

**.UBA**derecho



**IALAB**

# ¿Cuál es el costo de trabajar con inteligencia artificial generativa?

---

Juan G. Corvalán y Mariana Sanchez Caparros



**Juan Gustavo Corvalán y Mariana Sanchez Caparrós**

## **¿Cuál es el costo de trabajar con los grandes modelos de IAGen?**

<b>I. Introducción.....</b>	<b>2</b>
<b>II. ¿Cuánto cuesta utilizar un modelo de lenguaje grande?.....</b>	<b>3</b>
II.1. La importancia de los tokens y las ventanas de contexto.....	3
II.1.1. La variable: token.....	3
II.1.2. La variable: ventana de contexto.....	6
II.2. Servicios de OpenAI (ChatGPT).....	6
II.2.1. Servicios para usuarios individuales.....	7
II.2.1.1. ChatGPT. Gratuidad y una ventana de contexto limitada.....	7
II.2.1.2. GPT-4. El doble de ventana de contexto pero mensajes limitados.....	7
II.2.2. Servicios para clientes comerciales.....	7
II.2.2.1. GPT-4 turbo.....	8
II.2.2.2. GPT-4.....	8
II.2.2.3. GPT-3.5 turbo.....	8
II.2.3. ¿Cuánto cuesta hacer un ajuste fino? Pago por entrenamiento + pago por uso del modelo personalizado.....	9
II.3. Gemini (ex-Bard) de Google.....	9
II.4. LLaMa2 de Meta.....	10
II.4.1. Precio bajo demanda y por lotes (sin ajuste fino):.....	10
II.4.1.1. LLama2 Chat 13 billones de parámetros:.....	10
II.4.1.2. LLama2 Chat 70 billones de parámetros:.....	11
II.4.2. Precio por la personalización del modelo (ajuste fino):.....	11
II.4.3. Precio del rendimiento aprovisionado:.....	11
<b>III. Conclusiones .....</b>	<b>12</b>

## **I. Introducción**

El empleo de IA generativa (IAG) de texto para asistir, complementar y sustituir a las personas ha aumentado exponencialmente. El aumento en la productividad y la optimización de tiempos en diversas tareas arroja resultados sorprendentes. Además, los grandes modelos de lenguaje generativos (LLM) se están integrando en diversas aplicaciones y desarrollos específicos, lo que ha demostrado ser un catalizador para la innovación y la eficientización del trabajo en distintos sectores.

Sin embargo, cuando se decide incorporar IAG en los procesos de trabajo, es imperativo tener presentes los costos asociados al uso de los distintos servicios disponibles en el mercado.

Entre ellos, el de acceso a plataformas y modelos especializados, pero también otros como el consumo de recursos computacionales, el almacenamiento de datos y la posible necesidad de personal especializado para optimizar y gestionar eficientemente la implementación.

Considerar los precios de los servicios de IA generativa de texto es crucial para garantizar la viabilidad económica y sostenibilidad a largo plazo de proyectos que se apoyen en esta tecnología. También para anticiparse a posibles incrementos en su utilización, para poder tomar decisiones informadas y evitar sorpresas inesperadas desde lo económico y financiero.

En resumen, cuando se decide incorporar IA generativa para usarla de manera directa o integral en un proyecto, es conveniente considerar los siguientes aspectos:

- a. Analizar los costos específicos de diferentes plataformas y modelos, como por ejemplo, Gemini, LLaMa2, ChatGPT y GPT-4, entre otros.
- b. Evaluar si se consume de manera individual como usuario final o como empresa o institución, o si se realiza con fines de investigación.
- c. Desglosar los costos por tipo de servicio: acceso a la plataforma, alojamiento, consumo de tokens, costo de fine tuning.
- d. Comparar los precios entre diferentes proveedores, con los distintos planes que ofrecen y sus características.
- e. Analizar el costo-beneficio de elegir la IA generativa de texto para solucionar el problema elegido o lograr el objetivo buscado, comparándola con otro tipo de soluciones que también permiten lograrlo.
- f. Analizar los costos asociados a la escalabilidad del proyecto, esto es, cómo los costos de la herramienta de IA generativa de texto elegida pueden aumentar a medida que se escala su uso, incluyendo el consumo de recursos computacionales, almacenamiento de datos y la necesidad de personal especializado.

