

Esteban Gaviria Zambrano

AI Engineer | LLMs, RAG & AI Agents

Cali, Colombia • Open to remote • estebangaviria82@gmail.com • +57 310 612 7486

LinkedIn • GitHub • Portfolio

Summary

AI Engineer with one year of professional experience building and shipping a multimodal conversational AI SaaS platform (real-time, voice-cloned *digital humans*) to production. Specialized in LLM orchestration across OpenAI, Anthropic, Google Gemini and open-source models, retrieval-augmented generation (RAG) over private knowledge bases, agents and tool use (MCP), and low-latency voice AI pipelines. Builds production-ready AI services in Python and TypeScript and uses agent-assisted development workflows to ship faster without compromising quality. Final-semester student of a double degree in Telematics and Systems Engineering.

Experience

Brain AIX (Remote)

AI Software Developer

September 2025 – Present

Multimodal conversational AI SaaS platform: real-time agents with cloned voices and avatars (brainaix.online).

- Designed and implemented the product's LLM orchestration layer —context and conversational memory management, tool use, response streaming and per-session cost control— across OpenAI, Anthropic, Gemini and open-source models, using LangChain/LangGraph, LlamaIndex and native provider SDKs.
- Built the end-to-end RAG pipeline (PDF/Word ingestion, chunking, embeddings and vector search) and fine-tuned the embedding model on domain data to improve retrieval, letting each user ground agent answers in their own knowledge base and cutting unsupported answers by 40% on an in-house evaluation set.
- Engineered the low-latency voice pipeline (STT → LLM → TTS with voice cloning) on Deepgram, Whisper, ElevenLabs and Azure/Google Speech, sustaining natural conversations in 10+ languages at under 3s end-to-end latency; also configured a real-time speech-to-speech agent on Alibaba Cloud (Qwen-Omni, Qwen-TTS voice cloning).
- Co-designed the platform's microservices architecture and delivered 9+ services in Python (FastAPI) and Node.js (NestJS) —authentication, multi-tenancy and third-party API integrations—, shipping them to production with Docker and GitHub Actions CI/CD for automated, zero-downtime releases.
- Introduced AI agent-assisted development to the team (Claude Code, Cursor, Codex, Antigravity) and built custom *skills* and MCP servers that automate service scaffolding, testing and documentation.

Universidad Icesi (Cali, Colombia)

Teaching Assistant — Mathematics & Computational Thinking

2023 – 2026

- Supported engineering students in Applied Math I–III and Multivariable Calculus, and coached 25+ students in programming fundamentals and algorithmic problem-solving.

Technical Skills

- LLM Engineering:** orchestration, prompt engineering and evaluation, RAG, embeddings and embedding fine-tuning, vector databases (Weaviate), agents and tool use (MCP), streaming, cost control
- Models & Frameworks:** OpenAI, Anthropic Claude, Google Gemini, Alibaba Qwen, open-source LLMs; LangChain, LangGraph, LlamaIndex, Hugging Face
- Voice AI:** STT/TTS, voice cloning, real-time speech-to-speech; Deepgram, Whisper, ElevenLabs, Azure/Google Speech, Qwen-Omni, Qwen-TTS
- ML & Data:** Python (NumPy, pandas, scikit-learn), PostgreSQL, MongoDB
- Backend & Full-Stack:** Python (FastAPI), TypeScript/Node.js (NestJS), React, Next.js, Java; microservices, REST APIs, multi-tenancy, OAuth2/JWT
- AI Deployment:** Docker, Kubernetes, GitHub Actions, Azure, GCP, AWS
- AI-Assisted Engineering:** Claude Code, Cursor, Codex, Antigravity; custom *skills*, MCP servers and agents

Education

Universidad Icesi — Cali, Colombia

Expected graduation: August 2027

B.Sc. in Telematics Engineering & B.Sc. in Systems (Software) Engineering (double degree).

Additional

- Languages:** Spanish (native), English (C1 Advanced — EF SET certificate, 2025)
- Athletics:** Bronze medal (3rd place), Judo –90 kg — ASCUN 2026 National University Games, representing Universidad Icesi
- Interests:** generative AI and agents, open-source communities, judo, running and cycling