mdt.

Ex Datis, Claritas

La carrera silenciosa por la eficiencia en IA: hacer más con menos

La IA está atrapada en una crisis silenciosa: no es rentable y choca con un techo físico. Los próximos meses serán de eficiencia, no de fuerza bruta.

Manuel Delgado Tenorio · 12 de mayo de 2026

Publicado originalmente en www.manueldelgado.com/blog/carrera-eficiencia-ia



Hablamos a todas horas del avance imparable de la inteligencia artificial. Detrás de esa imagen de progreso sin freno, sin embargo, la industria está atrapada en un problema enorme: la inteligencia artificial, sencillamente, no le sale rentable (hoy) a casi nadie y se está topando con un techo físico para seguir creciendo.

Tomemos una cifra para situarnos. El gasto en inversiones de capital (CAPEX) de las cinco grandes tecnológicas norteamericanas en el primer trimestre de este año fue más de tres veces el coste total del Proyecto Manhattan, según el repaso de Ben Thompson a los resultados trimestrales. Tres veces lo que costó construir la primera bomba atómica. Y aun así, no es suficiente.

Cuando ni siquiera Anthropic puede contener su propio éxito

La demanda va por delante de la capacidad instalada. Dario Amodei, consejero delegado de Anthropic, ha admitido que la empresa había planificado **crecer diez veces este año respecto al anterior... y va por ochenta veces.** Ocho veces lo que ya era una previsión ambiciosísima.

Los usuarios lo notamos. En los últimos meses, Anthropic ha endurecido los límites de uso semanal en Claude, ha ajustado los precios efectivos al alza, y ha tenido que recortar la capacidad disponible para sus suscriptores de pago. No por una decisión comercial deliberada, sino por una cuestión aritmética: tienen que repartir la capacidad que poseen entre toda su base de usuarios, que crece más rápido que su infraestructura.

La situación les ha llevado a movimientos verdaderamente excepcionales. Anthropic ha alquilado a la competencia (literalmente a la competencia, a la empresa xAI de Elon Musk) un centro de datos entero, el llamado Colossus 1, con 300 megavatios de capacidad eléctrica, equivalente al consumo de una ciudad pequeña. ¿La razón? Que xAI, según reporta The Information (de pago), construyó ese centro de datos con tantas prisas y con una mezcla tan heterogénea de chips, que no le sirve para entrenar modelos nuevos, pero sí para servir inferencia. Es decir, para responder a las consultas de los usuarios. Y eso es exactamente lo que Anthropic necesita.

Del lado de los clientes empresariales, la historia se repite. Uber dio acceso a Claude Code a sus cinco mil ingenieros y en cuatro meses se gastó todo el presupuesto de tokens previsto para el año entero. A Nvidia, internamente, le ocurrió algo parecido. Empresas serias, con millones encima de la mesa, agotan en semanas lo presupuestado para todo un ejercicio.

El muro físico: electricidad, hormigón y cuarzo

La respuesta lógica sería construir más centros de datos. Pero ahí nos topamos con el segundo cuello de botella, que es el mundo físico.

La Agencia Internacional de la Energía calcula que los centros de datos van a consumir este año mil teravatios hora de electricidad. Para hacernos una idea: es aproximadamente toda la electricidad que consume Japón en un año. Un país entero, sólo para alimentar centros de datos. Y McKinsey proyecta que la demanda de capacidad para centros de datos se va a casi triplicar de aquí a 2030, con un 70% de ese crecimiento empujado directamente por la inteligencia artificial.

Cada centro nuevo necesita una pequeña ciudad de hormigón, miles de kilómetros de fibra óptica, miles de toneladas de cuarzo de altísima pureza para fabricar los chips. Y ese cuarzo, dato curioso, sale casi todo de un único pueblo de los Apalaches, de dos minas, en Spruce Pine (Carolina del Norte). Cuando el huracán Helene inundó esas minas en septiembre de 2024, durante tres semanas la industria global de semiconductores estuvo conteniendo la respiración.

Lo que cualquiera que siga esta industria ya sabe es que el principal cuello de botella actual no es el silicio, sino la electricidad. No hay megavatios suficientes. La capacidad de muchas redes de generación y distribución eléctrica está al límite. Y, encima, súmale la oposición vecinal: comunidades que no quieren un centro de datos del tamaño de un polígono industrial al lado de su casa, consumiendo agua para refrigeración y tensionando el precio de la luz.

