Qwen y el futuro de la IA en el software libre

Leandro Poli - Flisol 2026

Definiciones iniciales

Los *pesos (weights)* son los valores numéricos que la red neuronal aprendió durante el entrenamiento.

Los *pesos abiertos (open weights)* sonos parámetros entrenados de un modelo de IA que están disponibles públicamente para descarga y uso local.

Lo que SÍ puedes hacer con pesos abiertos

Descargar los pesos (.safetensors, .gguf) ✅ Sí

Ejecutar el modelo en tu máquina (local) ✅ Sí

Usarlo sin conexión a internet ✅ Sí

Hacer fine-tuning para tu caso de uso ✅ Sí

Auditar el comportamiento del modelo ✅ Parcialmente

Integrarlo en flujos offline o privados ✅ Sí

Lo que NO garantizan los pesos abiertos

Código de entrenamiento ❌ Cerrado

Datos de entrenamiento ❌ No divulgados

Proceso de curación de datos ❌ Opaco

Libertad para uso comercial ⚠️ Puede tener restricciones

Auditoría completa del pipeline ❌ Parcial

Open Weights ≠ Open Source ≠ Software Libre

Tener los pesos es un paso hacia la soberanía tecnológica, pero no garantiza libertad. Para la comunidad de software libre, exigimos transparencia total.

Qwen

(通义千问, Tōngyì Qiānwèn) es una familia de modelos de lenguaje de gran escala (LLMs) desarrollada por Alibaba Cloud y su laboratorio de investigación DAMO Academy. **Apache License 2.0**

🌐 Multilingüe: Soporta más de 100 idiomas, con fuerte optimización para chino e inglés

💻 Especialización en código: Variantes como Qwen-Coder / Qwen Code para asistencia en programación

🔧 Versátil: Desde modelos pequeños (0.6B) hasta gigantes (235B+ parámetros)

🆓 Pesos abiertos: Muchos modelos están disponibles para descarga pública en Hugging Face y ModelScope

🧠 Capacidades avanzadas: Razonamiento, multimodalidad (texto+imagen), y agentes autónomos

https://huggingface.co/Qwen

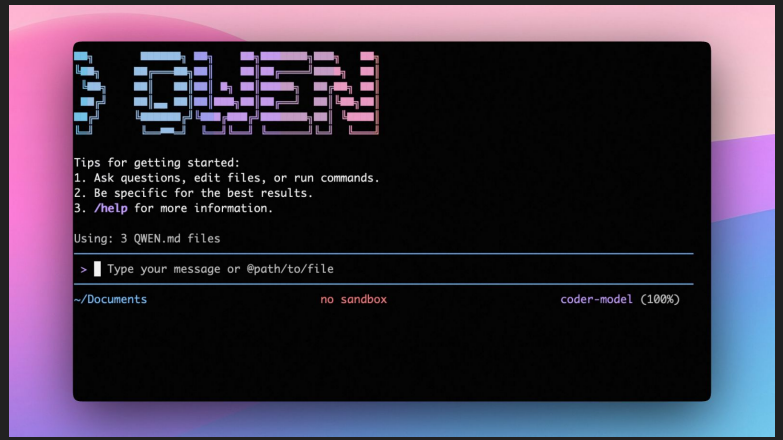
- Abril 2023: Lanzamiento interno de Qwen en Alibaba. Primer uso en productos como Taobao y DingTalk
- Agosto 2023: Lanzamiento público de Qwen-7B. Primer modelo abierto; compite con Llama 2
- Marzo 2025: Qwen3: arquitectura MoE nativa.
- Agosto 2025: Qwen3-Coder
- Enero 2026: Qwen3.5: soporte para contexto de 256K tokens
- Abril 2026: Fin del tier gratuito con OAuth

Qwen-code

Qwen Code is an open-source AI agent for the terminal, optimized for Qwen series models. It helps you understand large codebases, automate tedious work, and ship faster.

https://github.com/QwenLM/qwen-code

Apache License 2.0



Qwen-code

🎉 News

- **2026-04-15**: Qwen OAuth free tier has been discontinued. To continue using Qwen Code, switch to Alibaba Cloud Coding Plan, OpenRouter, Fireworks AI, or bring your own API key. Run `qwen auth` to configure.
- **2026-04-13**: Qwen OAuth free tier policy update: daily quota adjusted to 100 requests/day (from 1,000).
- **2026-04-02**: Qwen3.6-Plus is now live! Get an API key from Alibaba Cloud ModelStudio to access it through the OpenAI-compatible API.
- **2026-02-16**: Qwen3.5-Plus is now live!