## II. ¿Cuánto cuesta utilizar un modelo de lenguaje grande?

Para hacer más gráfico lo que expresamos en la introducción, a continuación analizaremos cuánto cuesta utilizar algunos de los principales modelos de lenguaje grandes que hay disponibles en el mercado.

### II.1. La importancia de los tokens y las ventanas de contexto.

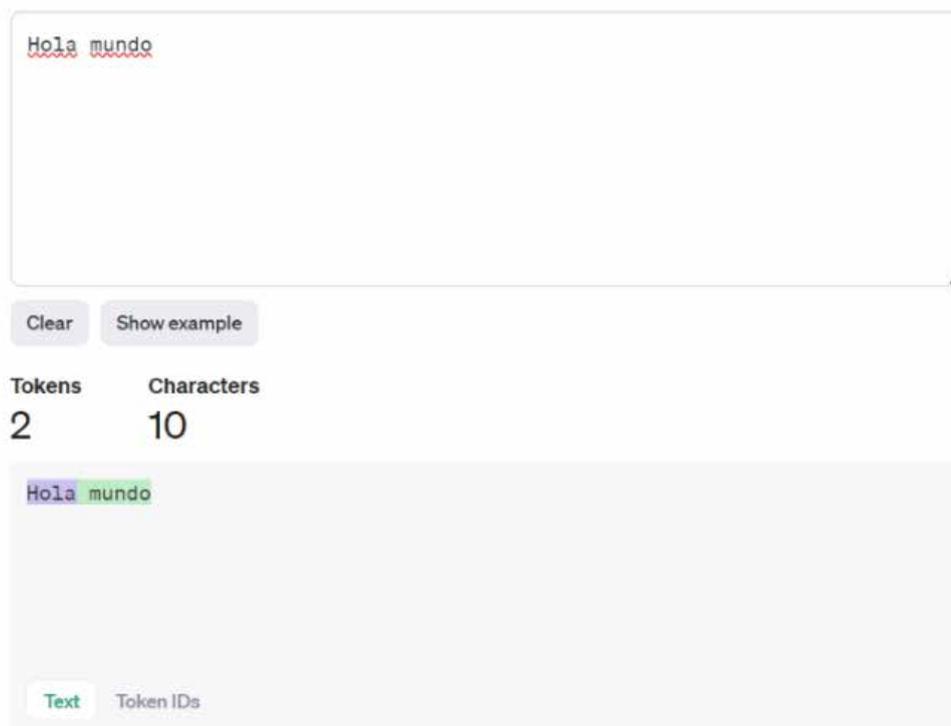
#### II.1.1. La variable: token

Una variable muy importante al estimar los costos de utilización de los modelos de OpenAI es el **token**.

Los modelos de lenguaje grandes procesan texto utilizando tokens, que es la unidad más pequeña de procesamiento en el modelo. Los tokens son secuencias comunes de caracteres que se encuentran en un conjunto de texto.

OpenAI explica que una regla de utilidad es pensar en el token como una cadena de aproximadamente 4 caracteres de texto común en inglés<sup>1</sup>. No obstante, un token puede ser una palabra individual, una parte de una palabra o, incluso, un carácter. Cada token se convierte luego en un vector numérico único (TokenID) que el modelo puede procesar para representar información.

A continuación, te dejamos algunos ejemplos de tokens (texto) y tokenIDs (vector numérico único) para una misma oración:



The screenshot shows a web interface for tokenization. At the top, there is a text input field containing "Hola mundo". Below the input field are two buttons: "Clear" and "Show example". Below the buttons, there is a table with two columns: "Tokens" and "Characters". The table shows that "Hola mundo" is split into 2 tokens and has a total of 10 characters. Below the table, there is a preview of the text "Hola mundo" with a green highlight under "Hola" and a purple highlight under "mundo". At the bottom, there are two tabs: "Text" (selected) and "Token IDs".

Tokens	Characters
2	10

<sup>1</sup>En una hoja A4, tipo de letra Times New Roman, 12 ptos., Interlineado 1 y espacio entre párrafos.

