

Ricardo Alejandro Corral García

Applied AI Engineer · GenAI Automation · LLM & RAG Systems

- Chihuahua, México
- alejandrorcorral27@gmail.com
- +52 614 344 3936
- [LinkedIn](#)
- [Descargar PDF](#)
- github.com/Aleco127

01 PROFESSIONAL SUMMARY

Applied AI engineer with an architecture and construction-operations background, currently completing a **Master's in Applied Artificial Intelligence at Tecnológico de Monterrey**. I design and ship **production GenAI systems**: multi-agent LLM pipelines, document-intelligence extractors, RAG-style retrieval, and end-to-end automation that connects models to real business workflows.

I don't stop at notebooks. Every project here is **containerized, deployed, and running** — FastAPI / Node backends, Supabase / SQL data layers, Claude API orchestration, and n8n automation, shipped on a self-managed Docker VPS. My domain edge is turning messy, real-world operational problems into reliable AI-driven products.

What I bring to a GenAI / AI Automation team: LLM orchestration & prompt engineering, retrieval pipelines, ML modeling, API integration, and the full-stack discipline to take an AI feature from prototype to a deployed, monitored, secured service.

AREA	TECHNOLOGIES
GenAI / LLMs	Claude API (Opus/Sonnet 4.x), agent orchestration, prompt engineering, prompt caching, vision-assisted extraction, RAG, embeddings / semantic search
Machine Learning	PyTorch, TensorFlow/Keras, Scikit-learn, XGBoost, LSTM, CNNs, ensemble methods, feature engineering, NumPy from-scratch backprop
Backend / APIs	FastAPI, Express/Node, Python 3.11/3.12, REST, Server-Sent Events (SSE), webhooks, auth & RBAC
Automation	n8n (20+ workflows), Twilio WhatsApp Business, Meta WhatsApp Cloud API, Telegram, scheduled pipelines, CRM/ops integrations
Data / Databases	PostgreSQL (Supabase + Row Level Security), SQLite / SQLCipher, MongoDB, SQL, Parquet, vector stores
Cloud / DevOps	Docker & Docker Compose, self-managed VPS, Caddy reverse proxy, Prometheus + Grafana, Git/GitHub
Frontend / Dashboards	React 18 + TypeScript + Vite + shadcn/ui, Vanilla JS SPA, TailwindCSS, Chart.js, Streamlit

Zook

A Claude-powered sales agent ("Alex") that answers, qualifies and follows up with real customers 24/7 on WhatsApp and web chat.

Problem solved. SMBs lose leads that message outside business hours or go cold without follow-up. Zook runs a production sales agent over WhatsApp Business (Twilio) and web chat — with human hand-off, a CRM, and campaign tooling around it.

My role. Sole architect & engineer — agent design, backend, channels, CRM, deployment.

Key features.

- **Production WhatsApp agent loop:** incoming bursts are grouped in an 8-second SQLite inbox buffer (state shared across 4 uvicorn workers — no in-process memory), answered with one Claude call, and split into human-like message bubbles at sentence boundaries.
- **Grounded answers:** retrieval over a curated knowledge base (system prompt + service/pricing docs stored in Supabase) keeps the agent on-script for a real sales funnel.
- **Sales-stage control:** the agent emits structured stage/score tags per conversation, driving lead scoring, hot-lead email alerts and automated 24-hour follow-up touches.
- **Multi-channel:** Meta click-to-WhatsApp ads land directly on the agent; web chat runs through n8n; FB/IG DMs route through a shared webhook; clients manage it from a session-based portal.

Architecture. FastAPI (Python 3.12, 4 workers) + SQLite/Supabase state; Twilio WhatsApp webhooks; n8n for the web channel; Dockerized behind Caddy on a VPS.

ARCHITECTURE — PRODUCTION WHATSAPP SALES AGENT



FastAPI · Claude API · Twilio WhatsApp · Supabase · SQLite · n8n · Docker · Caddy

➔ [Live: zook.mx](https://zook.mx)

Demonstrates: Production LLM-agent engineering (buffering, multi-worker state, reply shaping), grounded retrieval, channel integrations, end-to-end ownership. *Code: private repository — walkthrough available on request.*



Agentes de IA — Listos para trabajar

Automatiza tu negocio con agentes de IA que venden, responden y operan por ti 24/7

Más clientes. Menos trabajo. Sin contratar más personal.