Qwen-code

- *API Key (recommended): use an API key from Alibaba Cloud Model Studio (Beijing / intl) or any supported provider (OpenAI, Anthropic, Google GenAI, and other compatible endpoints).*
- *Coding Plan: subscribe to the Alibaba Cloud Coding Plan (Beijing / intl) for a fixed monthly fee with higher quotas.*

- ***Local Model Setup (Ollama / vLLM)***

Qwen-code

- ***Local Model Setup (Ollama / vLLM)*:** Ollama es una herramienta de código abierto que permite descargar, ejecutar y gestionar modelos de lenguaje (LLMs) directamente en tu máquina, sin conexión a internet ni dependencia de APIs en la nube (un “docker para LLMs”).

¿Cómo puedo usar mi claude-code con Qwen?

No se puede, es claude-code privativo y sólo funciona con Claude (api cloud).

Qwen-code se puede usar con clouds (Alibaba, Claude, etc) o corriendo libre y gratuitamente en nuestro equipo.

El fenómeno DeepSeek

euronews. Lo último Europa Mundo Business Viajes Next Cultura Earth Salud Videos M

Business > Mercados

La IA de DeepSeek sacude los mercados mundiales: Nvidia pierde 600.000 millones de dólares

Derechos de autor Andy Wong / AP

Por **Tina Teng**

Publicado 28/01/2025 - 11:20 CET • Ultima actualización 14:19

Compartir | Comentarios

La presentación del último modelo de inteligencia artificial de la empresa china, que supera a sus rivales en eficiencia, ha provocado una fuerte caída del mayor valor bursátil estadounidense, arrastrando al resto de tecnológicas y a las bolsas mundiales.

638 mil M

USD

Argentina: Producto interior bruto (2024)

Fuente: datatopics.worldbank.org • Mostrar metadatos • API code

¿Amenaza DeepSeek al dominio estadounidense de la IA?

DeepSeek, fundada en 2023 por el fondo de cobertura High Flyer, presentó a finales de diciembre **su último gran modelo lingüístico (LLM) gratuito y de código abierto, R1.** Según el comunicado de la empresa, el modelo se desarrolló por menos de 6 millones de dólares (5,7 millones de euros) en solo dos meses. Se trata de una cifra sensiblemente **más barata que las inversiones estadounidenses** en infraestructuras de inteligencia artificial.

El LLM chino superó a algunos de los programas más destacados del mercado como GPT-4o de OpenAI, Gemini 2.0 Flash de Google, Claude 3.5 Sonnet de Anthropic y Llama 3.1 de Meta. DeepSeek **consigue ser más eficiente** en la resolución de problemas matemáticos y de codificación complejos mediante cambios en su programación o en la forma en la que procesa información, **el cual abarata además el coste del proceso.**

