

VISTAGE · SANTIAGO · 22 DE OCTUBRE DE 2026

IA abierta y local: ya no es un sacrificio

Qué hay que decidir sobre modelos, proveedores y la frontera física de la IA

Marcelo José Rovai · UNIFEI · AIEng4D



Dos trayectorias, una misma pregunta

HASTA 2017

En la industria

- CEO y VP Latinoamérica en IGT/GTECH
- AT&T, NCR, Dell y Compaq
- Gestión de tecnología: P&L, contratos y proveedores



HOY

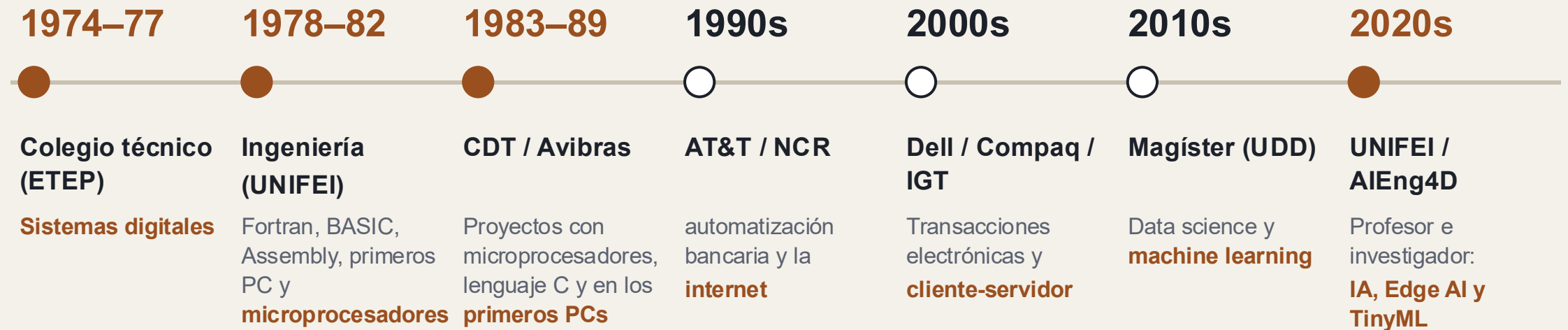
En la academia

- Profesor Honoris Causa, UNIFEI (Brasil)
- Edge AI: IA en chips de bajo costo, fuera del data center
- Co-chair AIEng4D y EDGE AI Foundation IAP



No dirigí una empresa de IA. Conozco los dos lados de la mesa.

Cinco décadas, siete olas tecnológicas

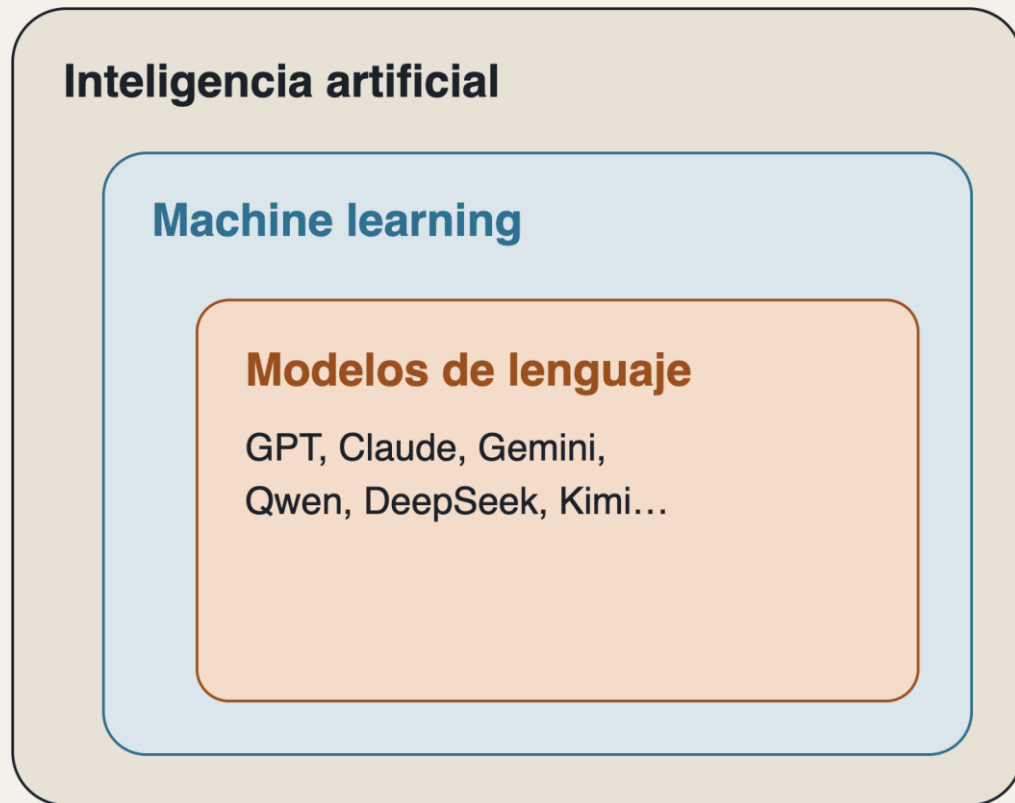


Empecé programando microprocesadores; hoy pongo IA en microcontroladores. El círculo se cierra.

LA PREGUNTA DE HOY

Si su proveedor de IA se apaga el lunes, ¿qué deja de funcionar en su empresa el martes?

Tres términos, uno dentro del otro



IA

Sistemas que hacen tareas que asociamos con inteligencia humana.

