

# Esteban Gaviria Zambrano

AI Engineer | LLMs, RAG y Agentes de IA

Cali, Colombia • Disponible para trabajo remoto • estebangaviria82@gmail.com • +57 310 612 7486

LinkedIn • GitHub • Portafolio

## Perfil Profesional

AI Engineer con un año de experiencia construyendo y llevando a producción una plataforma SaaS de IA conversacional multimodal (*humanos digitales* con voz clonada en tiempo real). Especializado en orquestación de LLMs (OpenAI, Anthropic, Gemini y modelos open-source), RAG sobre bases de conocimiento privadas, agentes y *tool use* (MCP) y pipelines de voz de baja latencia. Desarrolla servicios de IA en Python y TypeScript con flujos asistidos por agentes para entregar más rápido sin sacrificar calidad. Estudiante de último semestre de Ingeniería Telemática e Ingeniería de Sistemas.

## Experiencia

### Brain AIX (Remoto)

AI Software Developer

Septiembre 2025 – Presente

Plataforma SaaS de IA conversacional multimodal: agentes con voz clonada y avatar en tiempo real (brainaix.online).

- Diseñé e implementé la capa de orquestación de LLMs del producto —gestión de contexto y memoria conversacional, uso de herramientas, respuesta en streaming y control de costo por sesión— sobre OpenAI, Anthropic, Gemini y modelos open-source, con LangChain/LangGraph, LlamaIndex y SDKs nativos.
- Construí el pipeline completo de RAG (ingesta de PDF/Word, segmentación, embeddings y búsqueda vectorial) e hice fine-tuning del modelo de embeddings con datos del dominio para mejorar la recuperación, permitiendo que cada usuario fundamente las respuestas del agente en su propia base de conocimiento y reduciendo en 40% las respuestas sin sustento sobre un set de evaluación propio.
- Implementé el pipeline de voz de baja latencia (STT → LLM → TTS con clonación de voz) sobre Deepgram, Whisper, ElevenLabs y Azure/Google Speech, sosteniendo conversaciones naturales en más de 10 idiomas con latencia extremo a extremo menor a 3s; además, configuré un agente de voz a voz en tiempo real sobre Alibaba Cloud (Qwen-Omni, clonación con Qwen-TTS).
- Co-diseñé la arquitectura de microservicios y desarrollé más de 9 servicios en Python (FastAPI) y Node.js (NestJS) —autenticación, multi-tenancy e integraciones con APIs de terceros—, llevándolos a producción con Docker y CI/CD en GitHub Actions mediante releases automatizados y sin interrupción.
- Introduje en el equipo el desarrollo asistido por agentes de IA (Claude Code, Cursor, Codex, Antigravity) y construí *skills* y servidores MCP propios que automatizan el scaffolding de servicios, las pruebas y la documentación.

### Universidad Icesi (Cali, Colombia)

Monitor Académico — Matemáticas y Pensamiento Computacional

2023 – 2026

- Acompañé a estudiantes de ingeniería en Matemáticas Aplicadas I–III y Cálculo Multivariable, y formé a más de 25 estudiantes en fundamentos de programación y resolución algorítmica de problemas.

## Habilidades Técnicas

- **Ingeniería de LLMs:** orquestación, ingeniería y evaluación de prompts, RAG, embeddings y fine-tuning de embeddings, bases de datos vectoriales (Weaviate), agentes y *tool use* (MCP), streaming, control de costos
- **Modelos & Frameworks:** OpenAI, Anthropic Claude, Google Gemini, Alibaba Qwen, LLMs open-source; LangChain, LangGraph, LlamaIndex, Hugging Face
- **IA de Voz:** STT/TTS, clonación de voz, voz a voz en tiempo real; Deepgram, Whisper, ElevenLabs, Azure/Google Speech, Qwen-Omni, Qwen-TTS
- **ML & Datos:** Python (NumPy, pandas, scikit-learn), PostgreSQL, MongoDB
- **Backend & Full-Stack:** Python (FastAPI), TypeScript/Node.js (NestJS), React, Next.js, Java; microservicios, APIs REST, multi-tenancy, OAuth2/JWT
- **Despliegue de IA:** Docker, Kubernetes, GitHub Actions, Azure, GCP, AWS
- **Desarrollo Asistido por IA:** Claude Code, Cursor, Codex, Antigravity; *skills*, servidores MCP y agentes propios

## Educación

Universidad Icesi — Cali, Colombia

Grado estimado: agosto 2027

Ingeniería Telemática e Ingeniería de Sistemas (Software), doble programa.

## Información Adicional

- **Idiomas:** español (nativo), inglés (C1 Avanzado — certificado EF SET, 2025)
- **Deporte:** medalla de bronce (tercer lugar), judo –90 kg — Juegos Universitarios Nacionales ASCUN 2026, en representación de la Universidad Icesi
- **Intereses:** IA generativa y agentes, comunidades open-source, judo, running y ciclismo