

**XLIX Congreso del Instituto Argentino de Profesores Universitarios de
Costos**

EL COSTO DE LA IA

Categoría: Aporte a la disciplina

Autores:

**Julio Marchione (Socio Activo)
Marcelo Podmoguilnye (Socio Activo)
Gustavo Metilli (Socio Activo)**

República Argentina, La Punta, San Luis, Septiembre 2026

INDICE

Resumen

- 1. Introducción**
- 2. Antes de calcular costos es importante entender cómo funciona el ecosistema de negocios.**
- 3. Los factores a tener en cuenta en el cálculo del costo de producción de los cómputos o tokens generados por la IA**
- 4. Los factores que influyen en los costos de producción de aplicaciones de IA dentro de las empresas**
- 5. A pesar de la vorágine evolutiva del sector, ¿qué conceptos claves podemos rescatar?**
- 6. Conclusiones**

Bibliografía

El costo de la IA

RESUMEN

Hace un poco más de dos años atrás, el CEO de NVIDIA Jensen Huang, declaró que el costo del cómputo era de “aproximadamente cero”. Obviamente, sólo es posible sostener esta afirmación en base al costo incremental directo de cada cómputo de procesamiento, en función del tiempo que requiere procesarlo.

Pero, no está incluyendo toda la estructura de recursos puestos detrás de estos cómputos, materializados en “tokens” y basados en costos de capacidad desproporcionados.

Más allá de los costos de procesamiento que tienden a cero, hay que distinguir el costo de fabricación de las GPU (unidades de procesamiento de datos) que consumen minerales con alta concentración de proveedores y elevados costos de extracción y los costos de la fase de entrenamiento de la IA basado en el consumo energético y alto impacto ambiental.

Los objetivos de la ponencia son los siguientes:

- Analizar los recursos necesarios para la producción de los cómputos, contemplando los propios de la infraestructura para disponer de las GPU y su entrenamiento.
- Esquematizar la ecuación económica de costos de procesamiento de la IA.
- Dimensionar el impacto ambiental, social y energético de la producción de los cómputos.
- Evaluar la ecuación económica del negocio de procesamiento de la IA.

Palabras clave: cómputos de IA, costo de la GPU, inteligencia artificial generativa; costos de tokens; LLM; inferência; IA local; costos ocultos; gestión estratégica de costos

El costo de la IA

1. Introducción

La inteligencia artificial generativa se ha incorporado aceleradamente a organizaciones, profesiones y actividades empresariales bajo promesas de productividad, automatización e innovación. Sin embargo, buena parte de su adopción se produce sin una medición sistemática de los costos que la hacen posible.

El precio visible de una suscripción o de un servicio de API no agota el costo económico de la inteligencia artificial: detrás de cada interacción existen consumos de tokens, infraestructura computacional, energía, almacenamiento, mantenimiento, validación humana, privacidad, capacitación y dependencia tecnológica.

El presente trabajo, propone analizar la inteligencia artificial generativa, y agéntica también, como un nuevo objeto de costeo y gestión, considerando al TOKEN como unidad mínima de consumo económico-operativo.

A partir de literatura reciente sobre costos de inferencia, energía, modelos como servicio, despliegue local y eficiencia en el uso de tokens, se plantea una matriz conceptual para comparar alternativas de adopción: suscripción, API por consumo, infraestructura local y modelos híbridos.

La adopción de IA generativa y la llamada agéntica, en las organizaciones, no debe evaluarse sólo por productividad, automatización o innovación, sino por su estructura de costos: consumo de tokens, costos de inferencia, dependencia de proveedores, privacidad, infraestructura, mantenimiento, amortización y riesgo de submedición.

Éste puede ser un enfoque muy interesante para tener en cuenta desde nuestros roles de Gestión en distintos tipos de organizaciones y, obviamente, desde nuestras experticias en Costos y Gestión porque permite trabajar con: costo por unidad de uso, costos variables, costos fijos, punto de equilibrio, tercerización, producir o comprar, costos ocultos, costos de no calidad, costos de oportunidad, presupuestación, control de gestión y toma de decisiones.

Para ello intentaremos diferenciar en el trabajo 2 grandes encuadres sobre los costos de la IA :

- 1) proponer o distinguir una metodología de cálculo del costo del cómputo para quien es el productor de tokens (lo planteamos por batch productivo de 1 MM de tokens); y
- 2) Proponer el cálculo del costo de una aplicación de IA en una empresa, desarrollada o adquirida en forma total o parcial.

En ese sentido podríamos descubrir la existencia de una correlación entre el costo de producción del token con el costo de desarrollo de una aplicación de IA en una empresa., o al menos dejarlo planteado como desafío para próximas investigaciones.

2. Antes de calcular costos es importante entender cómo funciona el ecosistema de negocios.

Está claro que resulta difícil tomar una foto del ecosistema de negocios de la producción y consumo de la IA. La dinámica al límite de la velocidad de los cambios desnaturaliza la caracterización que uno pudo haber construido ayer. De todos modos, la película que estamos

visualizando nos presenta algunos participantes que -claramente- están adquiriendo alguno de los papeles protagónicos en esta dinámica.

Es por esto, que resulta importante identificarlos, para comprender quiénes son los sujetos de decisión, para luego enfocar el objeto de esas decisiones y definir las unidades de costeo en función de esto. En la siguiente figura, las diferencias de escala quedan claramente expuestas y servirá para identificar a los participantes en el ecosistema de negocios de la IA.