Hay quien está empezando a probar cosas verdaderamente extravagantes. La startup Panthalassa, según informa AI Secret, ha levantado 140 millones de dólares de Peter Thiel para construir centros de datos autónomos flotantes, refrigerados con agua de mar en el Pacífico Norte. Otra empresa, Span, respaldada por Nvidia, está probando colgar mini-centros de datos en la pared exterior de las casas, conectados al cuadro eléctrico doméstico. Y no olvidemos los centros de datos orbitando en el espacio que tiene en mente Elon Musk.

La reacción: tres frentes de mejora

Cuando una industria se enfrenta a un muro, espabila. Lo que estamos viendo estos meses es una verdadera explosión de creatividad para hacer lo mismo (o más), pero con muchísimos menos recursos. Las llamamos mejoras de eficiencia y se están dando en tres frentes a la vez.

Primer frente: cambios de paradigma (y un protagonista europeo)

El frente más fascinante es el de los cambios de paradigma. Replantearse cómo funcionan estos modelos desde la base. Y aquí, atención, la noticia más importante de las últimas semanas viene de Europa.

Yann LeCun es uno de los padres de la inteligencia artificial moderna. Es premio Turing (el equivalente al Nobel en informática) y es de origen francés. Hasta hace pocos meses dirigía la investigación en IA de Meta, la empresa de Facebook. Pues bien: dejó Meta en noviembre de 2025, después de doce años, y el pasado marzo **levantó la mayor ronda semilla de financiación de la historia de Europa, más de mil millones de dólares,** para fundar una compañía con sede principal en París, AMI Labs. Entre los inversores aparecen Jeff Bezos, Nvidia, Samsung y Toyota.

La apuesta de LeCun es radical. Lleva años defendiendo que el camino de los modelos actuales (los grandes modelos de lenguaje tipo ChatGPT) es un callejón sin salida: demasiado caro, demasiado consumidor de recursos, demasiado limitado para alcanzar lo que él considera una verdadera inteligencia. Su tesis es que hay que cambiar de paradigma y construir lo que él llama modelos del mundo (*world models*), que aprenden viendo (procesando imágenes, vídeo, sensores), como aprende un bebé, en lugar de memorizando millones de textos.

Tres días después de cerrar la ronda, sus colaboradores publicaron una prueba de concepto: LeWorldModel. Quince millones de parámetros (ochenta veces menos que un modelo de tamaño medio actual), entrenable en horas en una única

tarjeta gráfica (no en miles, como los modelos actuales), capaz de planificar tareas hasta cuarenta y ocho veces más rápido que los modelos basados en lenguaje. Manteniendo, eso sí, un rendimiento competitivo en los benchmarks.

Es la primera vez que vemos, con datos encima de la mesa, que el paradigma dominante quizá no sea el único posible. Y es muy relevante para Europa, porque significa que la batalla de la inteligencia artificial no está ganada todavía por los gigantes americanos y chinos. Hay otra forma de hacer las cosas, y AMI Labs es nuestro mejor caballo en esa carrera.

Segundo frente: mejoras de software (la ciencia de no repetir el trabajo)

El segundo frente son las mejoras de software, las menos llamativas pero las más efectivas a corto plazo. La idea común a todas ellas es muy sencilla: exprimir al máximo cada chip que ya tenemos.

Permíteme una imagen. Imagina que cada vez que el cajero del banco tiene que hacer una operación contigo, baja al sótano, abre la caja fuerte, saca todos los expedientes del archivo, los pone sobre su mesa, mira un único expediente, y luego los vuelve a guardar todos. Por absurdo que suene, así funciona la inteligencia artificial generativa hoy: por cada palabra[1] que genera de respuesta, tiene que leer miles de millones de parámetros del modelo, moverlos entre los componentes internos del servidor, hacer un cálculo diminuto, y empezar otra vez con la siguiente palabra.

