



## **Whitepaper**

Infraestructura de IA para América Latina

## Introducción

Este documento está dirigido a los responsables técnicos, institucionales y regulatorios que toman decisiones sobre cómo y dónde se despliega la inteligencia artificial en la región.

Está orientado a CTOs y líderes de ingeniería de empresas reguladas, equipos de transformación digital del sector público, equipos de alianzas tecnológicas de NVIDIA y proveedores de nube, así como a actores de política pública y regulación, y propone un marco arquitectónico para evaluar, diseñar y operar infraestructuras de IA que permitan control sobre datos, modelos y ejecución en entornos híbridos y regulados.

## 1. Resumen Ejecutivo

América Latina enfrenta una ventana de oportunidad que no superará los siete años. El mercado regional de inteligencia artificial crecerá de \$21 mil millones en 2024 a \$368 mil millones para 2033. Más de 650 millones de personas adoptarán tecnologías de IA en la próxima década, y lo harán con la infraestructura que esté disponible cuando tomen esa decisión.

Hoy, esa infraestructura no existe en la región. El 80% de la capacidad global de IA corre fuera de América Latina. Las empresas que quieren adoptar inteligencia artificial enfrentan una elección imposible: enviar sus datos a servidores extranjeros bajo jurisdicciones que no controlan, o no adoptar IA en absoluto. Para industrias reguladas (finanzas, salud, gobierno), la segunda opción es frecuentemente la única viable.

Saptiva AI existe para resolver este problema. Construimos la capa de infraestructura que permite a las empresas latinoamericanas ejecutar IA en sus propios términos: segura, compliant, sin dependencias. Cualquier modelo, en cualquier entorno (nube pública, nube privada, on-premise, air-gapped), con control total sobre datos e infraestructura. Sin dependencia de hyperscalers extranjeros. Sin vendor lock-in.

Hace una década construimos Quiubas Mobile, la primera plataforma de mensajería carrier-grade en México. Twilio nos adquirió. Ahora construimos la infraestructura para la siguiente era.

La carrera global por la infraestructura de IA ya comenzó. India, Arabia Saudita y Europa están invirtiendo cientos de miles de millones en independencia tecnológica. América Latina no tiene un jugador comparable. Ese vacío es nuestra oportunidad.

## 2. El Momento Decisivo

Existen momentos en la historia donde la confluencia de factores tecnológicos, geopolíticos y de capital genera ventanas de oportunidad que, una vez cerradas, no vuelven a abrirse. La revolución industrial benefició a quienes controlaron el carbón y el acero; la era digital premió a quienes dominaron el software y las plataformas. La era de la inteligencia artificial está redefiniendo esas reglas, y la infraestructura computacional se ha convertido en el recurso estratégico que determinará quién participa como arquitecto de esta transformación y quién queda relegado al papel de consumidor dependiente.

El año 2025 marcó un punto de inflexión. Amazon, Microsoft, Google y Meta invirtieron combinados más de \$370 mil millones en infraestructura de IA, un incremento del 60%

respecto a 2024. OpenAI lideró el proyecto Stargate con \$500 mil millones comprometidos. Los compromisos de inversión superaron \$1 billón en solo ocho proyectos principales anunciados ese año. Desde los desiertos de Texas hasta las costas del Golfo Pérsico, una carrera silenciosa pero decisiva rediseñó la geografía del poder tecnológico.

América Latina observó esta transformación desde una posición paradójica. Poseía 650 millones de personas con creciente digitalización y talento técnico probado globalmente. Sin embargo, la región representaba apenas el 5% de la capacidad global de data centers. Las inversiones de Big Tech en la región, aproximadamente \$12 mil millones distribuidos en 15 años, equivalían a menos de lo que Amazon invirtió globalmente en un solo trimestre de 2025.

Dos fuerzas opuestas definen la urgencia. Por un lado, el costo de usar IA se ha democratizado: la inferencia para sistemas equivalentes a GPT-3.5 se redujo más de 280 veces entre 2022 y 2024. Por otro, la capacidad para crear IA propia se ha concentrado: el procesamiento requerido para entrenar modelos de frontera se duplica cada cinco meses. Esta paradoja (acceso barato a modelos existentes, costo prohibitivo para crear los propios) es precisamente lo que hace urgente la acción.

La demora no resultará en un simple retraso tecnológico. Resultará en dependencia estructural permanente.

### **3. Lo Que Está en Juego**

La dependencia tecnológica no es una abstracción. Tiene consecuencias concretas que afectan la vida cotidiana de cientos de millones de personas.

#### **El sesgo de los modelos dominantes**

Los modelos de lenguaje más utilizados no son neutrales. Una evaluación publicada en PNAS Nexus encontró que todos los modelos de OpenAI exhiben valores culturales que se asemejan a los de países anglófonos y europeos protestantes. El Centro Nacional de Inteligencia Artificial de Chile construyó un benchmark para medir qué tan bien los modelos representan conocimiento latinoamericano. Los resultados: todos los modelos identificaron en qué país está Buenos Aires, pero fallaron sistemáticamente en preguntas sobre cultura local. Como señala Stanford, comunidades enteras están siendo excluidas de la revolución de IA y pierden oportunidades económicas y educativas que los hablantes de inglés obtienen.

