Carlos Hernández

Senior AI Platform Engineer

Observable, version-controlled, costed, deletable LLM systems.

San Salvador, El Salvador · Bilingual ES/EN · me@carloshdez.com · +503 7372 5925 · carloshdez.com · LinkedIn · GitHub

SUMMARY

Senior AI Platform Engineer with 8+ years building production cloud systems and growing the engineering organizations behind them. Now focused on building and operating production LLM systems end-to-end — RAG architectures, multi-agent systems, and AI-native backends with the discipline of platform engineering. Track record shipping LLM products 0→1 in enterprise contexts, with hands-on cloud platform leadership delivering 20%+ cost reduction and 95%→99.5% availability uplift. Currently researching comparative NLP techniques for text classification as part of Master's thesis at Universitat de Barcelona.

CORE TECHNICAL SKILLS

AI / LLM Engineering · LLM application development · RAG systems · Multi-agent architectures · Prompt engineering & versioning · LLM evaluation & observability (Langfuse) · Vector databases (pgvector) · Agent Development Kit · Google Gemini · Anthropic

Machine Learning (academic) · NLP text classification · Transformer fine-tuning (BETO, DistilBERT) · Classical ML benchmarks (TF-IDF + SVM, XGBoost) · Word embeddings · Evaluation methodology

Platform Engineering & SRE · Site Reliability · CI/CD design · Progressive Delivery · Infrastructure as Code · FinOps & cost optimization · Observability (OpenTelemetry, Grafana)

Cloud & Infrastructure · GCP (Cloud Run, Cloud SQL, Pub/Sub, API Gateway) · AWS · Kubernetes · Docker · Terraform

Languages & Tooling · Python (async, FastAPI, Pydantic) · TypeScript / Next.js · SQL / PostgreSQL · GitLab CI · pytest

PROFESSIONAL EXPERIENCE

Head of AI Platform Engineering · EsePlus Group (EsePlus & Alilo) — Apr 2019 – Present

San Salvador, SV

Joined as Tech Lead in 2019 at company formation — built the team from zero, hiring 30+ across web, mobile, DevOps/SRE, QA, design, and support, and establishing the technical foundation, patterns, and operational practices that scaled with the group. Transitioned to CTO in Oct 2022 (strategic leadership) and into Head of AI Platform Engineering in Oct 2025, with end-to-end ownership of the AI engines and APIs that the product team consumes to deliver end-user experiences.

AI Platform & Product

- Architected and shipped **Alilo LXD AI**, part of Alilo's AI suite — async generative-AI backend that produces complete corporate microlearning experiences from uploaded documents. Multi-stage pipeline coordinates document analysis, content generation, and activity design with backpressure handling and client-side streaming feedback.
- Architected and shipped **Alibot**, a multi-tenant RAG conversational backend in Alilo's AI suite — hybrid retrieval combining dense HNSW vector search, full-text search, and fuzzy matching via Reciprocal Rank Fusion; Matryoshka embeddings; Google ADK agent runtime with multimodal tool calling. API designed agnostic and composable — invokable iteratively or as a sub-agent from higher-level orchestrators.
- Built **defense-in-depth LLM safety stack** for Alibot — input sanitization, output groundedness classifier, structured refusal pools, and automated adversarial regression as a merge gate; zero jailbreaks at test time across the adversarial suite.
- Established dual observability for AI workloads: OpenTelemetry for infrastructure combined with Langfuse for AI quality (tokens, cost, prompt/response), unified via shared trace identifiers — enabling prompt iteration based on production evals, not intuition.
- Built versioned prompt management with 3-level progressive retry and AI-assisted field sanitization, **cutting token costs ~90%** on validation failures vs. full prompt retries.
- Designed a custom rate limiter for Gemini API orchestration (concurrency control with exponential backoff), eliminating 429 errors at scale.
- Migrated heavy document processing from synchronous HTTP to async architecture (Cloud Run Jobs with client-side streaming feedback), decoupling failure modes and unblocking long-running operations.

