

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), GPT-4, multi-agent orchestration, prompt engineering, prompt caching, vision-assisted extraction, RAG, embeddings / semantic search, LangChain
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), Make/Zapier, 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, Hostinger VPS, Caddy reverse proxy, AWS (SageMaker, EC2), GCP (Vertex AI, Colab), Prometheus + Grafana, Git/GitHub
Frontend / Dashboards	React 18 + TypeScript + Vite + shadcn/ui, Vanilla JS SPA, TailwindCSS, Chart.js, Streamlit

Zook

A platform where Claude-powered AI agents sell, automate, and analyze for SMBs 24/7.

Problem solved. SMBs can't afford full sales/ops/analytics teams. Zook deploys configurable AI agents (Alex – sales, Nova – operations, Vega – analytics) that work continuously across web and WhatsApp.

My role. Sole architect & engineer — agent pipeline, backend, frontend, deployment.

Key features.

- A **sequential 4-agent LLM pipeline** (Strategy → Design → Content → Builder) orchestrated through the Claude API.
- **Real-time streaming** of agent reasoning to the UI via Server-Sent Events.
- Project persistence, one-command export, interactive problem-selector that recommends the right agent mix.

Architecture. FastAPI (Python 3.12) orchestrator + SQLite; each agent a discrete module returning structured JSON; SSE event bus to a Vanilla JS SPA; Dockerized behind Caddy on a VPS.

ARCHITECTURE — MULTI-AGENT LLM PIPELINE



FastAPI · Claude API · SSE · Vanilla JS · Tailwind · Docker · Caddy

↗ [Live: zook.mx](https://zook.mx) ↗ github.com/Aleco127/ai-web-builder

Demonstrates: Multi-agent design, LLM orchestration, streaming UX, end-to-end product ownership and deployment.



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

Real capture — Zook production landing at zook.mx (AI agents platform)

LicitaGen

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

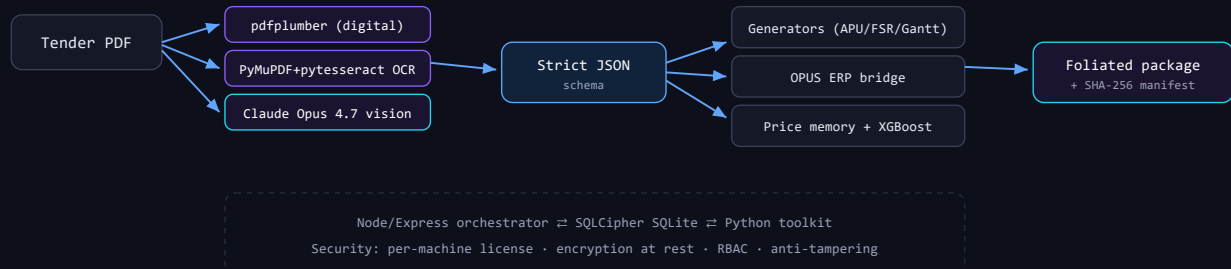
Problem solved. Assembling a public-works bid is a ~5-day manual marathon of reading hundreds of PDF pages, extracting requirements and pricing, and producing dozens of compliant annexes. LicitaGen targets ≤ 1 day.

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 Opus 4.7 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 Opus 4.7 (vision+extraction) · pdfplumber · pytesseract
 · SQLCipher · XGBoost · OPUS LocalDB

Status: v1 deployed locally (extraction + generators + OPUS bridge in production use); 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.

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.

