BLAM

# Por qué los bancos de LATAM no deberían construir IA desde cero

El argumento técnico, estratégico y económico para adoptar una plataforma especializada en lugar de construir capacidades conversacionales internamente, y cómo el BLAM cambia el punto de partida.

HERRAMIENTA
Whitepaper técnico-estratégico

LECTURA
14 minutos

FECHA
Abril 2026

**Hay una decisión que los equipos de tecnología e innovación de los bancos enfrentan cada vez con más frecuencia: ¿construimos internamente o adoptamos una plataforma especializada? Para la inteligencia artificial conversacional en banca, la respuesta es importante porque puede determinar muchas cosas, y construir desde cero tiene un costo que rara vez se calcula bien.**

---

Empecemos por un hecho que es incómodo reconocer: construir un agente conversacional bancario funcional, seguro, que cumpla normativa y que resuelva el 70% o más de los casos reales de clientes, es un proyecto de dos a tres años si se hace desde cero.

No porque la tecnología sea difícil de acceder, los modelos de lenguaje ya son accesibles a cualquier equipo técnico, sino porque los problemas difíciles en IA conversacional para banca no son tecnológicos. Son **de dominio, de datos, de normativa, de integración y de calibración continua.**

En este whitepaper explicamos por qué, y qué alternativa existe para los bancos que quieren moverse con velocidad sin sacrificar seguridad ni calidad.

PARTE 1

# El espejismo del "lo construimos nosotros"

Cuando un equipo técnico bancario evalúa construir capacidades conversacionales propias, el razonamiento inicial suena razonable: tenemos ingenieros, tenemos acceso a las APIs de los modelos de lenguaje más avanzados, conocemos nuestros propios sistemas mejor que cualquier proveedor externo, y queremos control total.

Cada uno de esos argumentos tiene algo de verdad, pero tiene una trampa. Tener ingenieros capacitados para usar APIs de LLMs no es lo mismo que tener el conocimiento acumulado sobre cómo esos modelos se comportan en conversaciones bancarias reales, a escala, en múltiples idiomas y dialectos del español latinoamericano, con clientes de diferentes perfiles de alfabetización digital, durante meses y miles de conversaciones por día.

Conocer los propios sistemas core no significa saber cómo integrarlos de manera que un agente conversacional pueda actuar sobre ellos en tiempo real sin generar riesgos operativos. La capa de integración es, en la práctica, uno de los problemas más costosos y lentos de resolver.

> **"La IA generativa hace que construir el 20% de la funcionalidad sea fácil. Lo que no cambió es que el 80% restante, el que importa en producción, sigue siendo igual de difícil que antes."**
>
> OBSERVACIÓN CONSISTENTE EN PROYECTOS DE IA CONVERSACIONAL EN BANCA, LATAM 2023–2025

Y el control total que prometía el desarrollo interno se convierte, en la práctica, en la **responsabilidad total:** mantenimiento del modelo, actualizaciones de seguridad, ajuste continuo por cambios regulatorios, monitoreo de alucinaciones, gobernanza de los outputs del agente. Eso requiere un equipo especializado permanente que pocos bancos tienen o quieren construir.

PARTE 2

# ¿Qué es el BLAM y por qué cambia el punto de partida?

El **Banking Large Action Model, BLAM,** es la respuesta a la pregunta: ¿qué pasa si en lugar de empezar desde cero, el banco empieza desde un modelo que ya fue entrenado, probado y calibrado específicamente para el contexto bancario latinoamericano?
El BLAM es una arquitectura que reúne más de 300 skills bancarias listas para usar. Una skill sería, por ejemplo, hacer una transferencia, responder una consulta de un préstamo, etc. No son respuestas predefinidas, automáticas, ni scripts, son habilidades especializadas que el agente puede encadenar para resolver situaciones completas: verificar identidad, consultar saldos, bloquear tarjetas, ejecutar pagos, ajustar límites, generar certificados, ofrecer productos según perfil, gestionar acuerdos de pago en cobranzas, y muchos casos más.

BLAM

BANKING LARGE ACTION MODEL · +300 SKILLS ESPECIALIZADAS

IDENTIDAD
Verificación biométrica

CUENTAS
Consulta de saldos

TARJETAS
Bloqueo y desbloqueo

PAGOS
Transferencias y pagos

COBRANZAS
Acuerdos de pago

CRÉDITO
Gestión de préstamos

INVERSIONES
Recomendación de perfil

CORPORATIVO
Tesorería y nómina

+300 Skills listas para implementar, validados en bancos de +15 países de Latinoamérica.

