

Diego Costa

AI Engineer — products with an LLM inside

diecoscai@gmail.com · diegocosta.dev · github.com/diecoscai · linkedin.com/in/diecoscai · Montevideo, Uruguay (UTC-3 – full East Coast day, West Coast mornings)

Six years in ad operations, where I automated my own job until automating became the job. Since mid-2025 I build products with an LLM inside, full time. The model is rarely the hard part — the hard part is grounding it, isolating tenants, and being able to tell what it actually did at 3am.

WHAT I FIXED, AND WHY IT WAS HARD

The timeout that wasn't there — AnswerAgent

A client's agent was cutting out at five seconds, intermittently. The timeout everyone had configured was a no-op: the MCP subprocess was spawned without inheriting its environment, so the setting never reached it. Behind that, the vectorstore ran one unindexed lookup per result document, so under load the tail crossed the limit. Fixed both.

Users couldn't see their own documents — AnswerAgent

Document chunks were being written with no owner id, so tenant-scoped queries matched nothing. Added the safeguard at write time and a migration to backfill the orphaned rows. Also put runtime guardrails on the public embed endpoint, which had been skipping them because embed requests carry no user.

A manual nobody had written, and the tool that answers from it — SBI Bids

Documented the product screen by screen across all eight workflows, checked against the code, then built the tool that answers "how do I do X" from that manual with a link to the exact section — and says the manual doesn't cover it instead of inventing UI.

Alarms that had been silent since March — internal desktop product

CPU, memory and disk alarms sat in INSUFFICIENT_DATA because nothing was installed to emit the metrics. Built the monitoring, baked the agent into host setup, and moved the alarms into CloudFormation. Also got the Electron app shipping again: signed builds had been failing notarization for a month, and auto-update published one manifest for two architectures, so arm64 users updated into the wrong binary.

EXPERIENCE

AI Engineer — AnswerAgentAI

Jun 2025 – Present

Remote · US team · contracting through Caivanode SAS · 180+ merged pull requests

- Three products in the same engagement: AnswerAgent, an agent orchestration platform (Flowise, AWS) for enterprise clients; SBI Bids, a construction-bidding product on Azure; and an internal Electron desktop product that shipped and was later discontinued.
- On SBI Bids I write the business rules inside the Schedule-of-Values export — the Excel workbook the whole bid rides on: insurance rates, GC/GR multipliers driven by project duration, net-outs between trades, accepted alternates and exclusions — and I run the acceptance pass the client signs off on.
- On AnswerAgent I work on the parts that decide whether an agent can be trusted in production: tenant isolation, MCP tooling, runtime guardrails, and tracing.

Full Stack Developer — Freelance

2024 · 2 months

Montevideo · Remote

- Built the Deal Management module for a DSP ad platform: Angular front end over a Node API on MySQL — the screens where an ad-ops team sets up deals, and the endpoints behind them.
- I had spent six years on the other side of a module like this one, filing the tickets against it. That was the whole advantage.

Ad Operations · Programmatic — Entravision

Jul 2017 – Dec 2023

Montevideo

- Ran campaigns on a programmatic advertising platform, and automated my own part of the job with custom scripts until automating became the work I wanted to do full time.
- Built the backend tooling the DSP ran on (Node.js, SQL).
- Built the interactive ad units that shipped in Spotify campaigns (React, HTML5, CSS3).
- Was the person who explained the platform to the regional teams that had to use it — the habit that later turned into writing the SBI Bids manual.

OPEN SOURCE

heavy-mcp — npm: @diecoscai/hevy-mcp · v0.5.0 · MIT · ~340 downloads/mo

A public MCP server exposing the Hevy fitness API as 23 strongly-typed tools over stdio for Claude Desktop, Cursor and VS Code. Zod edge-validation with self-correcting errors, dry-run-by-default writes, and schemas auto-generated from Hevy's OpenAPI spec via a weekly workflow that opens PRs on drift.

STACK

Daily	TypeScript · Node.js · React · PostgreSQL · MCP · RAG · Flowise · Langfuse · Fiddler · AWS
Shipped	Electron · Docker · Playwright · Vitest · GitHub Actions (OIDC) · Zod · Angular
Languages	Spanish (native) · English (working language — standups and calls, pull requests, docs, code)

EDUCATION

2022 – 2025	Programmer Analyst (Analista en Programación) — Universidad ORT Uruguay
2021	Web Development Diploma — CEIBAL · Jóvenes a Programar

Ask me about any number here. I'll show you where it comes from.