Enterprise Strategy Group™ by TechTarget

DOCUMENTO TÉCNICO ECONÓMICO

Descripción del coste total de la inferencia de modelos grandes de lenguaje

Cómo utilizar las soluciones en las instalaciones de Dell Technologies puede ser entre un 38 % y un 88 % más rentable para la inferencia de LLM con RAG en comparación con la cloud pública y las API basadas en tokens

Por Aviv Kaufmann, responsable de gestión de oportunidades y analista de validación principal
Enterprise Strategy Group

Abril de 2024

Contenido

Introducción

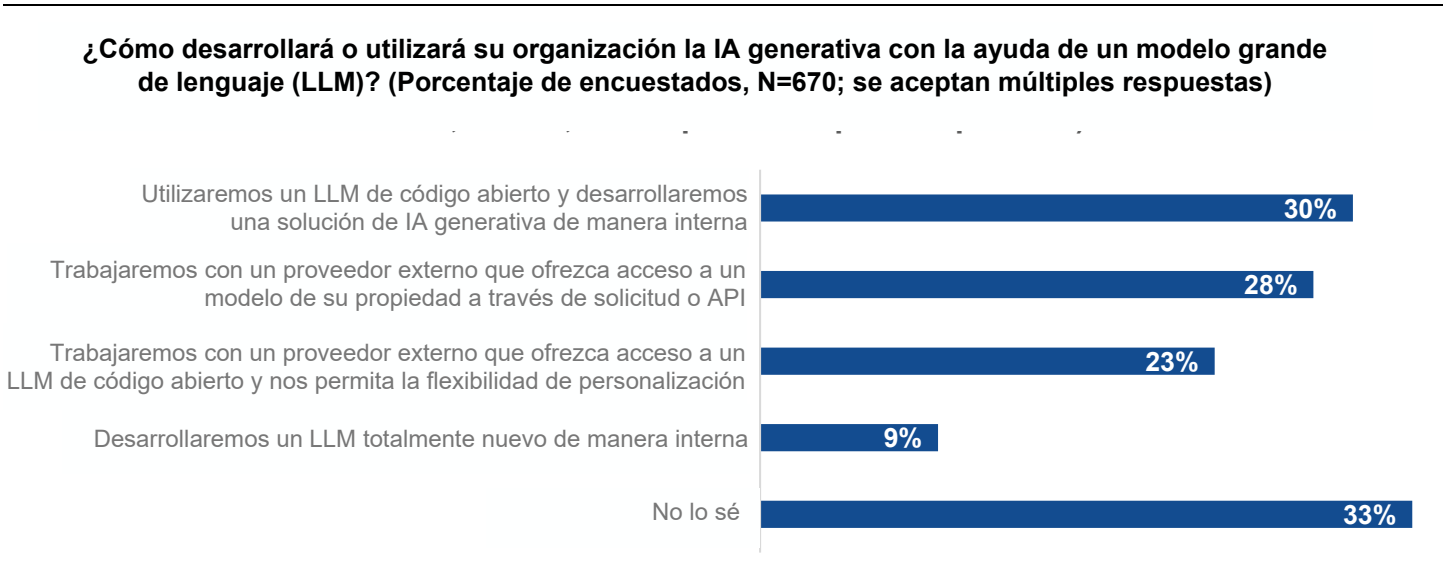
Este documento técnico económico presenta algunas de las opciones y consideraciones para la entrega de capacidades de IA generativa (GenAI) basada en texto a las organizaciones. Enterprise Strategy Group de TechTarget modeló y comparó los costes previstos de la inferencia de modelos grandes de lenguaje (LLM) utilizando generación aumentada por recuperación (RAG) en la infraestructura de Dell Technologies en las instalaciones frente al uso de una infraestructura como servicio (IaaS) de cloud pública nativa o el servicio de modelos de LLM OpenAI GPT-4 Turbo a través de una API. Determinamos que Dell Technologies podía ofrecer una inferencia de LLM hasta 4 veces más rentable que una IaaS y hasta 8 veces más rentable que con GPT-4 Turbo API.

