.UBAderecho IALAB

¿Cuál es el costo de trabajar con herramientas de IA generativa de texto?

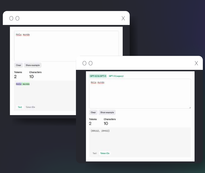
Juan G. **Corvalán** y Mariana **Sanchez Caparros**

El documento analiza los principales factores que influyen en el costo de utilizar modelos de lenguaje grandes como **ChatGPT**, **GPT-4**, **Gemini** y **LLaMa2**, tanto para uso personal como comercial.

Variables a considerar

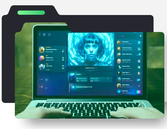
Tokens de **input** y **output** y ventana de contexto son dos de los factores más importantes a considerar al evaluar el costo de uso de grandes modelos de lenguaje.

El **token** es la unidad más pequeña de procesamiento en el modelo. Puede ser una palabra individual, una parte de una palabra o, incluso, un caracter. Cada token se convierte luego en un vector numérico único (TokenID) que el modelo puede procesar para representar información.



El costo de uso de grandes modelos de lenguaje se mide, fundamentalmente, en tokens de input y tokens de output.
1000 tokens = 750 palabras o un texto de 1 página y media.

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.



Los modelos que trabajan con ventanas de contexto más grandes, tienen un mayor costo de uso.

Chat GPT

- Modelo GPT-3.5. Unimodal.
- Trabaja con una ventana de contexto de 4096 tokens (aproximadamente unas 6 páginas).
- Gratuito.
- Límite de 75 mensajes por hora

Chat GPT4.0

- Multimodal.
- Trabaja con una ventana de contexto de 8192 tokens (aproximadamente unas 12 páginas).
- 20 USD por mes.
- Límite de 40 mensajes cada 3 horas.

Usuarios comerciales ejemplos

- Modelo GPT-4 turbo. Multimodal.
- Trabaja con una ventana de contexto de 128.000 tokens (aproximadamente unas 192 páginas).
- Costo: USD 0.01 por cada 1000 tokens de input y USD 0.03 por cada 1000 tokens output.
- Cada 20 páginas de input: **USD 0.08**
  Cada 20 páginas de output: **USD 0.24**

Ajuste fino y uso del modelo personalizado:

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.**

Costo de uso:
- **USD 0.024** cada 12 páginas de entrada
- **USD 0.048** cada 12 páginas de salida.

Servicios de Google y Meta

Gemini, la herramienta de Google, ofrece acceso gratuito a su modelo de IA generativa de texto.

LLaMa2, la herramienta de Meta, ofrece modelos de código abierto que se pueden operar en la nube.

En esos casos se paga costo de alojamiento, de ajuste fino (si hubiera) y de uso medido por consumo de tokens de entrada y salida.

Equipo de trabajo

Investigación
Juan G. **Corvalán**
Mariana **Sanchez Caparros**

Diseño Gráfico
Victoria **Mafud**

Resumen ejecutivo

Accedé acá

.UBAderecho IALAB