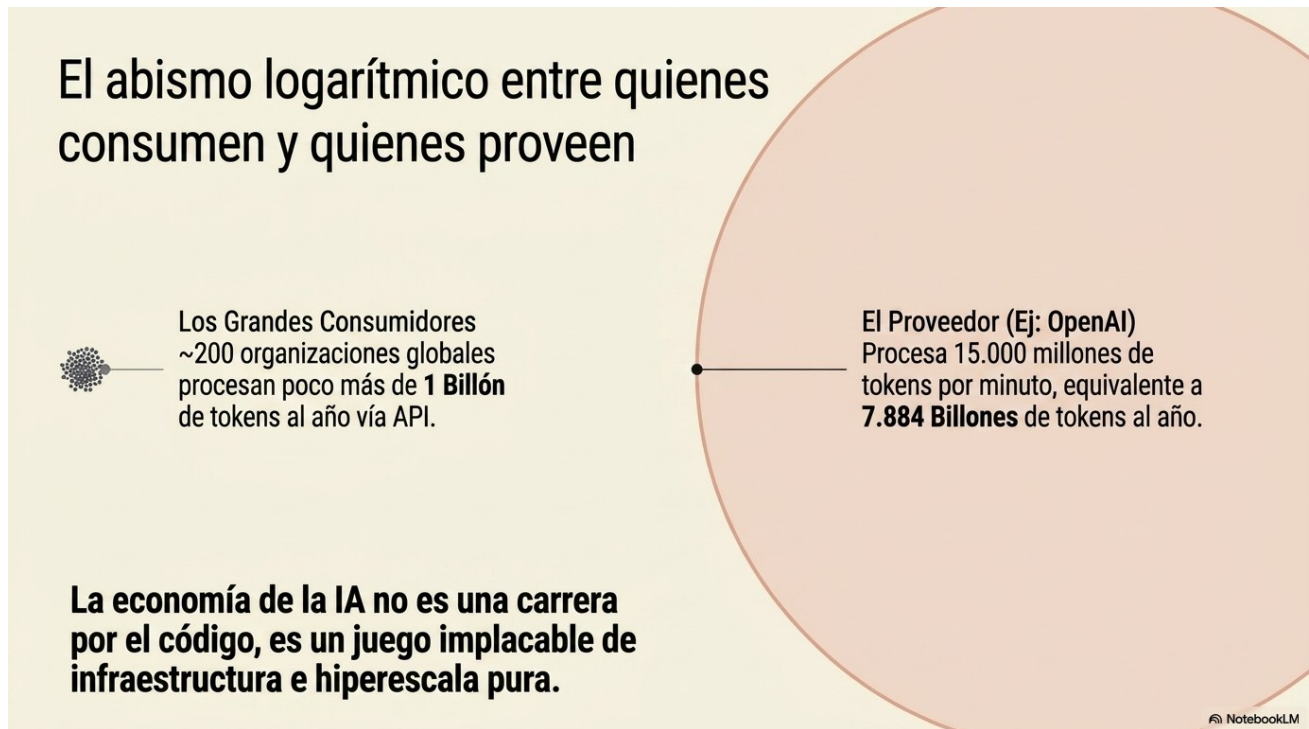


Figura 1 – Diferencias de magnitudes entre grandes consumidores de tokens y el principal proveedor (elaboración propia)

En la actualidad, se pueden identificar los siguientes participantes:

- Productores de plataformas de entrenamiento y desarrollo de programas de IA en grandes centros de datos refrigerados.
- Proveedores de herramientas de soporte para desarrolladores de IA y de nube para alojar los desarrollos.
- Proveedores de energía para las granjas de producción de IA.
- Fabricantes de semiconductores que diseñan chips especializados como unidades de procesamiento gráfico (GPU)
- Diseñadores de estructuras que sirven para entrenar los grandes modelos de lenguaje y redes neuronales base.
- Curadores de datos que proveen y limpian los conjuntos de datos masivos para entrenar los sistemas.
- Desarrolladores de aplicaciones de IA genéricas o verticales para una actividad económica en particular.

- Consumidores y usuarios de aplicaciones de IA (personas, instituciones y empresas).
- Gobiernos y entidades gubernamentales locales, regionales y globales como custodios de los avances, resultados y consecuencias de los desarrollos de IA.

En la dinámica de todo ecosistema de negocios, se destacan algunos participantes por sobre otros, sólo por brindar valor en la cadena de una manera diferente.

Un ejemplo de esto es la empresa Vercel que, a diferencia de OpenAI, Google, Microsoft o Meta, se enfocó en evitar competir con estas corporaciones y ayuda a los desarrolladores a construir, desplegar y alojar aplicaciones web. Los agentes de IA eligen el lenguaje desarrollado por Vercel ([Next.js](#) es el framework de código abierto de Vercel) para publicar sus desarrollos en la web.

Como este ejemplo de un unicornio argentino que rápidamente logró insertarse en el universo de los desarrollos de IA, es innegable que aparecerán otros a partir del comportamiento de los agentes de IA que eligen los recursos a utilizar orgánicamente, para optimizar su desempeño y avanzar a mayor velocidad. Estas circunstancias son claves en la configuración de un contexto que no sólo es dinámico, sino que además es violentamente cambiante.

