Alberto Rosas

AI & Agentic Systems Engineer
alberto.rsesc@protonmail.com | linkedin.com/in/albertorsesc | github.com/albertorsesc | alberto-rosas.dev
Mexico | English (Native) • Spanish (Native)

SUMMARY

AI & Agentic Systems Engineer with 12+ years in software engineering and 5+ years building production AI systems — multi-agent orchestration, RAG pipelines, NL2SQL engines, and LLM automation using LangGraph and MCP, with applied ML experience in fine-tuning, embedding model selection, and evaluation methodology. Career built through engineering management, software architecture, and technical leadership, with a cybersecurity automation background that brings security-first thinking to every AI system. Currently deploying agentic AI platforms across compliance, property management, and business formation SaaS.

TECHNICAL SKILLS

LLMs, Models & Applied ML: GPT-4o, Claude (Anthropic), Gemini, Llama 3, Mistral, Qwen3, ModernBERT; fine-tuning (Unsloth, HuggingFace Transformers); PyTorch, scikit-learn (applied); embedding model selection and benchmarking; MLflow experiment tracking
Agentic Systems: LangGraph, LangChain, MCP/tool-calling, multi-agent orchestration, multi-step reasoning, intent classification, function calling, memory systems (short/long-term, procedural)
RAG & Retrieval: Vector search (Qdrant, MongoDB Atlas, FAISS, Chroma), knowledge graphs (Neo4j), GraphRAG, hybrid retrieval (dense + sparse + graph), NL2SQL, embeddings, context engineering
Evaluation & LLMOps: RAGAS (context recall/precision, faithfulness, answer relevancy), hallucination detection, latency benchmarking, user satisfaction scoring, golden dataset evaluation, Langfuse, Opik, Langsmith, MLflow, prompt versioning, CI/CD-integrated eval pipelines
Engineering: Python, TypeScript, PHP/Laravel, FastAPI, Docker, Kubernetes, AWS (Bedrock, SageMaker, EC2, S3, Lambda), CI/CD, Clean Architecture, microservices, event-driven systems

KEY PROJECTS

- **UCF AI Engine:** Sole architect of the AI layer for a compliance SaaS — hybrid retrieval (Qdrant + Neo4j) over 91K+ regulatory records, LangGraph agentic chat, CRAG hallucination verification. Deployed on Docker/AWS.
- **SADIE:** Production NL2SQL agentic platform for property management — LangGraph + MCP tool-calling, MongoDB Atlas RAG pipeline, 93–97% query accuracy, Langfuse observability. Clean Architecture, FastAPI SSE.
- **TriageOps Framework:** 6-step AI adoption methodology for mid-market companies. Discovery Sprints recovered 70+ hrs/month ($42K annually) for one operations team.

PROFESSIONAL EXPERIENCE

IncFile — AI Engineer / Software Engineer / Technical Project Lead *2022 – Present*
Leading AI strategy and platform architecture at a business formation SaaS, driving the transition from legacy systems to AI-powered automation.

- Designed multi-agent system for document processing and decision support, handling 1,000+ daily requests across multiple data sources
- Built RAG pipeline with vector search and business documentation; experimented with GraphRAG/KAG for improved precision
- Prepared training datasets and ran fine-tuning experiments (Unsloth, HuggingFace Transformers) to adapt LLMs for domain-specific business context
- Implemented multi-layer memory systems (short/long-term, procedural) for cross-session context retention
- Reduced manual data entry by 65% through intelligent form processing and validation
- Led platform migration from legacy monolith to service-oriented stack; mentored 8 engineers on AI/ML practices and production deployment

Unified Compliance — Senior AI Engineer (Contract) *2025 – 2026*
Sole architect of the entire AI Engine for UCF's ControlSight compliance platform — 91K+ regulatory records.

- Built EEL pipeline: PostgreSQL → benchmarked and selected ModernBERT 768d + BM25 embeddings → dual-load into Qdrant and Neo4j with incremental sync
- Designed hybrid retrieval: Qdrant vector+BM25 search with Neo4j graph traversal across compliance hierarchies, plus CRAG grading
- Built LangGraph agentic chat workflow with intent classification, multi-turn context, query rewriting, and confidence tiers
- Modeled 8 entity types and 10+ relationships in Neo4j with Cypher queries, fulltext indexes, and cross-store validation
- Evaluated using RAGAS (context recall/precision, faithfulness, answer relevancy), latency tracking, and user satisfaction scoring
- Directed company AI strategy, aligning AI capabilities with product, GTM, and data privacy requirements

Storage360 — AI Engineer (Contract) *2024 – 2025*

Built SADIE — a production NL2SQL agentic platform for a property management SaaS.

- Built agentic workflow using LangGraph with MCP tool-calling, multi-step reasoning, and context retrieval
- Implemented RAG pipeline with MongoDB Atlas vector search, benchmarked and selected Qwen3 embeddings, and structured context library of business rules and metrics
- Achieved 93–97% accuracy via RAGAS evaluation, golden datasets, and CI/CD-compatible regression testing; prepared historical query data for fine-tuning
- Built MCP server for secure read-only database access with schema caching and FK detection
- Deployed via Docker Compose with Langfuse observability, health monitoring, and FastAPI SSE streaming

Global Cybersec — Engineering Manager *2017 – 2021*

Led engineering for a cybersecurity firm building security automation and incident response platforms.

- Designed event-driven architecture processing millions of daily security events for automated pattern recognition
- Integrated SIEM, IDS/IPS, firewalls, and SOAR for automated threat response — 60% reduction in incident response time
- Built 4 microservice applications and managed a team of 5 engineers across all projects

Multiple Companies — Software Engineer *2014 – 2017*

Full-stack roles across logistics, healthcare, and proptech.

- Built logistics, healthcare, and proptech platforms from the ground up with full-stack development, API design, and security practices

EDUCATION & CERTIFICATIONS

Universidad Politécnica de Baja California — Information Technology Engineering (2014–2016)

Laravel Certified Developer (2020)

Certifications: The Complete LangChain & LLMs Guide • Business Process Modeling with AI • Generative AI & LLMs: Architecture and Data Preparation • Gen AI Foundational Models for NLP & Language Understanding

ADDITIONAL

Technical Writing: Active blog on AI/ML engineering, agentic system design, and production AI architecture

AI Consulting: Co-founded TriageOps — AI adoption consultancy with a 6-step deployment methodology for mid-market companies