



NARSIL

INTELLIGENCE

ESTUDIOS NARSIL · 03

¿Qué equipo necesitas para tener IA en casa?

Guía de hardware para ejecutar modelos de lenguaje en local en octubre de 2026: cuánta memoria hace falta, cinco escalones de equipo con precios en España y el efecto de la crisis de la memoria

OCTUBRE DE 2026

Resumen

1. **La memoria manda.** Para ejecutar un modelo en local, el modelo tiene que caber entero en la memoria, y la velocidad depende sobre todo de lo rápido que se lea. Regla práctica: parámetros por bytes por parámetro, más un 20 %, más el contexto.
2. **Hay cinco escalones de equipo.** Portátil de trabajo, Apple Silicon con mucha memoria, miniordenadores AMD de 128 GB, equipos NVIDIA GB10 como el DGX Spark y estaciones con gráfica dedicada.
3. **Un portátil de 32 GB ya sirve para modelos pequeños,** como gpt-oss-20b, que OpenAI hace funcionar con 16 GB.
4. **Con 80 a 128 GB caben modelos grandes** como gpt-oss-120b. Equipos de este tipo cuestan hoy en España entre unos 3.900 y 6.700 euros con IVA.
5. **La crisis de la memoria ha cambiado los precios.** El DGX Spark de 128 GB cuesta hoy un 74 % más que en su lanzamiento, la RTX PRO 6000 ha duplicado su precio y Gartner prevé que DRAM y SSD suban un 130 % hasta final de 2026.
6. **El equipo se elige al final.** Primero el modelo que exige la tarea, el número de usuarios y el contexto; después, la máquina.

01

La memoria manda

Cuando se habla de ejecutar inteligencia artificial en local, la conversación suele empezar por la potencia de cálculo. Para los modelos de lenguaje, la pregunta importante es otra: **¿cabe el modelo en la memoria?** Un modelo tiene que cargarse entero para funcionar, y si no cabe en la memoria de la gráfica o en la memoria unificada del equipo, o no funciona o lo hace con una lentitud que lo vuelve inútil.

La segunda cuestión es la **velocidad de esa memoria**. Para generar cada palabra, el equipo tiene que leer una gran parte del modelo. Por eso, a igualdad de capacidad, un equipo que lee la memoria más rápido responde más rápido. Esa cifra, el ancho de banda, se mide en gigabytes por segundo, y es la que más separa unos equipos de otros.

Cuánta memoria pide un modelo

La cuenta básica es sencilla. Cada parámetro del modelo ocupa un número de bytes que depende de su precisión: dos bytes en la precisión habitual de entrenamiento, uno si se reduce a 8 bits y medio byte si se reduce a 4 bits, lo que se conoce como cuantización. Según la guía técnica de EleutherAI, que recoge también la documentación de Hugging Face, para ejecutar el modelo hay que sumar alrededor de un **20**

% adicional. A eso se añade la memoria que ocupa el contexto, es decir, el texto que el modelo tiene presente en cada conversación, que crece con la longitud de los documentos y con el número de usuarios simultáneos.

EJEMPLO

Un modelo de 120.000 millones de parámetros cuantizado a 4 bits ocupa unos 60 GB. Con el 20 % adicional, unos 72 GB, a los que hay que sumar el contexto. Es la razón por la que los equipos de 128 GB se han convertido en la referencia para los modelos grandes.

02

Cinco escalones de equipo

1. El portátil de trabajo

Microsoft exige a los equipos con la etiqueta Copilot+ una unidad de procesamiento neuronal (NPU) de más de 40 billones de operaciones por segundo. Esas NPU son útiles para tareas de IA ligeras y continuas, pero para los modelos de lenguaje lo que limita es la memoria. Un portátil con 32 GB puede ejecutar modelos pequeños: gpt-oss-20b, publicado por OpenAI con licencia Apache 2.0, funciona según su autor con 16 GB de memoria. Es suficiente para un asistente personal, resúmenes y pruebas, no para dar servicio a un equipo.

2. Apple Silicon con mucha memoria

Los Mac usan memoria unificada, compartida entre el procesador y la gráfica, lo que permite cargar modelos grandes en equipos compactos y silenciosos. Según Apple, el MacBook Pro con M5 Max admite hasta 128 GB con un ancho de banda de hasta 614 GB/s, y el M5 Pro hasta 64 GB. En la tienda española, el MacBook Pro de 16 pulgadas parte de 3.449 € con M5 Pro y de 5.149 € con M5 Max, en configuraciones base. El Mac Studio presentado en agosto de 2026 parte de 3.029 € con M5 Max y de 6.649 € con M5 Ultra (96 GB). Apple ha anunciado configuraciones de hasta 512 GB para el M5 Ultra que, según la prensa especializada, llegarán a finales de octubre; a 2 de octubre, la tienda española ofrece hasta 256 GB.