Por esta razón, sabemos que los cálculos y ponderaciones de recursos utilizados en los próximos apartados van a sufrir cambios en el corto plazo y que pueden llegar a quedar desactualizados o reconfigurados en términos de procesos y recursos.

Con el objetivo de identificar a los posibles usuarios de la información sobre costo de producción de la IA, hemos seleccionado como sujetos de decisión a las empresas procesadoras de tokens de salida y entrada a partir de los pedidos de los desarrolladores y de los usuarios finales, y por otro lado, a los desarrolladores de aplicaciones de IA genéricas o verticales, como así también a las empresas consumidoras estos desarrollos.

En el primer grupo de sujetos de decisión, se puede identificar la provisión del servicio de inferencia de tokens de salida y entrada para quienes necesitan este insumo para desarrollar aplicaciones verticales de IA y para los usuarios finales que consumen aplicaciones de IA genéricas (por ejemplo, OpenAI o Anthropic).

En el segundo grupo, se encuentran los desarrolladores de aplicaciones verticales de IA, quienes se orientan a las empresas que los contratan para aplicar esta herramienta en los procesos internos de los negocios y a la dinámica para elegir hacia dónde enfocar la estrategia de precios, en un mercado en expansión violenta.

A continuación, en los puntos 3 y 4, se profundizará en la ecuación económica de costos en ambos universos de decisores y los recursos necesarios para avanzar con aplicaciones y herramientas de IA.

3. Los factores a tener en cuenta en el cálculo del costo de producción de los cómputos o tokens generados por la IA

A partir de los principios básicos de Costos, es posible plantear una unidad de costeo estandarizada: el batch productivo de 1 millón de tokens. Esto permite la determinación del costo unitario sin reparar en el volumen total procesado y delimitando la asociación de costos de estructura y de entrenamiento.

Antes de avanzar, resulta fundamental entender el proceso de inferencia, que es el que produce los tokens. Esta macroactividad convierte la consulta del usuario en una secuencia de tokens de salida. Dentro de este proceso se desarrollan una serie de actividades encadenadas que deben ser comprendidas, para proceder a una determinación de costos y que se detallan a continuación::

- Recepción del requerimiento: en esta actividad el usuario ingresa un prompt. Este texto se convierte en un recurso clave para la generación del proceso.
- Tokenización: el texto del prompt se descompone en unidades denominadas tokens. En esta etapa comienza el consumo de los recursos productivos.
- Carga del modelo: el modelo previamente entrenado se encuentra almacenado en memoria GPU. Los miles de millones de parámetros del modelo son accedidos para realizar el cálculo.
- Procesamiento neuronal: La GPU ejecuta millones de operaciones matriciales. El modelo calcula probabilidades por cada token para cada posible palabra o fragmento.
- Generación del token: se selecciona el token más probable y se lo genera como token de salida.
- Retroalimentación secuencial: el token producido vuelve a ingresar al modelo y se recicla cientos o miles de veces, según sea necesario.
- Entrega de la respuesta: la secuencia completa de tokens se transforma nuevamente en un texto legible, y es aquí donde el proceso secuencial termina.

En este proceso de inferencia, es posible identificar los factores asociados al mismo y que resultan claves para la determinación posterior del costo de producción:

- GPU utilizadas en el proceso de inferencia.
- CPU de apoyo.
- Memoria RAM.
- Almacenamiento temporal.
- Energía eléctrica.
- Licencias de software de los servidores.
- Personal operativo afectado al proceso secuencial.

Teniendo en cuenta esta información, es posible arribar a un costo de inferencia por hora de USD 4.20. El mismo se encuentra integrado como muestra el siguiente cuadro:

Conceptos	USD/hora
GPU H100	3.50
CPU	0.30
RAM	0.15

Red	0.05
Energía	0.20
TOTAL	4.20

Cuadro I - Costo del proceso de inferencia por hora de ejecución (elaboración propia)

Ahora bien, sabiendo que una GPU H100 genera técnicamente 10.000 tokens por segundo, resulta en una generación de 36.000.000 de tokens por hora, alcanzando un costo directo por 1 millón de tokens de la siguiente forma:

$$\text{Costo directo} = \text{USD } 4.20 / 36.000.000 \text{ tokens hora} \times 1.000.000 \text{ tokens} = 0.117 \text{ USD/MM tokens}$$

Sin embargo, el proceso de generación de tokens también se encuentra afectado por otras dos actividades que son parte del proceso y que no han sido incluidos en el cálculo de los recursos aplicados al proceso de producción : a) el entrenamiento de la IA y su ajuste progresivo; y b) la infraestructura general para el funcionamiento de todo el proceso.

Haciendo un análisis genérico y por aproximación promedio podríamos suponer los siguientes parámetros a considerar en los cálculos::

- a) Sobre un costo total de entrenamiento de USD 4.000.000, se proyecta -en la actualidad- obtener una vida útil de 10.000.000 millones de tokens. Esto resulta en un costo por millón de tokens de USD 0,40
- b) Al considerar la infraestructura en general, es posible estimar -en la actualidad- que por cada USD 5.000.000 de inversión se proyecta una producción total de 10.000.000 MM de tokens. Esto arrojaría un aproximado de USD 0,50 de costo por millón de tokens.