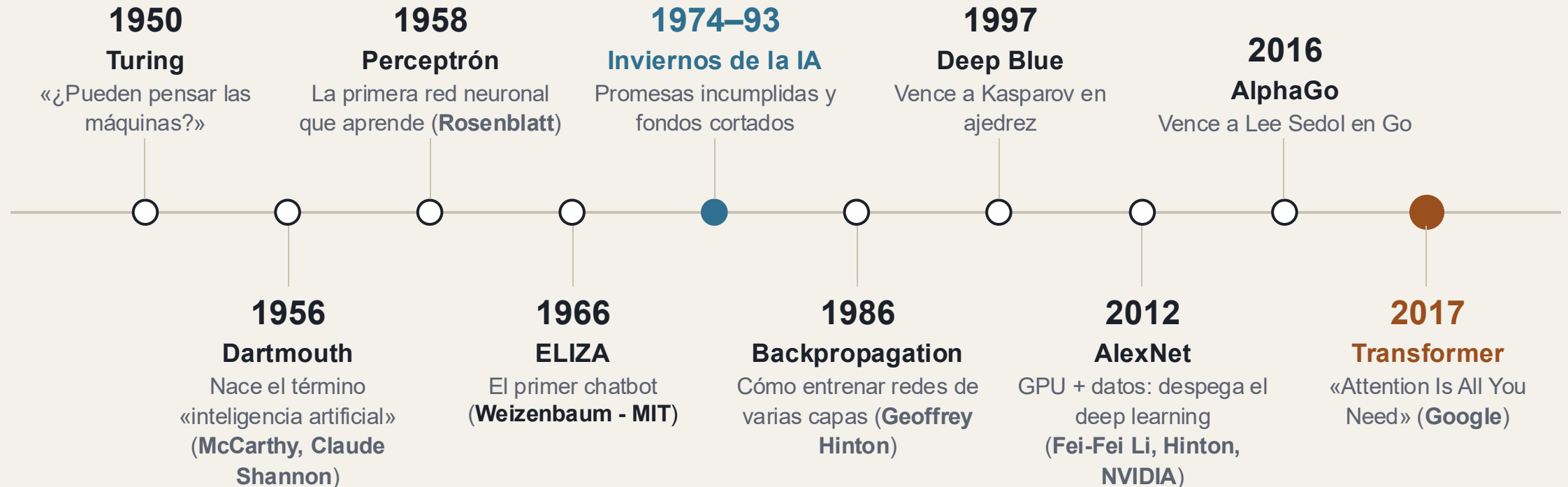
Machine learning

Aprende patrones de los datos. Ya está en su empresa: fraude, demanda, mantención.

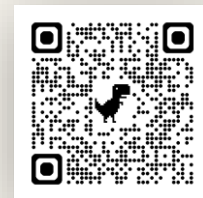
Modelos de lenguaje (LLM)

Entrenados con texto a escala masiva: leen, escriben, programan y usan herramientas. Lo que cambió desde 2022.

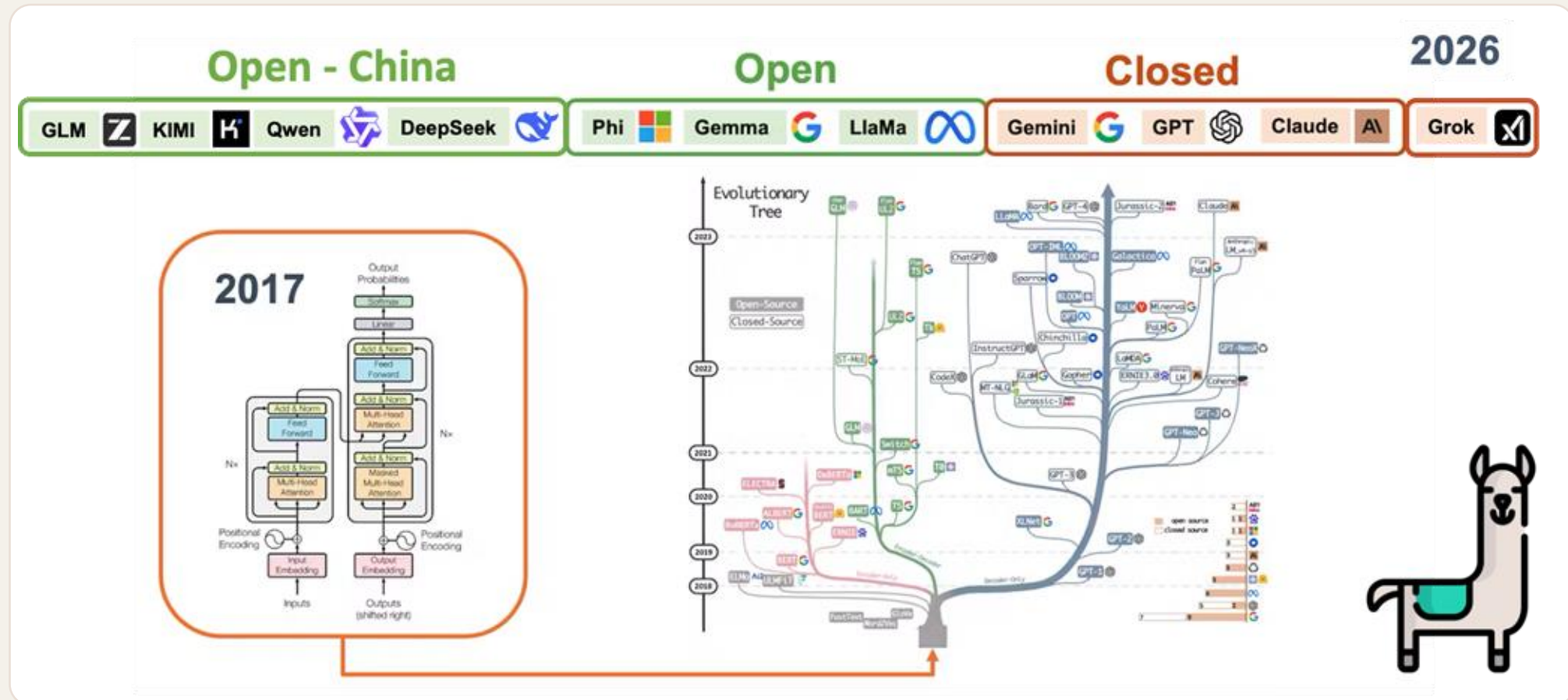
De Turing al Transformer, en diez hitos



Tres ingredientes lo cambiaron todo: datos, cómputo (GPU) y una arquitectura que escala.