[↩ Descubre qué agente necesitas](#)

[▶ Ver demo](#)

• Lead respondido en 3s



ALEX

⚡ Tarea automatizada



NOVA

📊 +40% ventas este mes



VEGA

Zook landing at zook.mx — front door of the sales agent in production

LicitaGen

Turns a government tender PDF into a structured, folio-numbered, ERP-integrated bid package.

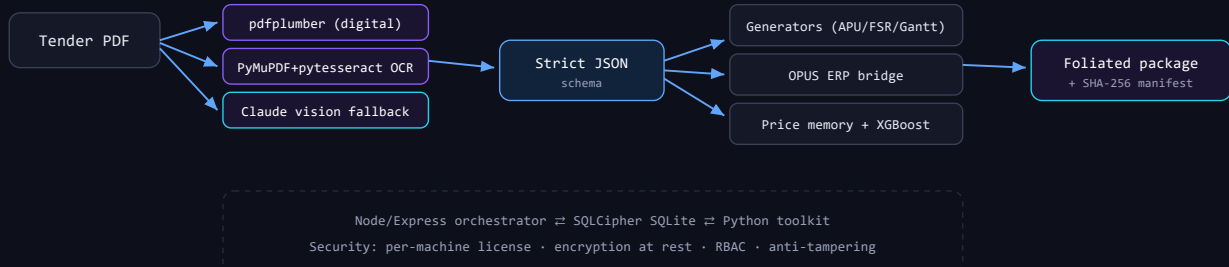
Problem solved. Assembling a public-works bid is a ~5-day manual marathon of reading hundreds of PDF pages, extracting requirements and pricing, and producing dozens of compliant annexes. LicitaGen targets ≤ 1 day.

My role. Product owner & full-stack/AI engineer — extraction pipeline, generators, OPUS ERP bridge, security architecture.

Key features.

- **LLM-assisted extraction pipeline:** digital text → OCR → **Claude vision fallback** on low-confidence pages, emitting a strict structured JSON schema.
- **Historical price memory** with normalized concepts + embeddings for hybrid semantic + exact retrieval; XGBoost regression baseline.
- **End-to-end generators:** APU cost cards, indirect-cost analysis, financial factor, critical-path schedule, cover letters, Excel/PDF — wired into the OPUS construction ERP.
- **Hardened deployment:** per-machine licensing, SQLCipher at rest, RBAC, anti-tampering.

ARCHITECTURE — DEGRADING-CONFIDENCE EXTRACTION + GENERATION



Node/Express · Python 3.11 · Claude API (vision + extraction) · pdfplumber · pytesseract · SQLCipher · XGBoost · OPUS LocalDB

Status: v1 in daily internal use at an active construction firm (desktop deployment: extraction + generators + OPUS bridge); v2 platform (LLM extractor, price ML, security hardening) specified and in iterative build.

Demonstrates: Real-world document intelligence, multimodal LLM fallback design, RAG/embedding retrieval, ML dataset engineering, security-conscious deployment. *Code: private repository — walkthrough available on request.*

WhisperWind

Multi-tenant platform to deploy and manage WhatsApp Business bots with live agent hand-off.

Problem solved. Businesses need automated WhatsApp conversations plus human oversight, multi-bot management, and secure multi-user access.

My role. Full-stack engineer — React dashboard, Supabase data/security model, WhatsApp webhook, n8n automation.