GPT-3.5 & GPT-4 GPT-3 (Legacy)

Hola mundo

Clear

Show example

Tokens

2

Characters

10

[69112, 29452]

Text

Token IDs

GPT-3.5 & GPT-4 GPT-3 (Legacy)

Hola mundo, ¡buen día!

Clear

Show example

Tokens

8

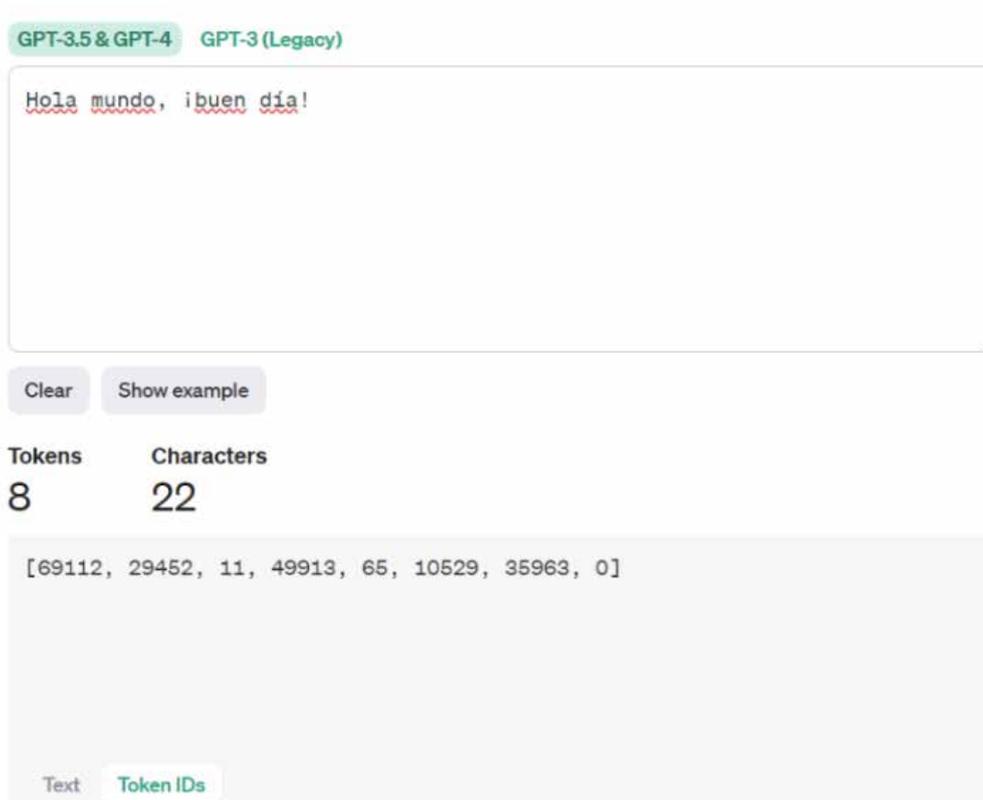
Characters

21

Hola mundo, ¡buen día!

Text

Token IDs



La tokenización es clave para que los grandes modelos de lenguaje puedan procesar el texto de manera eficiente, ya que logran capturar las relaciones estadísticas presentes en ellos para generar contenido a partir de predecir el siguiente token dada la secuencia de tokens anterior<sup>2</sup>.

En la web de OpenAI se explica que **100 tokens equivalen, aproximadamente, a unas 75 palabras en inglés.**

También podríamos exponerlo de la siguiente manera: **1000 tokens = 750 palabras = un texto de 1 página y media<sup>3</sup>.**

Existe una herramienta para el cálculo de tokens disponible en la plataforma: <https://platform.openai.com/tokenizer>. Allí podés cortar y pegar un texto para calcular la cantidad de tokens que lo conforman y comprender cómo funciona este concepto.

El costo de uso de grandes modelos de lenguaje se mide, fundamentalmente, en tokens de input y tokens de output. Veamos algunos ejemplos que van a ser de utilidad para representar esta información:

- a. Cuando se menciona un “costo x por cada 1000 tokens de input”, a lo que se hace referencia es al costo de procesar el contenido —medido en tokens— de la entrada de texto propuesta por el usuario al modelo a través del prompt.