En términos prácticos, esto significa que cuando un banco adopta el BLAM, no empieza desde cero. Empieza desde un modelo que ya sabe que una consulta de "¿cuánto tengo disponible en mi tarjeta?" puede ser parte de una conversación más larga donde el cliente quiere hacer un pago, que esa conversación puede ocurrir a las 11 de la noche por WhatsApp, que el cliente puede estar frustrado porque ya intentó resolverlo antes, y que la respuesta correcta no es solo dar un número sino también anticipar la acción siguiente.
Eso no se construye en un sprint, se acumula en miles de conversaciones reales, en múltiples mercados, con feedback sistemático.

PARTE 3

## El argumento económico: construir vs. adoptar

El análisis de costo de construcción propia rara vez incluye todos los componentes reales. La comparación más común es "el costo de licencia de la plataforma vs. el costo de mis ingenieros". Eso es comparar el precio del menú con el costo de un ingrediente.
El costo real de construir internamente incluye:

**Ingeniería especializada**
No solo desarrollo inicial: mantenimiento continuo, actualizaciones por cambios en los modelos base, ajuste por cambios regulatorios, monitoreo de calidad. Esto es un equipo permanente de al menos 4 a 6 personas con conocimiento muy específico.

**Datos y calibración**
Para que un modelo funcione bien en conversaciones bancarias, necesita ser ajustado con datos reales de conversaciones bancarias. Generar esos datos, etiquetarlos, usarlos para calibrar el modelo y validar los resultados es un proceso que lleva meses y no termina.

**Tiempo de mercado**
El costo de estar 18 o 24 meses construyendo lo que ya existe no es solo el presupuesto de ingeniería: es el costo de oportunidad de no tener la capacidad mientras la competencia la despliega.

**Riesgo regulatorio**
Si el modelo produce outputs incorrectos en operaciones financieras reales, como una tasa equivocada, una instrucción de pago mal ejecutada, o un dato de compliance incorrecto, el costo reputacional y regulatorio puede ser significativamente mayor que cualquier ahorro en licencias.

| Semanas | 99% | +15 |
| --- | --- | --- |
| Tiempo al primer agente en producción con plataforma especializada | de cobertura de casos cliente–banco con el BLAM desde el inicio | países de LATAM con normativa local resuelta en la plataforma |

PARTE 4

# Casos de uso donde el impacto es más rápido

No todos los casos de uso generan el mismo retorno en el mismo tiempo. Después de implementaciones en bancos de múltiples países, identificamos tres áreas donde el impacto es más rápido y más medible.

### Atención al cliente retail

Consultas de saldo, movimientos, bloqueos de tarjeta, cambios de PIN, certificados. Alta frecuencia, alta estandarización. Primer caso de uso en el 80% de las implementaciones por velocidad de impacto.

### Cobranzas y recupero

Contacto proactivo, negociación de acuerdos de pago, recordatorios personalizados. ROI directo en tasa de contacto efectivo y costo por caso. Resultado visible en 60–90 días de implementación.

### Cobranzas y recupero

Contacto proactivo, negociación de acuerdos de pago, recordatorios personalizados. ROI directo en tasa de contacto efectivo y costo por caso. Resultado visible en 60–90 días de implementación.

CASO · BANCA CORPORATIVA

Un cliente corporativo puede ir desde "¿Cuál es mi saldo consolidado hoy?" a "Cotízame tipo de cambio para transferir USD 50.000 al exterior" y "Necesito ayuda con la carga de la nómina de sueldos" en la misma conversación, con el mismo agente que mantiene el contexto de su empresa, sus productos contratados y sus últimas operaciones. Sin transferencias, sin repetición de información.

PARTE 5

# La pregunta de la seguridad y el cumplimiento

En cada conversación con equipos de innovación bancaria, la pregunta de seguridad y cumplimiento aparece como debe aparecer. Y la respuesta que escuchamos con más frecuencia es: "preferimos construir internamente para tener control total sobre los datos".
Ese razonamiento tiene un supuesto implícito que vale la pena cuestionar: que la construcción interna garantiza mayor seguridad que una plataforma especializada que opera sobre infraestructura enterprise (como Azure) con estricta gobernanza de datos.
La realidad es que la seguridad no es proporcional al control sobre el código fuente. Es proporcional a la **madurez de las prácticas de seguridad, la velocidad de respuesta ante vulnerabilidades, la profundidad de las auditorías y la consistencia de los controles.** Esas capacidades son difíciles de construir internamente y relativamente fáciles de exigir y auditar en un proveedor especializado.

PRINCIPIO DE ARQUITECTURA SEGURA

Privacidad de datos, cumplimiento normativo por país, auditoría de interacciones, gobernanza de los modelos de IA y monitoreo de outputs en tiempo real no son features que se agregan: son restricciones de diseño que determinan la arquitectura completa. Una plataforma que no fue diseñada con estas restricciones desde el inicio no puede cumplirlas bien después.