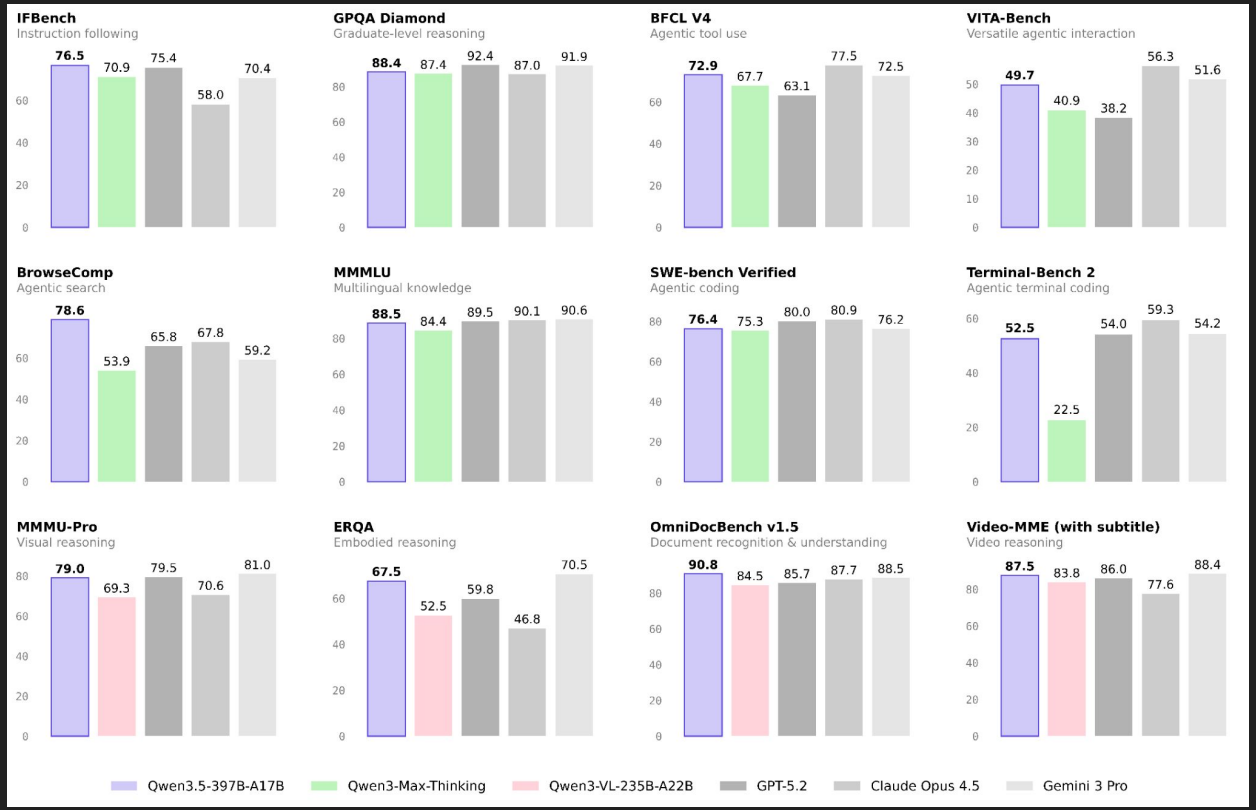
DeepSeek es una empresa china de tamaño chico a medio (~200 emp.) de inteligencia artificial especializada en modelos de lenguaje de código abierto, con foco en razonamiento, matemáticas y programación.

- 2023: Fundación de DeepSeek AI en Hangzhou, China, en el ecosistema de IA chino post-sanciones occidentales
- 2023: Primeras inversiones de fondos locales y alianzas con universidades. Enfoque inicial: investigación en razonamiento y matemáticas
- 2024: Lanzamiento de DeepSeek-Coder (1.3B/6.7B/33B). Primeros modelos especializados en código con licencia MIT
- Sept 2024: DeepSeek-V3. Arquitectura MoE (671B total / 37B activos)