- **Reutilizar lo ya calculado.** Cuando varios usuarios hacen preguntas parecidas, o cuando una misma conversación se prolonga en el tiempo, no tiene sentido recalcular el contexto cada vez. Las llamadas KV caches almacenan ese contexto y lo reutilizan. Frameworks como vLLM o SGLang (este último detrás de una startup llamada RadixArk, en la que han invertido conjuntamente Nvidia, AMD e Intel) son los grandes responsables de los saltos recientes en este terreno.
- **Generar en bloques.** Las técnicas de *speculative decoding* permiten que un modelo pequeño y rápido vaya proponiendo varias palabras por delante, y que el modelo grande sólo intervenga para validar. Los modelos de difusión llevan esta idea al extremo: en lugar de generar palabra por palabra, refinan en paralelo bloques enteros de texto.
- **Cuantización.** Probablemente la técnica más popular. Consiste en bajar la precisión numérica de los cálculos del modelo. Es como pasar imprimir en borrador en lugar de en calidad alta: los modelos ocupan menos, se mueven más rápido y, sorprendentemente, los resultados apenas pierden calidad. Con estas técnicas se consiguen modelos cuatro veces más pequeños, dos veces más rápidos, perdiendo sólo entre un cinco y un diez por ciento de precisión.
- **Steering en tiempo de inferencia.** Una vía aún experimental, pero prometedora: en lugar de reentrenar un modelo entero para corregir su comportamiento (lo que cuesta miles de dólares por iteración), se le aplican empujones matemáticos durante la inferencia que cuestan, literalmente, cero céntimos.

El efecto combinado de todas estas técnicas está siendo tan importante que la competencia se ha vuelto colaboración en algunos puntos clave. Que Nvidia, AMD e Intel coinviertan en la misma startup es algo prácticamente sin precedentes. Lo

[1] Sí, he dicho palabra y no *token* porque, cada vez que digo *token*, tengo que explicar qué es y así no se puede avanzar en el texto. No te des de baja de la newsletter por esa simplificación, por favor.

que vale para uno, vale para los demás. Nadie quiere quedarse fuera de la "capa neutra de inferencia".

Tercer frente: hardware especializado

El tercer frente es el del hardware. La fabricación de chips diseñados específicamente para que la inteligencia artificial funcione más rápido y consumiendo menos.

La estrella de las últimas semanas es Cerebras que, animada por las perspectivas del mercado, se ha permitido el lujo de elevar el precio de su salida a bolsa. Cerebras fabrica lo que se llaman chips a escala de oblea. Un chip normal mide unos tres centímetros por lado; un chip de Cerebras mide treinta, es del tamaño de un plato llano grandecito. Tiene seis mil veces más velocidad de acceso a su memoria que un chip Nvidia H100 (el más popular hasta ahora). Lo paga en flexibilidad: el chip de Cerebras sirve casi exclusivamente para una cosa muy concreta: ofrecer respuestas inmediatas a usuarios humanos.

Hay muchas otras empresas en esta carrera, cada una con una apuesta distinta. Groq y la propia Cerebras compiten por la velocidad de respuesta. Etched ha cableado directamente la arquitectura Transformer en su silicio, una apuesta arriesgada (si los modelos cambian de arquitectura, su chip queda obsoleto, pero si no cambian, es imbatible). d-Matrix y Furiosa están detrás de la eficiencia energética. Lightmatter trabaja en interconexiones ópticas en lugar de eléctricas. AMD juega la baza de la capacidad de memoria. Y los hyperscalers (Google con sus TPU, Amazon con Trainium e Inferentia, Microsoft con Maia, Meta con MTIA) llevan años invirtiendo en sus propios chips para reducir su dependencia de Nvidia.

Tras analizar la fiebre de inversión en redes y silicio, la diferencia entre el actor que termine llevándose el premio y el que se quede fuera no estará tanto en la potencia bruta, sino en cómo orquesta jerarquías de memoria, capas de inferencia y patrones de tráfico. La industria de la inteligencia artificial, que como ya dije hace un par de semanas se ha convertido en una industria pesada, completamente dependiente de la inversión en centros de datos, se está dando cuenta de algo: para que la inteligencia artificial sea sostenible (económica, energética y socialmente), no se puede contar solo con hacerla más grande. **Hay que hacerla más lista. Más eficiente.** Saber qué tirar por la borda y qué conservar, qué calcular y qué reutilizar, qué resolver con un modelo gigantesco y qué con un modelo pequeño hiperespecializado.

Y eso me lleva a un caso muy cercano, muy práctico y muy ilustrativo, que ha contado recientemente el equipo de Mercadona Tech.

El caso Mercadona Tech: refinamiento, no fuerza bruta

José Ramón Pérez Agüera, responsable del equipo de Mercadona Tech, ha publicado hace un par de semanas en su blog personal, Gemba, un artículo en el que cuenta, con un nivel de detalle y transparencia muy poco habitual (y muy loable), cómo han construido el buscador de la aplicación de Mercadona[2]. Incluye,

[2]No, los de Mercadona no me pagan por mencionarles aquí ni en la radio ni nada. Ni siquiera me pagan por haberme convertido en el primer y principal apostol de sus carrilleras al vino. Si no las has probado, no sabes lo que te estás perdiendo.

además, archivos descargables con las claves necesarias para replicar el proyecto en cualquier otra empresa. Un acto de generosidad técnica que conviene aplaudir.

