# Carlos Hernández

## Head of AI Platform Engineering

*Sistemas LLM observables, versionados, costeados y borrables.*

San Salvador, El Salvador · Bilingüe ES/EN · me@carloshdez.com · carloshdez.com · linkedin.com/in/carlosantoniohdz · github.com/carloshdez90

## RESUMEN

Ingeniero de plataforma con 8+ años construyendo sistemas cloud en producción y las organizaciones de ingeniería detrás de ellos, hoy construyendo las plataformas sobre las que corren sistemas LLM en producción — RAG multi-tenant, runtimes de agentes, arneses de evaluación y la infraestructura de entrega que va debajo. **Dos carriles por diseño:** ingeniería de plataforma por contrato como base estable, y propiedad end-to-end de la plataforma de AI en una empresa que ayudé a construir desde su formación — ahora consolidando ambos en un solo rol que es dueño de la plataforma de punta a punta: cimiento de cloud e identidad, runtime de agentes y plano de integración, observabilidad y FinOps. **El costo es una restricción de diseño, no una fase de limpieza** — el gate entra antes de que la cosa escale, no después de que llegue la factura. Historial de llevar productos LLM de 0→1 en contextos enterprise, con 20%+ de reducción de costo cloud y mejora de disponibilidad de 95%→99.5%. Cursando el TFM en NLP comparativo en la Universitat de Barcelona.

## COMPETENCIAS TÉCNICAS

**AI / LLM Engineering** · Arquitecturas RAG en producción · Retrieval híbrido y Reciprocal Rank Fusion · Runtimes de agentes y tool calling · Aislamiento multi-tenant · Servidores MCP y planos de integración de agentes · Prompt engineering, versionado y reintento progresivo · Evaluación de LLMs y red-teaming adversarial · Observabilidad de LLMs (Langfuse) · Búsqueda vectorial (pgvector, HNSW) · Google ADK · Gemini · Claude API

**Platform Engineering y SRE** · CI/CD a escala organizacional (GitLab CI) · Bibliotecas de componentes de pipeline reutilizables · Infraestructura como código (Terraform, módulos versionados) · Entrega progresiva · Métricas DORA · Fiabilidad y SLOs · FinOps sobre cloud e inferencia de AI — atribución de costo por tenant, presupuesto de tokens, control de consumo · OpenTelemetry · Vault

**Machine Learning (académico)** · Clasificación de texto NLP · Fine-tuning de transformers (DistilBERT) · Baselines clásicos (TF-IDF + SVM, XGBoost) · Diseño de benchmarks y detección de fuga de datos · Pruebas de significancia estadística (McNemar) · Despliegue en ONNX

**Cloud, identidad e infraestructura** · GCP (Cloud Run, Cloud Run Jobs, Vertex AI, Cloud SQL, Pub/Sub, API Gateway) · AWS (multi-cuenta) · Identidad y acceso: Keycloak, federación OIDC con role-hop, JWT, identidades de servicio de mínimo privilegio, Vault · Kubernetes · Docker

**Tecnología corporativa y gestión de proveedores** · Ciclo completo de proveedores en alilo — selección, negociación y renovación anual sobre cloud, DevOps, colaboración y herramienta de privacidad · SME técnico en decisiones de proveedor en Deckers · Programas de design partner y beta con proveedores de plataforma · Herramienta de privacidad y gestión de consentimiento · Entrega en entorno regulado: alcance PCI y GDPR en NA, EMEA y JP

**Lenguajes y herramientas** · Python (async, FastAPI, Pydantic, SQLAlchemy) · TypeScript / Next.js · SQL / PostgreSQL · pytest

## EXPERIENCIA PROFESIONAL

### EsePlus Group · EsePlus & Alilo

Abr 2019 – Presente

*San Salvador, SV*

Entré como Tech Lead en la formación de la empresa y construí la organización de ingeniería desde cero — 30+ contrataciones entre web, mobile, DevOps/SRE, QA, diseño y soporte — estableciendo los cimientos técnicos y las prácticas operativas que escalaron con el grupo.