#### **La vulnerabilidad de los datos**

La CLOUD Act estadounidense de 2018 permite a las autoridades de ese país exigir datos almacenados por empresas americanas sin importar la ubicación física de los servidores. Expedientes médicos, registros educativos y datos financieros de ciudadanos latinoamericanos en servidores de AWS, Microsoft o Google pueden ser requeridos por tribunales estadounidenses sin notificación al país de origen.

En 2025, la empresa holandesa Solvinity, elegida específicamente por el gobierno de Países Bajos para mitigar riesgos de la CLOUD Act, fue adquirida por una entidad estadounidense. De un día para otro, infraestructura crítica nacional quedó bajo alcance de

autoridades extranjeras. Si esto puede ocurrir en Europa, la vulnerabilidad de América Latina es mayor.

### **El sesgo en aplicaciones críticas**

Más de la mitad de los datasets para entrenar IA provienen de Estados Unidos o China. Los modelos de diagnóstico y tratamiento están calibrados para poblaciones que no representan la diversidad de América Latina. Cuando estos sistemas se despliegan en contextos diferentes, traen supuestos que no viajan bien.

### **La extracción de valor**

El valor generado por los datos de 650 millones de latinoamericanos fluye hacia corporaciones extranjeras. Los trabajadores de la región participan en la economía de IA principalmente como proveedores de anotación de datos mal remunerada, no como creadores de los sistemas. La región exporta datos crudos e importa servicios procesados, replicando el patrón extractivo que ha caracterizado su inserción en la economía global durante siglos.

La alternativa no es aislamiento tecnológico ni proteccionismo ingenuo. Es construir capacidad propia que permita a América Latina participar como creador, no solo como consumidor.

## **4. La Tesis de Saptiva AI**

Ejecutar IA en tus propios términos no significa desconexión del ecosistema global. Significa la capacidad de elegir: dónde corren los datos, qué modelos se usan, bajo qué jurisdicción opera la infraestructura. Significa independencia estratégica, no aislamiento.

Silicon Valley optimiza para extracción de valor. Beijing optimiza para control. Nosotros optimizamos para empoderamiento: distribuir agencia en lugar de concentrarla. El control opera en tres capas: física (infraestructura, centros de datos), código (modelos, estándares, diseño de sistemas) y datos (propiedad, flujos, uso). La debilidad en cualquiera compromete las demás. América Latina tiene déficits en las tres. Saptiva AI las aborda de manera integrada.

Nuestra tesis es simple: quien controle la infraestructura de IA controlará las condiciones bajo las cuales opera la economía digital. América Latina puede construir esa infraestructura o puede alquilarla a quienes la construyen para otros propósitos.

Elegimos construir.

## **5. Requisitos Mínimos de IA Soberana (Sovereign AI)**

Antes de describir qué construimos, es necesario establecer qué significa ejecutar IA en tus propios términos. No basta con declararlo; hay criterios objetivos y verificables que cualquier solución debe cumplir.

**Definimos IA Soberana (Sovereign AI) como cualquier infraestructura de inteligencia artificial que cumpla los siguientes siete requisitos.** Estos criterios son no negociables. Si una plataforma no los cumple, no permite control real sobre datos, modelos e infraestructura:

### **1. Control plane bajo jurisdicción local**

El cerebro que decide dónde y cómo se ejecutan las cargas de trabajo debe operar bajo jurisdicción del país o región del cliente. Si el control plane está en Estados Unidos, la infraestructura está sujeta a la CLOUD Act, sin importar dónde residan los datos.

### **2. Data plane sin dependencia legal extranjera**

Los datos en reposo y en tránsito deben estar protegidos de requerimientos legales extraterritoriales. Esto requiere que tanto el almacenamiento como el procesamiento ocurran en entidades legales no sujetas a jurisdicciones externas.

### **3. Observabilidad auditabile por el cliente**

El cliente debe poder verificar, de manera independiente, qué ocurre con sus datos: quién accedió, cuándo, qué modelos procesaron qué información. Sin observabilidad auditabile, el control es una promesa no verificable.

### **4. Portabilidad verifiable**

El cliente debe poder migrar sus datos, configuraciones y cargas de trabajo a otro proveedor sin barreras técnicas artificiales. Formatos propietarios y dependencias de APIs no estándar son formas de lock-in que comprometen el control.

### **5. Capacidad de salida (data e inference exit)**

En cualquier momento, el cliente debe poder extraer la totalidad de sus datos y dejar de usar el servicio. Esto incluye logs, configuraciones, modelos fine-tuneados, y cualquier artefacto generado en la plataforma.

### **6. Cifrado con control de llaves por el cliente**

En despliegues de alta sensibilidad, el cliente debe poder gestionar sus propias llaves de cifrado. Si el proveedor controla las llaves, controla el acceso a los datos.

### **7. Aislamiento verifiable entre clientes**

En arquitecturas multi-tenant, debe existir segregación completa entre organizaciones, verifiable mediante auditoría técnica. Un cliente no debe poder acceder, ni siquiera teóricamente, a datos de otro.

