ALEJANDRO ANTONY DE LA CRUZ SUÁREZ

Research Engineer — LLM External Data, Agentic Systems & Information Architecture

Chiclayo, Peru (Available Remote / Open to Relocation) | alejantony.22@hotmail.com | linkedin.com/in/devlacruz | github.com/devlacruz |

ORCID: 0009-0009-7303-685X

Systems Engineer and AI Infrastructure Builder specializing in **rethinking information representation and retrieval architectures for frontier LLMs**. Creator of machine-native knowledge formats, including DAG-based cryptographic atomic blocks (CirIAco) designed to eliminate hallucination vectors at the data layer, and dual-ring agentic execution runtimes leveraging the Model Context Protocol (FastMCP) and Rust FFI. Proven track record deploying fault-tolerant distributed telemetry, pgvector indexing, and high-frequency inference gateways under sub-500ms budgets, with hands-on empirical ML research across deep learning pipelines.

CORE SYSTEMS & RESEARCH ARCHITECTURES

CirIAco — Novel Knowledge Architecture for Non-Human (LLM) Ingestion 2024 – Present

Lead Architect & Research Author • Cryptographic Knowledge Graph Engine Personal / Independent Research

- **Rethinking Human Document Paradigms:** Architected a machine-native knowledge representation paradigm replacing monolithic, human-centric text documents with cryptographically verifiable atomic information units (`Atmo` blocks) organized in a Directed Acyclic Graph (DAG).
- **Hallucination Defense at the Data Layer:** Designed a zero-trust provenance mechanism where each factual proposition is SHA-256 hashed and parent-linked, enabling language models to traverse verifiable ground-truth subgraphs, systematically eliminating reference drift and hallucinated citations.
- **Attribution & Royalty Trees:** Implemented recursive attribution traversal algorithms allowing inference engines to compute the exact informational lineage of multi-hop answers back to foundational truth nodes.
- **Knowledge Indexing & Traversal:** Built deterministic graph traversal operators and structured state serializations designed for fast context-window packing and low-overhead token consumption.

SumrAI — Zero-Trust Dual-Ring Agentic Gateway & FastMCP Execution Runtime 2024 – Present

Core Systems & AI Infrastructure Engineer Axtorch / SumrAI

- **Dual-Ring Agentic Architecture:** Designed and implemented a heterogeneous agent architecture separating autonomous, long-running ReAct reasoning loops (Ring 1: OpenFang/Rust) from sub-500ms token-streaming tool-execution pipelines (Ring 2: Django Ninja + Granian ASGI).
- **Model Context Protocol (FastMCP) Tooling:** Engineered typed synchronous and asynchronous FastMCP servers enabling frontier models to safely execute parameterized operations, query vector databases, and invoke system tools across HTTP SSE and JSON-RPC.
- **Zero-Trust Vault for LLM Reasoning:** Formulated an information security architecture where sensitive transactional and operational entities are replaced with opaque cryptographic tokens (Rust FFI AES-256-GCM), preventing raw PII leakage to model context windows while preserving reasoning coherence.
- **Vector Retrieval & Semantic Indexing:** Implemented hybrid retrieval pipelines over PostgreSQL 16 + `pgvector`, combining dense vector embeddings with metadata-filtered queries for context injection under strict token budgets.

Empirical ML Research: Deep Learning for High-Variance Signal Processing 2023 – 2024

Principal Investigator & ML Engineer Academic Research • UNPRG

- **Multimodal Brain MRI Lesion Segmentation:** Engineered an end-to-end computer vision pipeline using ResNet50 and ResUNet backbones, optimizing with custom Focal Tversky Loss to overcome severe class imbalance; achieved Dice Similarity Coefficient >0.88 and IoU >0.80.
- **Edge AI Robotics & Low-Latency Inference:** Developed edge vision inference pipelines deploying quantized EfficientNet models on resource-constrained hardware (Raspberry Pi 4B), sustaining real-time processing (<200ms latency) without cloud backhaul.
- **Systematic Research Methodology:** Conducted empirical PRISMA-compliant systematic reviews analyzing security and auditing frameworks across public cloud infrastructures (AWS/GCP).