Retos

Las organizaciones están adoptando el poder de la IA generativa y los LLM para utilizar los datos específicos de la empresa y otra propiedad intelectual para automatizar la generación de contenidos, responder preguntas y generar información disponible para los responsables de la toma de decisiones. Junto con otros muchos beneficios, los participantes en un estudio de investigación de Enterprise Strategy Group informaron de que los principales beneficios de usar la IA generativa en su organización incluyen la mejora o automatización de los procesos y flujos de trabajo, la capacidad de análisis de datos e inteligencia empresarial, el aumento de la productividad de los empleados y la mejora de la eficiencia operativa.[1]

Los LLM pueden ser costosos y complejos de desarrollar, pero las organizaciones pueden ampliar, ajustar con precisión y personalizar fácilmente los LLM de código abierto existentes para satisfacer sus necesidades. Los servicios basados en API ya preparados, como OpenAI GPT, ofrecen una solución más sencilla, pero los costes de inferencia (es decir, consulta) pueden aumentar con rapidez, especialmente en el caso de las organizaciones de mayor tamaño y los LLM más complejos. Como alternativa, las organizaciones pueden crear y controlar su propia solución de inferencia de LLM en potentes servidores empresariales habilitados por GPU o instancias de cloud habilitadas por GPU equivalentes y una plataforma de aprendizaje automático como AI Enterprise de NVIDIA que ejecute LLM de código abierto. No es de extrañar que Enterprise Strategy Group determinara que la estrategia más popular para que las organizaciones desarrollen y utilicen la IA generativa con la ayuda de un LLM era utilizar un LLM de código abierto y desarrollar una solución de IA generativa de manera interna.[2]

Figura 1. La mayoría de las organizaciones tienen previsto desarrollar su propia solución de IA generativa de manera interna



Fuente: Enterprise Strategy Group, una división de TechTarget, Inc.

[1] Fuente: informe de investigación de Enterprise Strategy Group, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, agosto de 2023.
[2] Ibidem

Consideraciones clave sobre la inferencia de LLM

Los LLM basados en texto se basan en aprender, comprender y producir contenidos, respuestas, resúmenes y preguntas que se basan en texto y que se pueden personalizar para un sector, un caso de uso o una organización en concreto. RAG aumenta los resultados de los modelos de IA generativa con datos personalizados extraídos de fuentes adicionales, lo que hace que los modelos sean más precisos. Estos son los LLM más implementados para empresas y se pueden usar para chatbots, asistentes de preguntas y respuestas, mejora y automatización de procesos o como capacidades integradas en herramientas y aplicaciones personalizadas, además de para otros muchos casos de uso. Al distribuir modelos de LLM, las organizaciones deben tener en cuenta la infraestructura para entrenamiento (es decir, los análisis con uso intensivo de datos y computación necesarios para crear un modelo eficaz), inferencia (es decir, entrega de interacciones con los usuarios en un modelo entrenado) y ajuste preciso (es decir, actualización y optimización continuas del modelo). Este informe se centra en la infraestructura necesaria para facilitar la inferencia de cargas de trabajo. Hay varios métodos de implementación que se pueden usar para la inferencia de LLM, entre ellos:

- **Infraestructura tradicional.** La infraestructura tradicional adquirida o arrendada se compone de computación, memoria, GPU y almacenamiento que se pueden implementar y gestionar junto con una plataforma de IA comercial o de código abierto, lo que otorga a la organización un control de todos los aspectos de la implementación. Este método puede ser el más rentable para las cargas de trabajo más grandes y predecibles.
- **IaaS de cloud pública.** Del mismo modo, las organizaciones pueden implementar instancias de computación en la cloud equivalentes con GPU y almacenamiento, junto con una plataforma de IA comercial o de código abierto. Este método proporciona un control similar de la plataforma, con agilidad e integración fácil con las herramientas existentes. Este método puede ser el más rentable para las pequeñas implementaciones y aquellas con requisitos impredecibles o estacionales.
- **Servicios de API de LLM.** Se pueden utilizar servicios establecidos, como OpenAI GPT, para proporcionar rápidamente las capacidades sin tener que gestionar la infraestructura o una plataforma de IA. Este método puede ser el mejor para explorar y dar los primeros pasos, para implementaciones más pequeñas y para aquellas que no requieren un alto nivel de personalización o control.

