

RICARDO ALBERTO VEGA JR

AI Engineer · Full-Stack Architect · SaaS Founder

Northlake, TX (Dallas–Fort Worth) • (817) 601-7349 • rvegajr@noctusoft.com • ricardovega.dev • linkedin.com/in/rvegajr • github.com/rvegajr • US Citizen

PROFESSIONAL SUMMARY

34 years shipping production software across every platform wave — mainframe COBOL to multi-model AI pipelines. Operate a private **28-model AI relay** (LiteLLM + local Ollama on M4 Max, Azure VM, Tailscale VPN). Built a **7-component Gemini AI pipeline** converting live fire-dispatch audio to a scored, enriched lead in 8 seconds. Published **Model Context Protocol (MCP) servers** used daily by Claude Desktop and Cursor. Daily practitioner of **Claude Code and AI-assisted development** — including authoring the safety tooling others use to adopt it. Deep regulated-domain record: HIPAA-compliant clinical platform (100+ releases), statewide education-agency modernization, Fortune 500 energy and aviation. **Founder, Noctusoft, Inc.** (Texas S-Corp, 2011–present).

CORE TECHNICAL SKILLS

AI / ML: Gemini 2.5 Flash (JSON Schema structured outputs), GPT-4o / Whisper (real-time audio), Anthropic Claude, Ollama (local LLMs), LiteLLM multi-provider routing, MCP server architecture (published), RAG / LangChain, prompt & response evaluation, LLM observability, agentic tool-calling, AI governance controls

Languages & Data: C#/.NET (20+ yrs), TypeScript / Node.js, Python / FastAPI, Go, PHP, C++; SQL Server / T-SQL, PostgreSQL, MongoDB, CosmosDB, MySQL, Oracle PL-SQL, Redis

Cloud / DevOps / Frontend: Azure, AWS, Railway, Vercel, Docker, GitHub Actions, Fastlane, Turborepo; Next.js 15, React / React Native, Blazor, Angular, Swift / SwiftUI (iOS); Playwright E2E

PRODUCTION AI SYSTEMS (SELECTED)

- **Ember Lead — 7-Component Gemini Pipeline** — Whisper consumes live fire-scanner audio; Gemini 2.5 Flash extracts structured data with JSON Schema constraints; two-stage address verification gates paid enrichment; waterfall property enrichment (ATTOM → RentCast → Zillow → county scrapers) with AI fallback parsers; phone skip-trace waterfall; OccupantClassifier + LeadScorer. *Dispatch to scored lead in 8 seconds.*
- **Private 28-Model AI Relay** (api.noctusoft.com) — LiteLLM on Azure VM + Ollama on M4 Max, Tailscale VPN; 9 OpenAI + 9 Anthropic + 4 Gemini + 6 local models; per-tenant rate limiting, upstream keys server-side only
- **Browser MCP** — 33-tool MCP server for agentic tool-calling over live browser state; used daily by Claude Desktop and Cursor
- **LLM Gateway / Enterprise Shield** (Go) — model-routing proxy with sensitive-data masking; 24-test regression suite for prompt/response evaluation and guardrails in production
- **EZ-OData-API** (C#/.NET 10) — governed data-access platform with MCP endpoint; RBAC, rate limiting, audit logging, Prometheus metrics
- **StuPath AI Tutor** (Python/LangChain) — production RAG backend; 99.9% uptime multi-tenant platform with Playwright evaluation coverage
- **AI Financial Advisor** — Plaid API → AI analysis → automated daily S-Corp reports via SMS/email; no manual intervention

LIVE SAAS PRODUCTS

Ember Lead / Tankroom — Real-time fire-damage lead gen for restoration contractors (Gemini AI, Whisper, Next.js, React Native, MongoDB, Railway) • **Scholarmancy** — EdTech SaaS, live on Railway • **FieldView.Live** — Live streaming commerce SaaS • **YOLOVibeCode AI** — AI dev training workshops & consulting • **CaptionFlow** — EdTech accessibility tooling • **Noctusoft API Relay** — Multi-tenant relay for 28 AI models + SendGrid / Twilio / Square

PROFESSIONAL EXPERIENCE

Founder & Principal Solutions Architect — Noctusoft, Inc.