---

<sup>2</sup>Para profundizar sobre el funcionamiento de los grandes modelos de lenguaje como ChatGPT y GPT4 ver Corvalán Juan, Elsa Estevez, Enzo Le Fevre, Debora Schapira y Gerardo Simari (Directores), ChatGPT vs. GPT4. ¿Imperfecto por diseño?, UBA-IALAB, 2023, en <https://ialab.com.ar/libros/>.

<sup>3</sup> En una hoja A4, tipo de letra Times New Roman, 12 ptos., Interlineado 1 y espacio entre párrafos.

Consideremos la siguiente instrucción: “traducime al inglés un texto que te voy a proporcionar”. Luego se copia y pega el texto junto a esa instrucción en Gpt-4 y ello tendrá una determinada cantidad de tokens. Por ejemplo, 8.000 tokens que son 6 páginas de texto. Estos determinarán el precio a pagar por procesar la instrucción que se le dio al modelo.

Es importante tener presente que este costo se computa cada vez que se pide algo al modelo. Por tanto, si una vez obtenida una respuesta, se le piden agregados, aclaraciones o mejoras, este texto se computará como un nuevo input y se abonará por su procesamiento.

- b. Cuando hablamos de costo “x por cada 1000 tokens output”, se hace referencia al costo de la respuesta generada por el modelo elegido medido en tokens. Mientras más larga la respuesta, es decir, a mayor cantidad de tokens, más se abonará por ella.

En el caso del ejemplo, se pagará por las 6 páginas de texto traducidas que representan unos 8000 tokens aproximadamente. Además, se abonarán los tokens que ocupen la oración con la que el modelo nos responde: por ejemplo, cuando nos contesta “Claro, aquí tienes el texto traducido al inglés”.

### **II.1.2. La variable: ventana de contexto**

Otra variable relevante a considerar cuando se elige un modelo de lenguaje grande es la ventana de contexto con la que trabaja.

La **ventana de contexto** se refiere al rango de tokens que el modelo tiene en cuenta al generar una respuesta para una determinada solicitud del usuario.

Este factor es importante porque si la entrada o input supera el límite de tokens de la ventana de contexto, lo que puede suceder es que, al generar la respuesta, el modelo no tenga en cuenta parte de la información que se le ha provisto.

Generalmente los modelos que trabajan con ventanas de contexto más grandes, tienen un mayor costo de utilización.

### **II.2. Servicios de OpenAI (ChatGPT)**

Conforme surge de la web de OpenAI, la firma ofrece múltiples modelos, cada uno de ellos con diferentes capacidades y precios. La propuesta general de la empresa para clientes comerciales, es que sólo se paguen los recursos que se utilizan con el modelo elegido. Los clientes individuales en cambio, pueden usar ciertos modelos (GPT-3.5/GPT-4) gratuitamente o mediante el pago de una tarifa fija, como se verá seguidamente.

No es sencillo encontrar la información porque se encuentra dispersa en distintas URLs y, además, cambia con bastante frecuencia. No obstante, del relevamiento realizado en febrero de 2024 surge lo siguiente:

## II.2.1. Servicios para usuarios individuales

### II.2.1.1. ChatGPT. Gratuidad y una ventana de contexto limitada

OpenAI ofrece ChatGPT de forma gratuita, con acceso al modelo GPT-3.5 a través de la web, Android e iOS, con mensajes de texto, interacciones e historial ilimitados.

Este modelo trabaja con una ventana de contexto de 4096 tokens<sup>4</sup>. Esto es, de aproximadamente unas 6 páginas (o unas 3000 palabras). A modo de ejemplo, si se hace un copy paste de un texto de 20 páginas, es probable que el sistema no procese todo ese texto de entrada porque se excede largamente de la ventana de contexto con la que trabaja (las 6 páginas), con lo cual la respuesta no será del todo precisa y puede que se omita información relevante para construirla.

Pero además, la gratuidad del uso tiene un límite de 1600 palabras de input por hora<sup>5</sup>. Por lo tanto, el límite de consumo es de 6400 tokens por hora.

### II.2.1.2. GPT-4. El doble de ventana de contexto pero mensajes limitados

Por una tarifa fija de 20 USD mensuales, se puede acceder GPT-4, el modelo más avanzado de OpenAI.

