Cómo evaluar si tu negocio necesita IA privada

Una guía para decidir con criterios, no con entusiasmo

Para dueños y equipos que atienden clientes todos los días

> *Tus datos no salen de tu empresa. Ni siquiera cuando la IA trabaja para ti.*

Edición 2026 · Versión 1.0
contacto@iaprivada.mx
iaprivada.mx

Documento de criterios. No sustituye asesoría legal ni una propuesta comercial.

Contenido

01 Señales de que tu equipo paga el costo de responder lo mismo

02 Checklist de 10 puntos

03 Matriz de decisión: nube vs. IA privada local

04 Costos típicos y errores comunes

05 Plan piloto de 30 días

06 Cinco preguntas para cualquier proveedor

07 Preguntas frecuentes

08 Glosario breve

→ Resumen y contacto

LA GUÍA EN 30 SEGUNDOS

- Si no puedes medir cuántas veces se repite una pregunta, todavía no hay caso.
- Si tus datos son sensibles o la LFPDPPP no está documentada, resuelve eso primero.
- Nube para lo público y de bajo volumen; local para lo sensible y de alto volumen.
- Prueba 30 días con 40 preguntas reales y criterios de aceptación; luego decides.
- Antes de firmar: costo a 12 meses, borrado de datos y forma de salir.

Edición 2026 · Versión 1.0 · Se actualiza cuando cambian los costos de referencia o la normativa. Sugerencias y dudas: contacto@iaprivada.mx

Antes de empezar

En iaprivada.mx la conversación empieza casi siempre igual: alguien pregunta si ya necesita IA, si es carísima o si sus datos van a terminar en un servidor ajeno. La respuesta honesta es que depende: de tu volumen de consultas, de qué tan ordenada está tu información y de qué tan sensibles son los datos que atiendes. Esta guía no vende una solución; ordena los criterios para que decidas con números y no con moda.

Está pensada para dueños, gerentes de operación y responsables de atención en negocios que viven de documentos, catálogos, trámites o consultas repetitivas: despachos, clínicas, inmobiliarias, escuelas, talleres, distribuidores. Si tu operación se parece a esa lista, vale leerla con tu equipo antes de pedir cotizaciones.

CÓMO USAR ESTA GUÍA

- **Señales (1):** comprobar si el problema existe y cuánto cuesta hoy.
- **Checklist (2):** saber si tu caso está listo para intentarlo.
- **Matriz (3):** elegir nube o local sin dogmas.
- **Piloto (5) y proveedores (6):** cómo probar y cómo no comprar de más.
- **Preguntas frecuentes (7):** las dudas que recibimos todos los días.

01

Señales de que tu equipo paga el costo de responder siempre las mismas preguntas

El costo de responder lo mismo no aparece en el estado de resultados. Aparece en horas, en clientes que se van y en empleados que se cansan. Estas seis señales lo vuelven visible.

- **Las mismas preguntas se repiten.** "¿Tienen X en negro?", "¿cuánto cubre mi plan?", "¿qué requisitos lleva esta compraventa?", "¿a qué hora abren?". Si tu equipo contesta lo mismo cinco veces al día, cada respuesta cuesta tiempo que nadie está vendiendo. No es un problema de actitud: es un problema de diseño del servicio.
- **Cada persona contesta distinto.** Cuando cada quien responde su versión, el cliente recibe tres verdades distintas. La marca pierde credibilidad sin que nadie lo note.
- **El conocimiento vive en una sola cabeza.** Si la persona que sabe se va de vacaciones —o se va, punto— la operación pierde velocidad y el equipo improvisa.
- **Las consultas fuera de horario se pierden.** Cada mensaje de las 21:00 que se contesta a las 9:00 del día siguiente es una venta que decidió el horario, no el precio.
- **El equipo se desgasta.** Contestar lo mismo todo el día baja la moral, y el desgaste se nota en la calidad de la atención. Rotar gente en puestos de atención es carísimo.
- **La incorporación de personal nuevo tarda semanas.** Un empleado nuevo aprende a fuerza de preguntar lo que otros saben de memoria. Cada contratación paga dos veces el costo del conocimiento.

