AI Congress Barcelona

20 - 21 - 22 OCT 2026

Auditori L'illa

Un esdeveniment de:

CIDAI Centre of Innovation for Data tech and Artificial Intelligence

Organitzat per:

eurecat

Alineat amb l'Estratègia Catalunya IA 2030:

Generalitat de Catalunya

Segueix-nos a:

#AICongressBCN

aicongress.barcelona

Patrocinadors OR:

CaixaBank TECH

conversia compliance is our business

forvis mazars & infini DIGITAL EFFICIENCY

orange Empresas

raona coding corporate futures

softeng Maximize the cloud. Securely.

Patrocinadors PLATA:

apdcat Autoritat Catalana de Protecció de Dades

LABORAL kutxa

Col·laboradors:

Barcelona

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CVC Centre de Visió per Computador

i2cat

UPC Universitat Politècnica de Catalunya Barcelona Tech

IDEAI Intelligent Data Science and Artificial Intelligence Research Center

AI Congress Barcelona

Un esdeveniment de: CIDAI Centre of Innovation for Data tech and Artificial Intelligence

Organitzat per: eurecat

Alineat amb l'Estratègia Catalunya IA 2030: Generalitat de Catalunya

20 - 21 - 22 OCT 2026
#AICongressBCN
aicongress.barcelona



Jaume Clotet
Partner d'AI & Digital Solutions
Forvis Mazars - Infini

PONÈNCIA:

LA IA SOTA CONTROL: COM DESPLEGAR IA GENERATIVA AMB PRIVACITAT, TRANSPARÈNCIA I CONTROL DE COSTOS



Presenta:

Mireya Saludes
Responsable de desenvolupament de negoci i d'ecossistema
CIDAI

Patrocinadors OR: CaixaBank TECH · conversia compliance is our business · forvis mazars & infini DIGITAL EFFICIENCY · orange Empresas · raona coding corporate futures · softeng Maximize the cloud. Securely.

Patrocinadors PLATA: apdcat Autoritat Catalana de Protecció de Dades · LABORAL kutxa



La IA bajo control

Cómo desplegar IA Generativa con privacidad, transparencia y control de costes

Jaume Clotet · AI Congress Barcelona · 20 de octubre de 2026

forvis mazars & infini DIGITAL EFFICIENCY

Cada día, tus equipos envían datos a una IA.

¿Sabes **dónde** se procesan, **quién** puede verlos y **cuánto** te cuestan?



Foto: Pexels

Tus datos ya están saliendo

El 26,4 % de los ficheros que los empleados suben a herramientas de IA generativa contiene datos sensibles

26,4 %

de los ficheros subidos a IA generativa contiene datos sensibles



22 % en el trimestre anterior: **+4,4 puntos en 3 meses**



% del contenido sensible detectado

57 % Negocio y legal

35 % contratos y políticas · 35 % fusiones, adquisiciones y previsiones financieras

25 % Datos técnicos

65 % es código fuente propietario pegado para depurar o refactorizar

15 % Datos personales

Nombres, direcciones, expedientes de RR. HH. y nóminas

12 % **de las exposiciones llega desde cuentas personales de IA** (versiones gratuitas de ChatGPT, Gemini, Claude o Meta AI). El historial se queda en la cuenta del empleado, aunque deje la empresa.

Fuente: Harmonic Security (nov-2025). >3M prompts y subidas en 300 herramientas de IA, EE. UU. y Reino Unido, jul–sep 2025.