Antes de decidirse por una plataforma de LLM, las organizaciones deben dedicar tiempo a conocer sus requisitos y capacidades, además de hablar sobre algunas de las siguientes consideraciones en torno a la elección de una plataforma para la inferencia de LLM, como:

- **Coste/ROI.** Las organizaciones deben plantearse el coste y los beneficios de implementar y usar cada inversión en tecnología. Según un estudio de investigación de Enterprise Strategy Group, los ahorros de costes y el ROI fueron las métricas más comunes que las organizaciones afirman que utilizan para medir la eficacia de sus iniciativas de IA.[3]
- **Rendimiento y capacidad de ampliación.** El dimensionamiento de la infraestructura con recursos suficientes de procesadores, GPU, memoria y almacenamiento es importante para garantizar que sea capaz de gestionar la concurrencia inesperada de inferencia en cargas normales y máximas y que la latencia de inferencia media sea lo suficientemente baja para proporcionar a los usuarios una experiencia positiva. Las organizaciones también deben determinar si el entrenamiento con un uso intensivo de computación del LLM se producirá en la misma plataforma o en una plataforma de entrenamiento específica con mayor rendimiento antes de trasladarlo a la plataforma de inferencia.
- **Gestión sencilla.** Al comparar cualquier infraestructura en las instalaciones con la infraestructura y los servicios de cloud, es importante que una organización tenga en cuenta sus capacidades internas y conozca los costes de funcionamiento de la infraestructura y las plataformas (por ejemplo, administración, asistencia y mantenimiento y alimentación/refrigeración). Las opciones de housing también permiten a las organizaciones obtener muchos de los beneficios de este tipo de modelo en sus propios centros de datos y liberan, al mismo tiempo, los recursos y las habilidades que se necesitan para utilizar la infraestructura y la plataforma.
- **Cargas de trabajo de usuarios previstas.** Conocer y predecir cuántos usuarios accederán a la herramienta y con qué frecuencia plantearán preguntas al día es una métrica importante que tener en cuenta al elegir una solución. Si la demanda es pequeña, un servicio de API puede ser suficiente, pero cuando una organización ofrezca asistencia a más usuarios e inferencias, será más rentable crear una plataforma de su propiedad. Es importante que las organizaciones tengan en cuenta el crecimiento previsto en la adopción y la frecuencia de uso a lo largo del tiempo para asegurarse de que la infraestructura tiene el tamaño adecuado y puede crecer con las necesidades del negocio.
- **Gobernanza de datos.** Las organizaciones deben tener en cuenta los requisitos de ubicación y gobierno de datos de las fuentes de datos que se requieren para entrenar y mantener el modelo. La infraestructura de cloud híbrida funcionará mejor cuando los datos residan localmente o se puedan recuperar fácilmente cuando sea necesario, mientras que la

[3] Fuente: informe de investigación de Enterprise Strategy Group, *Navigating the Evolving AI Infrastructure Landscape*, septiembre de 2023.

cloud pública puede llevar a cabo la recopilación y centralización de los datos de forma más sencilla en algunos casos. Las instancias en las instalaciones también permiten a las organizaciones controlar mejor la seguridad y garantizar el cumplimiento normativo en lo que a seguridad de los datos confidenciales respecta. El entrenamiento y el mantenimiento de datos actualizados, completos y no sesgados producirán un mejor LLM e información más precisa derivada de la inferencia.

Análisis económico de Enterprise Strategy Group

Enterprise Strategy Group creó un análisis económico que comparaba los costes previstos de la entrega de inferencia para varios LLM de código abierto que utilizan RAG de varias complejidades (con el número de parámetros, incluidos 7000 millones y 70 000 millones) y para organizaciones de diferentes tamaños (con el número de usuarios entre 5000 y 50 000). Asumimos que el modelo ofrecía preguntas y respuestas internas basadas en texto y que la inferencia se producía donde estaban ubicados los datos, por lo que la migración de datos no tenía un coste elevado. En el análisis se observaron todos los costes asociados con la ejecución e inferencia de los modelos a lo largo de un periodo de tres años, lo que incluye proporcionar y ejecutar la infraestructura, administrar los sistemas y pagar por los servicios de cloud, si es necesario.

