Modelos disponibles y coste

En este capítulo, exploraremos los diferentes modelos disponibles en la API de OpenAI, sus funcionalidades y los costes asociados a su uso. Es fundamental comprender las características y precios de cada modelo para elegir el más adecuado para nuestras necesidades.

Modelos disponibles

- **GPT-4O**
  - **Descripción:** Nuestro modelo insignia de alta inteligencia, diseñado para tareas complejas que requieren múltiples pasos. Es capaz de procesar texto e imágenes y generar salidas de texto. Se destaca por la alta calidad de sus respuestas y su capacidad para seguir instrucciones complejas.
  - **Ventajas:** Alta capacidad de razonamiento, generación de texto de alta calidad, procesamiento de texto e imágenes, gran tamaño de contexto.
  - **Desventajas:** Coste elevado.
- **GPT-4O mini**
  - **Descripción:** Nuestro modelo pequeño y asequible, ideal para tareas rápidas y sencillas. Procesa texto y genera texto como salida. Es una opción eficiente para tareas que no requieren una gran capacidad de razonamiento.
  - **Ventajas:** Bajo coste, rapidez de respuesta, tamaño de contexto adecuado.
  - **Desventajas:** Menor capacidad de razonamiento que GPT-4O.
- **GPT-4 vista previa y mini**
  - **Descripción:** Esta es la versión preliminar de nuestro nuevo modelo de razonamiento, diseñado para tareas complejas que requieren un amplio conocimiento general. El modelo grande (preview) tiene un contexto de 8192 tokens, mientras que el mini un contexto de 1024. Ambos se encuentran en fase beta y solo disponibles para cuentas "Tier 5".
  - **Ventajas:** Mayor capacidad de razonamiento, aprendizaje por refuerzo, tamaño de contexto amplio.
  - **Desventajas:** Acceso limitado, mayor tiempo de respuesta.
- **GPT-4 Turbo y GPT-4**
  - **Descripción:** Estos son el conjunto anterior de modelos de inteligencia artificial de alto rendimiento.
  - **Ventajas:** Rapidez, económicos para tareas sencillas.
  - **Desventajas:** Menos avanzados que GPT-4O y sus variantes.
- **GPT 3.5 Turbo**
  - **Descripción:** Este modelo, al igual que GPT-4 Turbo y GPT-4, son modelos un poco más antiguos, pero aún válidos para algunas aplicaciones.
  - **Ventajas:** Rápido y económico para tareas sencillas.
  - **Desventajas:** Menos potente y con un coste similar a GPT-4O mini.

- **DALL-E**
  - **Descripción:** Un modelo que puede generar y editar imágenes a partir de una indicación en lenguaje natural.
  - **Ventajas:** Creación e edición de imágenes a partir de texto.
  - **Desventajas:** No se utiliza frecuentemente en API Assistant.
- **TTS**
  - **Descripción:** Un conjunto de modelos que pueden convertir texto en audio hablado con un sonido natural.
  - **Ventajas:** Generación de audio a partir de texto, sonido natural.
  - **Desventajas:** Existen alternativas con mayor calidad y menor coste.
- **Susurro (Whisper)**
  - **Descripción:** Un modelo que puede convertir audio en texto de forma muy precisa.
  - **Ventajas:** Alta precisión en la transcripción de audio.
  - **Desventajas:** -
- **Incrustaciones (Embeddings)**
  - **Descripción:** Un conjunto de modelos que pueden convertir texto en forma numérica (vectores). Sirve para hacer "embeddings" de palabras o frases.
  - **Ventajas:** Representación numérica de texto.
  - **Desventajas:** -
- **Moderaciones (Moderation)**
  - **Descripción:** Un modelo especializado en detectar contenido sensible o inseguro en el texto. Es útil para moderar el contenido generado por otros modelos.
  - **Ventajas:** Filtro de contenido inapropiado.
  - **Desventajas:** -
- **Base GPT**
  - **Descripción:** Un conjunto de modelos básicos que pueden comprender y generar lenguaje natural o código. Son modelos sin ajuste fino y menos potentes que GPT-4O.
  - **Ventajas:** Comprensión y generación de lenguaje natural y código.
  - **Desventajas:** Menos potente que GPT-4O.
- **Obsoleto**
  - **Descripción:** Lista completa de modelos que han quedado obsoletos junto con el modelo de reemplazo sugerido.
  - **Ventajas:** Ayuda a mantener los sistemas actualizados.
  - **Desventajas:** -

Costes de los Modelos

Los precios de los modelos se basan en el número de tokens procesados, tanto de entrada como de salida. Puedes consultar la página de precios de OpenAI para obtener información actualizada sobre los costes.

Ejemplos de costes (Mayo 2024)

- **GPT-4O:**
  - $0.05 por 1000 tokens de entrada.
  - $0.15 por 1000 tokens de salida.

- **GPT-4O mini:**
  - $0.0015 por 1000 tokens de entrada.
  - $0.0060 por 1000 tokens de salida.

- **O1:**
  - $0.015 por 1000 tokens de entrada.
  - $0.060 por 1000 tokens de salida.

- **O1 mini:**
  - $0.003 por 1000 tokens de entrada.
  - $0.012 por 1000 tokens de salida.

Recuerda:

- GPT-4O es el modelo más caro pero ofrece la mayor capacidad de razonamiento y calidad de salida.
- GPT-4O mini es una alternativa más económica para tareas rápidas y sencillas.
- Los precios de los modelos pueden variar con el tiempo, así que consulta la página de precios de OpenAI para obtener la información más actualizada.

Costes de las herramientas de la API de asistentes

- **Intérprete de código:** $0.003 por sesión (1 hora).
- **Búsqueda de archivos:**
  - Primer GB: gratuito.
  - GB adicionales: $0.001 por GB y día.

Recordatorio clave:

- El coste de File Search se basa en el almacenamiento de los datos de contexto.
- El primer GB es gratuito, pero los GB adicionales tienen un coste de $0.001 por GB y día.

Tips para el uso de los modelos:

- Elige el modelo más adecuado para tus necesidades, considerando la complejidad de la tarea, el coste y la velocidad de respuesta.
- Utiliza GPT-4O mini para tareas inmediatas y sencillas, y GPT-4O para tareas que requieran mayor potencia y calidad.

- Controla el coste de los tokens, especialmente al utilizar File Search.