaScript, SQL, C#, Java

PROFESSIONAL EXPERIENCE

Capmation Inc — Staff Engineer Twin Cities, MN, USA (Remote) | 2020 – Present

- Designed and deployed LLM-powered applications and agentic AI workflows that automated internal engineering processes, reducing manual work by approximately 60%.
- Built and shipped production MCP (Model Context Protocol) servers and custom tool surfaces that expose business logic to AI agents and non-technical teammates, with explicit tool boundaries and autonomy controls.
- Built tool-using AI agents and focused sub-agents with LangChain, LangGraph, and CrewAI, and implemented RAG pipelines over enterprise data using vector databases (Pinecone, Weaviate, FAISS) and embedding models to ground responses and reduce hallucinations.
- Established evaluation frameworks — offline eval sets, LLM-as-judge with human calibration, and CI regression gates — with tracing and observability (LangSmith, Langfuse) to keep agent systems reliable in production.
- Built internal tools and admin UIs (dashboards, configuration panels) that let non-engineers make routine changes safely, with preview and audit logging.
- Deployed services on Google Cloud Vertex AI with Docker, Kubernetes, and CI/CD, and engineered Microsoft Fabric and Databricks pipelines (Medallion architecture) powering Power BI; led teams of up to 5 and mentored 6.

SELECTED PROJECTS

Production MCP Server & Agent Harness

- Designed and shipped an MCP server exposing business tools to AI agents and internal clients, with focused sub-agents and clearly defined tool boundaries.
- Defined autonomy boundaries (always-do, ask-first, never-do) and guardrails, and instrumented traces, tool-call logs, and cost/latency budgets in dashboards a non-engineer can read.

Agentic AI System for Autonomous Workflow Execution (Multi-Agent Architecture)

- Designed a multi-agent system that reasons, plans, and executes multi-step tasks autonomously, with memory and context management for long-running workflows.
- Orchestrated tool-using agents with LangChain, CrewAI, and n8n so they coordinate actions and adapt to intermediate results, reducing failure cases in complex workflows.

Production-Grade RAG System for Contextual Knowledge Retrieval

- Built a RAG system over structured and unstructured data with embedding pipelines and vector search on Pinecone, FAISS, and Weaviate.
- Optimized chunking and retrieval to reduce hallucinations, and evaluated groundedness and answer quality against offline benchmark sets.

EDUCATION

LaSalle College — Program in Artificial Intelligence & Machine Learning Montreal, Canada | 2024 – 2025

Universidad San Francisco de Quito — B.Sc. in Computer Engineering Quito, Ecuador | 2014 – 2020