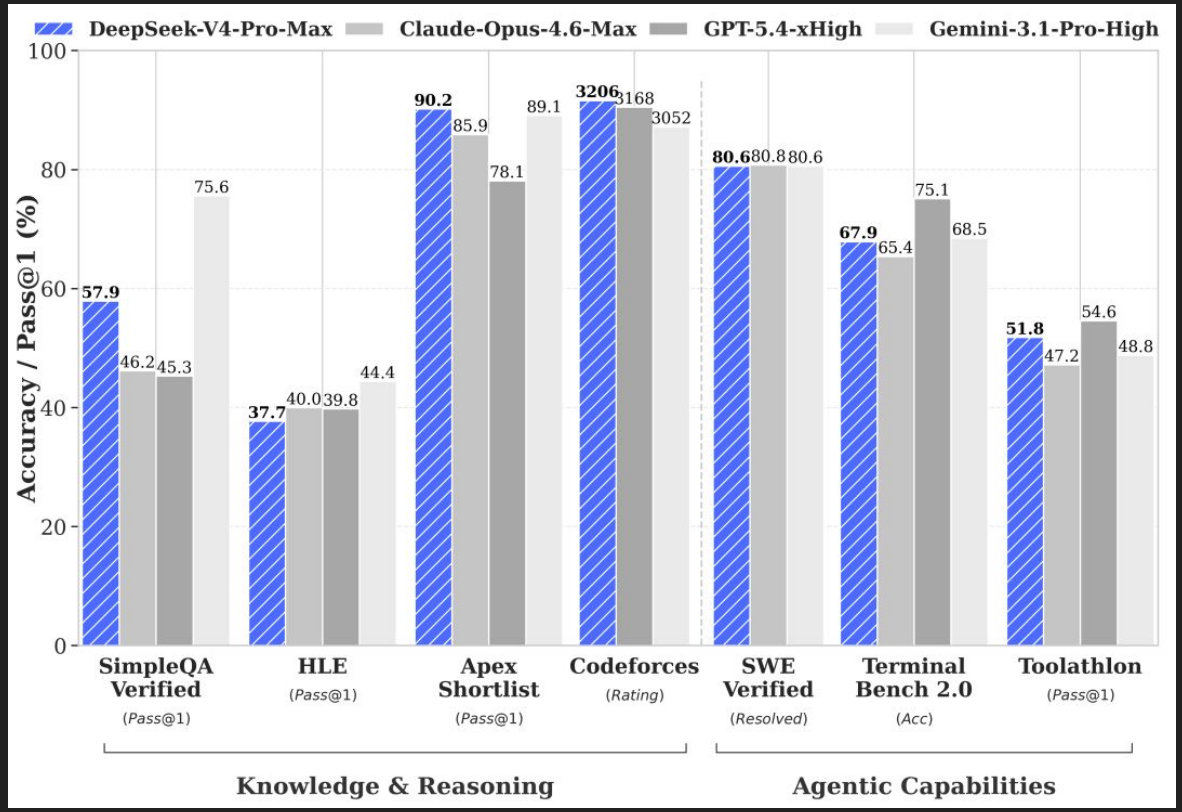
- Enero 2025: DeepSeek-R1. Razonamiento chain-of-thought nativo
- Marzo 2025: DeepSeek-Math. Especializado en resolución de problemas matemáticos
- Enero 2026: DeepSeek-V4-Lite: Versión optimizada para hardware accesible
- Abril 2026: DeepSeek-V4: 1T parámetros, 1M contexto, multimodal nativo
- Abril 2026: Integración con Huawei Ascend. Entrenamiento e inferencia en chips chinos open source

Los modelos de DeepSeek se pueden usar con qwen-code por ollama o API (paga).

Comparativa técnica



Comparativa técnica



Resumen técnico

Los LLMs y sus software clientes cuando son libres nos permiten

- Reducir costos operativos
- Aumentar la privacidad de la información
- Eliminar dependencia funcional con un proveedor