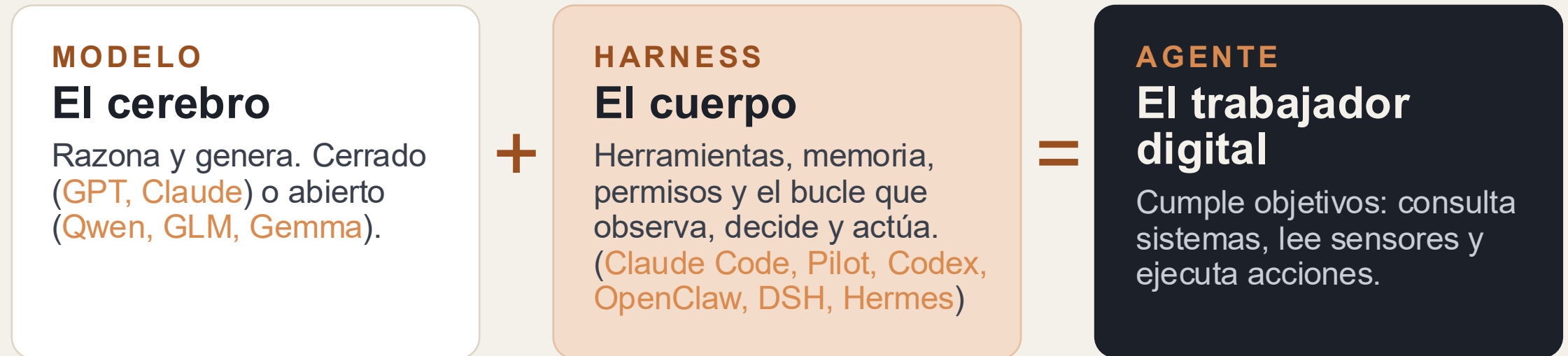


Una arquitectura, un paisaje entero



Modelo + harness = agente

Antes, asistentes: pregunta → respuesta. Hoy, agentes: objetivo → loop de acciones.



Quien controla el harness controla el agente. Con un harness propio, el modelo pasa a ser una pieza intercambiable.

Por dentro de un agente que uso todos los días

DeepSeek Harness (DSH): el modelo razona; el harness controla

YO Pido en lenguaje natural · apruebo los pasos riesgosos · evalúo el resultado



HARNESS · el software que rodea al modelo

Contexto de la sesión

- ✓ En qué carpeta trabaja
- ✓ Qué puede modificar
- ✓ Qué acciones necesitan mi aprobación

El ciclo del agente



Se repite hasta terminar la tarea

MODELO

Deepseek V4.1 flash

- ✓ Razona y elige las herramientas
- ✓ Nunca toca un archivo por sí mismo

Ejecutor de herramientas cada acción pasa por las reglas de seguridad, y cada resultado se revisa

Archivos

Shell

Búsqueda

Web

Subagentes

Entregables

El harness es donde la empresa pone sus reglas.

deepseekHARNESST

New Session

PluginsAutomation tasks

Workspaces

xiao_ips_display01-ebooksJules_VerneClass_15Clour-PaletteProject-teste

Introduce yourselfnowExplicar Transformer para ...1hSimple Greeting1hAplicativo 3D do míssil FO...1dEntão, bora pra próxima aq...7dShow 38 more sessions

Signed in to DeepSeek

Introduce yourselfStandard mode

ChatTrajectory

Introduce yourself10:32

Completed in 2s

Hi! I'm a coding agent running in DeepSeek Harness, powered by the deepseek-flash model.

What that means in practice:

- I work on your files — reading, writing, and running commands inside this workspace (Project-teste), with network access for research when needed.
- I can plan and delegate — multi-step tasks get a todo list, and large parallel work can be fanned out to subagents.
- I can build deliverables — code, plus Office documents, spreadsheets, presentations, and PDFs via the office skill family.
- I use your GUI — this session is served at http://127.0.0.1:49196; changes to the web shell or client plugins need a rebuild plus a page refresh to show up.

A few operating notes: I verify commands by exit code and inspect files before editing them; file writes are sandboxed to this workspace; and operations needing wider access will ask for your approval first.

What would you like to work on?

Usage 9.3K tok10:32

Message or run a task, / commands, @ files or sessions

+Workspace WriteDeepSeek-V4.1-Flash1017High

1 turns 1 steps · 202 tok/s9.3K tok · Cache hit 92%1%



Dos decisiones independientes

	NUBE · data center	BORDE · Edge AI
CERRADO solo vía API	GPT, Claude, Gemini Máxima capacidad. El proveedor controla precio, versiones y datos.	IA del fabricante Corre en el equipo (teléfono, PC), pero es una caja negra.
ABIERTO pesos descargables	Qwen, DeepSeek, GLM, Kimi, Muse En su nube o con el proveedor que usted elija.	Modelos locales En servidores, laptops, cámaras o máquinas. Sin internet, sin costo por token.