Estos siete requisitos constituyen el estándar mínimo de IA Soberana. Una infraestructura que no los cumpla puede ser útil, eficiente, o económica, pero no otorga control real al cliente.

*Saptiva AI ha sido diseñada para cumplir con los siete requisitos de IA Soberana y actualmente los opera en entornos de producción.*



## 6. Qué Construimos

Saptiva AI es la capa de infraestructura que permite a las empresas latinoamericanas ejecutar IA en sus propios términos. En una oración: tu IA, tu stack, 100% bajo tu control.

### El problema que resolvemos

Las empresas quieren implementar IA. No saben cómo. Y cuando intentan hacerlo, enfrentan obstáculos que los proveedores globales no resuelven: modelos cerrados que operan como cajas negras en nubes extranjeras, creando riesgos de compliance; imposibilidad de trazar quién hizo qué, cuándo, con qué modelo y qué datos; dependencia total de proveedores externos sin control sobre dónde corren las cargas de trabajo; y plataformas fragmentadas que hacen imposible gestionar seguridad, rendimiento y observabilidad de manera unificada.

El caso concreto: un banco mexicano quería desplegar IA para clasificación de documentos. No podía enviar archivos a servidores de OpenAI o AWS en Estados Unidos. No existía opción compliant. Ese banco no es una excepción; es la norma en industrias reguladas.

### Nuestra solución

Saptiva AI permite correr cualquier modelo, abierto o propietario, en cualquier entorno: nube pública, nube privada, on-premise, o air-gapped. Ofrecemos control completo sobre datos, modelos e infraestructura, garantizando cumplimiento con regulación latinoamericana, sin vendor lock-in.

La plataforma opera en tres capas integradas:

**Capa Aplicativa (Business Layer):** Entrega valor inmediato a través de un constructor de agentes, recetas preconfiguradas para RAG (RAGster), agentes SQL (Bank Advisor), y casos de uso listos para producción. Esta capa abstrae la complejidad técnica y permite a equipos de negocio desplegar soluciones de IA sin expertise en infraestructura.

**Capa de Orquestación (Platform Layer, frldA):** El sistema operativo de IA que permite desplegar donde el cliente decida, con observabilidad, portabilidad y trazabilidad completas. frldA gestiona modelos y aplicaciones de manera unificada, orquesta despliegues agnósticos de entorno, y garantiza compliance por diseño.

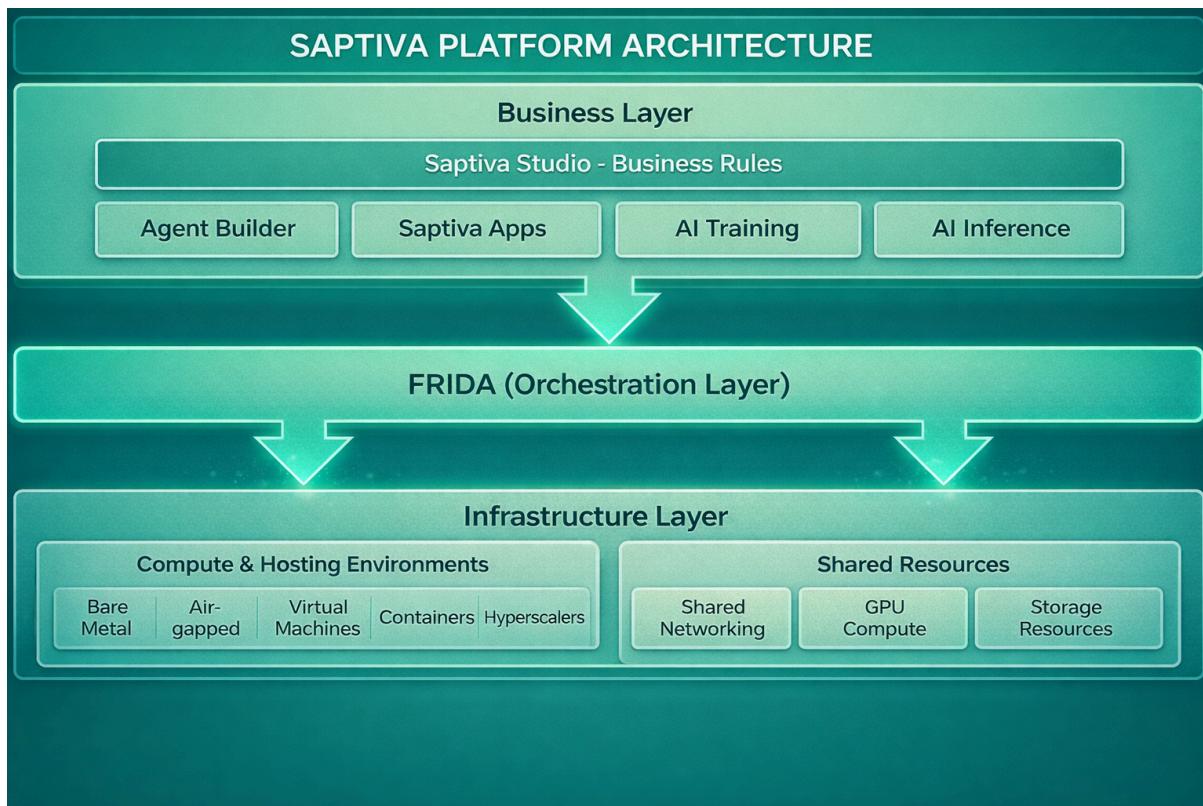
**Capa de Infraestructura (Infra Layer):** Instancias para el cerebro orquestador del clúster de cómputo y el cómputo de ejecución de cargas de trabajo bajo jurisdicción regional. Los datos permanecen donde el cliente decide.