Infraestructura en las instalaciones de Dell Technologies frente a IaaS de cloud pública

En primer lugar, nuestros modelos compararon el coste previsto de ejecutar la inferencia de LLM en una infraestructura tradicional (en las instalaciones, en entornos de housing, en ubicaciones perimetrales, etc.) con la ejecución en una IaaS de cloud pública con una configuración similar en instancias de Amazon EC2. Se dimensionaron los requisitos de configuraciones del servidor de nodos de inferencia y las GPU NVIDIA H100 para cada carga de trabajo en función de los resultados de las pruebas de base de referencia de inferencia para garantizar que pudieran gestionar los requisitos de concurrencia en cargas normales y máximas (en función de las solicitudes máximas y el número de instancias del modelo), además de proporcionar una latencia y un rendimiento adecuados para cada carga de trabajo prevista. A continuación, modelamos cada uno de los costes descritos en la Tabla 1 para la infraestructura de Dell Technologies y para la configuración equivalente de EC2.

Tabla 1. Costes y supuestos modelados para los requisitos de cada carga de trabajo de inferencia de LLM

Categoría de costes	Dell Technologies (en las instalaciones)	IaaS de cloud pública (Amazon EC2)
Coste inicial de adquisición (hardware y software)	Precio proporcionado por Dell Technologies para Dell PowerEdge R760xa y R660 con ProDeploy y ProSupport	N/A
Coste adicional de capital (intereses) y depreciación (beneficio)	Incluido en el modelo (8 % CMPC, 6 % beneficio de depreciación anual)	N/A
Coste de alimentación y refrigeración	Calculado en función de las especificaciones del sistema (0,173 $/kWh)	N/A
Gasto mensual en cloud	N/A	Costes de la instancia de EC2 calculados en función de descuentos por reserva de 3 años
Licencia/GPU NVIDIA AI Enterprise	Se basa en licencia de 5 años (prorrateados)	Por instancia/h, basados en 16 h/día, 5 días a la semana para limitar los costes
Administración de infraestructura/instancias	Modelado (10 %-100 % de administración del sistema en función del número de nodos)	66 % menos que el modelo en las instalaciones
Administración de modelos y plataformas de aprendizaje automático	Modelado (20 %-100 % de ingeniería de aprendizaje automático en función de la cantidad de instancias de modelo)	Igual que el modelo en las instalaciones

Fuente: Enterprise Strategy Group, una división de TechTarget, Inc.

Modelo de tamaño más pequeño: LLM Mistral 7B con 7000 millones de parámetros

Para la primera comparación, modelamos los costes de entrega de un modelo más pequeño que contiene aproximadamente 7000 millones de parámetros, similar al LLM Mistral 7B de código abierto. Para dimensionar los requisitos, utilizamos una herramienta de dimensionamiento basándonos en los resultados de las pruebas que predecían las configuraciones de servidor y GPU que serían capaces de ofrecer una latencia media por solicitud de aproximadamente 0,4 segundos y un rendimiento estimado de entre 2,29 y 6,86 inferencias por segundo. En la Tabla 2, se muestran los supuestos de alto nivel de recuentos de instancias y GPU.