01

El contexto



IA pública vs IA privada

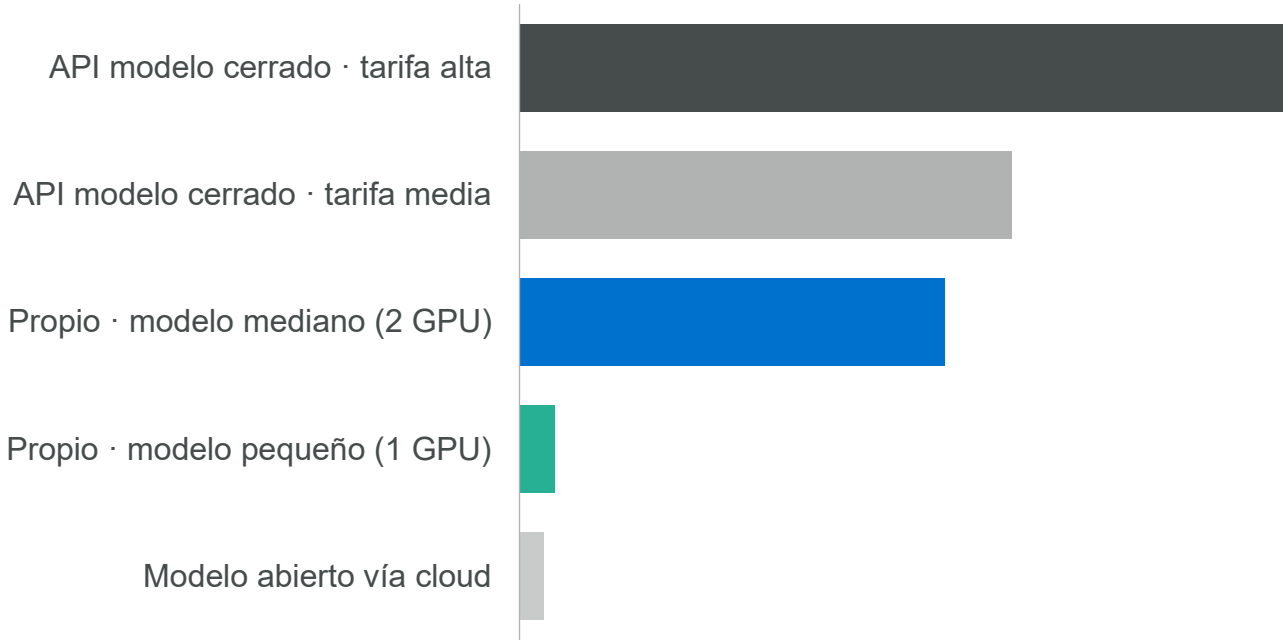
Misma tecnología, distinto lugar de ejecución y distinto dueño del control

IA pública	IA privada
En los servidores del proveedor	Dentro del perímetro de la organización
El proveedor actualiza o retira versiones	Versión fijada, auditable y reproducible
Modelo tokens OPEX	CAPEX + OPEX
Alta dependencia en precios y condiciones del proveedor	Baja dependencia
Modelos más potentes	A unos meses de distancia
GPT, Gemini, Claude, Copilot	Gemma 4, Mistral, Qwen, gpt-oss, ALIA

Modelos de costes

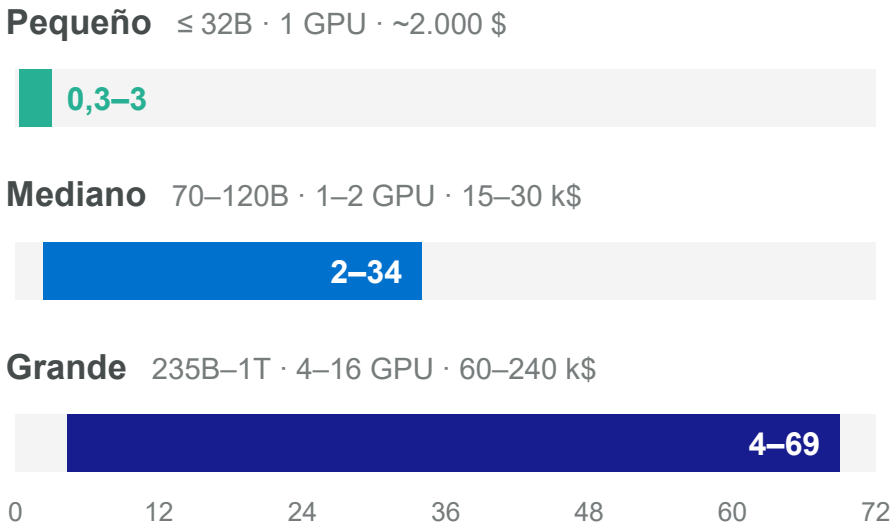
Pagar por token o invertir en capacidad: depende del volumen y de con qué lo comparas

Coste mensual relativo a igual volumen (orientativo)



- Azul y verde: los datos no salen de la organización
- Grises: los datos se procesan en un proveedor externo

Meses para amortizar el hardware propio frente a APIs de modelos cerrados (54 escenarios)



Frente a un modelo abierto vía cloud, el hardware propio no se amortiza por coste, se justifica por privacidad, soberanía y control.