Se ofrece a través de la web, Android e iOS, y permite crear, buscar y usar GPT's; interactuar con imágenes; comunicarse con la herramienta mediante el uso de la voz, y acceder a herramientas adicionales como DALLE.

En este caso, el modelo tiene una **ventana de contexto de 8192 tokens**<sup>6</sup>. Esto es, de aproximadamente **unas 12 páginas** (unas 6000 palabras). Duplica así la del ChatGPT (GPT-3.5).

En este caso, como ejemplo, si se hace un copy paste de un texto de 30 páginas, es probable que el sistema no procese todo ese texto de entrada porque se excede largamente de la ventana de contexto con la que trabaja (las 12 páginas). En tal caso, la respuesta no será del todo precisa y puede que se omita información relevante para construirla.

Si bien en una parte de la web se señala que en esta suscripción encontrarás todo lo que está disponible en el servicio gratuito en ChatGPT<sup>7</sup>, lo cierto es que a partir del 5 de enero de 2024 GPT-4 tiene un límite de 40 mensajes cada 3 horas, respecto de ello se debe tener en cuenta que los mensajes no utilizados no se acumulan, es decir, si se espera 6 horas, no se obtendrán 80 mensajes disponibles para usar durante las próximas 3 horas. Incluso, en ciertos casos, no se indica cuándo OpenAI puede ajustar dinámicamente el límite de mensajes para priorizar que el modelo sea accesible para una mayor cantidad de personas<sup>8</sup>.

---

<sup>4</sup>Desde UBA IALAB, pusimos a prueba el modelo ChatGPT 3.5 hasta llegar a su capacidad máxima, registrando un total de 75 mensajes de input por hora. Así, en un período de 12 horas, se acumularían un total de 900 mensajes de input.

<sup>5</sup>Ver <https://platform.openai.com/docs/models/gpt-3-5-turbo>

<sup>6</sup>Ver <https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo>

<sup>7</sup><https://openai.com/chatgpt/pricing>

<sup>8</sup>Ver en <https://help.openai.com/en/articles/7102672-how-can-i-access-gpt-4>.

## II.2.2. Servicios para clientes comerciales

Para clientes comerciales los precios del uso de los distintos modelos disponibles se calculan cada 1000 tokens.

### II.2.2.1. GPT-4 turbo

Supongamos que elijo GPT-4 turbo, un modelo con una ventana de contexto de 128.000 tokens<sup>9</sup>, lo que equivale a unas 190 páginas o 95.000 palabras. Según señala OpenAI, esta es una herramienta más potente que GPT-4 que se ofrece a un menor precio.

Tiene un costo USD 0.01 por cada 1000 tokens de input y USD 0.03 por cada 1000 tokens output. Este precio aplica para los modelos: GPT-4-0125-preview, GPT-4-1106-preview y GPT-4-1106-vision-preview<sup>10</sup>.

Si eligiéramos este modelo:

Largo del texto (aproximado)	Cantidad de tokens (aproximado)	Costo input (contenido del prompt)	Costo output (respuesta)
6 páginas (3.000 palabras)	4000	USD 0.04	USD 0.12
20 páginas	8000	USD 0.08	USD 0.24

### II.2.2.2. GPT-4

GPT-4 tiene un costo un poco más elevado, de USD 0.03 por cada 1000 tokens para el input y USD 0.06 por cada 1000 tokens para el output.

Si eligiéramos este modelo:

Largo del texto (aproximado)	Cantidad de tokens (aproximado)	Costo input (contenido del prompt)	Costo output (respuesta)
6 páginas (3.000 palabras)	4000	USD 0.12	USD 0.24
12 páginas	8000	USD 0.24	USD 0.48

### II.2.2.3. GPT-3.5 turbo

En el caso de elegir GPT-3.5, esta versión cuenta con una ventana de contexto de 16.385 tokens, lo que equivale a 12.000 palabras o 24 páginas. La versión GPT-3.5 turbo-0125 se ofrece a un precio de USD 0.0005 por cada 1000 tokens de input y USD 0.0015 por cada 1000 tokens de output.