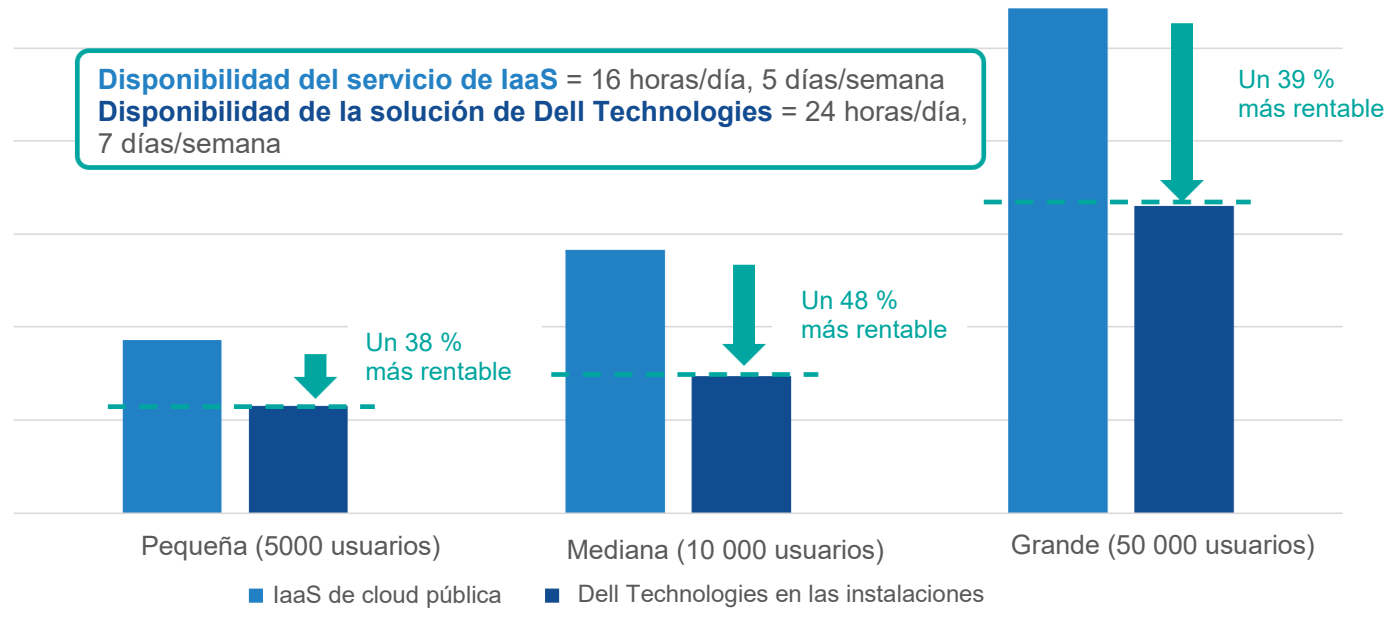
Tabla 2. Supuestos de configuración para la inferencia del modelo Mistral con 7000 millones de parámetros

Modelo de LLM (número de parámetros)	Número de usuarios	Número de nodos/instancias de inferencia	Número de GPU H100
Mistral (7B)	5000	1	1
	10 000	1	2
	50 000	1	4

Fuente: Enterprise Strategy Group, una división de TechTarget, Inc.

A continuación, modelamos todos los costes resumidos en la Tabla 1 para cada configuración. Como se muestra en la Figura 3, la infraestructura de Dell Technologies fue entre 1,6 y 1,9 veces (entre un 38 % y un 48 %) más rentable en la entrega de inferencia para la organización. Además, estaba disponible para la organización 24x7.

Figura 2. Coste previsto de la entrega de inferencia para un LLM Mistral con 7000 millones de parámetros con RAG



Fuente: Enterprise Strategy Group, una división de TechTarget, Inc.

Modelo de mayor tamaño: LLM Llama 2 con 70 000 millones de parámetros

A continuación, modelamos los costes previstos de la entrega de un modelo de mayor tamaño con 70 000 millones de parámetros, similar al LLM Llama 2 70B de código abierto. Volvimos a dimensionar los requisitos con la misma herramienta de dimensionamiento para predecir las configuraciones de servidor y GPU que serían capaces de ofrecer una latencia media por solicitud ligeramente superior de aproximadamente 1,8 segundos y un rendimiento estimado de entre 2,29 y 22,86 inferencias por segundo. En la Tabla 3, se muestran los supuestos de alto nivel de recuentos de instancias y GPU.