Platform & Reliability

- Improved group availability from **95% → 99.5%** through SLO-driven strategy, standardized golden paths, and progressive delivery with canary deployments.
- Achieved **20%+ reduction in cloud spend** through FinOps practices, rightsizing, and automated resource optimization.
- Standardized CI/CD pipelines and Infrastructure-as-Code patterns across services, eliminating manual deployment errors.

Stack: GCP · Vertex AI / Gemini · pgvector · GitLab CI · OpenTelemetry · Langfuse

Senior DevOps Engineer (Contract) · Deckers Outdoor — Oct 2025 – Present
Remote

- Lead technical workstreams for migration to GitLab CI/CD platform across multiple product teams.
- Apply progressive delivery and observability practices to enterprise-scale platform modernization.

Stack: GitLab CI · AWS · Kubernetes · Terraform

DevOps Engineer (Contract) · FullStack Labs — Oct 2022 – Feb 2025
Remote

- Built and maintained GitLab CI/CD pipelines for Kubernetes workloads, reducing deployment time and eliminating manual deployment errors.
- Automated infrastructure provisioning with Terraform on AWS, enabling consistent environments across teams.
- Partnered with product teams to reduce release friction and introduce platform best practices, improving developer velocity.

Systems & Database Administrator · Defensoría del Consumidor — Jul 2018 – Apr 2019
San Salvador, SV · National government agency (consumer protection)

- Managed server infrastructure for government consumer protection agency, maintaining **99.9% uptime** with automated secure backup systems; led internal improvements aligned with ISO 9000.

SELECTED PROJECTS

Alilo AI Suite · Production generative-AI platform (EsePlus / Alilo)

Two production AI engines designed and built end-to-end, exposed as agnostic APIs for the product team to consume and compose into end-user experiences. **Alilo LXD AI** — async backend that produces complete corporate microlearning experiences from uploaded documents, coordinating document analysis, content generation, and activity design with backpressure and streaming. **Alibot** — multi-tenant RAG conversational backend; hybrid retrieval fused via Reciprocal Rank Fusion, Matryoshka embeddings, Google ADK agent with multimodal tool calling, defense-in-depth safety stack with automated adversarial regression. Both APIs are deliberately composable — invokable iteratively or as sub-agents from higher-level orchestrators. **Shared engineering**: versioned prompts with progressive retry, dual observability (infrastructure + AI quality unified via shared trace identifiers), custom rate limiting, and sub-3s p50 latency in production.

Master's Thesis (TFM) · Comparative NLP for Transaction Classification · defense Sep 2026

Comparative evaluation of four NLP approaches for automatic categorization of bank transactions — classical ML baselines (TF-IDF + SVM, XGBoost), embedding-based neural networks (FastText/Word2Vec + Dense NN), and transformer fine-tuning (BETO / DistilBERT). Full evaluation pipeline measuring precision, recall, F1 (macro and weighted), confusion matrices, and inference latency across 68k+ labeled transactions from HuggingFace public datasets. Final deliverable includes an interactive dashboard integrating the winning model with anomaly detection.

Public Speaking · *Cursor + MCP* — technical talk on AI-assisted developer workflows and Model Context Protocol tool integration patterns (2026).

Personal Builds · Meal-planner web app and personal fitness/composition dashboard — React / Next.js / Tailwind, maintained for daily use. Workshop at *carloshdez.com*.

EDUCATION

Master's in Machine Learning & AI (60 ECTS) · OBS Business School / Universitat de Barcelona, Spain. — Oct 2025 – Oct 2026
In progress · Thesis on comparative NLP techniques for text classification

Executive Master's in Artificial Intelligence · Instituto de Inteligencia Artificial, Alicante, Spain. — Apr 2024

B.Sc. in Systems Engineering · University of El Salvador, El Salvador. — Jun 2017