3. Miniordenadores con AMD Ryzen AI Max+

El procesador AMD Ryzen AI Max+ 395 admite hasta 128 GB de memoria LPDDR5x y, según AMD, permite asignar hasta 96 GB a su gráfica integrada, suficiente para ejecutar modelos de hasta 128.000 millones de parámetros. AMD ha mostrado Llama 4 Scout, de 109.000 millones, funcionando en uno de estos equipos. El Framework Desktop con 128 GB cuesta 3.919 € con IVA en su tienda española, aunque a 2 de octubre figura como agotado.

4. NVIDIA DGX Spark y equipos GB10

El DGX Spark de NVIDIA, y los equipos con el mismo chip GB10 de fabricantes como ASUS, Dell, HP, Lenovo o Gigabyte, reúnen 128 GB de memoria unificada con un ancho de banda de 273 GB/s y, según NVIDIA, permiten ejecutar modelos de hasta 200.000 millones de parámetros. Su gran ventaja es el ecosistema de software de NVIDIA, el más extendido para IA. En España, el ASUS Ascent GX10 con 128 GB y 1 TB se vende a 6.051 € con IVA. Su fuente de alimentación es de 240 W.

5. Estación de trabajo con gráfica dedicada

Cuando hay que dar servicio a muchas personas a la vez, la opción es una estación con gráfica profesional. La NVIDIA RTX PRO 6000 Blackwell tiene 96 GB de memoria y un ancho de banda de 1.792 GB/s, más de seis veces el de un DGX Spark, con un consumo de hasta 600 W solo la tarjeta. La GeForce RTX 5090, pensada para consumo, tiene 32 GB y llega a 575 W. Son equipos más ruidosos, más caros y con más exigencias de ubicación, pero mucho más rápidos.

03

La crisis de la memoria de 2026

Cualquier recomendación de compra en 2026 tiene que contar con un hecho: la memoria se ha encarecido como nunca. La demanda de los centros de datos de IA, que consumen enormes cantidades de memoria de alta velocidad, ha desplazado la producción y ha disparado los precios del resto del mercado.

Los datos lo muestran con claridad. En febrero de 2026, Gartner previó una subida del 130 % de los precios combinados de DRAM y SSD hasta final de año y un encarecimiento de los ordenadores del 17 %. TrendForce espera nuevas subidas de la memoria convencional de entre el 10 % y el 15 % en el cuarto trimestre. En marzo, Apple dejó de ofrecer la configuración de 512 GB de su anterior Mac Studio, en un contexto que la prensa atribuyó a la escasez de memoria. La RTX PRO 6000, que salió en preventa por debajo de 8.000 dólares, cuesta ya 16.000 dólares en la tienda de NVIDIA.

Y hoy mismo, 2 de octubre, NVIDIA ha presentado una versión del DGX Spark con 64 GB a 4.999 dólares y ha subido la de 128 GB a 6.950 dólares. El equipo salió a 3.999 dólares en octubre de 2025 y ya había subido a 4.699 en febrero. Es un 74 % más caro que en su lanzamiento.

QUÉ IMPLICA

Comprar hoy exige previsión: los precios y la disponibilidad cambian de un mes a otro, y conviene dimensionar con margen en lugar de ir ampliando después.

04

Qué modelos caben

Los modelos abiertos de 2025 y 2026 han hecho viable la IA local de calidad. Muchos usan una arquitectura de mezcla de expertos: tienen muchos parámetros en total, pero solo activan una parte para cada palabra, lo que los hace mucho más rápidos de lo que su tamaño sugiere, aunque siguen necesitando la memoria para cargarlos enteros.

Modelo	Tamaño	Memoria orientativa	Licencia
gpt-oss-20b (OpenAI)	21.000 millones, 3.600 activos	16 GB, según OpenAI	Apache 2.0
gpt-oss-120b (OpenAI)	117.000 millones, 5.100 activos	Cabe en una gráfica de 80 GB, según OpenAI	Apache 2.0
Qwen3.5-122B-A10B (Alibaba)	122.000 millones, 10.000 activos	Unos 73 GB a 4 bits, según la regla del apartado 01	Apache 2.0
Mistral Small 4	119.000 millones, 6.500 activos	Unos 71 GB a 4 bits, según la misma regla	Apache 2.0
Llama 4 Scout (Meta)	109.000 millones, 17.000 activos	Cabe en una gráfica de 80 GB a 4 bits, según Meta	Licencia propia de Meta

Datos de las fichas de cada modelo. La memoria «según la regla» es una estimación propia con el cálculo del apartado 01, sin contar el contexto.