### La visión

Convertir a Saptiva AI en el sistema operativo estándar de IA en América Latina. Una red de infraestructura que permite a cualquier empresa ejecutar IA donde decida, sin ceder control, sin fricción, sin dependencia de hyperscalers extranjeros.

## 7. Arquitectura Soberana

Esta sección detalla cómo se implementa soberanía a nivel de arquitectura. Los principios establecidos en la sección anterior no son aspiracionales; son características operativas de la plataforma Saptiva AI.



### 7.1 frldA: El Sistema Operativo de IA para América Latina

frldA es el núcleo de la plataforma Saptiva AI. Opera como Sistema Operativo de IA, agnóstico al entorno computacional (nube pública, nube privada, on-premise o air-gapped) funcionando tanto como Platform as a Service (PaaS) como Infrastructure as a Service (IaaS) según el escenario de despliegue.

#### Por qué frldA existe

Las herramientas de orquestación estándar resuelven cómo distribuir cargas de trabajo eficientemente entre recursos de cómputo. Pero no fueron diseñadas para resolver control jurisdiccional.

Un orquestador genérico decide dónde correr una carga basándose en disponibilidad de recursos. frldA decide basándose en políticas de control: jurisdicción legal del dato, requisitos regulatorios del cliente, restricciones de localización, y nivel de aislamiento requerido.

#### Lo que frldA agrega sobre orquestación estándar

1. Validar que una carga de trabajo cumpla políticas de residencia de datos antes de ejecutarla
2. Enrutar inferencias a modelos específicos según clasificación de sensibilidad de la empresa
3. Garantizar que logs y métricas permanezcan en jurisdicción incluso cuando el cómputo es federado
4. Aplicar políticas de compliance como precondición de despliegue, no como auditoría posterior

### **El problema que resuelve**

Los hyperscalers construyen infraestructura excelente optimizada para escala global. Esta optimización no prioriza requisitos de control regional porque su mercado principal no lo demanda. frldA llena ese vacío: permite a las empresas latinoamericanas aprovechar infraestructura de clase mundial, incluyendo hyperscalers cuando es apropiado, mientras mantienen control sobre políticas, datos y compliance.

### **frldA como punto de control**

frldA no compite con la infraestructura de los hyperscalers; la complementa agregando la capa de control que hace posible ejecutar IA en tus propios términos. Sin esa capa, un cliente puede pedir control. Con frldA, el control es una precondición que el sistema valida antes de ejecutar cualquier operación.

### **Operación como PaaS**

frldA actúa en dos capas coordinadas. Hacia la capa de Negocio, valida políticas de configuración: nivel de cumplimiento regulatorio, preferencias de despliegue, restricciones de localización de datos y criticidad de servicios. Hacia la capa de Plataforma, orquesta la asignación y liberación de cómputo de manera automatizada.

Este enfoque ofrece capacidades críticas: abstracción de la capa computacional física o virtual (portabilidad), orquestación de recursos, escalamiento horizontal y vertical automático, portabilidad entre proveedores de nube (agnosticismo del entorno computacional), resiliencia de operación mediante recuperación automática de fallos, y compatibilidad nativa con herramientas de observabilidad estándar de la industria.

### **Operación como IaaS**

Para escenarios que requieren GPU pools con Servidores Virtuales Privados (GPU as a Service, Compute as a Service, VPS), frldA gestiona hardware bare metal y servicios de nube a través de una capa de virtualización propietaria optimizada para cargas de trabajo de IA. Sobre esta infraestructura, frldA mantiene todas las capacidades del modo PaaS.

## **7.2 Control del Dato: Mecanismos y Garantías**

El control del dato no es una declaración de principios; es un conjunto de mecanismos técnicos y contractuales verificables:

**Residencia de datos regional:** Los datos permanecen físicamente en la jurisdicción elegida por el cliente. El servidor de control para orquestación con frldA opera en infraestructura bajo jurisdicción latinoamericana, eliminando exposición a leyes extraterritoriales como la CLOUD Act.

**Cifrado por defecto:** Cifrado en reposo (AES-256) y cifrado en tránsito (TLS 1.3) para toda comunicación y almacenamiento. Las llaves de cifrado pueden ser gestionadas por el cliente (BYOK, Bring Your Own Key) en despliegues on-premise.

**Acceso y control del cliente:** Cada cliente, con los privilegios adecuados, puede acceder a sus logs transaccionales y almacenes de datos. Puede solicitar volcado completo de su información y ejecutar borrado seguro certificado de todos sus datos.