> **MEDIRLO EN UNA SEMANA**
>
> **Paso 1.** Abre una hoja con tres columnas: pregunta, minutos que tardó y dónde buscaste la respuesta.
>
> **Paso 2.** Llena una fila cada vez que alguien contesta algo que ya se contestó antes.
>
> **Paso 3.** El viernes haz la cuenta: consultas repetitivas por semana × minutos promedio ÷ 60 = horas por semana. Esas horas × costo por hora = lo que cuesta hoy.
>
> **Ejemplo ilustrativo:** 35 consultas repetitivas al día de 6 minutos son 3.5 horas diarias, cerca de 77 horas al mes. A 180 pesos por hora, unos 13,800 pesos mensuales en solo repetir.

Cuando el número aparece, la conversación cambia: de "¿deberíamos usar IA?" a "¿cuánto nos cuesta no hacerlo?". La señal real es una combinación: consultas que se repiten, un costo por respuesta que no baja y una operación que no puede delegar. Si las tres coinciden, hay un caso que vale medir con seriedad.

LO QUE NO ES SEÑAL DE NADA

- **Tener muchos documentos.** Sin orden, no sirven de nada. Más PDFs desordenados solo hacen más lento el problema.
- **Que un competidor lo anunció.** Su operación no es la tuya; su volumen y sus riesgos tampoco.
- **Que todos hablan de IA.** La moda no es criterio de compra. El criterio es el número que calculaste arriba.

02

Checklist de 10 puntos

Diez preguntas, cada una con lectura verde, amarillo o rojo. No pesan igual: tres son eliminatorias y están marcadas en la columna de lectura.

#	PUNTO	QUÉ REVISAR	LECTURA
1	Datos sensibles	¿Hay expedientes, datos de salud, contratos, precios no públicos o datos personales de clientes?	**Eliminatorio.** Si es el corazón del servicio, la privacidad no es opcional
2	LFPDPPP	Base legal, aviso de privacidad, consentimiento y quién queda como encargado de datos.	**Eliminatorio.** Sin esto documentado, rojo. Confirma con tu especialista legal
3	Volumen repetitivo	Consultas al mes y porcentaje que se repite.	Menos de 100 al mes rara vez justifica la inversión
4	Disponibilidad 24/7	¿Cuántas consultas llegan fuera de horario y cuántas se pierden?	Si la pérdida es real, pesa mucho en la decisión
5	Calidad documental	¿Hay manuales, políticas y catálogo? ¿Están actualizados? ¿Quién los mantiene?	**Eliminatorio.** Rojo si el conocimiento está en la cabeza de alguien
6	Criticidad del error	¿Qué pasa si el asistente se equivoca? ¿Se pierde dinero, se daña a alguien, hay consecuencia legal?	Define cuánta verificación y cuántas citas exigir
7	Dueño interno	¿Quién valida respuestas y actualiza documentos?	Sin dueño, el asistente envejece en meses
8	Integración	¿WhatsApp, web, correo, CRM? ¿Hay API o forma de conectarlo?	Un asistente que nadie consulta no sirve
9	Presupuesto y mantenimiento	¿Hay disposición a pagar piloto y mantenimiento mensual?	El mantenimiento no desaparece aunque el piloto termine
10	Medición	¿Hay métricas y alguien dispuesto a leerlas?	Sin medición no hay decisión, solo opiniones

CÓMO CALIFICAR

VERDES	LECTURA
7 a 8	Vale explorar con seriedad: tu caso tiene volumen, datos y dueño.
4 a 6	Piloto acotado a un solo caso de uso, con alcance escrito.
0 a 3	Primero ordena el conocimiento: ningún modelo arregla documentos que no existen.

Los tres eliminatorios son datos sensibles sin protección definida, LFPDPPP sin base legal documentada y documentación que no existe o está desactualizada. Si uno está en rojo, el problema no es la IA: todavía no hay nada que leer. En iaprivada.mx el diagnóstico revisa exactamente estos puntos antes de cotizar.