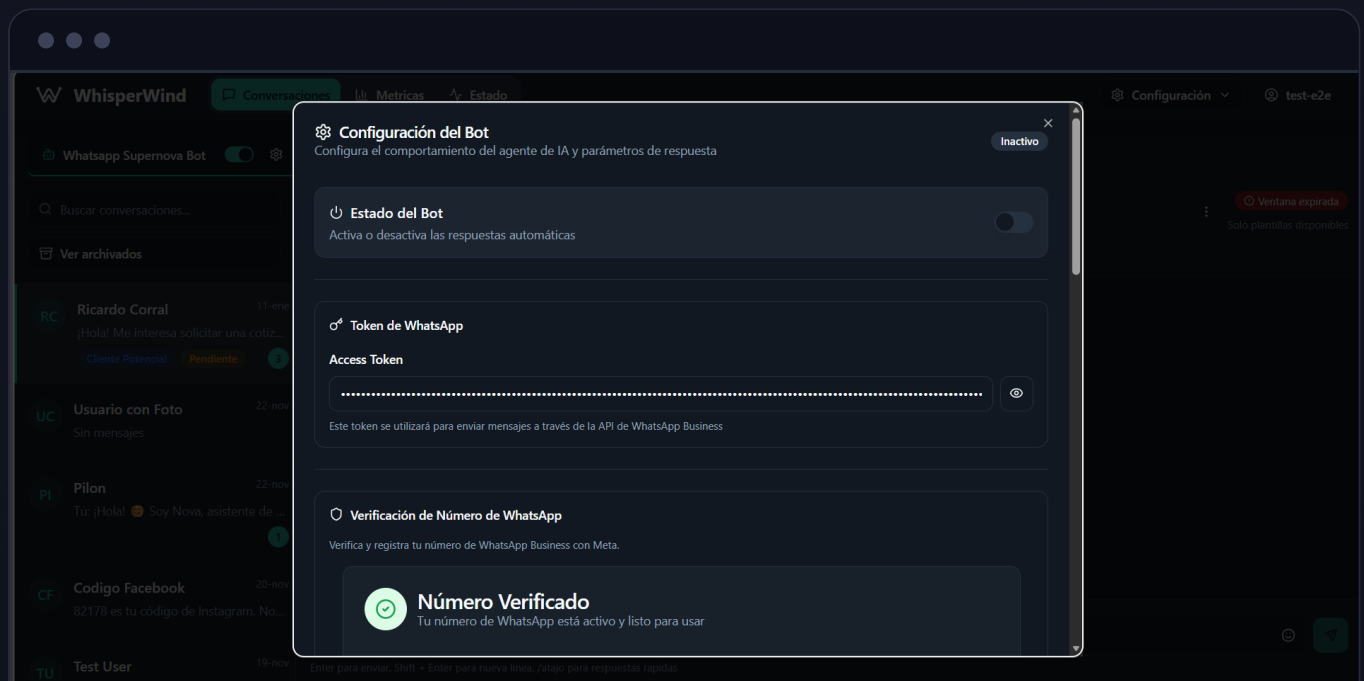
Key features. AI-agent behavior configuration panel (response parameters, bot enable/disable); role-based auth + admin panel; multi-bot management; real-time conversation inbox with tags/notes/quick-replies; Meta WhatsApp Cloud API webhook; n8n automated responses; **12-table PostgreSQL schema with non-recursive 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 ➤ github.com/Aleco127/whisperwind-dashboard

Demonstrates: Production full-stack AI automation, secure multi-tenant data modeling, third-party API integration, workflow orchestration.



Real capture — WhisperWind production app: AI-agent behavior configuration & live conversation inbox (API token masked by the app)

Gold Price Predictor

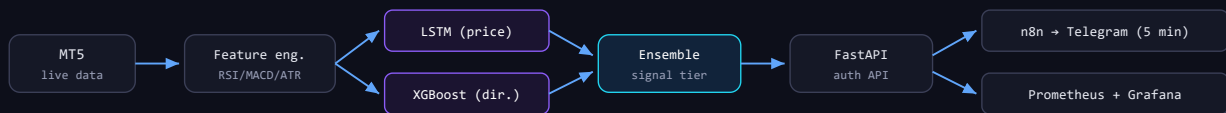
LSTM + XGBoost ensemble that predicts XAUUSD movement and pushes automated trading signals.

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

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

Control de Obra

Multi-tenant SaaS managing the full lifecycle of construction projects — in real production use.

Problem solved. Construction firms juggle budgets, expenses, estimates, schedules, RFIs, payroll and accounting across disconnected tools. Control de Obra unifies 25+ modules — used to manage **\$5M+ in real construction projects**.