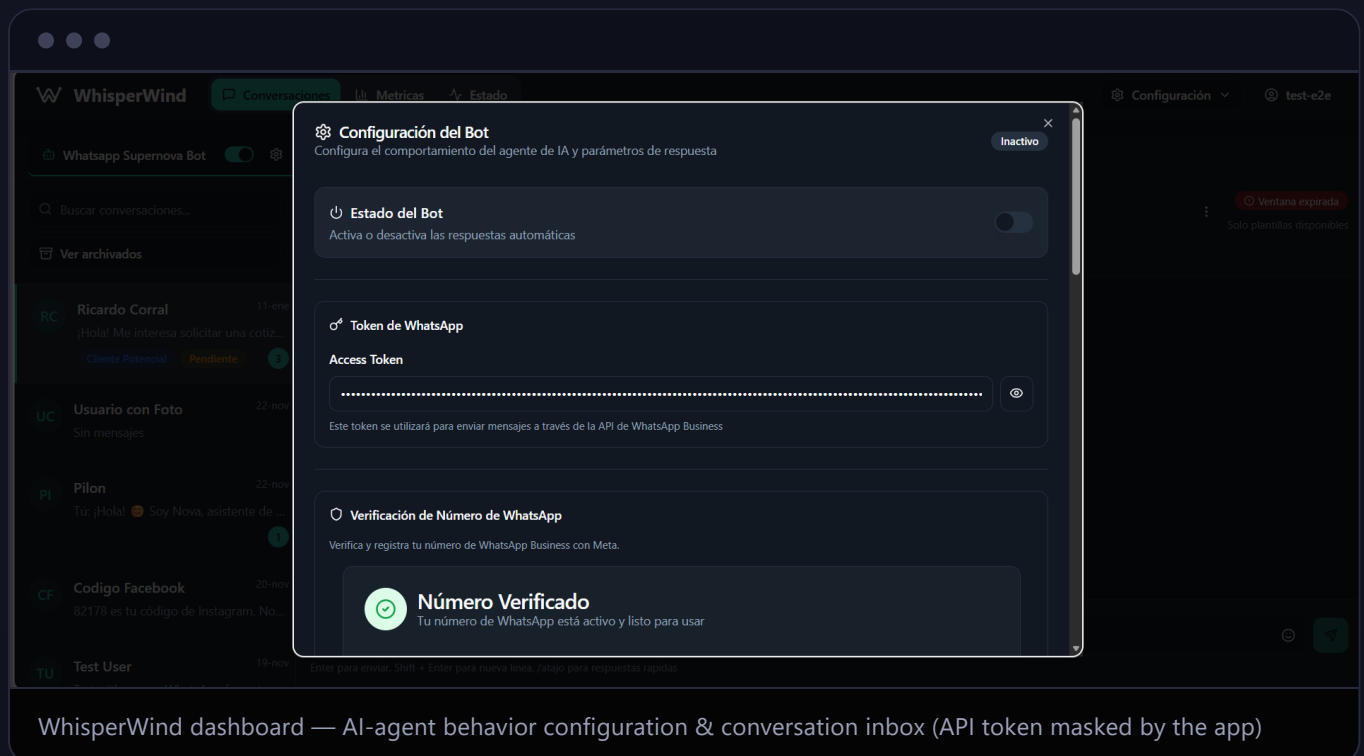
Key features. AI-agent behavior configuration panel (response parameters, bot enable/disable); role-based auth + admin panel; multi-bot management; real-time conversation inbox with tags/notes/quick-replies; Meta WhatsApp Cloud API webhook; n8n automated responses; **12-table PostgreSQL schema with non-recursive Row Level Security.**

Architecture. React 18 + TypeScript + Vite + shadcn/ui frontend; Supabase (PostgreSQL + Auth + RLS) backend; Express webhook server; n8n automation layer; Dockerized on VPS behind Caddy.

React/TS · Vite · shadcn/ui · Supabase (PostgreSQL + RLS) · Express · Meta WhatsApp API · n8n · Docker

➔ [Live: dashboard.srv1090924.hstgr.cloud](https://dashboard.srv1090924.hstgr.cloud)

Demonstrates: Full-stack AI automation, secure multi-tenant data modeling, third-party API integration, workflow orchestration. *Code: private repository — walkthrough available on request.*



Gold Price Predictor

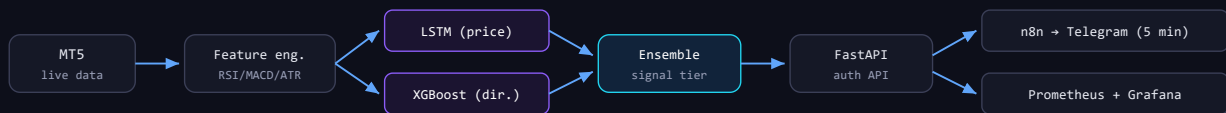
LSTM + XGBoost ensemble that predicts XAUUSD movement and pushes automated trading signals.

Problem solved. Converts noisy market data into actionable, confidence-scored directional signals delivered automatically.

My role. ML engineer — data pipeline, modeling, API, monitoring, automation.