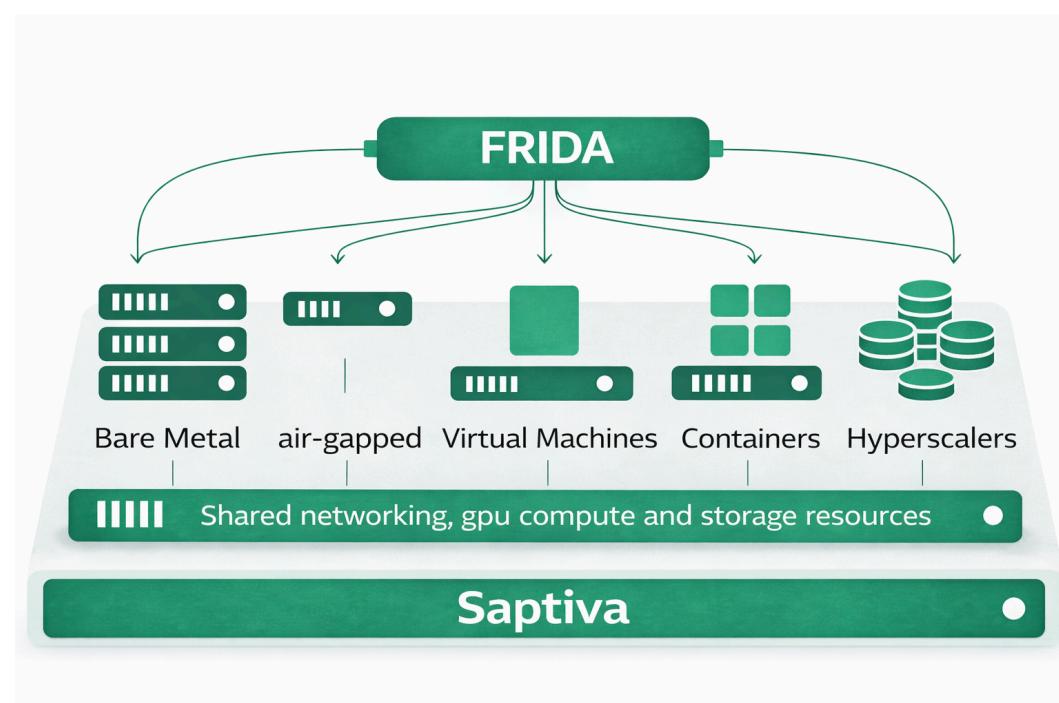
**Aislamiento multi-tenant:** Cada cliente opera en un entorno aislado, con políticas de red que garantizan segregación completa del tráfico entre organizaciones.

**Retención configurable:** Retención de datos predeterminada de 30 días, extensible según contrato. Los logs de conversaciones y uso de servicios se almacenan separados por cliente con cifrado en reposo.

**Garantías contractuales:** Contratos de confidencialidad que especifican jurisdicción, mecanismos de auditoría, y procedimientos de respuesta ante requerimientos legales.

### 7.3 Despliegues Multi-Entorno

La arquitectura de frldA es agnóstica al entorno de ejecución. Soportamos cuatro modalidades de despliegue:





**Nube Pública:** Despliegue en hyperscalers (AWS, Azure, GCP, OCI), o proveedores regionales, con frIdA gestionando la orquestación y aplicando políticas de soberanía sobre infraestructura de terceros.

**Nube Privada:** Infraestructura dedicada gestionada por Saptiva AI o el cliente, con control total sobre hardware y software. frIdA opera como el plano de control unificado.

**On-Premise:** Despliegue en data centers del cliente. frIdA se instala como licencia y gestiona recursos locales con las mismas capacidades que en nube.

**Air-Gapped:** Para entornos de máxima seguridad (gobierno, defensa, infraestructura crítica), despliegue completamente aislado de internet. Actualizaciones mediante canales seguros offline.

### Conectividad Multi-Cloud y Híbrida

En escenarios que requieren multi-cloud o combinaciones híbridas (Hyperscaler + On-Premise), frIdA establece conectividad segura entre entornos con cifrado punto a punto, segmentación de red, y políticas de acceso zero-trust.

## 7.4 Trazabilidad y Observabilidad

Un sistema con control real debe ser auditabile. frIdA implementa trazabilidad completa:

**Registro de entradas y salidas:** Cada interacción con modelos de IA queda registrada con timestamp, usuario, modelo utilizado, tokens consumidos, y namespace de origen.

**Audit trails:** Logs inmutables de todas las acciones administrativas, cambios de configuración, y accesos a datos sensibles.

**Observabilidad integrada:** Compatibilidad con plataformas de monitoreo estándar de la industria para métricas, logs y trazas distribuidas. El cliente puede conectar sus herramientas de observabilidad existentes.

**Alerting y anomaly detection:** Configuración de alertas sobre patrones de uso anómalos, consumo de recursos, y eventos de seguridad.