#### Head of AI Platform Engineering

Oct 2025 – Presente

- Diseñé y entregué **Alilo LXD AI** — backend generativo asíncrono que convierte documentos subidos en experiencias completas de microlearning corporativo. Pipeline multi-etapa que coordina análisis de documento, generación de contenido y diseño de actividades, con manejo de backpressure y feedback en streaming.
- Diseñé y entregué **Alibot** — backend conversacional RAG multi-tenant: retrieval híbrido que fusiona búsqueda vectorial HNSW densa, full-text y fuzzy matching vía Reciprocal Rank Fusion; embeddings Matryoshka; runtime de agente Google ADK con tool calling multimodal. Ambas APIs diseñadas agnósticas y componibles — invocables iterativamente o como sub-agente desde orquestadores superiores.
- Construí un **stack de seguridad LLM en profundidad**: sanitización de entrada, clasificación de groundedness en salida, pools de rechazo estructurado y regresión adversarial automatizada como merge gate — cero jailbreaks en la suite adversarial.
- Establecí **observabilidad dual para cargas de AI**: OpenTelemetry para infraestructura combinado con Langfuse para calidad de AI (tokens, costo, prompt/respuesta), unificados por trace IDs compartidos — permitiendo iterar prompts con evals de producción, no con intuición.
- Construí gestión de prompts versionada con reintento progresivo de 3 niveles y saneamiento de campos asistido por AI, **recortando el costo en tokens ~90%** en fallos de validación frente a reintentar el prompt completo; reduje el contexto conversacional de 70K→16K tokens por truncado de historial.
- Diseñé un rate limiter propio para orquestar la API de Gemini (control de concurrencia con backoff exponencial), eliminando errores 429 a escala, y migré el procesamiento pesado de documentos de HTTP síncrono a arquitectura asíncrona sobre Cloud Run Jobs.
- Soy dueño del **ciclo completo de proveedores y tecnología corporativa** — selección, negociación y renovación anual con los proveedores de cloud, DevOps, colaboración y privacidad del grupo, incluida su plataforma de gestión de consentimiento. La decisión de proveedor se toma contra los estándares de la plataforma y contra su costo anual, no por separado.

#### Chief Technology Officer

Oct 2022 – Oct 2025

- Elevé la disponibilidad del grupo de **95% → 99.5%** con estrategia guiada por SLOs, golden paths estandarizados y entrega progresiva con despliegues canary.
- Logré **20%+ de reducción en gasto cloud** mediante prácticas FinOps, rightsizing y optimización automatizada de recursos.
- Estandaricé pipelines de CI/CD y patrones de infraestructura como código entre servicios, eliminando errores de despliegue manual.

**Tech Lead** Abr 2019 – Oct 2022

- Definí la arquitectura inicial, las prácticas de entrega y el estándar de contratación del equipo de ingeniería mientras la empresa pasaba de su formación a una organización multi-producto.

***Stack:*** *GCP · Vertex AI / Gemini · pgvector · Google ADK · FastAPI · GitLab CI · OpenTelemetry · Langfuse · Keycloak*

### Deckers Brands · DevOps Engineer · Contrato — Oct 2025 – Presente

*Remoto*

- Diseñé y entregué una **plataforma compartida de entrega CI/CD** — biblioteca de componentes reutilizables (instalación, linting, testing, build, deploy) que cualquier repositorio de la organización consume con tres líneas de configuración, con tres repositorios consumidores en producción, despliegue ruteado por rama a sandboxes por equipo y métricas DORA instrumentadas de punta a punta.
- Recorté el **wall clock del pipeline 52%, el tiempo de runner 54% y los jobs por pipeline 62%** — medido p50 sobre una ventana de 14 días y 295 pipelines, no una corrida suelta. La mayor ganancia fue estructural: el build nunca consumía la salida del lint, así que el quality gate se movió aguas abajo para bloquear la publicación en vez de la compilación.
- Construí una **plataforma self-service de agentes SRE** que provisiona un agente de respuesta a incidentes con alcance por equipo — sus integraciones, permisos y comportamiento — desde un solo bloque de configuración; dar de alta un equipo es un merge request, no un proyecto.
- Hice revisable el comportamiento del agente: playbooks e instrucciones permanentes se entregan como **paquetes versionados pineados por agente**, así que cambiar cómo se comporta es un diff revisado con aprobador en vez de una edición de consola no auditable. Toda integración de solo lectura por construcción; el agente recomienda, los humanos actúan.
- Implementé autenticación cross-account por OIDC con role-hop y sin credenciales permanentes en CI, con identidades de servicio de mínimo privilegio por integración.
- Participo como **SME técnico en decisiones de proveedor** — aporto la mirada de plataforma a la selección de herramientas y a las conversaciones de renovación anual con los managers, que llevan el lado comercial. Participo en programas de design partner y beta con proveedores de plataforma, devolviendo requisitos de producción antes de que la feature salga.