Finalmente se deberían estimar los costos de administración y soporte, que son aplicados a los procesos de generación y comercialización de los servicios de IA. En este sentido, se consideran anualmente los siguientes costos totales anuales:

Dirección	USD 3.000.000
Finanzas	USD 1.000.000
RRHH	USD 500.000
Legales	USD 500.000
Marketing	USD 2.000.000
TOTAL	USD 7.000.000

Cuadro II - Costo de comercialización, administración y soporte (elaboración propia)

Suponiendo una producción de 10.000.000 MM de tokens anuales, con este soporte apoyando a la infraestructura estándar, estaría arrojando un costo administrativo y de marketing de USD 0,70 por millón de tokens.

Los costos de soporte tecnológico y técnico están vinculados directamente con los recursos afectados para brindar seguridad y soporte a los clientes, en general, segmentados por usuarios domésticos y corporativos, y por desarrolladores de aplicaciones de IA. Estos servicios de soporte se realizan con los recursos que se detallan a continuación:

- Equipos DevOPS que realizan tareas de mantenimiento de servidores y servicios en funcionamiento, automatización de despliegue de nuevas versiones, gestión de infraestructura en la nube, monitoreo de disponibilidad y rendimiento, resolución de incidentes técnicos y gestión de copias de seguridad y recuperación.
- Equipos MFLOps que practican la implementación de modelos en producción, controlan las versiones, monitorean precisión y desempeño, detectan degradación del modelo, gestionan reentrenamientos, administran experimentos y supervisan el consumo de las GPU

Suponiendo costos totales anuales de estas actividades, aproximados a un tamaño de empresa con capacidad para generar 10.000.000 MM de tokens en ese período de tiempo, se obtendrían los siguientes valores promedio aproximados:

DevOps	USD	2.000.000
MLOps	USD	1.500.000
Seguridad	USD	500.000
Soporte clientes	USD	500.000
TOTAL	USD	5.000.000

Cuadro III – Costos de soporte tecnológico y técnico para el servicio de seguridad y soporte a los clientes (elaboración propia)

Aplicando este costo total a la capacidad productiva nos lleva a un costo de USD 0.5/MM de tokens.

Realizando un cuadro final de costos podríamos aproximar los siguientes siempre considerando que para llegar a una capacidad de 10.000.000 MM de tokens anuales se necesitarán :

Concepto	USD/MM
Inferencia directa	0.12
Entrenamiento amortizado	0.40
Infraestructura productiva	0.50
Administración	0.70
Soporte	0.50
TOTAL	2.22

Cuadro IV - Costo integral del proceso de inferencia por 1 MM tokens (elaboración propia)

La idea de lo expuesto en este apartado es contemplar todos los recursos y actividades necesarias para generar tokens en una empresa generadora de IA. Para la confección del costo se trabaja con una serie de supuestos genéricos y para un modelo de empresa ubicada en el límite de los segmentos de pequeña y mediana producción. Las empresas más conocidas trabajan en el tamaño denominado hiperescala con una producción de más de 100.000.000 MM de tokens anuales. El volumen en estos casos provoca una disminución del costo medio de USD 1 por MM de tokens aproximadamente, es decir USD 1.20 por MM de tokens.

La siguiente infografía permite clarificar la magnitud y relevancia de los costos y precios a los que se negocia y comercializa el servicio de inferencia de la IA sobre el lote mínimo de 1 MM de tokens.



Figura II – Recursos y ecuación económica para generar 1 millón de tokens (elaboración propia)

Las empresas que suministran el servicio de IA cobran por separado tokens de entrada (input) y los tokens de salida (output). Algunos ejemplos actuales nos llevan a calcular un precio promedio de USD 2.50 de input por MM de tokens y hasta USD 10.00 por MM de tokens de output.

Esto nos hace pensar que si las empresas están cobrando estos precios y los costos en hiperescala oscilan entre USD 1.20 y USD 1.50 tienden a tener grandes márgenes de utilidades.

Sin embargo en estos cálculos no se expresan adecuadamente los costos previos de experimentación, diseño y grandes inversiones que estas empresas a hiperescala (OpenAI, Anthropic, Google Gemini, Claude) han realizado, y que le han dado el verdadero valor económico al servicio que prestan hoy en día.

De hecho y, a modo de conclusión de este análisis, podemos afirmar que el token no se vende por su costo marginal de producción, sino por el valor económico incorporado en el modelo finalmente entrenado. Un caso muy parecido al que se da con las patentes farmacéuticas, las cuales se comercializan muy por encima de sus costos de producción.

Por otra parte, la principal ventaja competitiva de los grandes proveedores de IA no reside únicamente en la calidad del modelo, sino en las economías de escala derivadas de producir cantidades masivas de tokens sobre una infraestructura ya instalada, desarrollada y entrenada.

Una última reflexión nos lleva a conmensurar la capacidad de producción anual en millones de tokens. Esto dependerá del tamaño de la empresa. Al hablar -en la actualidad- de volúmenes observamos valores de miles de millones hasta billones de tokens por año. Existen algunas publicaciones de la empresa OpenAi que nos permite diferenciar al menos dos niveles de empresas consumidores:

- Empresa que consume IA vía API: más de 9.000 organizaciones ya han procesado más de 10.000 millones de tokens y cerca de 200 organizaciones han superado el billón de tokens.
- Empresa proveedora directa de IA: la empresa OpenAi declaró que ha procesado más de 15.000 millones de tokens por minuto. Lo que nos lleva a la cifra de 7.884.000.000 MM de tokens/año.