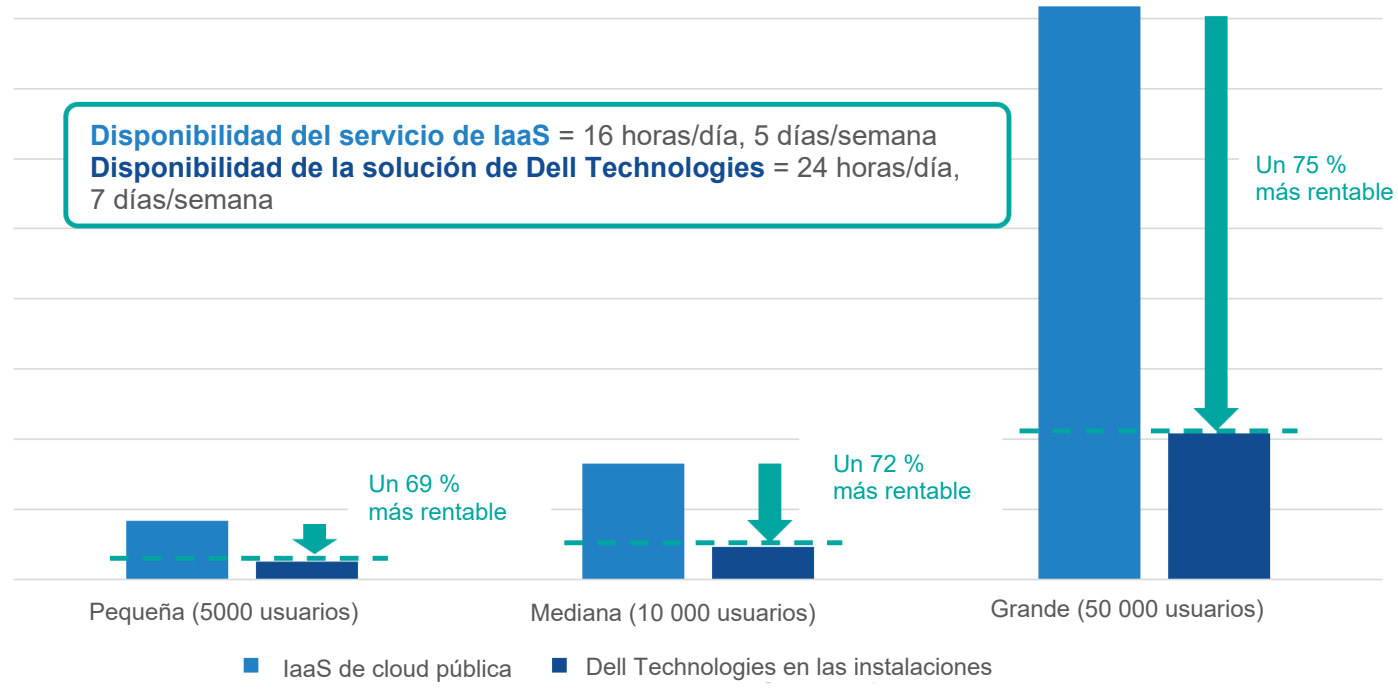
Tabla 3. Supuestos de configuración para la inferencia del modelo Llama 2 con 70 000 millones de parámetros

Modelo de LLM (número de parámetros)	Número de usuarios	Número de nodos/instancias de inferencia	Número de GPU H100
Llama 2 (70B)	5000	2	8
	10 000	4	16
	50 000	20	80

Fuente: Enterprise Strategy Group, una división de TechTarget, Inc.

Después de modelar de nuevo todos los costes resumidos en la Tabla 1 para cada configuración de las que se muestran más arriba, determinamos que la infraestructura de Dell Technologies era entre 3,3 y 4 veces (entre un 69 % y un 75 %) más rentable en la entrega de inferencia para la organización, además de estar disponible para la organización 24x7.