Fuentes: Precios públicos de API 2025–2026. Proporciones orientativas: el coste real depende de modelo, proveedor y contrato.

De consumir IA a operarla

La pregunta ya no es si usar IA generativa, sino dónde se ejecuta y quién la controla

96,7 %

coincidencia de rendimiento entre el mejor modelo cerrado y el mejor modelo abierto

≈ 4 meses

es lo que tarda un modelo abierto en alcanzar el nivel de la frontera cerrada

1/2

de las grandes empresas ya ejecuta o planea ejecutar su IA en producción en nube privada

Fuentes: Stanford HAI (2026) AI Index Report; Epoch AI (2026); Harmonic Security (2025), GenAI Data Exposure Report Q3.

02

La solución



Plataforma corporativa de IA on-prem

LLMs, agentes y aplicaciones ejecutados dentro del perímetro de la organización

Perímetro de la organización

Aplicaciones	Chat corporativo, copilotos y automatización de procesos
Agentes y orquestación	Flujos con herramientas (MCP) y supervisión humana
Conocimiento (RAG)	Búsqueda vectorial sobre documentos propios, con permisos
Modelos	Gemma 4 · Mistral · Qwen · gpt-oss · ALIA + embeddings
Inferencia	vLLM · SGLang · llama.cpp: batching y PagedAttention
Hardware	GPUs en CPD propio o colocation

Gobierno transversal

- Identidad y roles
- Registro auditable
- Guardrails
- Medición de costes

API externa (opcional)

Solo datos no sensibles o picos de demanda, vía gateway

Microsoft Foundry: modelos abiertos como servicio

Las mismas capas que el on-prem, pero operadas por Microsoft en Azure



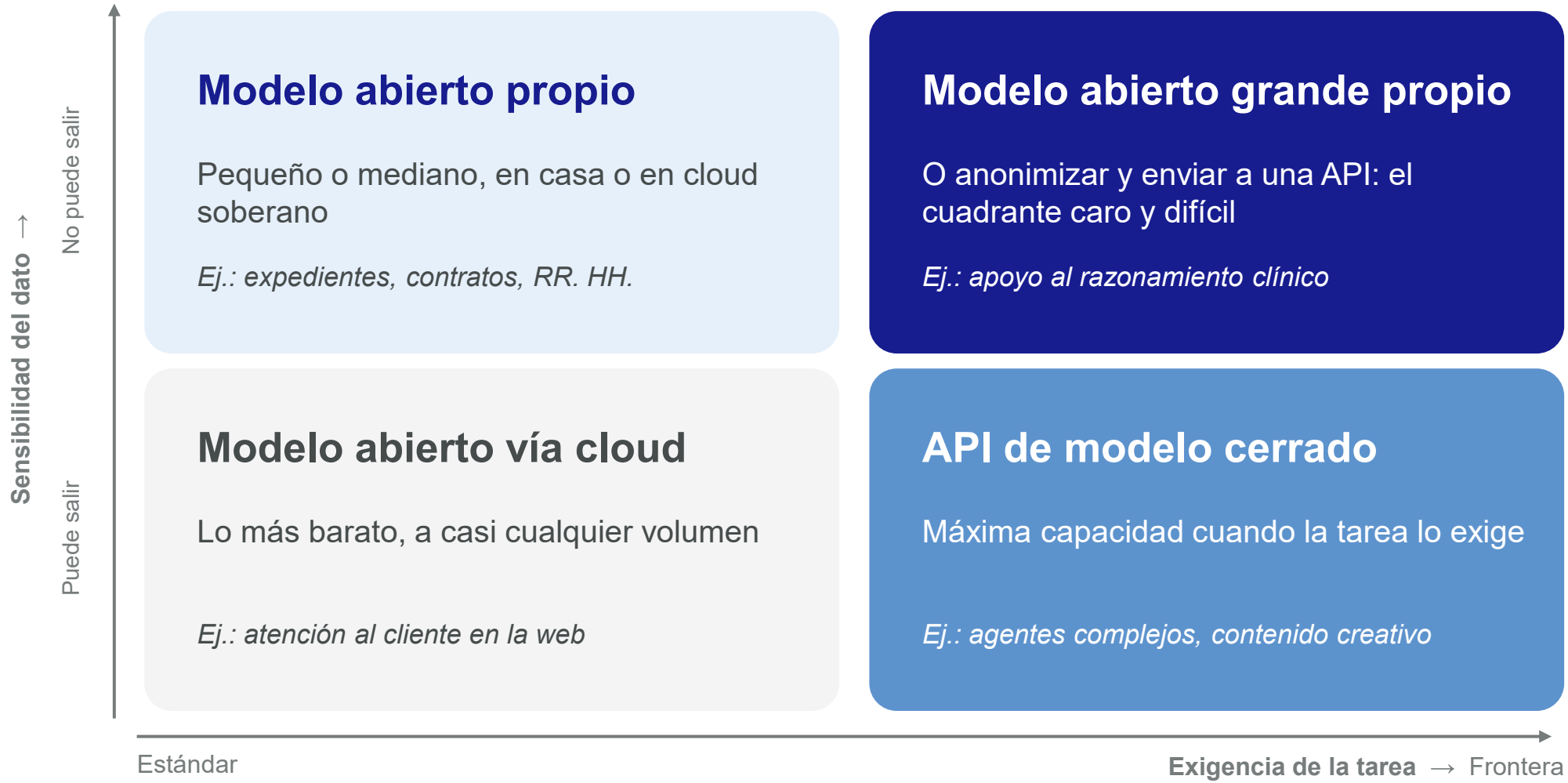
Lo que garantiza Microsoft: en los modelos vendidos por Azure (p. ej. DeepSeek), tus prompts no se comparten con el proveedor ni se usan para entrenar. Con Data Zone UE se procesan solo en la UE.