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

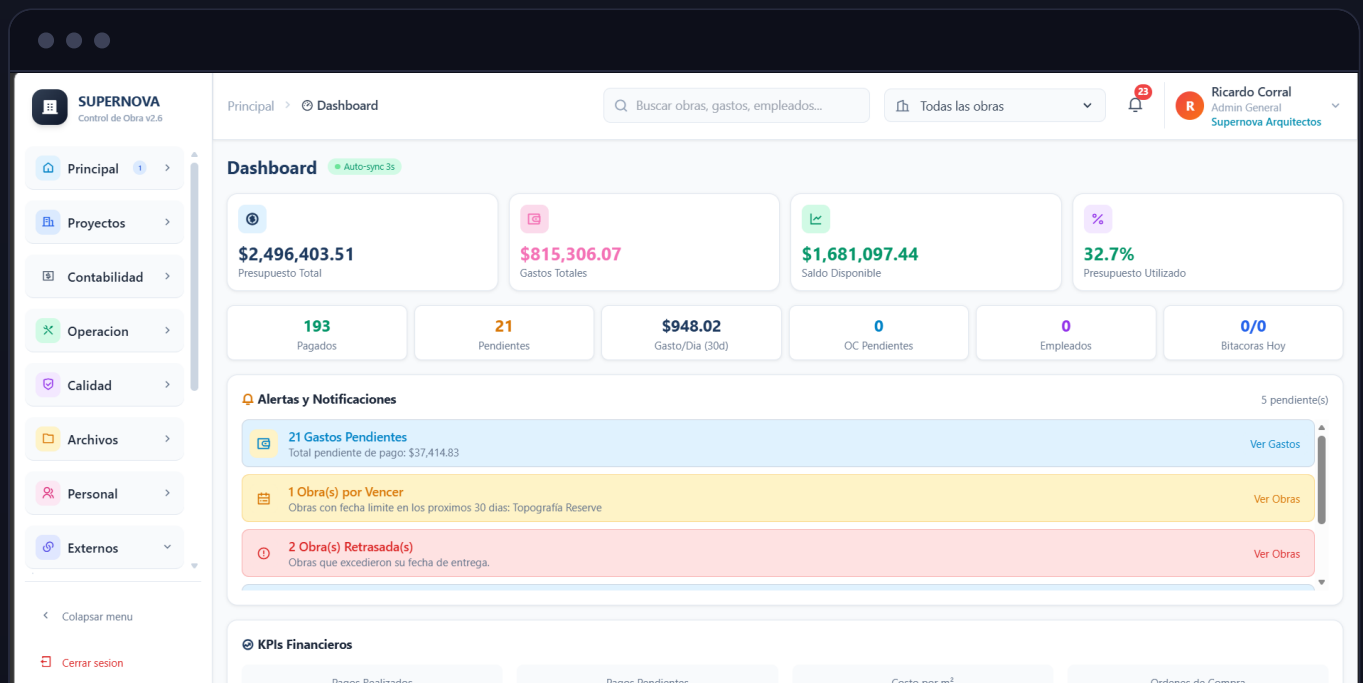
Key features. 25+ operational modules (budget, Gantt, expenses, POs, estimates, RFIs, punch list, accounting/CFDI, payroll); custom bcrypt auth; multi-tenant isolation; subscription tiers; centralized client-side error monitoring with fingerprinting; PDF/Excel export.

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.



Real capture — Control de Obra production dashboard: live financial KPIs & alerts across active construction projects

A – Zook: Orchestrating a Multi-Agent LLM Pipeline

CHALLENGE

A single LLM prompt can't reliably produce a coherent, multi-faceted business deliverable. Work must be decomposed, each step specialized, and the user kept informed while a long generation runs.

SOLUTION

A sequential multi-agent architecture where four specialized Claude agents each own one concern — Strategy, Design, Content, Builder — each consuming the previous agent's structured JSON output.

ARCHITECTURE & IMPLEMENTATION

A FastAPI orchestrator coordinates the pipeline and emits Server-Sent Events (`{step, status, message}`) so the frontend streams live progress instead of blocking. Agents are isolated modules with strict JSON contracts — debuggable and extensible. State persists in SQLite; containerized behind Caddy, live at zook.mx.

RESULTS / VALUE

A reproducible LLM workflow productized into an SMB-facing platform — prompt engineering turned into a maintainable, streaming, deployed system.

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 Opus 4.7 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	Multi-agent GenAI platform	Claude 4-agent orchestration, SSE streaming	FastAPI · Claude API · JS · Docker	Live (zook.mx)	Productized AI workforce for SMBs
LicitaGen	AI document-intelligence	Claude Opus vision extraction, embeddings/RAG, XGBoost	Node · Python · SQLCIPHER · OPUS	Production (local/hardened)	~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)	Manages \$5M+ real projects

Supporting work: VecinosPro (HOA SaaS, OpenPay), Inventario MDV (inventory SaaS), 20+ n8n automation workflows, FGCONNECT lead-gen, and graduate deep-learning work (CNNs on CIFAR-10 reaching 78.7% test accuracy with 2.5× fewer parameters than an FC baseline; NumPy-only backprop networks).

- ▶ **GenAI in production, not demos.** Zook is a live multi-agent Claude pipeline with streaming UX; 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 knowledge base (LicitaGen), plus RAG for context-aware support responses.
- ▶ **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.

Built from a live workspace audit · Screenshots are real captures of deployed production systems · No credentials or secrets are exposed
— API tokens appear masked by the application.