En función de todo lo analizado, es comprensible la estrategia de precios en estas empresas proveedoras de inferencia de tokens de IA y la proyección de un futuro en el cual, no se vislumbran cambios significativos en los precios del servicio a los usuarios domésticos, desarrolladores y empresas.

La diferencia estará en la velocidad de procesamiento y en la amplitud de la calidad en función de las condiciones y variedad tecnológica en la que se presenten los tokens de entrada para la inferencia.

Existe en la actualidad una disyuntiva planteada en torno a las condiciones de contratación de recursos de IA por parte de las empresas. Por un lado, existen modalidades contractuales con algunos LLM que brindan acceso a sus plataformas a través de un canon mensual directo y fijo (tarifa Flat), escalando en sus tarifas a medida que las versiones son mejoradas.

Otras organizaciones, prefieren desarrollar -como ya se ha visto- sus propias aplicaciones de IA, para lo cual deben tener en cuenta algunas cuestiones fundamentales:

- Más allá que Jensen Huang, desde NVIDIA declaró a principios de 2024 que el costo del cómputo es “aproximadamente cero”, sólo está considerando los costos directos de inferencia, evitando profundizar en los costos de fabricación de las unidades de procesamiento de gráficos o GPU y el costo asociado con la fase de entrenamiento, así también como el costo energético para mantener en condiciones de temperatura y humedad a las GPU en operación.
- Para la fabricación de GPU se utilizan minerales como el cobalto, el litio y otros minerales raros, existentes en cantidad finita y difíciles de obtener. Estos materiales son esenciales para la fabricación de los semiconductores y otros componentes electrónicos avanzados. Su disponibilidad limitada plantea desafíos importantes para la sostenibilidad a largo plazo de la industria tecnológica en términos de acceso a recursos naturales, de impacto ambiental y de financiamiento para poder producirlos.
- Según Infobae (2026) “la concentración de estos minerales en ciertas regiones del mundo puede generar tensiones y dependencia de proveedores específicos, lo que puede afectar la cadena de suministro global”.
- La extracción y el procesamiento de estos materiales pueden causar contaminación del agua, degradación del suelo y problemas de salud en las comunidades locales. El reciclaje

de componentes electrónicos es limitado, lo que encarece el proceso, tanto en costo de recursos como en impacto ambiental.

- Finalmente, la falta de capacidad para producir este tipo de hardware por cualquiera de las razones antes mencionadas, obliga a su adquisición. El costo de adquisición de una GPU A100 de Nvidia se estima entre 10 mil y 15 mil dólares. Teniendo en cuenta que ChatGPT-4 utilizó 25 mil de estas GPU, el costo de adquisición del hardware para entrenar al modelo superó los 250 millones de dólares.

Complementado lo planteado hasta el momento, según surge de la publicación del 26/06/2026 en en "[Blogweb.com](https://blogweb.com)", la economía de la IA tiene un problema de base que hace insostenible su ecuación económica: los costos de entrenamiento son extremadamente altos y hoy no se están trasladando a los precios de uso de la LLM (Large Language Model).

Como ejemplos, basta con saber que el entrenamiento de GPT-4 tuvo un costo de USD 79.000.000, mientras que el de Gemini Ultra llegó a los USD 191.000.000 y Llama 3.1-405B requirió USD 170.000.000. De este modo, sólo las gigantes tecnológicas están en condiciones de generar nuevos LLM, mientras que los restantes desarrolladores de asistentes de IA optan por alquilar la inferencia de tokens.

Ahora bien, además de estos números de inversión en entrenamiento de los LLM, el consumo de energía está en niveles descontrolados para una actividad económica sostenible. El entrenamiento de GPT-3 consumió unos 1.287 MWh y generó aproximadamente 552 toneladas de CO₂, según datos relevados por el Observatorio de la Universidad Tecnológica de Monterrey. Además, la misma fuente informa que se estima que desarrollos de lenguajes posteriores escalaron su generación de CO₂ a 8.930 toneladas.

Por otra parte, la demanda de energía de los centros de datos pasó de unos 49 GW en 2023 a una proyección de 96 GW para todo el año 2026, casi el doble en 3 años. Las proyecciones en el consumo de agua en el año 2027, para mantener los niveles de temperatura que requieren estos centros para no colapsar por calor, hablan de 6.600 millones de metros cúbicos.

Más allá del impacto de los costos en la estrategia de precios de los grandes proveedores del servicio de inferencia de tokens, es importante orientar ahora la mirada a la ecuación económica de costos de otros de los sujetos en el ecosistema de negocios de la IA: los usuarios de estas inferencias, tanto sea en forma doméstica directa para uso personal o a partir de desarrolladores de aplicaciones de IA, que utilizan como insumo primario la secuencia de tokens de entrada y de salida para activar la interacción necesaria para los desarrollos en empresas, para sofisticar, automatizar y robotizar sus procesos internos.

4. Los factores que influyen en los costos de producción de aplicaciones de IA dentro de las empresas

Como en todo cálculo de costos de producción, lo más importante es definir la unidad de costeo y sus variantes en términos de magnitud, escala y adopción de su uso por parte de los consumidores.

En el caso de las aplicaciones y desarrollos de IA, es factible encontrar tamaños y niveles de prestación de los desarrollos, que alteran significativamente los costos vinculados a su producción, entrenamiento, adaptación y consumo.