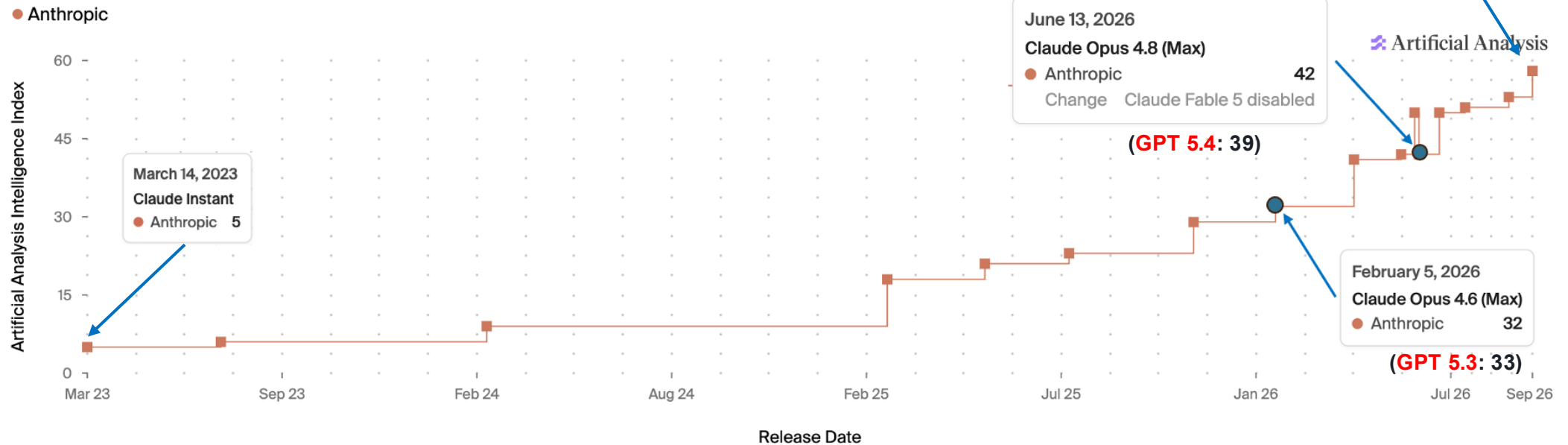
Abierto vs. cerrado: qué cambia para la empresa

	Cerrado (solo API)	Abierto (pesos descargables)
Dónde corre	Servidores del proveedor	Donde usted decida
Quién cambia el modelo	El proveedor: precio, versión, retiro	Usted congela la versión que funciona
Sus datos	Salen de la empresa, bajo contrato	Pueden no salir nunca
Capacidad	La frontera	Unos meses detrás de la frontera
Lo que exige	Gobierno de proveedores	Equipo técnico propio o un integrador

La frontera de Anthropic, mes a mes

Frontier Language Model Intelligence, Over Time

Artificial Analysis Intelligence Index v4.3.2 incorporates 10 evaluations: AA-Briefcase v1.1, GDPval-AA v2.1, AutomationBench-AA, Terminal-Bench 4.0, SciCode, Humanity's Last Exam, GDP.pdf, CritPt, AA-Omniscience, AA-LCR v1.1

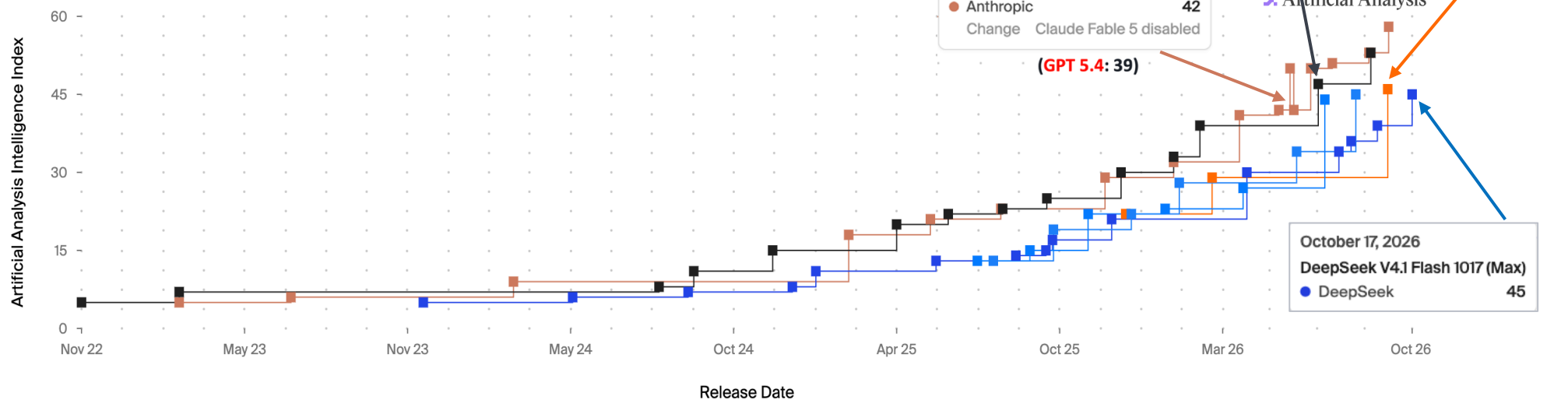


Los abiertos alcanzan a los cerrados en pocos meses

Frontier Language Model Intelligence, Over Time

Artificial Analysis Intelligence Index v4.3.2 incorporates 10 evaluations: AA-Briefcase v1.1, GDPval-AA v2.1, AutomationBench-AA, Terminal-Bench 4.0, SciCode, Humanity's Last Exam, GDP.pdf, CritPt, AA-Omniscience, AA-LCR v1.1