La licencia importa tanto como el tamaño. Apache 2.0 permite el uso comercial sin apenas restricciones; otras licencias imponen condiciones que conviene revisar antes de construir un servicio sobre ellas.

05

Más allá de la máquina

El precio del equipo es solo una parte. Hay que contar el **consumo eléctrico** y el calor: un equipo GB10 se alimenta con 240 W, mientras que una estación con una o dos gráficas profesionales puede superar el kilovatio. El **ruido** decide si el equipo puede estar en un despacho o necesita un cuarto técnico. La **concurencia** cambia el cálculo: no es lo mismo una persona consultando un modelo que veinte a la vez, cada una con su contexto. Y está el **mantenimiento**: actualizar modelos y software, vigilar el rendimiento y proteger el sistema, porque un modelo local no es seguro por el mero hecho de estar en casa, como explicamos en los estudios 01 y 02.

Cómo elegir

El orden correcto es el inverso al habitual: primero la tarea, después el modelo y por último el equipo. Antes de comprar nada, conviene responder a cuatro preguntas.

1. **¿Cuántas personas lo usarán a la vez?** Una persona puede trabajar con un equipo de escritorio; un departamento entero suele necesitar una estación con gráfica dedicada o varias máquinas.
2. **¿Qué tamaño de modelo exige la tarea?** Resumir y redactar se resuelve con modelos pequeños; analizar documentación compleja o razonar con precisión pide modelos grandes.
3. **¿Cuánto texto tiene que leer de una vez?** Una página o un expediente entero marcan una diferencia grande en memoria.
4. **¿Dónde va a estar, cuánto consume y quién lo mantiene?**

Lo más sensato es **probar antes de comprar**: ejecutar el modelo candidato con documentos reales de la organización y medir si la calidad y la velocidad son suficientes. Con esa prueba, la elección del equipo deja de ser una apuesta.

Para una pequeña organización que quiere un asistente interno con sus documentos, un equipo de 128 GB, sea un miniordenador AMD, un Mac o un equipo GB10, cubre hoy la mayoría de los casos. Para servicios que usan muchas personas a la vez, la estación con gráfica dedicada sigue siendo la referencia.

Anexo

Fuentes

Páginas consultadas el 2 de octubre de 2026, salvo las noticias, fechadas en su publicación.

1. Microsoft. *Copilot+ PCs developer guide*. learn.microsoft.com/en-us/windows/ai/npu-devices
2. Apple. *Apple debuts M5 Pro and M5 Max* (3-3-2026). apple.com/newsroom. Tienda Apple España, MacBook Pro y Mac Studio (consultada el 2-10-2026).
3. 9to5Mac. *Apple unveils next-generation Mac Studio with M5 Max and M5 Ultra* (25-8-2026) y *Apple no longer offers M3 Ultra Mac Studio with original highest RAM configuration* (5-3-2026).
4. AMD. Ficha del Ryzen AI Max+ 395 y *AMD Ryzen AI Max+ upgraded: run up to 128 billion parameter LLMs* (29-7-2025). amd.com
5. Framework. Framework Desktop, configurador para España (consultado el 2-10-2026). frame.work/es/es
6. NVIDIA. DGX Spark, RTX PRO 6000 Blackwell y GeForce RTX 5090, especificaciones. nvidia.com
7. Alternate España. ASUS Ascent GX10 (consultado el 2-10-2026). alternate.es
8. The Register. *Nvidia debuts \$4,999 DGX Spark with half the RAM and storage amid memory crunch* (2-10-2026).
9. Tom's Hardware. *Nvidia doubles RTX Pro 6000 Blackwell's MSRP to a staggering \$16,000* (2026).
10. Gartner. *Gartner Says Surging Memory Costs Will Reduce Global PC and Smartphone Shipments in 2026* (26-2-2026).
11. TrendForce. Previsión de precios de memoria para el cuarto trimestre de 2026 (30-9-2026). trendforce.com
12. EleutherAI. *Transformer Math 101*. blog.eleuther.ai/transformer-math; Hugging Face, *Model memory estimator*.
13. OpenAI. *gpt-oss-120b* y *gpt-oss-20b*, fichas de modelo. huggingface.co/openai
14. Qwen. *Qwen3.5-122B-A10B*; Mistral AI, *Mistral Small 4*, fichas de modelo. huggingface.co
15. Meta. *The Llama 4 herd* (5-4-2025). ai.meta.com/blog/llama-4-multimodal-intelligence



NARSIL INTELLIGENCE

Diseñamos, construimos e implantamos sistemas de inteligencia artificial en la infraestructura de cada organización, y formamos a los equipos que los usan.

NARSILINTELLIGENCE.COM