PROFESSIONAL SYSTEMS EXPERIENCE

NTT DATA — Enterprise Telemetry, Distributed Systems & Database Optimization 2021 – 2024

Software / Systems Operations Engineer • Mission-Critical Financial Squads Lima / Chiclayo, Peru

- **Large-Scale Financial Microservices Observability (Client: BCP):** Managed 24/7 cross-squad telemetry and anomaly detection for high-volume banking payment rails; monitored distributed service meshes using Dynatrace, Kibana, and Grafana, resolving production latency regressions in real time.
- **High-Throughput Relational Optimization (Client: Banco Pichincha):** Profiled, tuned, and refactored enterprise Oracle and SQL Server query plans across millions of records, eliminating deadlock conditions and reducing query latency by up to 60%.
- **Autonomous Operational Automation:** Engineered custom Python integration bots for enterprise ITSM (ServiceNow API) over long-polling Telegram architectures, replacing manual workflows with reliable event-driven automation.
- **Resilience & Circuit-Breaking:** Implemented defensive degradations, retry backoffs, and queue pacing during high-traffic transactional spikes across distributed banking integrations.

TECHNICAL EXPERTISE

LLM & Agentic Protocols:	Model Context Protocol (FastMCP), Function Calling / Tool Use, ReAct loops, Claude API, Structured Outputs, Hard Eval Dataset Curation, Multi-Agent Orchestration, Prompt Distillation.
Information Architecture:	DAG-based Knowledge Graphs, Hybrid Search (Dense Embeddings + BM25), `pgvector`, Vector Indexing (HNSW/IVFFlat), Chunking Strategies, Token-Budget Context Optimization, Cryptographic Fact Verification.
Programming & Runtimes:	Python 3.12+ (Advanced AsyncIO, Typing, Pydantic, FastAPI, Django Ninja, Granian ASGI), Rust (FFI, PyO3, Memory Safety), SQL (PostgreSQL, Oracle, SQL Server), TypeScript, Bash/Linux.
Systems & Observability:	Podman / Docker Containerization, Redis In-Memory Caching, Distributed Logging & Tracing (Grafana, Kibana, Dynatrace), High-Concurrency Async Pipelines, Zero-Trust Architecture.
ML Frameworks:	PyTorch, TensorFlow/Keras, Hugging Face Transformers, Scikit-learn, OpenCV, NumPy/Pandas, Model Quantization.

HONORS, RESEARCH LEADERSHIP & DISTINCTIONS

Global Nominee — NASA International Space Apps Challenge 2025

- Recognized internationally for architectural design and execution of an AI-driven scientific visualization and telemetry prototype under intense hackathon time constraints.

Research Mentor & Technical Speaker 2023 – 2024

- **Universidad Nacional Pedro Ruiz Gallo (UNPRG):** Appointed Mentor in Entrepreneurship, Innovation, and AI Systems (Official Faculty Resolution).
- **Universidad de San Martín de Porres (USMP):** Delivered advanced technical training workshops on Systems Engineering and Applied AI Architectures.

Professional Certifications 2024 – 2025

- **Oracle Cloud Infrastructure 2024 Certified Foundations Associate** • Oracle Certified Professional
- **Oracle Cloud Infrastructure 2024 Certified AI Foundations Associate** • Oracle Certified Professional
- **Advanced Applied Artificial Intelligence** • Universidad Católica de Salta (UCASAL) & USAT

EDUCATION

Universidad Nacional Pedro Ruiz Gallo (UNPRG) Chiclayo, Peru

Bachelor of Science in Systems Engineering (Ingeniero de Sistemas) Graduated

- Core Focus: Distributed Systems, Artificial Intelligence, Cryptographic Protocols, Database Architecture.