Key features. Ensemble of an LSTM price model and an XGBoost direction classifier; MetaTrader 5 live data; technical-indicator feature engineering (RSI, MACD, Bollinger, ATR); authenticated FastAPI; confidence-thresholded signal tiers; n8n → Telegram alerts every 5 min; **Prometheus + Grafana monitoring.**

ARCHITECTURE — ML SERVING + AUTOMATION + OBSERVABILITY



Python · PyTorch (LSTM) · XGBoost · FastAPI · pandas-ta · MT5 · n8n · Prometheus · Grafana · Docker

↗ github.com/Aleco127/gold-predictor-api

Demonstrates: MLOps end-to-end, ensemble modeling, time-series, API-first ML, automation integration.

Control de Obra

A 25-module construction-management platform in daily production use by an active construction firm.

Problem solved. Construction firms juggle budgets, expenses, estimates, schedules, RFIs, payroll and accounting across disconnected tools. Control de Obra unifies 25+ modules and runs the day-to-day of **Supernova Arquitectos**, an active construction company (its first tenant), on real budgets, estimates and payroll.

My role. Sole architect & engineer; also the SaaS platform layer (tenants, subscription plans, admin, error monitoring).

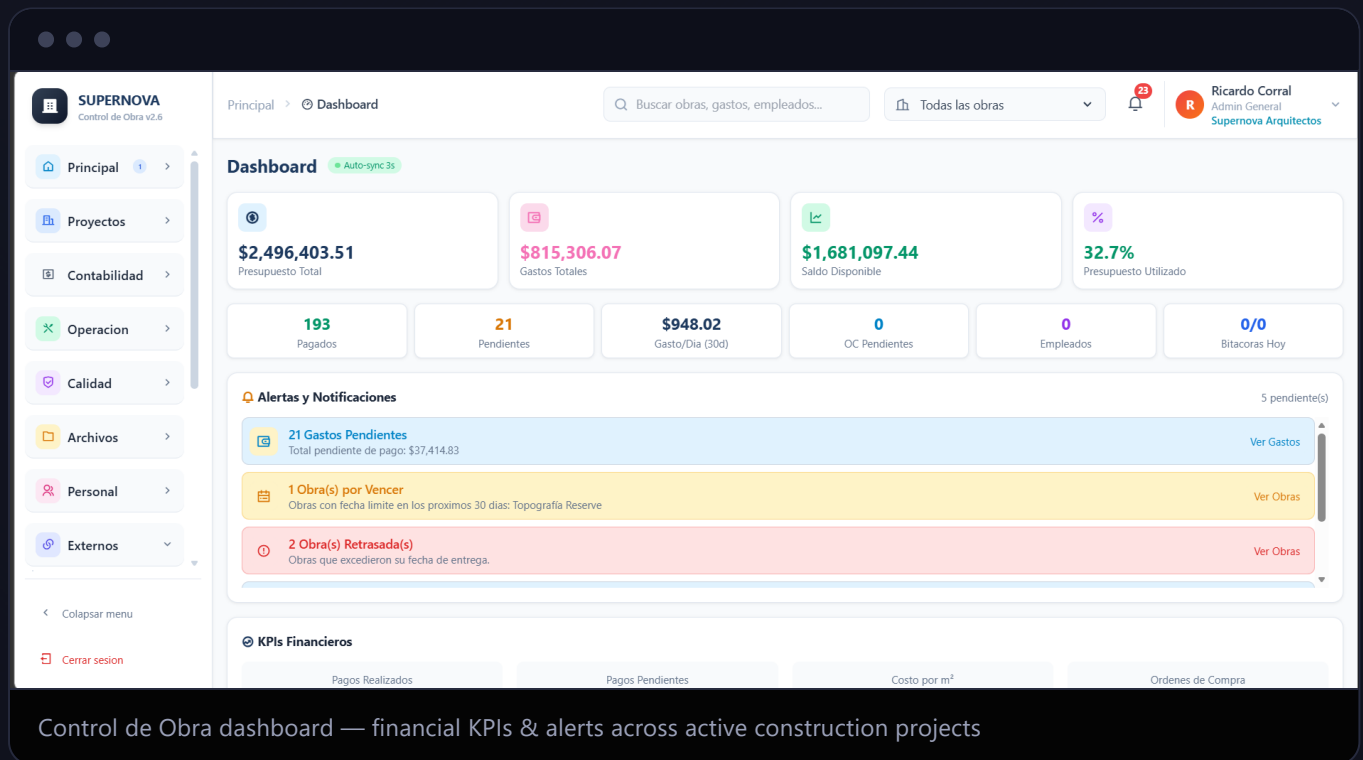
Key features. 25+ operational modules (budget, Gantt, expenses, POs, estimates, RFIs, punch list, accounting/CFDI, payroll); custom bcrypt auth; multi-tenant isolation; subscription tiers; centralized client-side error monitoring with fingerprinting; PDF/Excel export.