● Anthropic ● OpenAI ● Xiaomi ● DeepSeek ● Z AI ● Kimi



Lo que era frontera hace ocho meses, hoy corre en un Mac

■ **Cerrado (Anthropic)** ■ **Abierto** · Artificial Analysis Intelligence Index

Claude Opus 4.6

frontera en feb 2026

Qwen3.8 27B

abierto hoy · un Mac de 32 GB



32



34

Claude Opus 4.8

frontera en jun 2026

GLM-5.3-Flash

abierto hoy · un Mac Studio de 256 GB



42



42

Claude Opus 5.5

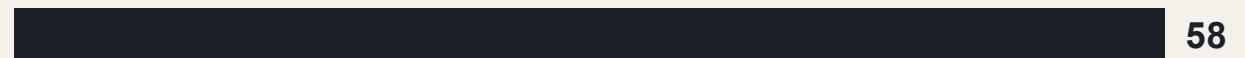
frontera hoy

MiMo-V2.6-Pro

el mejor abierto · necesita un clúster

DeepSeek V4.1 Flash

abierto · la mitad de tamaño · un Mac Studio de 512 GB



58



46



45

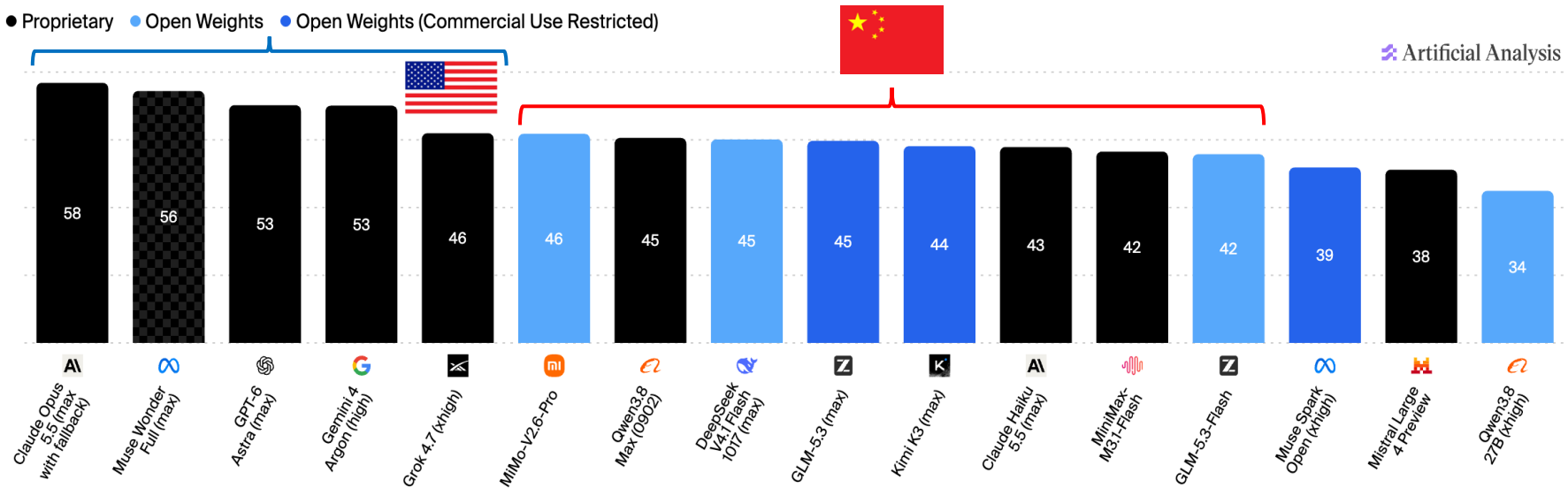
Los mejores cerrados son de **EE.UU.**; los mejores abiertos, de **China**

Artificial Analysis Intelligence Index by Open Weights / Proprietary

Artificial Analysis Intelligence Index v4.3.2 incorporates 10 evaluations: AA-Briefcase v1.1, GDPval-AA v2.1, AutomationBench-AA, Terminal-Bench 4.0, SciCode, Humanity's Last Exam, GDP.pdf, CritPt, AA-Omniscience, AA-LCR v1.1

■ No publicly available API

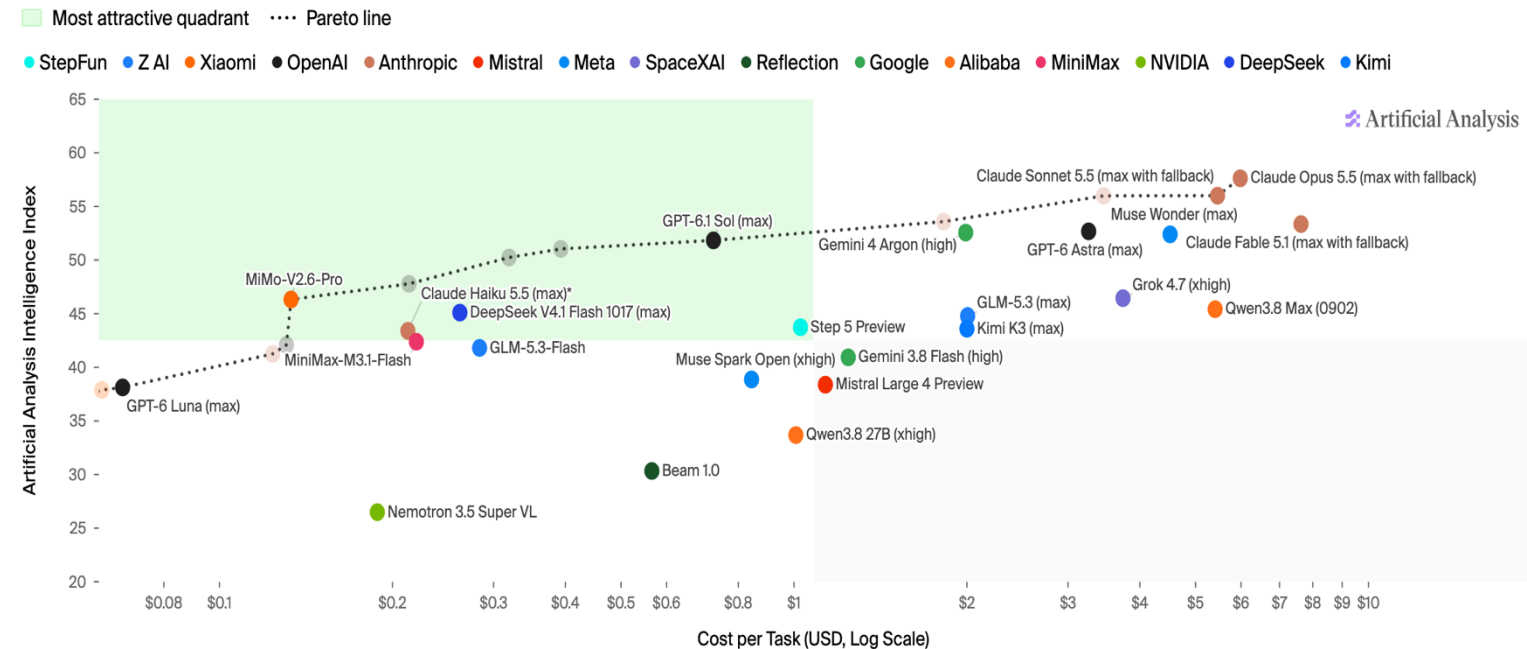
● Proprietary ● Open Weights ● Open Weights (Commercial Use Restricted)