Pero el hardware lo tenemos que poner nosotros

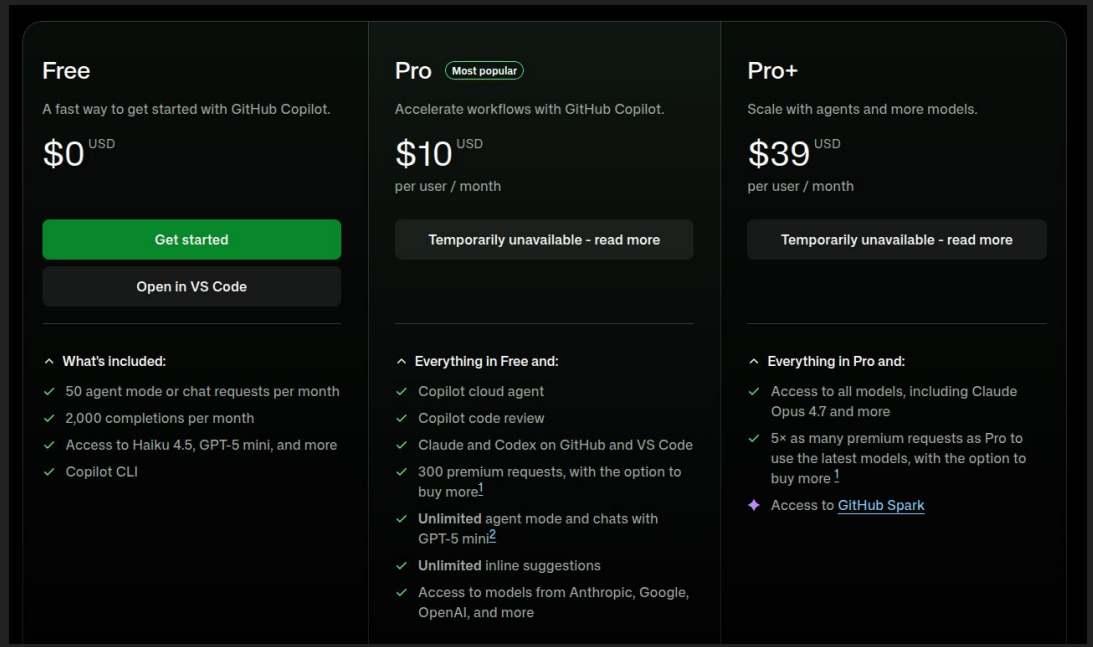
Comparativa económica

Modelo	Input (Cache Miss)	Input (Cache Hit)	Output	Contexto Máx.
DeepSeek V4-Flash	$0.14	$0.028	$0.28	1M tokens
DeepSeek V4-Pro	$0.14	$0.145	$0,87	1M tokens
Qwen3-Coder	$0.22	~$0.02	$0.90	262K tokens
Qwen3-Coder Plus	$0.65	~$0.07	$3.25	1M tokens
Qwen3 Max	$0.78	~$0.08	$3.90	262K tokens
GPT-5.4	$2.50	$0.25	$15.00	1M tokens
Claude Opus 4.6	$5.00	~$0.50	$25.00	200K

Pricing: Qwen coding plan

	Pro
Supported models	Recommended models: **qwen3.6-plus** (vision), **kimi-k2.5** (vision), **glm-5**, and **MiniMax-M2.5** More models: `qwen3.5-plus` (vision), `qwen3-max-2026-01-23`, `qwen3-coder-next`, `qwen3-coder-plus`, and `glm-4.7`
Price	**$ 50**/month
Quota	• **6,000** requests per 5 hours • **45,000** requests per week • **90,000** requests per month

Pricing: Github copilot



Free

A fast way to get started with GitHub Copilot.

$0 USD

Get started

Open in VS Code

What's included:

- 50 agent mode or chat requests per month
- 2,000 completions per month
- Access to Haiku 4.5, GPT-5 mini, and more
- Copilot CLI

Pro Most popular

Accelerate workflows with GitHub Copilot.

$10 USD

per user / month

Temporarily unavailable - read more

Everything in Free and:

- Copilot cloud agent
- Copilot code review
- Claude and Codex on GitHub and VS Code
- 300 premium requests, with the option to buy more[1]
- **Unlimited** agent mode and chats with GPT-5 mini[2]
- **Unlimited** inline suggestions
- Access to models from Anthropic, Google, OpenAI, and more

Pro+

Scale with agents and more models.