Según lo relevado en el blog "<https://appinventiv.com/blog/ai-development-cost/>" por el especialista Chirag Bhardwaj (2026), los costos totales de desarrollo de aplicaciones de IA oscilan entre los USD 40.000 y los USD 400.000 según la siguiente escala:

- Proyectos de IA de nivel básico (chatbots de IA, motores de recomendación, automatización basada en reglas ligeras de aprendizaje) tienen costos de desarrollo entre USD 40.000 y USD 80.000 ya que utilizan modelos pre-entrenados (denominados API o interfaz de programación de aplicaciones).
- Proyectos de IA de nivel intermedio (asistentes de IA personalizados, paneles de análisis predictivo y motores de personalización) tienen costos entre USD 80.000 y USD 200.000, al estar desarrollados sobre datos propios del negocio, integrándose con otros sistemas.
- Proyectos de IA empresariales de nivel avanzado (plataformas de IA generativas, sistemas de visión artificial, pipelines de automatización de IA) implican costos de desarrollo entre USD 200.000 y USD 400.000, al garantizar la escalabilidad exponencial en el manejo de datos complejos, requiriendo capacitación y actualización en forma continua.

En este estudio tomado como base de análisis, se presentan seis factores que influyen en las características de los recursos requeridos, según el tamaño del producto final, que se detallan a continuación:

1. Disponibilidad y calidad de los datos:

Los datos que van a ser utilizados deben pasar por un proceso de extracción desde diversas fuentes, recopilación, limpieza, estructuración y etiquetado. En general, esta etapa es la más recursos demanda, llegando a representar entre el 25% y el 40% de los costos del desarrollo. Obviamente, la accesibilidad ordenada a los datos primarios reduce drásticamente este componente de la ecuación de costos.

2. Complejidad del modelo y enfoque:

El uso de modelos de IA pre-entrenados implica un recurso de menor costo, que la aplicación de modelos perfeccionados y modelos personalizados base cero.

3. Infraestructura y capacidad de procesamiento:

La configuración adecuada del desarrollo está condicionada por la capacidad de entrenamiento con un volumen elevado de cómputos, la capacidad de almacenamiento de los conjuntos de datos procesados y el requerimiento de salidas de información procesada en tiempo real.

4. Grado de integración con los sistemas existentes:

Como todo desarrollo de una aplicación de IA requiere la interacción con otros sistemas existentes en las empresas (plataformas de CRM, paneles de control internos o aplicaciones orientadas al cliente), el nivel de complejidad de las interfaces encarece su costo y en mayor medida cuando los sistemas son más antiguos o de distintas plataformas de datos.

5. Configuración de los integrantes del equipo de desarrollo:

Los equipos de trabajo están, normalmente, integrados por especialistas de datos, ingenieros en analítica de datos y desarrollos de IA, y desarrolladores backend. Obviamente, surge siempre la posibilidad de estar conformados por recursos propios o tercerizados, impactando en la ecuación económica final.

6. Mantenimiento y actualizaciones continuas:

Una vez que el desarrollo entra en funcionamiento, requiere la programación del seguimiento de su rendimiento, la actualización de los modelos con nuevos conjuntos de datos y la adaptación al aumento de la frecuencia de uso de un mayor volumen de datos.

Según la fuente, esta instancia suele consumir entre un 15% y un 25% anual del costo del desarrollo.

La oportunidad y ocasión de uso, como así también el objetivo que la solución pretende, también son factores que condicionan la sofisticación de la aplicación y su costo de desarrollo asociado. En este caso, existen algunas diferencias que pueden ser consideradas como referencias de diferenciales de costos de desarrollo:

- Asistentes virtuales, automatización de respuestas y soporte a los clientes, implican un costo de desarrollo estimado entre USD 40.000 y USD 80.000
- Herramientas de escritura, generación de códigos y copilotos de IA, en tanto se trabajen como aplicaciones de IA generativa, generan un costo estimado entre USD 80.000 y USD 250.000
- Sistemas de recomendación en tiempo real, con sugerencias de productos y contenidos basadas en el comportamiento predictivo de los usuarios, requiere un costo de USD 60.000 hasta USD 180.000
- Soluciones de visión artificial de reconocimiento de imágenes, detección de objetos y análisis de contenidos en video, requiere procesar grandes volúmenes de datos, un intenso entrenamiento y una infraestructura de mayor nivel, impactando en un rango de costos de desarrollo entre USD 100.000 y hasta algo más de USD 300.000
- Aplicaciones predictivas del comportamiento de la demanda y análisis de riesgos, que requieren la integración de más de un sistema, oscilan entre un costo de USD 70.000 y USD 200.000

Finalmente, todo desarrollo de aplicaciones de IA está impactado por los “desempeños ocultos” que derivan de algunos de los siguientes factores, que están generando un número importante de aplicaciones a la deriva y sin un objetivo cercano:

- Es lento el proceso de preparación de los datos, a partir de inconsistencias en el formato, valores faltantes, registros duplicados y desactualizados, espacios que requieren un etiquetado manual específico.
- La deriva del modelo que genera una caída en la productividad y la calidad de la información al usuario, es normalmente consecuencia de la restricción de recursos para el seguimiento y mantenimiento de la aplicación en el mediano y largo plazo.
- La escalabilidad en el uso de las aplicaciones a partir de un mayor número de usuarios, impacta en un mayor número de solicitudes que genera un mayor costo de inferencia. Esto ocurre particularmente en los sistemas generativos que requieren la continua ampliación de la IA.
- La complejidad en la sincronización con otros sistemas dentro de la empresa es la consecuencia de la sincronización con bases de datos obsoletas, que -en muchos casos- requiere la personalización de la API contratada, encareciendo aún más el costo de mantenimiento del desarrollo.
- El manejo de datos cada vez más sensibles requiere agudizar las condiciones y requisitos de uso, en particular, respecto de la normativa sobre la privacidad de los datos, el control de accesos y los registros de auditoría, y la verificación de los sesgos del modelo.