### Qué ve un regulador:

La arquitectura de Saptiva AI está diseñada para que el cumplimiento sea demostrable, no declarativo. Un regulador puede solicitar:

1. Exportación de todos los logs de un período específico
2. Demostración de que los datos nunca salieron de jurisdicción
3. Evidencia de cifrado en reposo y tránsito
4. Registro de quién accedió a qué datos y cuándo
5. Prueba de que los controles estaban activos durante el período auditado

frIdA puede generar esta evidencia en horas, no semanas.

## 7.5 Escalamiento y Grid de GPUs

El escalamiento se realiza de forma horizontal, agregando nodos de GPUs de acuerdo a los límites establecidos por las capacidades físicas del entorno (on-premise) o la configuración del proveedor de nube, respetando siempre las políticas de billing y localización de datos del cliente.

**Escalamiento intra-cluster:** Ajuste automático de recursos de cómputo y memoria según demanda, tanto horizontal (más instancias) como vertical (más recursos por instancia).

**Escalamiento inter-cluster:** Capacidad de federar infraestructura entre múltiples proveedores de nube, permitiendo burst de capacidad sin comprometer control.

**GPU pooling:** Gestión unificada de pools de GPUs heterogéneos (NVIDIA A100, H100, L40S) con scheduling inteligente basado en tipo de carga de trabajo.

## 7.6 Mecanismos Anti Lock-in

La dependencia de un solo proveedor es una forma de pérdida de control. Saptiva AI implementa múltiples mecanismos para garantizar portabilidad:

**Protocolos abiertos:** Todas las APIs exponen interfaces estándar (REST, gRPC, GraphQL). No existen formatos propietarios para datos o configuraciones.

**Arquitectura basada en estándares:** El stack técnico se basa en tecnologías de orquestación ampliamente adoptadas por la industria, garantizando interoperabilidad y evitando dependencias de proveedores específicos.

**Modelos AI open source:** Soporte nativo para modelos abiertos (LLaMA, Mistral, Qwen) además de APIs propietarias. El cliente puede migrar entre modelos sin cambios de infraestructura.

**Portabilidad de workloads:** Las cargas de trabajo se empaquetan en formatos estándar de la industria, eliminando dependencia de entornos de ejecución propietarios.

**Exportación de datos:** En cualquier momento, el cliente puede exportar la totalidad de sus datos, configuraciones, y logs en formatos estándar.

## 7.7 Cumplimiento Normativo

La arquitectura de Saptiva AI está diseñada para facilitar el cumplimiento con marcos regulatorios latinoamericanos e internacionales:

**Ley Federal de Protección de Datos Personales (México):** Residencia de datos en territorio nacional, controles de acceso, y mecanismos de consentimiento.

**LGPD (Brasil):** Aislamiento de datos, reportes de tratamiento, y capacidad de respuesta a solicitudes de titulares.



**Regulaciones sectoriales:** Arquitectura compatible con requerimientos de CNBV (México), SBS (Perú), CMF (Chile) para instituciones financieras.

**Estándares internacionales:** Controles alineados con ISO 27001, SOC 2 Type II, y frameworks de gobierno de datos.

## 8. Modelo Económico

Saptiva AI opera con un modelo de negocio transparente y predecible, estructurado en componentes que se adaptan al nivel de control que requiere cada cliente.

### 8.1 Capa Aplicativa (SaaS)

Acceso a herramientas de productividad de IA listas para usar: constructor de agentes, recetas de RAG, agentes SQL, y casos de uso preconfigurados.

**Modelo de cobro:** Suscripción mensual por usuario activo por organización, con tiers basados en volumen de uso y funcionalidades habilitadas.

**Incluye:** Acceso a la interfaz de usuario, plantillas, integraciones estándar, y soporte técnico dependiendo del tier.

### 8.2 Infraestructura (Consumo)

Cómputo, almacenamiento, y recursos de inferencia consumidos en la plataforma Saptiva AI o en infraestructura gestionada por Saptiva AI.

**Modelo de cobro:** Pay-as-you-go basado en métricas de consumo: tokens procesados para inferencia, horas de GPU para entrenamiento/fine tuning, almacenamiento utilizado (GB/mes), y transferencia de datos.

**Opciones:** Contratos de capacidad reservada con descuentos para clientes con volúmenes predecibles.

### 8.3 Licenciamiento de frIdA (On-Premise / Nube Privada)

Para organizaciones que requieren desplegar frIdA en su propia infraestructura, ya sea on-premise o en nube privada.

**Modelo de cobro:** Licencia anual por GPU gestionada.

**Incluye:** Derecho de uso del software frIdA, actualizaciones de versión, parches de seguridad, y soporte técnico dedicado.

**Servicios adicionales:** Implementación, integración con sistemas existentes, capacitación, y soporte premium disponibles como servicios profesionales.

### 8.4 Servicios de Optimización (Fine-tuning)



Saptiva AI acompaña a sus clientes en el proceso de aprovechar sus datos propios para optimizar el rendimiento de los sistemas de IA.

**Fine-tuning de LLMs:** Ajuste de modelos de lenguaje con datos específicos del cliente para mejorar precisión, reducir alucinaciones, y adaptar el tono y vocabulario al dominio de negocio.

