

Flash Note: La IA abandona la nube y se ejecuta en tu escritorio. Implicaciones.

Hace unos días volví a escuchar las ponencias del CES 2026 (una de las mayores ferias tecnológicas del mundo, en donde las empresas presentan sus nuevas tecnologías de consumo). Y aunque escuché a un gran número de tecnólogos, puse especial atención en lo que diría Lisa Su (CEO de AMD) por ser esta una empresa que está aportando ahora algunos de los avances más importantes (junto con NVDA), ambas seguidas de cerca por mi equipo desde hace años.

Lisa Su (AMD) aprovechó el CES de este año para acercar a nuestro escritorio la ejecución local de modelos de IA, sin conexión a data centers, sin conexión a la nube. Lo que se conoce como el *edge computing*; y lo hizo mediante su AI box (llamado Ryzen AI Max+ 395). Un ordenador compacto que carga y ejecuta en local modelos de hasta 235.000 millones de parámetros, sin conexión a ningún data center y que ejecuta inferencia sin depender de la nube. Así, su velocidad y la calidad no dependen de factores externos, sino exclusivamente del modelo integrado y de cómo se ha comprimido.

Por supuesto, esta presentación suscitó una serie de preguntas que necesitan ser respondidas:

¿Es esta una IA propia de AMD que compite con la IA de Open AI o Anthropic?

No. AMD no vende un LLM. AMD vende el mini-PC que ejecuta modelos ya entrenados (por terceros, como Open AI o Anthropic) y que están descargados en esa caja por el usuario. No es, por lo tanto, un competidor de las empresas frontera que entrenan grandes modelos de IA.

¿Puede este mini-PC de AMD, y que ejecuta IA en local, amenazar el modelo de suscripción de las empresas frontera en la nube?

No. El modelo LLM que viene integrado en el mini-PC de AMD es más pequeño que los modelos a los que uno puede acceder por suscripción. AMD ha comunicado que el modelo de IA integrado en su mini-PC tiene 235.000 millones de parámetros; en cambio, los principales modelos propietarios se entrenan con un número muy superior de parámetros (¿4bn?), siendo estos últimos, por lo tanto, mucho más potentes. Eso se ve en el rendimiento en pruebas, velocidad y contexto. Por consiguiente, no deberíamos presentar este mini-PC de AMD con IA local como el equivalente a los modelos frontera por suscripción en la nube. La nube seguirá siendo más adecuada para aquellos

usuarios que requieran los modelos más potentes, grandes volúmenes y cargas que requieren escalar rápidamente.

¿Cuál sería entonces la utilidad de este mini-PC con IA en local de AMD?

Podría ser una alternativa parcial a la suscripción en la nube para ciertos clientes con ciertas tareas. El mini-PC de AMD tiene una IA local con capacidad suficiente para muchas de las tareas que realizamos hoy vía suscripción (y que debemos pagar a cada consulta). Pero se da el caso que los modelos grandes (como Qwen3) tienen muchos parámetros, pero solo ejecutan una pequeña fracción por consulta o tarea. Esas tareas que requieren una fracción de los parámetros podrían hacerse ahora en local (sin suscripción). Pongamos un ejemplo: un usuario intensivo de IA podría pagar por Claude Code Max (200 \$), ChatGPT Pro (200 \$), Cursor (20 \$) y Gemini (20 \$). Esto suma un total de 5.280 \$ al año. El mini-PC de AMD cuesta \$1.499 (un solo pago). El beneficio para el usuario sería sustituir parte del uso de la nube y encargarle al mini-PC de AMD tareas menos densas — de programación, búsqueda, análisis de documentos, o tareas programadas. Ello permitiría reducir algunas suscripciones, manteniendo el acceso a la nube para los trabajos más exigentes. ¿El ahorro sería automático? No lo sé. Tendría sentido si lo que deja de pagar por suscripción es superior al coste del mini-PC de AMD.

¿Qué TAM -Total Adressable Market- puede tener este mini-PC con IA de AMD?

Este dispositivo es potencialmente útil para un número muy amplio de funciones: desarrollar y probar asistentes de programación, agentes, búsquedas sobre documentos propios, etc. Y es especialmente útil cuando en estas tareas importa la latencia, la privacidad, o trabajar sin conexión. Su público inicial son empresas o personas con necesidades concretas de inferencia local: empresas de software que generan masivamente asistentes y agentes y quieren evitar tener que pagar por cada uno. Investigadores que experimenten en un entorno de privacidad durante el periodo de research. Bufetes que preparan documentación con IA, pero no quieren dar acceso a esos archivos a un proveedor externo. Etc.

No hay una estimación fiable del TAM de estas "cajas con AI embebida" (integrada) en particular. Como referencia amplia, la firma de análisis para el mercado tecnológico, International Data Corporation -IDC- proyecta un gasto global de 450.000 millones de dólares en edge computing para 2029, aunque la cifra incluye muchas más soluciones que estos mini-PC.

¿Es esto edge computing?

Sí. Es ejecución de la IA en un dispositivo próximo al usuario. NVDA y AMD cuentan ya con una cartera amplia de procesadores para edge industrial

Alex Fusté

*Chief Global Economist Andbank
Director of Wealth Management*

ANDBANK /
Private Bankers

(procesadores que se integran en máquinas o sistemas para procesar datos cerca de donde se generan). Estas empresas mantienen el liderazgo en todo el mercado del *edge computing*, con equipos que destacan en memoria unificada para IA local. ¡Y esto va rápido! Ayer mismo, 12 de octubre, AMD anunció HALO, una nueva plataforma de desarrollo para IA local con capacidad declarada para modelos de más de 300.000 millones de parámetros.

Significa que el *edge computing* (entendido como tu propia IA ejecutada cerca de ti, sin enviar cada consulta a la nube) será cada vez más potente, y más tuya.

¿Nuevo campo de inversión?

Todo esto abre una nueva etapa en el negocio de la informática personal. Si lo que queremos es ejecutar los mejores modelos de IA, pero en local (sin depender de fuentes externas), como así creo que queremos todos, se hacen necesarias arquitecturas más capaces. Con las actuales no basta para acometer ese gran paso. La apuesta de empresas como AMD y NVDA apunta en esa dirección, aunque aún es pronto para hablar de revolución consolidada en el ámbito de la informática personal. Su futuro dependerá de ver su precio, rendimiento en tareas reales, adopción, etc. Factores que nunca fueron obstáculos insalvables en el pasado.

Cordiales saludos,