- El monitoreo del desarrollo es una condición básica para visibilizar el rendimiento del modelo, las tasas de error y la latencia o estado del sistema. Esto demanda herramientas de infraestructura adicionales al proyecto original.
- La dependencia de una API desarrollada por terceros, conlleva a la consideración permanente del costo del cambio del proveedor y de las condiciones para la evaluación de su rendimiento.

Es posible encontrar datos respecto a desarrollos de aplicaciones de IA con un costo de desarrollo menor a los planteados, sobre los que es fundamental disponer de otros datos adicionales para comparar montos: el número de usuarios que acceden, la cantidad de inferencias promedio al mes que generan, las condiciones de ciberseguridad que cumplen y garantizan, como así también los costos de actualización de nuevas versiones que se generan a gran velocidad.

5. A pesar de la vorágine evolutiva del sector, ¿qué conceptos claves podemos rescatar?

La dinámica de este enorme rubro de la economía en la actualidad, puede llegar a presumir de la capacidad de bloquear o diluir la observación de sus números, para evaluar si la ecuación económica detrás de esta realidad es viable y sostenible en el tiempo. Pareciera ser que estamos llegando al límite de avanzar en los desarrollos sin reparar en los números.

De todos modos, es posible rescatar algunos conceptos claves que serán aplicables durante este proceso dinámico y evolutivo exponencialmente:

- El concepto de un lote estándar de 1 millón de tokens es la unidad de costeo, de análisis y de facturación.
- El ciclo de inferencia y generación de tokens no es el mismo en cualquier dimensión operativa de la empresa en la que se opera.
- Es clave definir el ciclo de uso de las aplicaciones de IA y los desarrollos de LLM, para lograr un cálculo razonable de amortización de la inversión en entrenamiento de la IA, en infraestructura energética y en operaciones de los grandes centros de procesamiento de las GPU.
- Existen grandes diferencias tarifarias en las contrataciones de servicios entre desarrolladores de tamaño PyME y empresas denominadas “de hiperescala”.
- Hoy se registran relaciones sin lógica alguna entre los precios de mercado de las aplicaciones de IA y los costos de desarrollo y mantenimiento de la inferencia de 1 millón de tokens.
- Los modelos de costeo no están incorporando el valor del conocimiento, la escala y la propiedad intelectual incorporada en los modelos. Pareciera que son recursos tan sustituibles y disponibles en cantidad, que no tienen un valor económico relevante.

En función de todo lo planeado, cabe una observación final.

De la lectura de esta ponencia, es posible que los lectores entiendan que se trata de un planteo coyuntural y que está expuesto a una próxima desactualización en los valores. Esto, seguramente, ocurrirá en los próximos meses y años, habida cuenta de la detección de una falta de lógica en la estrategia de precios de los grandes proveedores del servicio de inferencias de tokens, como así

también en la definición de escalas tarifarias asociadas a un volumen en crecimiento exponencial, tanto por el volumen de participantes como por el volumen de cálculos requeridos por ellos.

De todos modos, la vigencia de este trabajo no se basa en las cifras presentadas, sino en la metodología de cálculo, en base a determinados factores y técnicas que mantendrán su utilidad práctica a pesar del paso del tiempo.

La disciplina nos ofrece herramientas de toda índole para lograr sostener técnicamente una mirada crítica, en este caso, sobre unos pocos sujetos de decisión dentro del ecosistema de producción y consumo de aplicaciones de IA. El desafío está planteado. Es posible extender este estudio al resto de los participantes de la cadena de valor extendida. Quizás encontremos una lógica detrás de la dinámica de valor del ecosistema que, en la actualidad, no es posible estructurar y comprender.

6. Conclusiones:

La inteligencia artificial se ha incorporado a la economía y a la vida de las empresas y personas con una velocidad de adopción inédita en la historia de nuestra civilización. Sin embargo, esta expansión exponencial se ha producido sobre una contradicción aparente: mientras que la percepción generalizada sostiene que el costo del cómputo tiende a cero, la estructura económica que permite generar, entrenar y operar modelos de inteligencia artificial demanda inversiones crecientes en infraestructura, energía, talento especializado y recursos naturales estratégicos.

Desde la perspectiva de los costos y la gestión, esta situación plantea la necesidad de desarrollar nuevas formas de análisis y prácticas de medición que permitan comprender la verdadera ecuación de costos de la IA. En este trabajo se propuso considerar al millón de tokens como una unidad normalizada de costeo para la producción de inferencias, permitiendo asociar a dicha unidad los recursos directos de procesamiento, los costos de entrenamiento, la infraestructura tecnológica, las actividades de soporte y los costos de administración.