**Fine-tuning de Guardrails:** Entrenamiento de sistemas de seguridad y filtrado personalizados para detectar contenido sensible, aplicar políticas de uso, y garantizar que las respuestas cumplan con lineamientos internos y regulatorios.

**Modelo de cobro:** Proyectos de alcance definido o retainer mensual según volumen de iteración requerido.

## 8.5 Por qué este modelo funciona

**Economía de escala regional:** Saptiva AI opera infraestructura compartida optimizada para América Latina. En lugar de que cada empresa construya su propia capacidad (capex prohibitivo), múltiples clientes comparten infraestructura con aislamiento garantizado. El costo de control se distribuye.

**Compliance incluido, no adicional:** Para empresas reguladas, el costo total de usar un hyperscaler incluye consultoría de compliance, auditorías de seguridad, y gestión de riesgo regulatorio. Saptiva AI integra estos controles por diseño, simplificando la ecuación económica.

## 9. Por Qué Nosotros

### Experiencia en infraestructura regional

Hemos construido infraestructura de misión crítica antes. Hace una década creamos Quiubas Mobile, la primera plataforma de mensajería carrier-grade en México. Conectamos millones de dispositivos con operadores de telecomunicaciones que no toleran downtime. Twilio nos adquirió. Esa experiencia (confiabilidad, compliance regulatorio, integración con sistemas legacy, operación donde el downtime tiene consecuencias severas) es directamente relevante para desplegar IA en industrias reguladas.

### Conocimiento profundo de IA

No somos integradores revendiendo APIs. Entendemos la tecnología a nivel fundamental. Lanzamos KAL, el primer LLM mexicanizado. Nuestro equipo aprende constantemente, iterando sobre lo que funciona y descartando lo que no. En un campo que evoluciona semanalmente, la capacidad de aprendizaje continuo es más valiosa que cualquier ventaja técnica estática.

### Densidad de talento



El 50% de nuestro headcount ha sido fundador. Ronald Escalona (SVP of Engineering) fue VP de Ingeniería en Platzi y fundador de OpenSinergia. Carlos Lara (CPO) fundó Point CRM. Cris Huertas (CRO) fundó Morgana y lideró Latam para Bnext. Este no es un equipo de operadores ejecutando el playbook de alguien más. Es un equipo de constructores.

## Tracción temprana

Nueve clientes pagando. Pilotos con instituciones financieras, intermediarios regulatorios, contratistas de seguridad nacional. La validación no es teórica.

## First mover en el espacio correcto

OpenAI y Anthropic ofrecen APIs de modelos, pero no control ni localización para América Latina. AWS, Azure y Google ofrecen infraestructura bajo CLOUD Act. HuggingFace ofrece modelos abiertos sin orquestación empresarial. Nadie combina infraestructura + control jurisdiccional + localización regional + cero lock-in.

El mercado latinoamericano carece hoy de una plataforma de infraestructura de IA que combine control real, gobernanza y portabilidad a nivel de arquitectura.

Ese vacío es nuestra oportunidad.

## 10. El Camino Adelante

La carrera global ya comenzó. India entrena modelos soberanos con 4,000 GPUs H100 del gobierno. Arabia Saudita comprometió \$100 mil millones y rediseña el currículo educativo para seis millones de estudiantes. Europa respalda a Mistral con €1.3 mil millones de ASML.

América Latina no tiene un proyecto comparable. Todavía.

## Lo que creemos

Creemos que América Latina merece infraestructura de IA construida para sus necesidades, no adaptada como afterthought de soluciones diseñadas para otros mercados.

Creemos que la soberanía tecnológica no es un lujo de países ricos, sino una necesidad estratégica para cualquier región que aspire a participar en la economía del siglo XXI como creador.

Creemos que la ventana de oportunidad es real, finita, y se está cerrando.

Creemos que este equipo, con esta experiencia, en este momento, puede construir la infraestructura que América Latina necesita.

*Conectamos a Latinoamérica a través de mensajería móvil. Ahora vamos a potenciar su inteligencia.*

## 11. Marco de Referencia

Este documento establece el estándar de IA Soberana (Sovereign AI) para América Latina.

Los siete requisitos mínimos definidos en la Sección 5 no son solo características de producto de Saptiva AI. Son el criterio objetivo contra el cual debe evaluarse cualquier solución que pretenda ofrecer control real:

1. Control plane bajo jurisdicción local
2. Data plane sin dependencia legal extranjera
3. Observabilidad auditable por el cliente
4. Portabilidad verificable
5. Capacidad de salida (data e inference exit)
6. Cifrado con control de llaves por el cliente
7. Aislamiento verificable entre clientes

Una plataforma que no cumpla estos siete requisitos puede ser útil, eficiente, o económica. No califica como IA Soberana.

Saptiva AI cumple los siete. Los opera en producción con nueve clientes. Y publica este marco para que la región tenga un estándar claro contra el cual medir alternativas.

*La infraestructura de IA en tus propios términos para América Latina se construye así. Y nosotros ya la estamos operando.*

## Glosario Técnico

Este glosario proporciona definiciones de términos técnicos utilizados en el documento, orientado tanto a CTOs como a reguladores y tomadores de decisiones no técnicos.