January 2011 – Present | Dallas–Fort Worth, TX • Enterprise consulting & SaaS products for 20+ clients: healthcare, energy, financial services, aviation, government

Texas Education Agency — Statewide Systems Modernization (Apr–Jul 2026) — C#, .NET 4.8 → .NET 10, SQL Server, Playwright, xUnit, GitHub Copilot / Claude Code

- Modernized 8 statewide systems; SQL dashboard query 4,254 ms → 550 ms (87% faster); list pages up to 6× faster
- Agency-wide reliability audit (~20 business classes); request-scoped SQL profiler and crash-interception diagnostics adopted portfolio-wide
- 127+ Playwright E2E tests across 4 applications; governed AI-assisted engineering practice with usage and cost tracking

Christensen Farms (2024–Present) — Azure VM, MobileFrame, Dynamics SL, SSIS, PowerBI

- MobileFrame app saves 15 hrs/week; Dynamics SL automation saves 10 hrs/week; unblocked on-premise → Azure migration; real-time PowerBI farm analytics

PnQ Health — HIPAA-Compliant Clinical Platform (2024–Present) — Swift/SwiftUI, TypeScript, Node.js, Azure Service Bus, MongoDB, Fastlane

- Architected iOS + TypeScript API + React dashboard with near-real-time Azure Service Bus sync, entity-level HIPAA encryption, EN/FR/ES internationalization
- 100+ TestFlight deployments, zero production incidents; Fastlane CI/CD cut deployment time 70–85%

StuPath — RubriqFlow (2024–Present) — Next.js 15, TypeScript, Python AI Tutor, LibSQL, Vercel, Playwright

- Full-stack multi-tenant education platform live at rubriqflow.com (99.9% uptime); RBAC across 4 roles; AI tutoring backend; comprehensive Playwright E2E

SJI — Flight Deck Pro (2024) — TypeScript, Turborepo, Azure DevOps, Docker, Next.js

- Resolved critical Azure pipeline failures; fixed TypeScript builds across 15+ packages; Docker–Azure parity system cut iteration 70–85%; 15+ PRs in one month restoring stability

Choctaw Nation — Tribal Benefits Administration (2022–2024) — .NET 6, Blazor, EF Core, SQL Server, CosmosDB

- Led 3-dev team; generic Blazor components +60% productivity; generics/DI refactor –70% code duplication; SQL optimization –80% processing time; Web Forms → Blazor .NET 6 modernization

Pioneer Natural Resources — Enterprise Integration (2016–2022) — C#, Angular, OData, SQL Server, Oracle, Azure, NuGet SDKs

- Parsley acquisition data migration 50% ahead of schedule; Unifier Conduit & P6 Conduit NuGet SDKs used company-wide; T4/RazorEngine code generator +40% developer efficiency; ArcGIS query tuning 90% faster

Earlier Engagements (2011–2016): *AAFES (Federal, 1,200+ locations) • American Airlines (10M+ user Facebook app; zero-downtime careers site on AWS) • T-Systems Healthcare (ER software, 100+ hospitals; 1,000× managed cache improvement) • Match.com / PeopleMedia • Live Energy • Summit Investment & Financial Services*

Principal Consultant — Vega And Associates, Ltd. | Programmer Analyst — Rapp Collins | Programmer — Lomas Information Systems

1992–2011 | Keller / Irving / Dallas, Texas

15 years independent consulting across financial services, banking, and mortgage (TD BankNorth, Bayview Funding, Interbay Funding, Amerifunds). Bell Canada award system redesign on MVS/ESA — 25% development time reduction. COBOL II loan-acquisition conversion on OS/MVS/ESA — 80% subsystem efficiency improvement, 60% run-time reduction via Assembler performance tuning.

EDUCATION

University of Texas at Arlington — Computer Science Engineering (1993–1995) • Northlake College — Information Technology, Irving, TX