Lo que no cambia: el dato sale de tu perímetro y queda bajo contrato y jurisdicción de Microsoft. En la escala de soberanía: cloud con residencia UE.

Fuentes: Microsoft Learn (2026), Deployment types for Foundry Models; Data, privacy and security for Foundry Models.

¿Cuándo tiene sentido?

Dos preguntas deciden: ¿puede salir el dato? y ¿qué exige la tarea?



Criterio práctico: una tarea es estándar si un modelo abierto la resuelve bien con tus propios datos de prueba.

Caso 1 · Hospital universitario

Caso tipo ilustrativo

Apoyo al razonamiento clínico sobre la documentación del propio hospital

El reto

Los profesionales quieren apoyo para razonar sobre casos complejos a partir de protocolos, guías y documentación clínica. Los datos de salud son categoría especial (art. 9 RGPD): no pueden salir a una API externa.



La solución

- Modelo abierto grande en un clúster de GPU, dentro del CPD del hospital
- RAG sobre protocolos y guías, respetando los permisos de cada servicio
- Filtro de datos personales y respuestas con cita a la fuente
- Profesional en el circuito: la IA propone, el médico decide

Lo que hemos visto

Privacidad

Ningún dato clínico sale del perímetro del hospital.

Transparencia

Cada respuesta cita su protocolo y queda registrada con modelo y versión.

Costes

Inversión alta: se justifica por privacidad y uso intensivo, no por precio.

Normativa: RGPD art. 9 · AI Act: alto riesgo si actúa como dispositivo médico.

Caso 2 · Ayuntamiento

Caso tipo ilustrativo

Un asistente para tramitar expedientes en castellano y catalán

El reto

El personal municipal tramita expedientes con datos de la ciudadanía: buscar normativa, redactar borradores, resumir. Tareas estándar, presupuesto ajustado y obligación de cumplir el Esquema Nacional de Seguridad.



La solución

- Modelo pequeño en 1 GPU, en el CPD municipal o en un cloud soberano
- Modelo con soporte de lenguas cooficiales (p. ej. familia ALIA del BSC)
- Búsqueda en normativa municipal y plantillas de resolución
- Aviso de interacción con IA y revisión humana de cada resolución

Lo que hemos visto

Privacidad

Datos de la ciudadanía bajo el ENS y sin transferencias fuera de la UE.

Transparencia

Aviso de IA y registro de qué documento se consultó en cada expediente.

Costes

Inversión mínima: para una tarea estándar basta un modelo pequeño en 1 GPU.

Normativa: ENS (RD 311/2022) · RGPD · AI Act art. 50 (transparencia).