> **UN PUNTO QUE CASI NADIE REVISA**
>
> - ¿Quién actualiza los documentos cuando cambia un precio, una política o un requisito? Si la respuesta es “nadie”, el asistente nacerá desactualizado el día uno. El contenido no es un archivo que se entrega una vez: es un proceso que alguien tiene que cuidar.

0 3

Matriz de decisión: IA en la nube vs. IA privada local

La pregunta correcta no es cuál es más barata, sino cuál soporta tus restricciones. Compara los dos caminos con los mismos criterios que usarías para elegir un proveedor de nómina o un sistema de facturación: no eliges el más moderno, eliges el que puedes mantener.

CRITERIO	IA EN LA NUBE	IA PRIVADA LOCAL
Datos sensibles	Viajan al proveedor y se procesan en sus servidores	Nunca salen de tu infraestructura
LFPDPPP	El proveedor queda como encargado externo; se documenta	Tú sigues siendo responsable; no hay encargado adicional
Costo	Por token o conversación: crece con el uso	Inversión inicial más energía; después, costo fijo
Volumen alto	Se encarece y te ata al proveedor	Se amortiza con el uso intensivo
Tono y dominio	Limitado; entrenar con tus datos es el problema	Natural: el modelo se ajusta con tus ejemplos
Disponibilidad	Depende del internet y del proveedor	Corre en tu red, 24/7, sin depender de afuera
Arranque	Días	Semanas
Mantenimiento	Lo lleva el proveedor	Requiere a alguien que lo mantenga
Control de cambios	Depende de las versiones que decida el proveedor	Tú decides qué modelo usar y cuándo actualizarlo

CUÁNDO LA NUBE SÍ JUSTIFICA

- Tu contenido es público o poco sensible: catálogo, horarios, preguntas frecuentes.
- El volumen es bajo —menos de 200 consultas al mes— y no tienes equipo de sistemas que pueda cuidar un servidor.
- Quieres validar el caso de uso antes de invertir en infraestructura propia.
- Puedes tolerar que el proveedor vea las conversaciones y lo tienes documentado.

CUÁNDO LA IA PRIVADA LOCAL SÍ JUSTIFICA

- Atiendes expedientes, contratos, datos de salud o información financiera de tus clientes.
- Tienes un requisito —legal, contractual o de negocio— de no designar un encargado externo.
- El volumen es constante y haría crecer la factura de nube sin techo.
- Necesitas que el asistente hable como tu equipo y cite tus documentos.
- Quieres que siga funcionando si falla el internet o si el proveedor cambia de precio.

EL HÍBRIDO, QUE CASI NADIE TE CUENTA

No es una indecisión: es una arquitectura. Lo público puede vivir en la nube y lo sensible quedarse en casa, con el mismo criterio de calidad y el mismo set de preguntas. Muchas operaciones terminan ahí, y no es una salida elegante: es una decisión de diseño.

TRES ESCENARIOS TÍPICOS

ESCENARIO	SEÑAL CLAVE	LECTURA
Clínica dental: 300 consultas al mes, datos de pacientes	Datos de salud más volumen medio	Local: el expediente no sale de la clínica
Inmobiliaria: 80 consultas al mes, catálogo público	Contenido público y volumen bajo	La nube puede bastar; local si el catálogo es confidencial
Despacho notarial: 600 consultas al mes, requisitos documentales	Volumen alto y respuestas que deben citar el documento	Local: se amortiza y se puede verificar cada respuesta

El error de tratarlo como religión: ni la nube es pecado ni lo local es santo. Un servidor sin respaldo, sin cifrado y sin control de accesos es peor que una nube bien configurada. La privacidad es una decisión de arquitectura, no una etiqueta. Y al revés: la nube no te exime de la LFPDPPP.

0 4

Costos típicos y errores comunes

Los costos se comparan mal casi siempre: se pone el precio del piloto junto a un cero. La comparación correcta es contra el costo de seguir como estás, que ya calculaste en la sección 1.