La app y la web de Mercadona reciben cuatro millones y medio de búsquedas a la semana. Eso significa una factura considerable y mucha infraestructura. Además, conseguir que el buscador esté ajustado a las características concretas del negocio (que entienda que "leche entera" debe priorizar marca propia, que un usuario que busca "papel" probablemente quiere papel higiénico antes que folios) y que no falle demasiado es complicado y fastidioso. Las soluciones genéricas de los grandes proveedores tienden a no estar lo suficientemente afinadas, y resultan caras y pesadas.

Pues bien, el equipo de Mercadona Tech aplicó precisamente algunas de las técnicas que acabamos de describir, en particular la cuantización, para reducir el peso de los cálculos. En algunas fases del proceso pasaron de tardar veinte milisegundos por búsqueda a tardar tres o cuatro. El modelo entero les cabe en algo más de cien megabytes de memoria, funciona sin necesidad de caras GPUs y han reducido enormemente la factura que pagaban a terceros, mejorando, al mismo tiempo, la calidad del buscador. Es un problema que ya no hay que resolver a base de fuerza bruta ni de pagarle una gran factura a un proveedor, sino que han resuelto con refinamiento.

Conviene, eso sí, matizar dos cosas.

La primera, que el buscador de Mercadona no usa inteligencia artificial generativa del tipo que se usa en un Claude o un GPT. Es un tipo de inteligencia artificial más convencional, más establecida, especializada en una tarea muy concreta (entender consultas en lenguaje natural y devolver productos relevantes). **Pero ilustra a la perfección lo que vengo contando: cuando uno se preocupa de verdad por la eficiencia, los resultados son espectaculares.**

La segunda, y aquí viene lo más bonito, es que el propio equipo de Mercadona Tech cuenta cómo han sacado adelante este proyecto en tiempo récord precisamente porque han usado inteligencia artificial generativa, en concreto Claude Code, para programarlo. Es decir, que la inteligencia artificial cara y poderosa de hoy les ha permitido construir, con poca gente y en poco tiempo, una pieza de tecnología que luego no necesita inteligencia artificial cara para funcionar.

Es una metáfora muy bonita del momento que vivimos: **usamos la inteligencia artificial cara y poderosa de hoy para construir la inteligencia artificial barata y sensata del mañana.**

Para ir acabando

Al próximo que te diga que la inteligencia artificial es una burbuja porque las cuentas no salen, ofrécele este matiz: las cuentas, efectivamente, no salen, pero precisamente por eso la industria está reinventándose.

Los próximos doce a veinticuatro meses no van a ir de modelos más grandes, sino de modelos más listos. De ingeniería de sistemas, de chips especializados, de paradigmas alternativos. **De gente como Yann LeCun apostando a que se puede hacer mucho con poco.** De equipos como el de Mercadona Tech demostrando, desde Valencia, que el refinamiento puede ser más valioso que la fuerza bruta.

Y, si me permites una predicción, **ese giro hacia la eficiencia es la mejor noticia que le podía pasar a Europa.** Porque, hasta ahora, la batalla se libraba en un

terreno (el del gasto masivo en infraestructura y entrenamiento) donde no teníamos manera de competir. El terreno de la eficiencia es otro juego, mucho más abierto.



Sobre el autor

Manuel Delgado Tenorio

Manuel Delgado Tenorio es asesor independiente de estrategia de datos e inteligencia artificial, profesor e investigador. Cofundó y vendió una *startup* de inteligencia artificial, fue socio de la práctica de Data & Analytics para servicios financieros en EY y responsable de Data Analytics para el sur de Europa en AWS. Es profesor asociado en IE University, donde imparte la primera asignatura de grado sobre Customer Lifetime Value en España, investiga sobre el CLV en su doctorado en la Universidad Complutense de Madrid y es autor del libro "Customer Lifetime Value (CLV): No todos tus clientes valen lo mismo".

manueldelgado.com · **LinkedIn** · **X**

Recibe cada artículo en tu correo con la newsletter Ex Datis, Claritas:
manueldelgado.com/newsletter/suscribete