Caso 3 · Empresa de servicios

Caso tipo ilustrativo

Tres necesidades, tres destinos: el router decide

El reto

Una misma empresa necesita atención al cliente en la web, revisión de contratos y contenido creativo de marketing. Cada caso combina una sensibilidad y una exigencia distintas.



La solución

1 · Atención al cliente: modelo abierto vía cloud, con filtro de datos personales

2 · Contratos: modelo abierto propio, los documentos no salen

3 · Marketing creativo: API de modelo cerrado, máxima calidad

Router y gobierno común: identidad, registro y coste por caso de uso

Lo que hemos visto

Privacidad

Cada dato va solo donde su sensibilidad lo permite.

Transparencia

Un único registro para los tres flujos, sea cual sea el modelo.

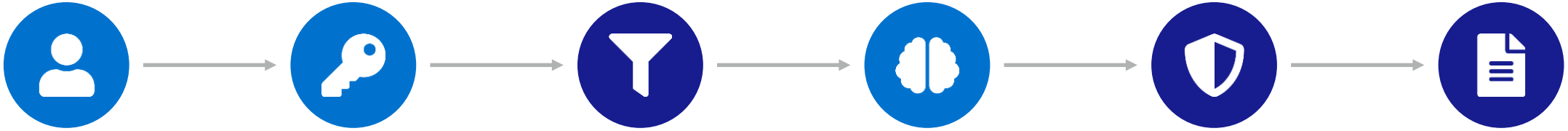
Costes

El volumen estándar va a la nube abierta; la API cerrada, solo donde aporta.

Normativa: AI Act art. 50 (chatbot con clientes) · RGPD art. 5.1.c (minimización).

Privacidad y transparencia por diseño

Cada petición recorre un circuito controlado y deja un rastro auditable



Usuario

Identificado con el SSO corporativo

Gateway

Control de acceso por rol y por caso de uso

Filtro de entrada

Detecta y anonimiza datos personales

Modelo + RAG

Responde solo con documentos que el usuario puede ver

Filtro de salida

Cita la fuente y filtra contenido

Registro

Guarda prompt, contexto, modelo y versión

Qué obligación cubre cada pieza

Trazabilidad

Registro de eventos (art. 12 AI Act) y responsabilidad proactiva (art. 5.2 RGPD)

Transparencia

Informar de que se interactúa con una IA (art. 50 AI Act)

Minimización

Solo salen del origen los datos necesarios (art. 5.1.c RGPD)

Supervisión humana

Revisión en decisiones relevantes (art. 14 AI Act y art. 22 RGPD)

Evidencia: la generación aumentada con recuperación reduce las alucinaciones frente a generar sin fuentes (Shuster et al., Findings of EMNLP 2021).

forvis mazars & infini DIGITAL EFFICIENCY

Conclusiones

La IA bajo control no es renunciar a la nube, es decidir dónde se ejecuta cada caso



Privacidad

Los datos sensibles no salen del perímetro de la organización.