**Air-gapped:** Entorno de cómputo completamente aislado de internet y redes externas. Utilizado para cargas de trabajo de máxima seguridad donde no se permite ninguna conexión exterior.

**API (Application Programming Interface):** Conjunto de protocolos y herramientas que permiten que diferentes aplicaciones de software se comuniquen entre sí de manera estandarizada.

**Bare metal:** Servidores físicos dedicados, sin capa de virtualización intermedia. Ofrecen máximo rendimiento y control sobre el hardware.

**BYOK (Bring Your Own Key):** Modelo de gestión de cifrado donde el cliente mantiene control sobre sus propias llaves de encriptación, en lugar de usar llaves gestionadas por el proveedor.

**CLOUD Act:** Ley estadounidense (Clarifying Lawful Overseas Use of Data Act, 2018) que permite a autoridades de EE.UU. exigir datos almacenados por empresas americanas independientemente de la ubicación física de los servidores.

**Compliance:** Cumplimiento de regulaciones, estándares y políticas aplicables. En contexto de datos, refiere a adherencia a leyes de protección de datos y regulaciones sectoriales.

**Cerebro Orquestador:** Componente de una plataforma que gestiona y coordina los recursos de cómputo. Toma decisiones sobre dónde y cómo ejecutar cargas de trabajo.

**Data Plane:** Componente de una plataforma donde se procesan y almacenan los datos. A diferencia del control plane, ejecuta las cargas de trabajo en lugar de coordinarlas.

**Fine-tuning:** Proceso de ajustar un modelo de IA pre-entrenado con datos específicos para mejorar su rendimiento en tareas o dominios particulares.

**GPU (Graphics Processing Unit):** Procesador especializado originalmente diseñado para gráficos, ahora ampliamente utilizado para entrenar y ejecutar modelos de IA debido a su capacidad de procesamiento paralelo.

**gRPC:** Framework de comunicación de alto rendimiento utilizado para conectar servicios en arquitecturas distribuidas.

**GraphQL:** Lenguaje de consulta para APIs que permite a los clientes solicitar exactamente los datos que necesitan.

**Guardrails:** Sistemas de seguridad y filtrado que controlan las entradas y salidas de modelos de IA para garantizar que cumplan con políticas de uso, detecten contenido sensible, y eviten respuestas inapropiadas.

**Hyperscaler:** Proveedor de nube a escala masiva (AWS, Microsoft Azure, Google Cloud, Oracle Cloud). El término refiere a su capacidad de escalar infraestructura rápidamente.

**IaaS (Infrastructure as a Service):** Modelo de servicio de nube donde el proveedor ofrece recursos de cómputo, almacenamiento y red como servicio bajo demanda.

**Inferencia:** Proceso de usar un modelo de IA entrenado para generar predicciones o respuestas a partir de nuevos datos de entrada.

**LLM (Large Language Model):** Modelo de inteligencia artificial entrenado con grandes cantidades de texto para entender y generar lenguaje natural.

**Multi-tenant:** Arquitectura donde múltiples clientes comparten la misma infraestructura pero mantienen sus datos y configuraciones completamente aislados.

**Namespace:** Mecanismo de aislamiento lógico que permite separar recursos y cargas de trabajo de diferentes usuarios o proyectos en el mismo entorno.

**Observabilidad:** Capacidad de entender el estado interno de un sistema a partir de sus outputs externos (métricas, logs, trazas).



**On-premise:** Infraestructura de TI ubicada físicamente en las instalaciones de la organización, en contraste con servicios de nube.

**PaaS (Platform as a Service):** Modelo de servicio de nube donde el proveedor ofrece una plataforma para desarrollar, ejecutar y gestionar aplicaciones sin la complejidad de mantener la infraestructura.

**RAG (Retrieval-Augmented Generation):** Técnica que combina búsqueda de información con generación de texto, permitiendo a modelos de IA responder con información específica de documentos o bases de datos.

**REST/RESTful API:** Estilo de arquitectura para diseñar APIs web que utiliza métodos HTTP estándar (GET, POST, PUT, DELETE).

**Self-healing:** Capacidad de un sistema para detectar y recuperarse automáticamente de fallos sin intervención humana.

**Tenant:** En arquitecturas multi-tenant, cada cliente u organización que comparte la infraestructura pero mantiene sus datos y configuraciones aislados.

**TLS (Transport Layer Security):** Protocolo criptográfico que proporciona comunicaciones seguras sobre redes.

**Token:** En contexto de LLMs, unidad básica de texto procesada por el modelo. Aproximadamente equivale a 3/4 de una palabra en inglés o español.

**VPS (Virtual Private Server):** Servidor virtual que opera como un servidor dedicado pero comparte hardware físico con otros VPS.

**WebSocket:** Protocolo de comunicación que permite conexiones bidireccionales persistentes entre cliente y servidor, útil para aplicaciones en tiempo real.

**Zero-trust:** Modelo de seguridad que no confía en ningún usuario o dispositivo por defecto, requiriendo verificación continua independientemente de la ubicación en la red.