$39 USD

per user / month

Temporarily unavailable - read more

Everything in Pro and:

- Access to all models, including Claude Opus 4.7 and more
- 5× as many premium requests as Pro to use the latest models, with the option to buy more [1]
- Access to GitHub Spark

Pricing: Claude Code (API)

Opus 4.7

Most intelligent model for agents and coding

Input	$5 / MTok
Output	$25 / MTok
Prompt caching	
Write	$6.25 / MTok
Read	$0.50 / MTok

Sonnet 4.6

Optimal balance of intelligence, cost, and speed

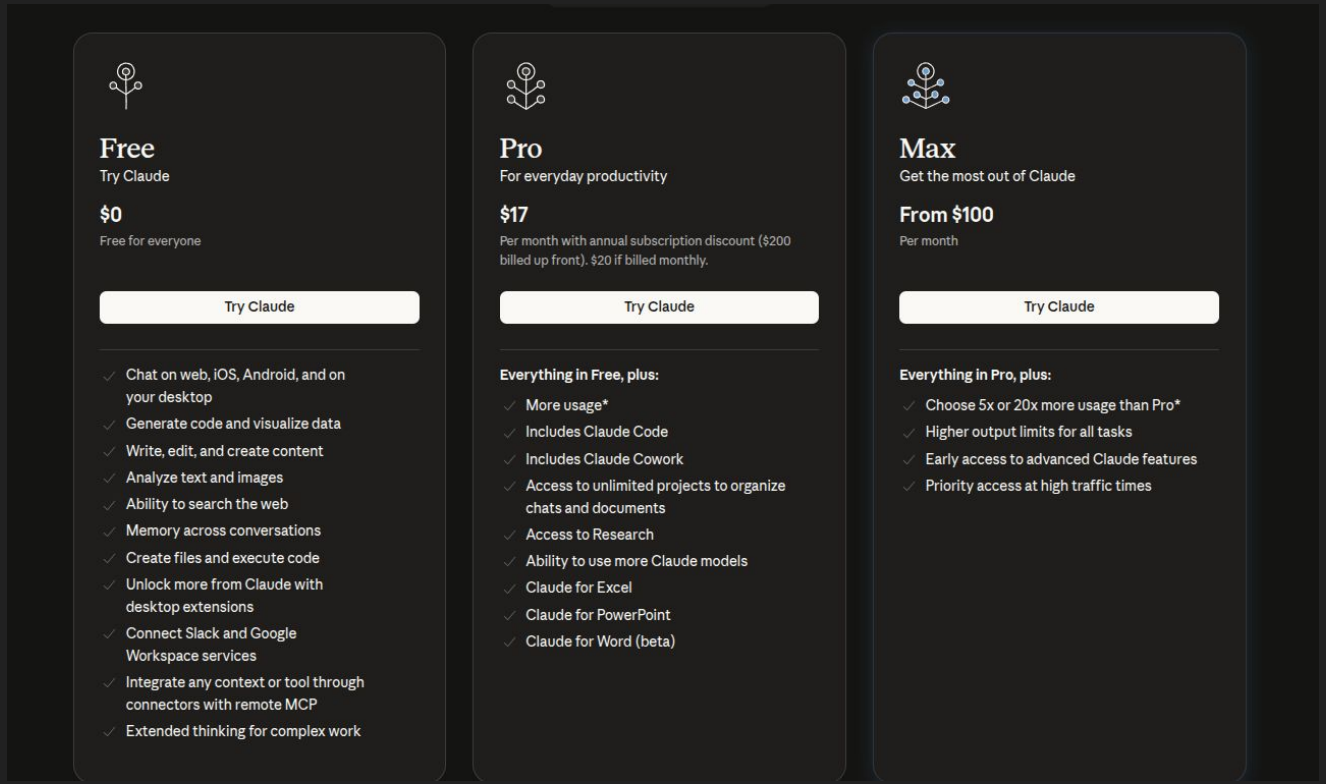
Input	$3 / MTok
Output	$15 / MTok
Prompt caching	
Write	$3.75 / MTok
Read	$0.30 / MTok

Haiku 4.5

Fastest, most cost-efficient model

Input	$1 / MTok
Output	$5 / MTok
Prompt caching	
Write	$1.25 / MTok
Read	$0.10 / MTok

Pricing: Claude Code (individual)



Free

Try Claude

$0

Free for everyone

Try Claude

- Chat on web, iOS, Android, and on your desktop
- Generate code and visualize data
- Write, edit, and create content
- Analyze text and images
- Ability to search the web
- Memory across conversations
- Create files and execute code
- Unlock more from Claude with desktop extensions
- Connect Slack and Google Workspace services
- Integrate any context or tool through connectors with remote MCP
- Extended thinking for complex work

Pro

For everyday productivity

$17

Per month with annual subscription discount ($200 billed up front). $20 if billed monthly.