***Stack:*** *GitLab CI · AWS (multi-cuenta, OIDC, Cloud Control API) · Terraform · Vault · Datadog · Kubernetes*

### FullStack Labs · DevOps Engineer · Contrato — Oct 2022 – Feb 2025

*Remoto*

- Construí y mantuve pipelines de GitLab CI/CD para cargas de Kubernetes, reduciendo el tiempo de despliegue y eliminando errores manuales.
- Automaticé el provisionamiento de infraestructura con Terraform sobre AWS, habilitando entornos consistentes entre equipos, y trabajé con equipos de producto para reducir la fricción de release.

### Defensoría del Consumidor · Administrador de sistemas y bases de datos — Jul 2018 – Abr 2019

*San Salvador, SV · Institución nacional de protección al consumidor*

Gestioné la infraestructura de servidores de una institución de gobierno, manteniendo **99.9% de uptime** con respaldos seguros automatizados; lideré mejoras internas alineadas a ISO 9000.

## PROYECTOS DESTACADOS

### Clarytia · Plataforma RAG multi-tenant · diseñada, construida y auditada en solitario — Beta privada

Asistente de conocimiento interno verificable: cada respuesta lleva su documento fuente, la página y un nivel de confianza, y devuelve “no lo sé” en vez de improvisar. Aislamiento por base de datos por tenant, así que borrar un cliente es un DROP y no una consulta que tenga que cubrir todas las tablas. Cinco carriles de ranking — HNSW denso, full-text, trigramas, sin acentos y fuzzy — fusionados por Reciprocal Rank Fusion en una sola consulta SQL, sin servicio de búsqueda externo. Control de acceso aplicado dentro de la ruta de recuperación, no como filtro posterior. La detección de vacíos de conocimiento agrupa preguntas sin responder en huecos nombrados y rankeados por usuarios distintos, con resolución condicionada a adjuntar un documento y un re-chequeo asíncrono que reabre el hueco si el documento sigue sin responder.

**2,814 tests en 232 archivos (>90% de cobertura) · 37 ADRs · 105 endpoints · red team generativo como gate en cada deploy con cero fugas · 189 commits en 5 meses, autor único.**

### TFM · NLP comparativo para clasificación de transacciones — Defensa Sep 2026

Evaluación comparativa de baselines clásicos (TF-IDF + SVM, XGBoost), redes neuronales basadas en embeddings y fine-tuning de transformer (DistilBERT) para categorización de transacciones bancarias, juzgados por precisión, robustez a texto ruidoso, significancia estadística y costo total de propiedad. **Encontré que el benchmark público tenía fuga de datos**: sus datasets se construyen con plantillas, así que un split aleatorio deja casi todo comercio en train y en test a la vez. Reconstruí la evaluación degradando el test set con cuatro tipos de ruido realista — y escribí la predicción antes de correrlo, para después reportar que me equivoqué. Al 30% de ruido el modelo lineal retuvo 89.9% de su score frente al 84.2% de DistilBERT, y McNemar pasó de $p=1.00$ a $p=1.3\times10^{-78}$ a su favor. El transformer costó 356× los parámetros y 15× la latencia de inferencia para empatar en texto limpio y perder en sucio. El modelo ganador exporta entero a un grafo ONNX de 1.3 MB que corre en el navegador: $0 de servir, y los datos nunca salen del dispositivo.

### Escritura de ingeniería · carloshdez.com

Field notes semanales sobre sistemas LLM en producción, publicadas en español e inglés: versionar prompts como migraciones de esquema, qué significa realmente “observable” en una app de LLM, la regresión adversarial como merge gate, y los resultados del TFM. Charla pública: *“Cursor + MCP — when your IDE knows your stack”* (2026).

## FORMACIÓN

**Máster en Machine Learning e IA (60 ECTS)** · OBS Business School / Universitat de Barcelona, España — Oct 2025 – Oct 2026

*En curso · TFM sobre NLP comparativo para clasificación de texto*

**Máster Executive en Inteligencia Artificial** · Instituto de Inteligencia Artificial, Alicante, España — Abr 2024

**Ingeniería de Sistemas** · Universidad de El Salvador, El Salvador Jun 2017