Ricardo Ledan

Senior AI Engineer · Full-Stack Architect · New York, NY

ricardoledan@proton.me · github.com/ricoledan · linkedin.com/in/ricardoledan

SUMMARY

Systems thinker and AI engineer who ships with business outcomes in mind. 10+ years across business intelligence, software engineering, and applied AI — bridging data, infrastructure, and business needs end to end.

EXPERIENCE

Consultant, Applied AI & Engineering | Deloitte Consulting — Jan 2021 – Present

Multi-Agent Intelligence Report System — AI Engineering Lead 2025

- Built a new AI capability within Deloitte's internal AI startup that landed a federal intelligence agency as a client. Recruited to R&D for a unique blend of fullstack engineering and AI/ML skills. Designed and pitched a multi-agent knowledge base for consuming SARs reports, automating analysis that previously required multiple analysts over days.
- Owned end-to-end system architecture for a production agentic AI platform serving a federal intelligence agency — from inference pipeline design and vector store configuration to deployment on AWS GovCloud. Built the foundational application template and led a 3-person team (data scientist + 2 AI engineers) from zero to production.
- Engineered a GraphRAG pipeline that discovered 67% more entities than traditional RAG through multi-hop graph traversal. Four-stage architecture: document ingestion, YAML-driven knowledge extraction via GPT-4, Neo4j graph analytics (centrality, community detection, entity resolution), and hybrid retrieval combining vector search with graph traversal.

Credit Card Rules Automation — Process Automation Engineer 2025

- Removed the biggest bottleneck in a top-tier bank's credit card merchant onboarding pipeline, where manual regex rule creation was capped at 226 rules/month. Led the integration of LLM capability into the rules engine, projected to reduce manual effort by 70–85%.
- Designed and presented a multi-option technical roadmap to client executives, evaluating regex enhancement, supervised ML, LLM integration with human-in-the-loop feedback, and a long-term unified merchant identifier strategy. The recommended LLM approach was approved for implementation.
- Shipped the production-ready application in three weeks — rewrote the Python codebase to internal quality standards, added test coverage, built CLI tooling for the rules team, and packaged the application for local distribution across team machines. Coordinated delivery across onshore and offshore teams.

Federal & Blockchain Systems — DevOps / Blockchain Engineer 2023 – 2024

- Contributed to a federal WMD detection system on AWS GovCloud, building infrastructure across a classified environment. Designed CI/CD pipelines with end-to-end encryption, RBAC, and NLP processing workflows using AWS Comprehend.
- Built TrustGen, a blockchain proof of concept for data provenance and access control that gained sponsorship from a Deloitte partner. Implemented on-chain dataset tracking and secured data retrieval on Hyperledger Besu.

Cloud-Agnostic Analytics Platform — Lead Services Engineer 2021 – 2024

Led infrastructure engineering for a multi-cloud analytics platform (AWS EKS, Azure, GCP) — designing Kubernetes-based deployment pipelines, Terraform IaC modules, and CI/CD standards adopted org-wide. Mentored 3 engineers and established trunk-based development and automated test coverage requirements.

Software Engineer | Red Ventures `Apr 2019 – Aug 2020`

- Optimized high-traffic React/TypeScript components serving 2M+ monthly users — driving a 15% conversion lift through systematic A/B testing and load performance improvements on Bankrate.com's rate comparison tables.
- Maintained 99.9% uptime on a distributed advertising platform handling 500K+ API requests/day — stack included GraphQL, Kafka, Golang microservices, and PostgreSQL.
- Reduced mean time to incident detection, enabling the team to resolve production issues before they affected users. Instrumented platform services with Datadog for proactive observability across all critical paths.

Business Intelligence Developer | Automotive Management Services Inc `Sep 2014 – Apr 2019`

- Delivered actionable insights that supported account executives and dealers across national markets in matching purchases to offers. Built reports, dashboards, and automations that turned messy CRM data into decision-ready intelligence.
- Wrangled and cleaned data pipelines from Reynolds and Reynolds and other large CRMs, resolving data quality issues that were undermining reporting accuracy. Transformed unreliable source data into trusted datasets the business could act on.

SELECTED PROJECTS

Rasin.ai — Self-hosted AI research platform

- Architected and operated a full production LLM inference stack on NVIDIA GB10 (CUDA 12.x) — running TensorRT-LLM with GPU-accelerated OCR, vector retrieval (Qdrant + BGE-M3 embeddings), and GraphRAG (Neo4j) across 228K+ pages at $50/month total infrastructure cost.

Harombe — Open-source distributed agent framework

- Built Harombe, an open-source multi-agent orchestration framework with container-isolated agent execution (Docker), MCP Gateway for tool access control and capability sandboxing, and mDNS-based cluster coordination for cloud-agnostic deployment across heterogeneous hardware — including mixed CPU/GPU node configurations. Motivated by published research on infrastructure-level AI security (Zenodo, 2026).

RESEARCH

The Capability-Container Pattern: Infrastructure-Level Security for Autonomous AI Agents — `Preprint · Zenodo, 2026 · DOI: 10.5281/zenodo.18614503`

TECHNICAL SKILLS

LLM Inference	TensorRT-LLM, vLLM, SGLang, Ollama, AWS Bedrock, Self-Hosted Infrastructure (NVIDIA GB10, CUDA 12.x)
AI / ML	LangGraph, LangChain, RAG, GraphRAG, Multi-Agent Systems, LLM Fine-Tuning, RLHF, LangSmith, Langfuse, GLiNER
Languages	Python, TypeScript, JavaScript, Rust, Go, SQL
Data & Graph	Neo4j, Qdrant, ChromaDB, PostgreSQL, MongoDB, Redis, BGE-M3 Embeddings
Infrastructure	Docker, Kubernetes, Terraform, GitHub Actions, AWS (EKS, GovCloud, SageMaker)

CERTIFICATIONS

- IBM Qiskit Quantum Excellence (2025)
- NVIDIA AI Supercomputer Certified (2021)
- Deloitte: Advanced AI, AI Academy XP, ML Foundation (2025)