Inteligencia vs. costo por tarea

Intelligence Index vs. Cost per Task

Artificial Analysis Intelligence Index - Weighted average cost (USD) per Artificial Analysis Intelligence Index task



El costo de la inteligencia se desploma

280×

más barato, a igual capacidad

Un modelo de nivel GPT-3.5 pasó de US\$20 a US\$0,07 por millón de tokens en dos años (nov 2022 – oct 2024).

Stanford AI Index 2025

15×

más barato, a igual inteligencia

Índice 42: GLM-5.3-Flash (abierto) ≈US\$0,28 por tarea; Claude Opus 4.8 (cerrado) ≈US\$4,10.

Artificial Analysis v4.3.2, oct 2026 · valores aprox.

Para el CEO, la pregunta deja de ser «¿podemos pagar la IA?» y pasa a ser: «¿Qué tarea merece el modelo más caro?».

De la placa de US\$79 al escritorio de US\$10.799

US\$79

Arduino UNO-Q

Un agente que usa
herramientas, sin internet.



~US\$1.300

Laptop

Asistente local para
documentos y código.



~US\$2.700

Mac mini 64 GB

Modelos medianos para
un equipo de trabajo.



US\$10.799

Mac Studio 256 GB

La inteligencia de la
frontera de junio.



Sin API key, sin límite de uso, sin cobro por token.

Precios EE.UU., ago 2026; en Chile sumar IVA e importación

Abierto no significa solo «en casa»

1 · VÍA API DE TERCEROS

Cambia de proveedor, no de modelo

El mismo modelo abierto lo sirven varios proveedores: si uno sube el precio o falla, se migra sin reescribir nada.

2 · EN SU NUBE PRIVADA

Datos dentro del perímetro

Capacidad elástica, con los datos y la versión del modelo bajo su control.

3 · EN CASA O EN EL BORDE

Control total, sin internet

Conviene si los datos no pueden salir, si el equipo ya existe o si el agente trabaja muchas horas al día.

Costo marginal de una misma tarea de agente de código (25 turnos)

≈US\$0,074

Modelo abierto (GLM-5.3-Flash)
hospedado por un tercero

≈US\$0,003

Mac Studio propio: solo electricidad

≈US\$0,003

API de DeepSeek con caché, en horario valle

Un modelo, muchos agentes: una cocina con un solo chef

DeepSeek V4.1 Flash en 4 × DGX Spark (~US\$29.000) frente a la API de DeepSeek

1 AGENTE

~50–90 tok/s

Se siente rápido. Con un solo agente, cualquier equipo serio sirve.

16 AGENTES A LA VEZ

~15 tok/s

Cada uno. Son 247 tok/s en total, repartidos entre todos.

API DE DEEPSEEK

~209 tok/s

Cada agente recibe la velocidad completa. En otros proveedores, entre 60 y 220.

~50 s vs. ~3 s

Una respuesta típica: en casa con 16 agentes, y en la API

124.000 → 760

Tokens que el agente lee para escribir cada respuesta

98%

De esa lectura debe salir de la caché; si no, cada turno tarda minutos

Lo difícil no es que el modelo quepa: es atender a muchos agentes al mismo tiempo.

¿Cuándo se paga un servidor propio?

Caso real: 111 horas de agentes de código en paralelo, en un equipo con cuatro tarjetas RTX PRO 6000

LA MISMA CARGA, VÍA API

US\$86

en la API de DeepSeek, horario
valle: unos US\$19 al día

≈US\$3.500

la misma carga en Claude Opus 5.5
(estimación propia)