Figura 3. Coste previsto de la entrega de inferencia para un LLM Llama 2 con 70 000 millones de parámetros con RAG

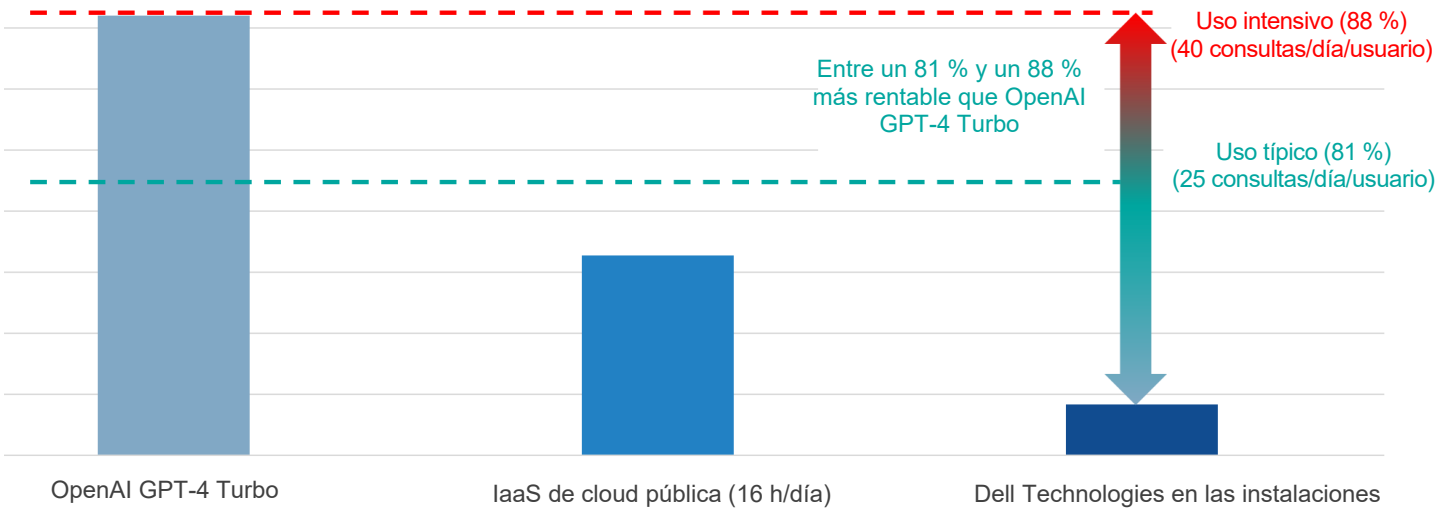


Fuente: Enterprise Strategy Group, una división de TechTarget, Inc.

Infraestructura en las instalaciones de Dell Technologies frente a servicio de IA generativa basado en API

A continuación, comparamos los costes previstos para que una gran organización proporcione un modelo de 70 000 millones de parámetros equivalente a sus 50 000 usuarios utilizando el servicio de IA generativa establecido GPT-4 Turbo, basado en la API OpenAI, que tiene un precio rentable por "token" de entrada y salida. Las preguntas y respuestas basadas en texto requieren una intensidad de tokens moderada por consulta, no tienen muchas variaciones en la carga máxima y producen un equilibrio relativamente uniforme entre el número de tokens de entrada y salida necesarios. Asumimos un total de 1500 tokens (entrada y salida) por consulta, con una media de unas 25 consultas por día y usuario, tanto para las soluciones en las instalaciones como para las basadas en API. Basándonos en nuestra investigación de las afirmaciones públicas, determinamos que este es un número moderado de consultas por usuario, ya que las organizaciones menos establecidas generan menos consultas por usuario y las más establecidas tienen una media de hasta 40 consultas por usuario y día. Nuestros cálculos de GPT-4 Turbo predijeron un coste de aproximadamente 12,50 $/usuario/mes, lo que se compara favorablemente con las herramientas de asistencia de IA basadas en conjuntos que pueden costar aproximadamente 30 $/usuario/mes. Partiendo de estos supuestos, determinamos que la infraestructura en las instalaciones de Dell Technologies podía proporcionar una inferencia entre 5,4 y 8,6 veces (entre un 81 % y un 88 %) más rentable que utilizando un servicio basado en API, con entrega de capacidades de IA generativa por solo aproximadamente 2,31 $/usuario/mes.

Figura 4. Coste previsto de la entrega de inferencia para un LLM Llama 2 con 70 000 millones de parámetros a 50 000 usuarios



Fuente: Enterprise Strategy Group, una división de TechTarget, Inc.