<sup>9</sup>Ver <https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo>

<sup>10</sup>Ver <https://openai.com/pricing>

En este caso:

Largo del texto (aproximado)	Cantidad de tokens (aproximado)	Costo input (contenido del prompt)	Costo output (respuesta)
6 páginas (3.000 palabras)	4000	USD 0.002	USD 0.006
12 páginas	8000	USD 0.004	USD 0.012

### II.2.3. ¿Cuánto cuesta hacer un ajuste fino? Pago por entrenamiento + pago por uso del modelo personalizado.

OpenAI permite que los clientes creen sus propios modelos personalizados mediante un ajuste fino de los modelos de base con sus propios datos de entrenamiento. Una vez que el cliente tenga entrenado el modelo, sólo pagará por los tokens que se usen en las solicitudes a dicho modelo<sup>11</sup>.

#### Costo de entrenamiento:

El ajuste fino del modelo GPT-3.5 turbo tiene un costo de entrenamiento de USD 0.0080 por cada 1000 tokens. Esto es por cada aproximadamente 1 página y ½ de texto o 750 palabras que se incorporan para el entrenamiento<sup>12</sup>.

Veamos algunos ejemplos:

- Hacer un ajuste fino con una base de 30 sentencias de unas 5 páginas cada una, equivale a 150 páginas o 100.000 tokens. El costo asciende aproximadamente a USD 0.80.
- Hacer el ajuste fino con una base de documentos de 300 contratos de unas 5 páginas cada uno equivale a 1.500 páginas o 1.000.000 tokens. El costo asciende aproximadamente a USD 8.
- Hacerlo con una base de 3.000 documentos de 5 páginas cada uno, equivale a 15.000 páginas o 2.000.000 tokens. Te costaría unos USD 16.

#### Costo de uso del modelo personalizado:

Concluido el ajuste fino, por el **uso del modelo** se pagarán USD 0.0030 por cada 1000 tokens de entrada y USD 0.0060 por cada 1000 tokens de salida.

Si elegimos esta opción y utilizamos el modelo con el ajuste fino:

Largo del texto (aproximado)	Cantidad de tokens (aproximado)	Costo input (contenido del prompt)	Costo output (respuesta)
6 páginas (3.000 palabras)	4000	USD 0.012	USD 0.24
12 páginas	8000	USD 0.024	USD 0.48

<sup>11</sup>Ampliar en: <https://platform.openai.com/docs/guides/fine-tuning>

<sup>12</sup>Ampliar en: <https://openai.com/pricing>



Largo del texto (aproximado)	Cantidad de tokens (aproximado)	Costo input (contenido del prompt)	Costo output (respuesta)
6 páginas (3.000 palabras)	4000	USD 0.0076	USD 0.010
12 páginas	8000	USD 0.015	USD 0.02

#### II.4.1.2. LLama2 Chat 70 billones de parámetros:

Para el modelo LLama2 Chat de 70 billones de parámetros, se abonan USD 0.0019 por 1000 tokens de input y USD 0.0025 por 1000 tokens de output, por lo que si eligiéramos este modelo:

Largo del texto (aproximado)	Cantidad de tokens (aproximado)	Costo input (contenido del prompt)	Costo output (respuesta)
6 páginas (3.000 palabras)	4000	USD 0.003	USD 0.004
12 páginas	8000	USD 0.006	USD 0.008

#### II.4.2. Precio por la personalización del modelo (ajuste fino):

En caso de apuntar a un modelo con ajuste fino, se debe sumar el precio de almacenamiento personalizado, fijado en USD 1.95, más el de personalización del modelo:

Modelos de Meta	Precio por entrenar 1000 tokens	Precio mensual por almacenamiento personalizado	Costo cómputo para entrenamiento (Precio a deducir de un modelo personalizado por 1 unidad de modelo por hora con precios de rendimiento aprovisionado sin compromiso)
Llama 2 Pre-trained (13B)	USD 0.00149	USD 1.95	USD 23.50
Llama 2 Pre-trained (70B)	USD 0.00799	USD 1.95	USD 23.50

#### II.4.3. Precio del rendimiento aprovisionado:

Modelos de Meta	Precio por hora por unidad de modelo con compromiso de 1 mes	Precio por hora por unidad de modelo con compromiso de 6 mes
Llama 2 Pre-trained (13B)	USD 21.18	USD 13.08
Llama 2 Pre-trained (70B)	USD 21.18	USD 13.08

Según surge de la web de AWS, los modelos de Llama 2 Pre-trained solo se encuentran disponibles con un rendimiento aprovisionado después de la personalización. Es decir, este precio es sólo para modelos con ajuste fino.