La aplicación de esta lógica permitió observar que el costo económico de producción de un token excede ampliamente el costo marginal de inferencia. En consecuencia, el precio de mercado de los servicios de IA no puede explicarse únicamente por el consumo de capacidad computacional, sino principalmente por la incorporación de conocimiento, propiedad intelectual, economías de escala y capacidad de entrenamiento acumulada en los modelos. En este sentido, el token se asemeja más a un bien intensivo en conocimiento que a un producto digital de reproducción ilimitada.

Asimismo, el trabajo permitió distinguir dos universos claramente diferenciados de decisión. Por un lado, las organizaciones que producen inferencias y comercializan capacidad de procesamiento; por otro, las empresas que utilizan la IA como insumo para desarrollar aplicaciones, automatizar procesos o asistir a la toma de decisiones. Cada uno de estos actores enfrenta estructuras de costos, riesgo y criterios de evaluación diferentes, por lo que requieren modelos específicos de análisis económico y de gestión.

Otro aspecto relevante tratado es que la ecuación de costos de la inteligencia artificial no se limita a recursos visibles. Los costos ocultos vinculados con la calidad de datos, el mantenimiento de modelos, la integración tecnológica, la dependencia de proveedores, la seguridad de la información, el cumplimiento normativo y la actualización permanente constituyen factores

determinantes para evaluar la sostenibilidad económica de los proyectos de IA en las organizaciones.

Del mismo modo, el análisis efectuado permite advertir que las ventajas competitivas de los grandes proveedores globales de IA no se sustentan exclusivamente en la calidad técnica de sus modelos, sino también en la posibilidad de operar a escalas masivas que reducen significativamente el costo medio de producción de tokens. Esta característica aproxima a la industria de la inteligencia artificial a otros sectores con elevados costos fijos y bajos costos marginales, donde las economías de escala se convierten en un factor decisivo de competitividad.

Finalmente, entendemos que la disciplina se encuentra frente a un nuevo objeto de estudio. Así como en otros momentos fue necesario desarrollar herramientas para comprender los costos de automatización, la digitalización o la economía del conocimiento, la expansión de la IA exigen construir marcos conceptuales capaces de medir, interpretar y gestionar sus recursos, capacidades y desempeños.

Como línea futura de investigación, se propone profundizar el análisis de la relación existente entre el costo de producción de tokens y el costo total de desarrollo de aplicaciones de IA, explorando la posible correlación entre ambas variables y su impacto en la formulación de estrategias empresariales, modelos de negocio y sistemas de gestión orientados a la creación de valor sustentable.

Bibliografía

- Anthropic. (2025). *Claude API documentation and pricing*. <https://docs.anthropic.com>
- Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton & Company.
- Campanario, S. (2026) *Tecnólogos versus economistas: la pelea por la velocidad de la IA y el eslabón más débil de la cadena*. www.lanacion.com.ar/economia.
- Chen, L.; Zhang, Ch.; y otros. *El fenómeno de la inversión de precios: cuando los modelos de razonamiento más baratos cuestan más*. arXIV. www.arxiv.org/abs/
- Cottler, B.; Snodin, B.; Owen, D. y Adamczewski, T. (2025) *Los precios de la inferencia LLM han caído rápidamente, pero de forma desigual, entre las distintas tareas*. Epoch AI. epoch.ai/data-insights/
- Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review*, 96(1), 108–116.
- Google. (2025). *Gemini API documentation*. <https://ai.google.dev>
- Lacort, J. (2026) *La “economía de los tokens” se ha roto: las tarifas planas de IA para programar son matemáticamente imposibles*. www.kataka.com/robotica

- Maslej, N. (2025). *Indice de IA 2025: El estado de la IA en 10 gráficos*. AI Stanford University. Human-Centered Artificial Intelligence.
- Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Portfolio.
- NVIDIA Corporation. (2025). *NVIDIA H100 Tensor Core GPU*. <https://www.nvidia.com/en-us/data-center/h100/>
- NVIDIA Corporation. (2025). *AI inference performance*. <https://developer.nvidia.com/deep-learning-performance-training-inference/ai-inference>
- OpenAI. (2025). *API pricing*. <https://openai.com/api/pricing>
- OpenAI. (2025). *OpenAI developer documentation*. <https://platform.openai.com/docs>
- Parra-Acosta, J. y Peña González, Y. (2014) *La teoría de los costos-desempeños ocultos: una aproximación teórica*. Cuadernos de Contabilidad, vol. 15, número 39, Pontificia Universidad Javeriana. (Bogotá, Colombia).
- Podmoguilyn, M., Marchione, J., Farré, D., Pott Godoy, J. y Metilli, G. (2023). *Costos y Gestión 5.0*. Thompson Reuters La Ley, Buenos Aires, Argentina.
- Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
- Savall, H. y Zardet, Véronique (1989) *Maitriser les couts et les performances cachés. Le contrat d'activité périodiquement négociable*. 2da edición (París, Éditions Économica).
- Ultima Milla (2026) *Tu propio ChatGPT, en tu servidor, en pesos*. www.ultimamilal.com.ar/blog/
- World Economic Forum. (2025). *The future of jobs report 2025*. <https://www.weforum.org>
- Wilhelm, P.; Wittkopp, T. y Kao, O. (2025) *Beyond Test-Time Compute Strategies: Advocating Energy-per-Token in LLM Inference*. www.euromlsys.eu/euromlsys25-27.pdf/
- Zhu, Q. (2026) *Tokenomics: The Economics of Tokens, Computation and Pricing in Foundation Models*. arXIV. www.arxiv.org