Architecture. ~20k-line Vanilla JS SPA + TailwindCSS; Supabase (PostgreSQL + RPC) with custom auth/session model; separate SaaS admin console; Dockerized on VPS behind Caddy.

Vanilla JS SPA · TailwindCSS · Supabase (PostgreSQL + RPC) · Chart.js · jsPDF/XLSX · Docker · Caddy

➔ [Live: obra.srv1090924.hstgr.cloud](https://obra.srv1090924.hstgr.cloud) ➔ github.com/Aleco127/control-obra-dashboard

Demonstrates: Large-scale full-stack ownership, multi-tenant SaaS architecture, data modeling, observability — the engineering maturity AI features must live inside.



A – Zook: Running an LLM Sales Agent in Real Production

CHALLENGE

Real customers on WhatsApp message in bursts, expect human-sounding replies, and go cold without follow-up. A naive one-request-per-message bot double-replies and loses context — and any in-process state breaks the moment the backend runs multiple workers.

SOLUTION

A production agent loop: incoming messages are grouped in an 8-second inbox buffer persisted in SQLite (shared across 4 uvicorn workers), answered with a single Claude call grounded in a curated knowledge base, and split into natural message bubbles at sentence boundaries. The agent emits structured stage/score tags that drive lead scoring, hot-lead email alerts and automated 24-hour follow-up touches.

ARCHITECTURE & IMPLEMENTATION

FastAPI webhooks for Twilio WhatsApp Business; prompt and knowledge documents versioned in Supabase; conversation state and buffers in SQLite (no process-local state); n8n powers the web-chat channel and a shared webhook routes FB/IG DMs. Meta click-to-WhatsApp ads land directly on the agent. Containerized behind Caddy, live at zook.mx.

RESULTS / VALUE

A paying acquisition channel runs end-to-end without a human first touch: ad → conversation → qualification → follow-up, with hand-off when a human is needed. The engineering lessons (buffering, multi-worker state, reply shaping, stage control) are the difference between an LLM demo and an agent customers actually talk to.

B – Licitagen: Multimodal Document Intelligence Under Real Constraints

CHALLENGE

Government tender PDFs are inconsistent — some digital, many scanned, all dense with binding requirements. A missed requirement or mispriced concept can disqualify a multi-million-peso bid. Manual assembly takes ~5 days.

SOLUTION

A degrading-confidence extraction pipeline: cheap digital text first, escalate to OCR, finally fall back to Claude vision on rendered pages — always emitting one strict JSON schema. A historical-price memory with embeddings enables semantic retrieval of comparable concepts and feeds an XGBoost baseline.

ARCHITECTURE & IMPLEMENTATION

Node/Express orchestrator + SQLCipher-encrypted SQLite $\rightleftharpoons$ a Python toolkit (pdfplumber, PyMuPDF, pytesseract, Claude, openpyxl, XGBoost) $\rightleftharpoons$ an OPUS construction-ERP automation bridge generating APUs, indirect costs, financial factor and a critical-path schedule. Security is first-class: per-machine licensing, encryption at rest, RBAC, anti-tampering.

RESULTS / VALUE

Designed to cut a ~5-day expert process to ≤ 1 day, catch compliance errors pre-submission, and accumulate a proprietary pricing dataset that improves estimates over time.

C – WhisperWind: Secure Multi-Tenant Conversational Automation

CHALLENGE

Automating WhatsApp at scale while keeping human oversight, isolating multiple businesses' data, and exposing a usable real-time operator interface.

SOLUTION

A multi-tenant platform pairing a Meta WhatsApp Cloud API webhook with n8n automation, and a React/TypeScript operator dashboard for live conversation, tagging and hand-off — on a 12-table PostgreSQL schema with non-recursive Row Level Security.

ARCHITECTURE & IMPLEMENTATION

React 18 + Vite + shadcn/ui frontend; Supabase (PostgreSQL + Auth + RLS) data layer; Express webhook ingest; n8n automation brain; Dockerized on a VPS behind Caddy with role-based access and an admin panel.