Años para recuperar el hardware frente a la API

4× RTX PRO 6000 · US\$64.000

un usuario intensivo

~9 años

4× DGX Spark + switch

~US\$35.000 · un usuario, 16 agentes

~5–6 años

4× DGX Spark compartido

un equipo, 24 horas al día

~4 años

2× DGX Spark · ~US\$14.000

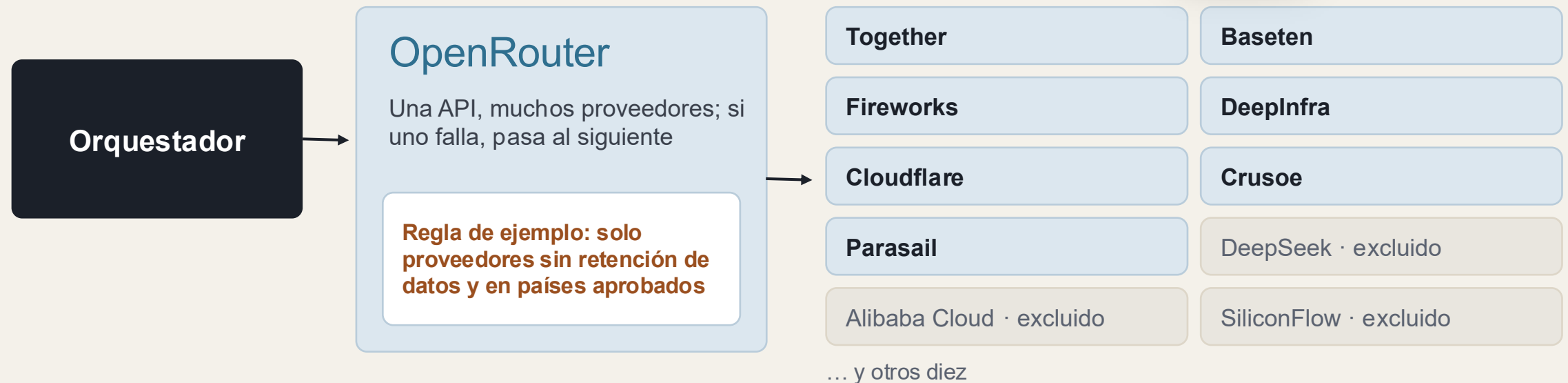
lo mismo que ~6 años de Claude Max

~6 años

No se compra por ahorro: se compra por soberanía de datos o porque un equipo lo mantiene ocupado. Antes de comprar, pruebe la API.

Vía API no significa vía China

DeepSeek V4.1 Flash 1017 en [OpenRouter](#): un modelo abierto, servido por 20 proveedores



20

proveedores sirven el mismo modelo

US\$0,07–0,30

por millón de tokens de entrada, según proveedor (salida: US\$0,18–1,20)

99,93%

de disponibilidad en 3 días con enrutamiento, contra 92,45% sin él

Abierto significa poder elegir también quién lo sirve y bajo qué reglas.

Los líderes de la frontera todavía pierden dinero

Datos de 2025, según documentos financieros reportados por Reuters y el Financial Times (aún no auditados)

Anthropic

2,76

US\$ gastados por cada US\$1 de ingreso

8,1

mil millones de US\$ de pérdida operacional

518

mil millones de US\$ en cómputo a 10 años, ~80% fijo

OpenAI

2,60

US\$ gastados por cada US\$1 de ingreso

20,9

mil millones de US\$ de pérdida operacional

600

mil millones de US\$ en cómputo hasta 2030

No es una predicción de quiebra: es una razón para tener un plan B. Si el proveedor sube los precios o cambia los modelos, su negocio no debería depender de esa decisión.

La dependencia que nadie ha medido

91% no comprende del todo sus dependencias de proveedores, modelos e infraestructura de IA

81% sufriría una interrupción severa o crítica si su proveedor cae durante 7 días

71% dice que cambiar de proveedor o modelo principal sería difícil

6 interrupciones de IA en promedio en dos años, la mayoría causadas por proveedores

Qué **riesgos** cambian con un modelo abierto propio

Riesgo	API cerrada	Modelo abierto propio
Caída del proveedor	El proceso se detiene	Sigue funcionando
Alza de precio o límites de uso	Lo absorbe usted	Costo fijo de infraestructura
Retiro o cambio de versión	Decide el proveedor	Decide usted
Datos sensibles	Salen bajo contrato	Se quedan en casa
Riesgos nuevos	—	Licencia, origen, operación propia

Muchos de los mejores modelos abiertos son chinos. Los ejecutados en casa no envían datos a nadie, pero hay que revisar la licencia y los sesgos.

En Chile, el dato personal pronto tendrá su propia agencia

LEY 21.719

1 dic 2026

Entrada en vigencia prevista. Hay un proyecto en el Senado para postergarla a 2027.

MULTAS

20.000 UTM

Monto máximo por infracciones gravísimas, con una nueva Agencia que fiscaliza.

Cada dato personal que se envía a una API externa es un tratamiento que habrá que poder justificar.

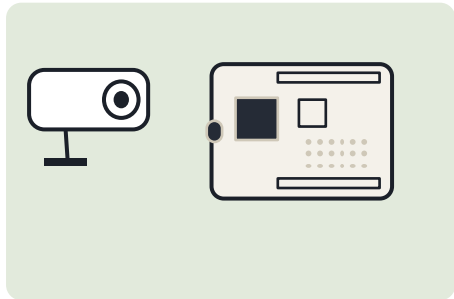
Primero, clasificar los datos

Qué capa puede recibir cada clase de dato

Capa	Confidencial código propio, clientes, RR.HH., contratos, video de personas	Interno manuales, correos de trabajo, catálogos	Público web, normas, documentos publicados
Borde · UNO Q	Sí	Sí	Sí
Local · Qwen3.8 27B	Sí	Sí	Sí
DeepSeek V4.1 Flash local	Sí	Sí	Sí
DeepSeek · API (China)	No	Con cautela	Sí
DeepSeek · API sin retención	Con contrato	Sí	Sí
Frontera vía API	No	Con contrato	Sí

DeepSeek es un modelo abierto: lo sirven decenas de proveedores, en China, EE.UU. y otros países. Elija según dónde procesa cada uno y qué hace con sus datos; con la Ley 21.719, cada envío de datos personales debe poder justificarse.