COMPONENTE	QUÉ INCLUYE	CÓMO ESTIMARLO
Nube	Plataforma y consumo de tokens	Consultas al mes × costo promedio por conversación
Local: equipo	Computadora de oficina o mini-PC con GPU dedicada	Cotización por caso; no compres antes de medir
Desarrollo del piloto	Base documental, ajuste del modelo, interfaz	Días de trabajo × tarifa, con alcance escrito
Mantenimiento	Actualizar documentos, monitorear, ajustar	Mensual y fijo; pregunta por escrito
Energía	Consumo del equipo local	Depende del nivel que necesites
Capacitación	Que el equipo lo use bien desde el día uno	Una o dos sesiones de tu gente

Como referencia del mercado: un uso moderado en la nube —varios miles de conversaciones al mes— suele ubicarse entre 200 y 400 dólares mensuales entre plataforma y consumo, y crece con el volumen. En local, la inversión inicial se amortiza con el uso intensivo: una computadora de oficina puede bastar para recepción; con varios usuarios y más documentos conviene un equipo con GPU dedicada. Las cifras exactas dependen de tu caso y se cotizan después del diagnóstico, nunca antes.

El costo de no hacer nada es el número de la sección 1. Si te sale mayor que el piloto más el mantenimiento anual, la decisión ya no es técnica: es de operación.

ERRORES QUE SE REPITEN

ERROR	POR QUÉ DUELE	CÓMO EVITARLO
Comprar el equipo antes de medir	Se invierte en máquina y el problema sigue siendo el desorden	Primero el set de preguntas y la línea base
Empezar por el caso más difícil	El piloto se eterniza y nadie ve resultados	Un caso repetitivo y acotado
Documentos desordenados	Basura entra, basura sale	Limpiar e indexar antes de conectar
No definir qué hace cuando no sabe	Inventar es peor que no saber	Regla explícita: “no está en mis documentos, te paso con alguien”
No medir antes ni después	No hay forma de saber si mejoró	Línea base en la semana 1 del piloto
Confundir el bot gratis de WhatsApp	No conoce tus documentos ni tu tono	Usar ambos con roles claros: público y privado
Subestimar el mantenimiento	Cada cambio de precio o política lo deja obsoleto	Dueño interno y calendario mensual de revisión
No prever la salida	Quedas atado sin plan ni exportación	Pedir borrado y exportación por contrato

Casi todos estos errores se evitan con la misma cosa: un piloto acotado con criterios escritos. No es burocracia; es lo que separa una compra informada de una apuesta.

05

Plan piloto de 30 días con entregables verificables

Un piloto de 30 días no es un sistema en producción: es una prueba con criterios de aceptación que puedes auditar. Si no tiene criterios, es una demostración con diapositivas. Estas son las cuatro semanas, con lo que debe quedar por escrito en cada una.

SEMANA 1 — ALCANCE Y LÍNEA BASE

- Un solo caso de uso, escrito y firmado por ambas partes.
- 40 preguntas reales con su respuesta esperada y el documento de donde sale.
- Inventario documental y métricas actuales de tiempo y volumen.
- **Entregables:** documento de alcance, set de evaluación y línea base.

SEMANA 2 — BASE DOCUMENTAL Y PRIMER PROTOTIPO

- Limpieza e indexación de los documentos del caso.
- Primer asistente respondiendo sobre una muestra real, no sobre ejemplos inventados.
- **Entregables:** demo funcional y bitácora de fallas.

SEMANA 3 — EVALUACIÓN CIEGA

- Dos revisores que no participaron en el desarrollo evalúan las 40 preguntas.
- **Entregables:** reporte con porcentaje de aciertos, respuestas con cita, “no sé” correctos, tiempo de respuesta y lista priorizada de errores.

SEMANA 4 — DESPLIEGUE CONTROLADO Y DECISIÓN

- Grupo piloto de 3 a 5 personas y registro de conversaciones.
- Capacitación de una hora para el grupo.
- **Entregables:** asistente operando en el grupo, informe contra la línea base y decisión go/no-go documentada.

