Alejandro Sánchez Yalí

AI Full Stack Engineer

Let's Connect

asanchezyali@gmail.com | asanchezyali | asanchezyali | asanchezyali.com

SUMMARY

AI Full Stack Engineer with 8+ years designing and shipping web applications end to end and 3+ years building LLM-powered products. Architect and lead engineer of three deployed systems – an AI WhatsApp support platform, an AI-assisted recruitment platform and a surgical supply chain system – spanning Python/FastAPI backends, Next.js/TypeScript frontends, PostgreSQL and cloud deployment. M.Sc. in Mathematics and 11 years teaching mathematics and physics, which shows up in design docs and code review.

TECHNICAL SKILLS

- **Languages:** Python (8+ years), TypeScript (6+ years), JavaScript (6+ years), SQL
- **Frontend:** React.js (7+ years), Next.js (6+ years), Tailwind CSS, Shadcn UI, Redux, Zustand, Zod, Effect, Three.js
- **Backend:** FastAPI, Django, Django REST Framework, Celery, NestJS, Node.js, Express, RESTful APIs, GraphQL, microservices
- **AI & LLMs:** LangChain, LangGraph, LiteLLM, RAG with pgvector, MCP, Claude (Anthropic API), OpenAI API, Gemini, Whisper, ElevenLabs, WhatsApp Cloud API
- **Databases & Infra:** PostgreSQL, pgvector, Redis, MongoDB, Firebase, Docker, AWS, GCP, Terraform, GitHub Actions, CI/CD
- **Engineering:** Clean architecture, DDD, TDD, Vitest, Jest, React Testing Library, code review

EXPERIENCE

Independent AI Consultant Oct 2026 – Present
Freelance *Remote*

– Lead engineer of Aluna, an AI recruitment platform for staffing agencies (aluna.works): a Next.js 15 application owning all domain state and a stateless FastAPI service owning model calls and the WhatsApp channel; 336 source files, 26 tables, 11 external providers, Ley 1581/2012 compliance in the data model
– Durable Inngest analysis pipeline (9 checkpointed steps, 3 retries) where deduplication runs before quota enforcement so a re-analysed CV reuses its cached score instead of paying for another model call; one screening engine serves WhatsApp, web chat and email

AI Specialist Aug 2025 – Sep 2026
Lapzo *Remote, Mexico*

– Designed and built a Digital Professor that holds real-time conversations with learners and explains course content through interactive slides
– Owned the architecture connecting LLM services, ElevenLabs voice synthesis and the real-time layer built with Next.js, NestJS and Firebase Realtime Database
– Built an HR agent (LangGraph, FastAPI, PostgreSQL) that answers policy questions with citations and executes HR requests with human confirmation
– Built the AI course generator (LangChain, LangGraph, n8n, Celery workers) that cut course creation from weeks to about 15 minutes, with videos and slides, leaving content curation as the only manual step

Independent Software & AI Consultant Aug 2024 – Jul 2025
Freelance *Remote*

– Lead engineer of CREARIA's multi-tenant AI support agent on WhatsApp, email and Instagram, in production with several client businesses including a real-estate agency. Support went from an 8-hour human shift to 24/7 and average response time from 15–20 minutes to seconds. Led a team of three across 5 FastAPI services and 2 Next.js dashboards on AWS ECS with Terraform, RAG over pgvector, 11 MCP tools scoped per tenant, LiteLLM fallback routing, ARQ (Redis) workers for email, Meta messaging and agent orchestration, and human escalation; load-tested to 250 virtual users. Also designed the architecture of an AI coaching agent for one of CREARIA's clients
– Delivered end-to-end web applications for business clients: React.js/Next.js/TypeScript on the front, Python/FastAPI/Django and Node.js/NestJS behind it, deployed on AWS with CI/CD and Stripe payments

– Ran technical design sessions with clients and made the technology decisions for each project; code reviewer for a client's development team

Full Stack Engineer Apr 2021 – Aug 2024
Monadical *Remote, Canada/US*

– Full-stack engineer on ZoHuddle, a video conferencing app for scientific conferences (Neuromatch) that matches participants by mutual interests and supports conferences of about 200 participants: React front end and Django back end, with Django Channels (WebSockets) and Twilio Video
– Full-stack engineer on Virtue Poker, a decentralized poker platform on Ethereum supporting up to 50 concurrent tables of 8 players: React front end, Django back end and smart-contract integration
– Built VC Tools, an LLM tool that retrieves and processes company information for investors (OpenAI, LangChain)
– Reviewed code on a distributed team across Canada, the US and Latin America

Full Stack Engineer Aug 2018 – Oct 2020
BCFort *Medellín, Colombia*

– Designed blockchain products from scratch; built dApp frontends (React, Web3.js) and analytics platforms (Django, Node.js) on AWS, GCP and IBM Cloud; mentored new developers and reviewed the team's code

Professor of Mathematics & Physics (part-time) Jan 2010 – Oct 2021
Universidad de Antioquia *Medellín, Colombia*

– Taught mathematics and physics; in the last year and a half also machine learning and data science with Python, NumPy, Pandas, TensorFlow and PyTorch
– Developed the AI/ML elective curriculum for the Mathematics Institute and mentored student research projects

FEATURED PROJECTS

- **Plixiq – AI Customer Support on WhatsApp** – *Python, FastAPI, LiteLLM, PostgreSQL, Redis, Next.js, WebSocket*
  - Architect and lead engineer of a multi-tenant SaaS where AI agents answer routine customer inquiries, book appointments and escalate the rest to human operators with full conversation context
  - Modular monolith (14 components plus a shared kernel) with boundaries enforced in CI; LiteLLM layer with a primary and fallback provider per agent, so an outage becomes a fallback
- **VitaStock – Surgical Supply Chain for Clinics** – *Python, FastAPI, Next.js, TypeScript, PostgreSQL, Docker, Railway*
  - Architect and lead engineer of a platform commissioned by a private surgical clinic in Colombia – 12 modules and 5 user roles designed to replace spreadsheet-based inventory and procurement; delivered, with rollout at the clinic pending
  - Surgery scheduling with supply baskets and real-time consumption tracking, lot-level traceability with expiry alerts, and true per-patient cost through a cardex

OPEN SOURCE

- **talking-avatar-with-ai** – 464 stars, 110 forks. Open-source conversational AI avatar: OpenAI GPT, Whisper, ElevenLabs, Rhubarb lip-sync, Three.js
- **Monadical-SAS/Morpheus** – 19 pull requests. Open-source AI image generation platform: FastAPI, React, Stable Diffusion

TALKS & WRITING

- **Scaling Learning Models: Parallelism Strategies in JAX with Flax** – PyCon Colombia, 2024
- **Engineering case studies:** Plixiq · CREARIA Agent · Aluna – architecture and trade-offs behind each system (2026)

EDUCATION

M.Sc. in Mathematics Aug 2010 - Sep 2013
Universidad de Antioquia *Medellín, Colombia*

B.Sc. in Teaching Mathematics and Physics Jan 2004 - Oct 2009
Universidad de Antioquia *Medellín, Colombia*