Cuatro capas, una regla: el dato decide dónde trabaja la IA



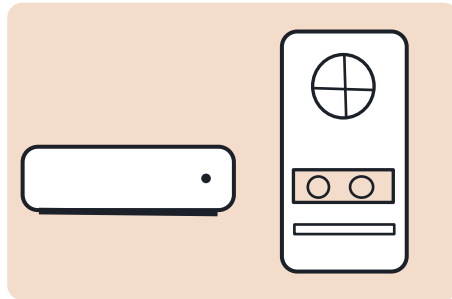
1 · BORDE

Qwen3.5 0.8B / 2B

en Arduino UNO Q

Cámaras y sensores,
sin internet

~US\$79 por punto



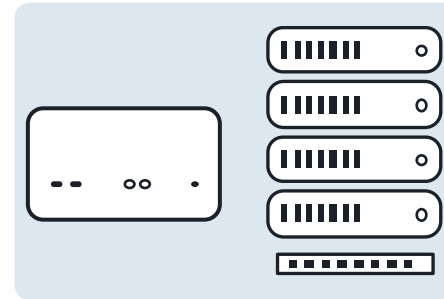
2 · LOCAL

Qwen3.8 27B

en un Mac o PC con GPU

Código y datos
sensibles

Costo fijo: energía



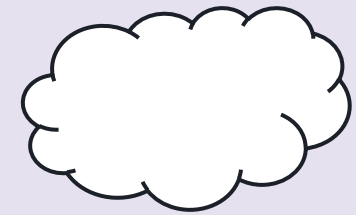
3 · ABIERTO GRANDE

DeepSeek V4.1 Flash

en casa o vía API

Lo masivo y los agentes

~US\$0,26 por tarea



4 · FRONTERA

Claude Opus 5.5

vía API, por excepción

Lo complejo y ocasional

~US\$6 por tarea

← más control y privacidad

más capacidad y más costo por uso →

Todo pasa por un orquestador

Un gateway de IA con reglas de la empresa entre las personas y los modelos



Una sola interfaz, compatible con OpenAI: cambiar de modelo es cambiar una regla, no reescribir aplicaciones.

La próxima ola vive en sensores, cámaras y máquinas

Minería

Faenas con poca conectividad: fatiga del operador, mantención predictiva.

Acuicultura (salmoneras)

Cámaras en el centro de cultivo que cuentan, miden y alertan.

Agro y fruta

Calidad en la línea de packing, a la velocidad de la cinta.

Retail

Analítica de tienda sin enviar video de clientes a la nube.

Energía

Monitoreo de activos distribuidos en terreno.

Salud

Apoyo diagnóstico donde el dato clínico no puede salir.

Cinco preguntas para decidir

Si la respuesta es sí...	...inclínese por
¿Los datos son sensibles o regulados?	Abierto, en infraestructura propia
¿Debe funcionar sin internet o en milisegundos?	Borde (Edge AI)
¿Es una tarea compleja, de bajo volumen o exploratoria?	Cerrado, en la nube
¿Es repetitiva y de alto volumen?	Abierto, propio o vía proveedor
¿Es un proceso crítico del negocio?	Dos proveedores o un respaldo abierto probado

PARA CERRAR

Una placa de US\$79 ejecuta un agente. Un escritorio de US\$10.799, la frontera de hace cuatro meses. **Ninguno necesita permiso de nadie.**

Preguntas

<https://mjrovai.com/articles/open-and-local-october-2026/>



ANEXO A

Modelos abiertos grandes: qué hardware piden

Modelo	Total / activos	Licencia	Índice	Memoria Q4	Hardware mínimo
MiMo-V2.6-Pro Xiaomi	1,02 T / 42 B	MIT	46	~570 GB	2× Mac Studio 512 GB en clúster, o 8× H200
GLM-5.3 Z.ai	744 B / 40 B	Restringida	45	~420 GB	Mac Studio 512 GB
Kimi K3 Moonshot	2,8 T / no divulgado	Restringida	44	~1,6 TB	Data center, varios nodos GPU
DeepSeek V4.1 Flash DeepSeek	552 B (+196 B Engram) / 16 B	MIT	45	~430–510 GB*	Mac Studio 512 GB (1 agente) o 4× DGX Spark
MiniMax-M3 MiniMax	428 B / 23 B	Restringida	29	~240 GB	Mac Studio 512 GB (256 GB muy justo)

*DeepSeek ya publica sus pesos en 4 y 8 bits: 510 GB oficiales, ~430 GB en versiones de la comunidad. Las tablas Engram pueden ir en un SSD NVMe.

ANEXO B

Modelos abiertos medianos y pequeños

Modelo	Total / activos	Licencia	Índice	Memoria Q4	Hardware mínimo
GLM-5.3-Flash	321 B / 18 B	MIT	42	~180 GB	Mac Studio 256 GB o 2× DGX Spark
Qwen3.8-Flash-Next	125 B (+51 B n-gram) / 6 B	Restringida	40	~96–114 GB	Mac 128 GB o 1× DGX Spark (justo)
gpt-oss-120b	117 B / 5,1 B	Apache 2.0	12	~63 GB	Mac 96 GB o 1× DGX Spark
Qwen3.8 27B	27 B, denso	Apache 2.0	34	~16–18 GB	Mac 32 GB o RTX 4090/5090
Muse Glimmer (Meta)	~30 B, denso	Apache 2.0	17	~17–20 GB	Mac 32 GB o RTX 4090
Gemma 4 26B A4B	26 B / ~4 B	Abierta	17*	~15 GB	Laptop de 24–32 GB

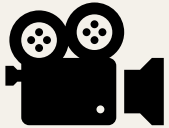
ANEXO C

Hardware de referencia para modelos locales

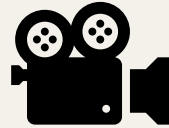
Equipo	Memoria	Ancho de banda	Precio EE.UU.	Hasta qué modelo
RTX 5090	32 GB	1.792 GB/s	~US\$4.300–5.000	Qwen3.8 27B, Muse Glimmer
Mac Studio M5 Max	128 GB	614 GB/s	~US\$4.800	Qwen3.8-Flash-Next, gpt-oss-120b
1 × DGX Spark	128 GB	273 GB/s	US\$6.950	gpt-oss-120b, Flash-Next (justo)
2 × DGX Spark	256 GB	273 GB/s c/u	~US\$13.900	GLM-5.3-Flash
Mac Studio M5 Ultra	256 GB	1,2 TB/s	US\$10.799	GLM-5.3-Flash, a buena velocidad
Mac Studio M5 Ultra	512 GB	1,2 TB/s	Fines de oct · n/d	GLM-5.3, DeepSeek V4.1 Flash
4 × DGX Spark	512 GB	273 GB/s c/u	US\$27.800 + switch	DeepSeek V4.1 Flash (full)

ANEXO D

Videos



IA abierta y local: Vistage Chile - 22/10/2026
<https://youtu.be/5RbSP-Vhu4c>



IA abierta híbrida _ implementacion
<https://youtu.be/0wf6o68XnuY>