PARTE 6

# Cómo implementar: el modelo de adopción por fases

La adopción de una plataforma conversacional especializada no implica una migración completa desde el día uno. El modelo que funciona en la práctica es progresivo, con impacto medible en cada fase.

FASE 1 · SEMANAS 1–8

## Primer agente en producción

Caso de uso de alto volumen y alta estandarización: consultas de tarjeta, saldo, movimientos. Integración con sistemas core. Primeras métricas de resolución y satisfacción. El objetivo es tener un agente en producción real, no en piloto cerrado, en menos de 8 semanas.

FASE 2 · SEMANAS 9–20

## Expansión de dominio y canales

Incorporación de casos de uso adyacentes: crédito, cobranzas, onboarding. Expansión a canales adicionales con contexto compartido. Las métricas de la fase 1 guían qué se prioriza en la fase 2.

FASE 3 · MES 6 EN ADELANTE

## Escala y personalización profunda

El agente tiene cobertura de +70% de los casos. Se incorporan capacidades de personalización basadas en perfil del cliente, ventas cruzadas, alertas proactivas y análisis predictivo de necesidades. La plataforma aprende de cada interacción.

LO QUE ACORTA LOS TIEMPOS

La velocidad de implementación depende más de la claridad en la definición del problema y la disponibilidad para integración con sistemas core que de cualquier factor tecnológico. Los bancos que llegan con un caso de uso definido, criterios de éxito claros y soporte ejecutivo para las integraciones técnicas reducen los tiempos de implementación entre un 40 y 60%.

PARTE 7

## El momento: ¿por qué importa actuar ahora?

Hay una tendencia que los datos de la industria validan consistentemente: la brecha entre los bancos que están construyendo capacidades conversacionales reales y los que están en modo evaluación **se amplía cada trimestre, no cada año.**

La razón es que las capacidades conversacionales generan datos propios que alimentan mejora continua. Un banco que tiene un agente en producción procesando 100.000 conversaciones por mes tiene un flujo de información sobre el comportamiento de sus clientes, sus puntos de fricción y sus necesidades que no tiene el banco que todavía está evaluando proveedores.

Esa asimetría de información se convierte en asimetría competitiva, y se acumula. La tecnología subyacente, los modelos de lenguaje, es accesible para todos. El conocimiento acumulado sobre cómo aplicarla en el contexto específico de la banca latinoamericana, a escala, con seguridad y con resultados medibles, no lo es.

CONCLUSIÓN

## La decisión de arquitectura define la velocidad estratégica

La pregunta "¿construimos o adoptamos?" para capacidades conversacionales en banca tiene una respuesta diferente en 2026 de la que habría tenido en 2020. En 2020, construir era la única opción viable si querías algo realmente especializado. En 2026, hay plataformas que resolvieron los problemas difíciles del dominio bancario y los tienen disponibles desde el primer día de implementación.

Lo que queda por resolver es siempre específico de cada institución: la integración con sus sistemas particulares, la adaptación al tono y cultura de su marca, la calibración a las necesidades de sus clientes. Eso sí requiere trabajo conjunto. Pero partir de una base de +300 skills bancarias probadas, con normativa latinoamericana resuelta, es completamente diferente que construir desde los fundamentos.

Para los CTOs, CIOs y líderes de transformación digital de bancos e instituciones financieras de la región, la pregunta estratégica no es si la IA conversacional va a ser parte del modelo operativo, eso ya está resuelto. La pregunta es **cuánto tiempo y capital de innovación van a invertir en construir lo que ya existe,** versus invertirlo en diferenciarse en lo que realmente importa: la experiencia única de sus clientes.

SIGUIENTE PASO

### ¿Cómo se vería implementar el BLAM en tu institución?

Podemos hacer un análisis de viabilidad específico para tu stack tecnológico, tus casos de uso prioritarios y tu contexto regulatorio local. Sin compromiso, con foco en lo que realmente importa para vos.

Hablemos ↗ Ver plataforma ↗

**Delto** es una plataforma SaaS especializada en agentes conversacionales de IA generativa para banca e instituciones financieras. El **Banking Large Action Model (BLAM)** es una arquitectura propietaria que reúne más de 300 skills bancarias diseñados específicamente para el contexto latinoamericano. Con más de 15 años de experiencia en la industria financiera y presencia en más de 15 países, Delto opera sobre infraestructura cloud enterprise (Azure) con estricta gobernanza de datos y cumplimiento normativo.

*Datos de implementación y benchmarks basados en proyectos operativos de clientes Delto, 2022–2026.*

delto | CONVERSATIONAL BANKING SUITE



A HAUSSA COMPANY

delto.com