CRITERIOS DE ACEPTACIÓN

CRITERIO	UMBRAL SUGERIDO	POR QUÉ IMPORTA
Aciertos en el set	85% o más	Mide utilidad real, no entusiasmo
Respuestas factuales con cita	100%	Se pueden verificar contra el documento
Preguntas trampa	Cero respuestas inventadas	Es la base de la confianza
Adopción del grupo piloto	70% o más lo usa	Si nadie lo usa, no sirve

LAS TRES DECISIONES POSIBLES

- **Avanzar:** los criterios se cumplen y el grupo lo adoptó. Se escala con un plan por escrito.
- **Avanzar con condiciones:** falta contenido, faltan integraciones o falta capacitación. Se corrige antes de escalar.
- **No avanzar:** el caso no era el correcto o la calidad no alcanza. Se documenta la razón y se ahorra el año siguiente.

Un no-go bien documentado también vale la pena. El piloto acotado alcanza para un caso de uso, no para toda la empresa. En iaprivada.mx el diagnóstico sin costo dura 60 minutos y termina en una propuesta con alcance escrito en 48 horas.

0 6

Cinco preguntas que deberías hacer a cualquier proveedor

No son para interrogar: son para comparar. La diferencia entre dos propuestas casi siempre está en lo que no dicen.

PREGUNTA	RESPUESTA QUE DA CONFIANZA	SEÑAL DE ALERTA
¿Dónde se procesan mis datos y quién más los toca?	Subprocesadores, retención, registros y si se usan para entrenar	Responde “en la nube” sin detalle
¿Qué pasa con mis documentos cuando termino el contrato?	Borrado, exportación y plazos por escrito	No hay respuesta o “se pueden borrar si lo pides”
¿Cómo evalúan la calidad y me dejan ver los resultados?	Set de prueba, criterios y resultados crudos	Solo dice “funciona muy bien”
¿De qué depende mi continuidad?	Modelo de código abierto, portabilidad y quién mantiene	Todo corre en su API y no hay salida
¿Cuál es el costo total a 12 meses, con mantenimiento y salida?	Desglose por escrito, no un rango verbal	Solo el precio del piloto

Pide una prueba con tus propios documentos antes de firmar. Es la única evaluación que no depende de lo que el proveedor diga de sí mismo.

UNA SEXTA PREGUNTA DE REGALO

- ¿Puedo ver el asistente funcionando con mis documentos y mis preguntas antes de pagar el piloto? Si la respuesta es no, no es un piloto: es una conferencia.

07

Preguntas frecuentes

Estas son las dudas que recibimos todos los días en iaprivada.mx. Las respuestas están en palabras simples, sin prometer lo que no se puede prometer. Si la tuya no está aquí, escríbenos a contacto@iaprivada.mx: respondemos con una pregunta de vuelta, no con un discurso.

¿Mis datos salen de mi empresa?

En IA privada local, no. El modelo corre en tu infraestructura: los documentos y las conversaciones se procesan en tu equipo o en la red que tú controlas, y nada viaja a una nube ajena. En la nube, sí: cada documento y cada conversación se procesa en servidores del proveedor, con la retención que él defina. Por eso la primera pregunta de cualquier diagnóstico es dónde viven tus datos.

¿Necesito saber de tecnología?

No. El proveedor instala, ajusta y mantiene; tú usas el asistente. Lo que sí necesitas es un dueño interno que valide respuestas y mantenga los documentos al día: sin esa persona, cualquier asistente envejece en meses.

¿Cuánto cuesta?

Depende del alcance. En la nube pagas por consumo y la cuenta crece con el volumen; en local pagas equipo, piloto y mantenimiento, y después un costo fijo. Por eso no damos un precio antes de un diagnóstico: la diferencia entre cotizar y adivinar es un piloto que no sirve.

¿Cuánto tarda?

Un piloto funcional en semanas, no en meses. Un despliegue completo depende del número de documentos, usuarios e integraciones; se plantea después del piloto, con números reales en la mano.