RESULTS / VALUE

Conversational automation with a human safety net and tenant isolation — the architecture pattern customer-facing AI products need to be trusted in production.

PROJECT	TYPE	AI / ML COMPONENT	MAIN STACK	DEPLOYMENT	BUSINESS VALUE
Zook	Production GenAI agent	Claude sales agent — RAG grounding, stage/score tags, follow-ups	FastAPI · Claude · Twilio · Docker	Live (zook.mx)	24/7 sales agent answering real customers
LicitaGen	AI document-intelligence	Claude vision extraction, embeddings/RAG, XGBoost	Node · Python · SQLCipher · OPUS	Internal daily use (desktop)	~5 days → ≤1 day bid assembly (target)
WhisperWind	AI automation / full-stack	LLM/n8n conversational automation	React/TS · Supabase · WhatsApp · n8n	Live (VPS)	Automated WhatsApp w/ human oversight
Gold Predictor	ML service + MLOps	LSTM + XGBoost ensemble	Python · PyTorch · FastAPI · n8n	Dockerized (VPS internal)	Automated, monitored ML signals
Control de Obra	Full-stack SaaS	Operational data platform (AI-ready)	Vanilla JS · Supabase · Docker	Live (VPS)	Runs an active construction firm daily

Supporting work: [fgconnect-whatsapp-bot](#) (public sanitized extract of a production WhatsApp sales bot for a US ISP retailer — debounced inbound buffer, multimodal ingest, live-editable prompt, funnel tracking, human hand-off), [fuel-route-api](#) (public Django REST API — route + cost-optimized fuel stops, single OSRM call, offline geocoding, 16 tests), [chatbot-rag-llm-cheatsheets](#) (public RAG chatbot — tabular extraction, FAISS, multilingual embeddings), VecinosPro (HOA SaaS, OpenPay), Inventario MDV (inventory SaaS), 20+ n8n automation workflows, FGCONNECT lead-gen, and graduate deep-learning work (a 620K-parameter CNN trained on CPU reaching 78.7% on CIFAR-10 with 2.5× fewer parameters than an FC baseline; NumPy-only backprop networks).

06 WHY THIS PORTFOLIO IS RELEVANT FOR GENAI / AI AUTOMATION ROLES

- ▶ **GenAI in production, not demos.** Zook's Claude agent answers real customers on WhatsApp every day; LicitaGen uses multimodal LLM extraction with a disciplined fallback strategy and strict output schemas — the core of applied GenAI engineering.
- ▶ **RAG & retrieval.** Embedding-based semantic retrieval over a historical price knowledge base (LicitaGen), plus curated-knowledge grounding that keeps Zook's sales agent on-script.
- ▶ **Automation at the center.** 20+ n8n workflows, WhatsApp/Telegram/CRM integrations, scheduled pipelines — exactly the AI-Automation-Engineer mandate.
- ▶ **ML depth.** Ensemble modeling, time-series, XGBoost, CNNs, and from-scratch backprop — I understand the models I deploy, not just the APIs.
- ▶ **Production discipline.** Containerized, deployed, secured (RLS, SQLCIPHER, RBAC), observable (Prometheus/Grafana, error monitoring). I ship and operate.
- ▶ **Domain leverage.** An architecture/construction background lets me extract real ROI from AI in underserved verticals.

I build AI systems that run in production, integrate with real business workflows, and create measurable operational value — and I own them end to end, from model to deployment.

07 BACKGROUND & AVAILABILITY

2017–2021 — B.Arch, Tecnológico de Monterrey (professional license 12957387).

2019–2023 — Construction & engineering roles (Bioecosfera, Grupo México, Axis Arquitectura, Quinven): site engineering, QA/QC and BIM coordination on projects up to **\$300M MXN** — the operational domain my AI work now automates.

2024–present — Founded Supernova Arquitectos (construction) and built its internal tooling — Control de Obra and LicitaGen — to run real operations, which is where my production engineering discipline comes from.

2025–present — Master's in Applied AI, Tecnológico de Monterrey; designed, shipped and operate Zook's production WhatsApp sales agent.

Availability: I'm looking to join an AI engineering team **full-time**. Supernova's operation is handled by my business partner (Ing. Daniel Loera), and Zook remains a nights-and-weekends side project. I worked independently because I had real operational problems worth automating — now I want to build with a team, at scale, with code review and production standards beyond my own.