Transparencia

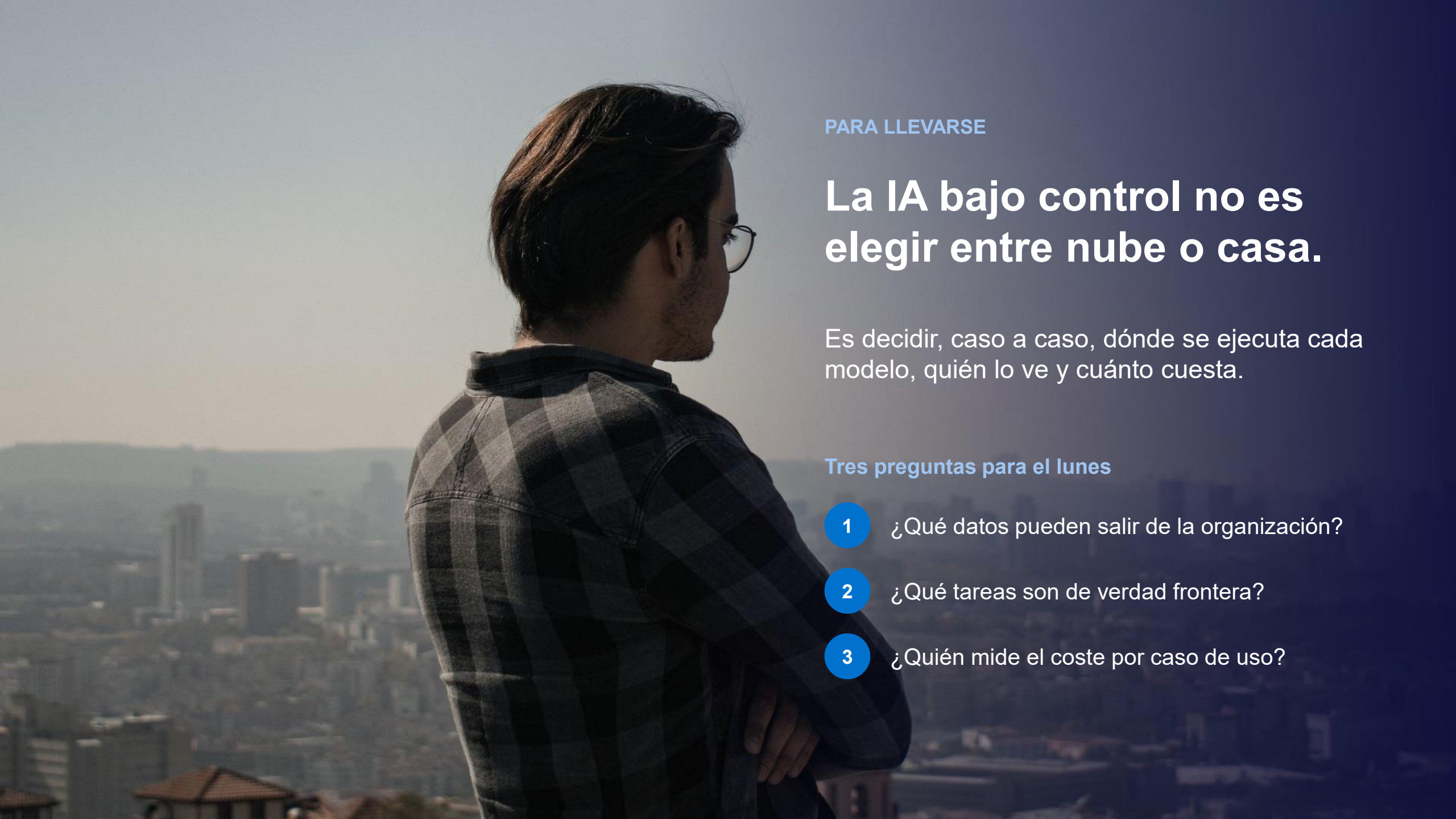
Cada respuesta es trazable: usuario, modelo, versión y fuentes.



Costes

Con volumen sostenido, pagar por capacidad sale más barato que pagar por token.

Los modelos abiertos ya están en la frontera. El cuello de botella ya no es la tecnología es el gobierno.



PARA LLEVARSE

La IA bajo control no es elegir entre nube o casa.

Es decidir, caso a caso, dónde se ejecuta cada modelo, quién lo ve y cuánto cuesta.

Tres preguntas para el lunes

1. ¿Qué datos pueden salir de la organización?
2. ¿Qué tareas son de verdad frontera?
3. ¿Quién mide el coste por caso de uso?

GRACIAS

Jaume Clotet
Infini - AI & Digital Solutions - Forvis Mazars

AI Congress Barcelona

20 - 21 - 22 OCT 2026

Auditori L'illa

Un esdeveniment de:

CIDAI Centre of Innovation for Data tech and Artificial Intelligence

Organitzat per:

eurecat

Alineat amb l'Estratègia Catalunya IA 2030:

Generalitat de Catalunya

Segueix-nos a:

#AICongressBCN

aicongress.barcelona

Patrocinadors OR:

CaixaBank TECH

conversia compliance is our business

forvis mazars & infini DIGITAL EFFICIENCY

orange Empresas

raona coding corporate futures

softeng Maximize the cloud. Securely.

Patrocinadors PLATA:

apdcat Autoritat Catalana de Protecció de Dades

LABORAL kutxa

Col·laboradors:

Barcelona

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CVC Centre de Visió per Computador

i2cat

UPC Universitat Politècnica de Catalunya Barcelona Tech

IDEAI Intelligent Data Science and Artificial Intelligence Research Center