Try Claude

Everything in Free, plus:

- More usage*
- Includes Claude Code
- Includes Claude Cowork
- Access to unlimited projects to organize chats and documents
- Access to Research
- Ability to use more Claude models
- Claude for Excel
- Claude for PowerPoint
- Claude for Word (beta)

Max

Get the most out of Claude

From $100

Per month

Try Claude

Everything in Pro, plus:

- Choose 5x or 20x more usage than Pro*
- Higher output limits for all tasks
- Early access to advanced Claude features
- Priority access at high traffic times

Resumen económico

- Las IAs libres como Qwen y DeepSeek con competitivas en términos de rendimiento con otras como Claude, Gemini y GPT 5.x (mismo performance o superior en algunos)

- Pero son sensiblemente más baratas que las anteriores, aumentando la tasa de ROI y reduciendo los costes operativos

La guerra interna de IA China

Ascend: Open for All to Build a Vibrant Ecosystem

Sep 20, 2025

[Shanghai, China, September 20, 2025] At HUAWEI CONNECT 2025, Zhang Dixuan, President of Huawei's Ascend Computing Business, delivered a keynote speech highlighting Ascend's commitment to driving developer-centric ecosystem growth. He announced the official establishment of the CANN Technical Steering Committee and emphasized Ascend's dedication to accelerate innovation through a strategy of architectural upgrades, layered decoupling, and full open-source collaboration.

Reuters World Business Markets Sustainability Legal Commentary Technology Investigations

DeepSeek unveils new AI model tailored for Huawei chips as China pushes for tech autonomy

By **Eduardo Baptista**, **Ethan Wang** and **Che Pan**

April 24, 2026 12:11 AM GMT-3 · Updated 10 hours ago

Reuters World Business Markets Sustainability Legal Commentary More

Alibaba shares leap on Nvidia partnership, data center plans

By **Liam Mo** and **Eduardo Baptista**

September 24, 2025 8:07 AM GMT-3 · Updated September 24, 2025

Summary Companies

- Alibaba shares surge 9.7% to four-year high on AI announcements
- Company partners with Nvidia, plans new data centers globally
- Unveils Qwen3-Max AI model with over one trillion parameters

BEIJING, Sept 24 (Reuters) - Alibaba (9988.HK) announced on Wednesday a partnership with Nvidia, global data center expansion plans and new artificial intelligence products, as it positions AI as a core business priority alongside its traditional e-commerce operation.

The announcement helped send Hong Kong-listed shares of the Chinese company up nearly 10% to a four-year high on Wednesday, as investors welcomed its plan to double down on AI amid grueling competition with local peers that include DeepSeek and Tencent (0700.HK).

Mi perspectiva para el futuro de la IA libre

1. No hay (casi y por ahora), modelos competitivos libres que no sean de origen chino. Creo que esta tendencia se va a mantener porque los modelos como Claude y GPT son menos eficientes.
2. Si los modelos son menos eficientes y más costosos (construcción y operación), es entendible que sean cerrados para mantener cautivo al cliente al no ofrecerle la posibilidad de usar su propio entorno (amortización por cautivero).
3. Por tanto, en contextos de limitantes económicas, el consumo de IAs como servicio de origen chino va a crecer.
4. Si el hardware libre se hace extensivo y asequible, esto va a fomentar más aún esta tendencia.

Gracias por su tiempo vital dedicado a escucharme.

Pueden hacer preguntas.