Cuestiones que tener en cuenta

Aunque los modelos de Enterprise Strategy Group se han diseñado de buena fe con supuestos conservadores, creíbles y validados, ninguna situación modelada representará nunca un potencial entorno. El ahorro de los clientes dependerá de su caso de uso concreto, la naturaleza de sus datos, su nivel de experiencia y los requisitos de su modelo e infraestructura. Enterprise Strategy Group recomienda que realice su propio análisis de los productos disponibles y consulte a Dell Technologies para entender y hablar de las diferencias entre las soluciones probadas a través de sus propias pruebas de concepto.

Dell Technologies para inferencia de LLM

Dell Technologies ayuda a las organizaciones a integrar la IA fácilmente en sus datos, sin importar dónde residan. Esto significa ofrecer la cartera de servicios de IA más amplia (del equipo de sobremesa al centro de datos y a la cloud), de modo que las organizaciones puedan dimensionar bien sus inversiones y aprovechar los datos para crear sus fábricas de IA y materializar los casos de uso de IA de forma eficiente, segura y sostenible. Para ello, Dell proporciona acceso a una completa cartera de servicios y un amplio ecosistema abierto de socios para ayudar a las organizaciones sin importar en qué parte de su transición a la IA se encuentren, ya estén desarrollando estrategias de IA o acelerando y ampliando sus inversiones en IA generativa.

Para las organizaciones que se enfrentan a amenazas de seguridad de datos, cuestiones de cumplimiento normativo, silos de datos y conjuntos de datos no validados, los Dell Professional Services para IA generativa pueden ayudar a generar consenso entre los líderes empresariales y de TI en torno a los casos de uso priorizados, proporcionar un plan de trabajo útil para alcanzar los objetivos, preparar los datos empresariales para la integración de LLM, desarrollar la madurez en ciberseguridad y establecer una plataforma de IA alineada con las necesidades específicas del negocio. Además, con Dell APEX, las organizaciones pueden suscribirse a soluciones de IA y optimizarlas para casos de uso multicloud.

Para obtener más información sobre las soluciones de Dell, visite la página web de IA de Dell.

Conclusión

El mayor uso de IA generativa en casi todas las áreas del negocio es un factor crucial para garantizar la mejora de las operaciones y el éxito futuro. La investigación de Enterprise Strategy Group revela que las principales áreas en las que las organizaciones están aplicando actualmente la IA generativa incluyen la investigación, el marketing, el desarrollo de software, el desarrollo de productos y las operaciones de TI, y se espera que el potencial de uso en cada área aumente.[4] Las organizaciones pueden obtener resultados más impactantes y significativos entrenando su propia versión personalizada de un LLM y realizando la inferencia con este.

Hay varios métodos de implementación que se pueden usar para la inferencia de LLM y cada uno proporciona ventajas para casos de uso y requisitos particulares. Para las organizaciones con miles de usuarios preparadas para aprovechar las capacidades que incluye un LLM personalizado, la infraestructura de Dell Technologies puede ofrecer inferencia de LLM de alto rendimiento de forma hasta 4 veces más rentable que una IaaS y hasta 8 veces más rentable que con OpenAI GPT-4 Turbo. Enterprise Strategy Group recomienda encarecidamente que las empresas que implementen LLM para impulsar sus organizaciones se planteen el uso de las rentables tecnologías y los servicios expertos que ofrece Dell Technologies para garantizar la obtención de resultados, acelerar sus iniciativas de IA generativa y reducir el tiempo hasta conseguir los ahorros previstos.

[4] Fuente: informe de investigación de Enterprise Strategy Group, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, agosto de 2023.

Acerca de Enterprise Strategy Group
Enterprise Strategy Group de TechTarget realiza informes específicos y viables de inteligencia de mercado dirigidos al sector de la demanda y ofrece servicios de asesoramiento por parte de analistas, orientación estratégica de GTM, validaciones de soluciones y contenido personalizado que justifican la compra y venta de tecnología empresarial.

contact@esg-global.com
www.esg-global.com