¿Y el agente de IA gratis que ya trae WhatsApp?

Sirve para lo básico: responder con guiones simples. No conoce tus documentos internos, no habla tu tono y las conversaciones se procesan en la nube de Meta. Puedes usar los dos: el gratis para lo público y el privado para lo que sí importa.

¿Qué equipo necesito? ¿Necesito una computadora cara?

No. Para recepción puede bastar una computadora de oficina sin tarjeta gráfica; con más usuarios y más documentos conviene un equipo compacto con GPU dedicada. El diagnóstico dice qué nivel necesita tu caso antes de que compres nada.

¿Puedo empezar pequeño y crecer?

Sí, y es lo recomendable. Un solo caso, un grupo piloto y una decisión documentada. Crecer es agregar usuarios y documentos, no rehacer el proyecto ni cambiar de proveedor.

¿Qué pasa si el asistente se equivoca?

Se mide y se corrige. Un asistente serio cita el documento de donde sacó la respuesta y, cuando no sabe algo, lo dice en vez de inventar. Las preguntas que fallan se revisan, se ajustan o se pasan a un humano. Ningún sistema debe contestar lo que no está en tus documentos.

¿Habla en español y con mi forma de hablar?

Sí. Un asistente privado se ajusta con ejemplos de cómo tu equipo trata a sus clientes: tus términos, tus saludos, lo que sí dices y lo que nunca dices. No es una traducción de un modelo genérico.

¿Qué pasa con la LFPDPPP y el aviso de privacidad?

La ley te sigue haciendo responsable de los datos de tus clientes. En local no designas un encargado externo porque nadie más los procesa; en la nube sí, y eso se documenta. En ambos casos el aviso de privacidad debe reflejar cómo se tratan los datos. Confirma el detalle con tu especialista legal: nosotros lo dejamos por escrito en la arquitectura.

¿Sirve si mis documentos están en PDF desordenados?

Sí, pero conviene limpiarlos primero. Un asistente no arregla un catálogo con precios contradictorios: los contradice con más velocidad y más confianza. Ordenar el material es parte del proyecto, no un trámite previo.

0 8

Glosario breve

Para que nadie se pierda en la junta. Si un proveedor usa una palabra de esta lista sin explicarla, pide la explicación: es parte de su trabajo.

TÉRMINO	EN PALABRAS SIMPLES
IA privada local	Un modelo que corre en tu equipo o en tu red, sin salir a internet.
RAG	El asistente busca en tus documentos antes de responder y cita de dónde sacó cada dato.
Fine-tuning (SFT)	Ajustar el modelo con ejemplos de tu tono y tu gremio para que suene a tu equipo.
DPO	Ajustar el comportamiento con preferencias: ser breve, no prometer de más y saber derivar a un humano.
Token	La unidad de texto que el modelo procesa; en la nube se cobra por ella.
On-prem	En tus instalaciones, no en la nube del proveedor.
Encargado de datos	El tercero que procesa datos personales por encargo tuyo, según la LFPDPPP.
Piloto	Prueba acotada con criterios de aceptación y una decisión documentada al final.

Resumen para llevar

- Mide antes de comprar: consultas repetitivas × minutos ÷ 60 × costo por hora.
- Si hay datos sensibles o LFPDPPP sin base legal, resuelve eso primero.
- Nube para lo público y de bajo volumen; local para lo sensible y de alto volumen.
- Un piloto de 30 días con 40 preguntas y criterios de aceptación vale más que una demo bonita.
- Antes de firmar: costo a 12 meses, borrado de datos y forma de salir.

¿Quieres revisarlo para tu caso?

Agenda un diagnóstico gratuito en iaprivada.mx

contacto@iaprivada.mx

Diagnóstico de 60 minutos, sin costo · propuesta con alcance escrito en 48 horas

Este documento es una guía de criterios y no sustituye asesoría legal, fiscal ni una propuesta comercial. Las cifras son ilustrativas y se cotizan caso por caso. Para más información y casos de uso, consulta iaprivada.mx o escribe a contacto@iaprivada.mx.