De acuerdo a las averiguaciones realizadas desde el IALAB, este último esquema de precios es sin límite de tokens, pero sólo es conveniente cuando se consumen más de 21.000 tokens/hora, lo que serían aproximadamente unas 31 páginas de texto o 15.000 palabras.

### III. Conclusiones.

Este documento es una guía para estimar los principales costos asociados a la incorporación de herramientas de IA generativa de texto, tanto a nivel de organización como para uso personal.

Para modelos específicos, como los ofrecidos en servicios comerciales de OpenAI y LLaMa2, el costo está influenciado por varios factores como:

el número de tokens del texto de entrada, que abarca tanto la instrucción como el texto destinado al análisis; los tokens del texto de salida, es decir, aquellos que conforman la respuesta generada por el modelo. En este contexto, la longitud de la instrucción y de la respuesta son elementos que impactarán de manera directa en el costo al utilizar modelos de lenguaje grandes.

Adicionalmente, se deben considerar otros factores como la ventana de contexto, el ajuste fino y el costo de alojamiento del modelo. Aunque este último no suele ser de impacto significativo en el precio final.

La ventana de contexto, en cambio, juega un papel crucial al definir la "memoria" del modelo para responder con mayor precisión a las solicitudes del usuario. Por ejemplo, modelos con ventanas de contexto más amplias tienden a ser más costosos que aquellos con ventanas más pequeñas.

El ajuste fino, por su parte, se agrega como costo cuando el modelo base necesita entrenamiento específico con datos particulares para mejorar su rendimiento en tareas específicas. Es decir, esto implica un costo adicional que varía según el proveedor de servicios elegido.

En síntesis, a lo largo del documento se ofrece una comparación detallada de precios y características de algunas de las herramientas más conocidas del mercado y proporcionan ejemplos de cálculos de costos para diversos escenarios de uso con el fin de facilitar la toma de decisiones. El uso de IA generativa de texto implica considerar diversos factores, como el costo, la calidad, la personalización y la escalabilidad de los modelos disponibles. Existen diferentes opciones en el mercado, como ChatGPT, GPT-4, Gemini y LLaMa2, cada una con sus ventajas y desventajas en cuanto a costos asociados.

Ahora bien, los servicios gratuitos o de bajo costo presentan limitaciones en la ventana de contexto y en el número de respuestas a procesar. También se puede

sumar una capa de entrenamiento específico o de integrarlas a otros sistemas o aplicaciones. Los servicios y versiones de IA generativa que ofrecen servicios más avanzados y personalizados, presentan mayores costos y en ciertos casos más complejidad.

En todos los casos, es importante analizar las necesidades y objetivos de cada proyecto, así como el costo-beneficio de cada herramienta, para elegir la más adecuada y sostenible. En esencia, la definición del tipo de servicio a utilizar depende de los casos de uso que se definan y seleccionen dentro de las organizaciones para aplicar IA Generativa; de la optimización que el uso de la IA Generativa puede generar calculada tanto por tarea como por semana, mes y año; así como de las diferencias en los resultados obtenidos a partir de los servicios gratuitos y los servicios pagos. Por ese motivo, la selección del servicio debe responder directamente a las necesidades que se presenten en el interior de las organizaciones.

La IA generativa de texto es una tecnología disruptiva que “cambia el paradigma de trabajo” con un gran potencial, pero con costos asociados que deben tenerse en cuenta. La decisión del tipo de servicio, debe tomarse como la elección de un traje a medida de las necesidades que se presenten.

# Equipo de trabajo

## Investigación

Juan G. Corvalán

Mariana Sanchez Caparros

## Diseño Gráfico

Victoria Mafud

Sofía Rolleri

## Infografía

[Accedé acá](#)

.UBAderecho



IALAB