**INFORMACIÓN SOBRE CHAT GPT**

### **1. Tarifa por tokens de entrada y salida**

* Los modelos de IA cobran en función de la cantidad de *tokens* procesados.
* **Tokens de entrada** = lo que envías en el prompt.
* **Tokens de salida** = lo que el modelo genera en la respuesta.
* El precio se expresa como **costo por 1 000 tokens**. Ejemplo: GPT-4o cobra $0,005 por 1 000 tokens de entrada y $0,015 por 1 000 de salida.

### **2. Ejemplo de cálculo por llamada**

* Fórmula:  
   **Costo = (tokens entrada ÷ 1 000 × precio entrada) + (tokens salida ÷ 1 000 × precio salida)**
* Sirve para saber cuánto cuesta una única petición (una interacción usuario ↔ modelo).

### **3. Estimación mensual con N peticiones**

* Para calcular el gasto mensual, se multiplica el costo promedio por llamada × el número de llamadas (**N**).
* Ejemplo: si cada llamada cuesta $0,02 y haces 10 000 al mes → $200.

### **4. Otros costos adicionales**

* **Embeddings (RAG)**: crear representaciones vectoriales de texto para búsquedas contextuales. Se cobran aparte, también por 1 000 tokens.
* **Fine-tuning**: ajustar un modelo con tus propios datos; implica costo adicional por entrenamiento y uso posterior.
* **Herramientas/servicios**: algunos modelos incluyen funciones extra (Copilot, extensiones, API batch, caching) con precios fijos o descuentos.

En resumen:

* **Tokens** son la unidad básica de cobro.
* Con la fórmula puedes estimar costos por llamada o mensuales.
* Además de tokens, hay costos opcionales (embeddings, fine-tuning, suscripciones).

VENTANA DE CONTEXTO

### **1. Límite total de tokens visibles por el modelo (entrada + salida)**

* Cada modelo tiene un **máximo de tokens** que puede manejar en una sola interacción.
* Ese límite incluye tanto el **prompt de entrada** como la **respuesta generada**.
* Ejemplo: si el límite es 8 000 tokens y envías 7 000 de entrada, el modelo solo podrá generar hasta ~1 000 de salida.

### **2. Impacto práctico del límite**

* **Truncamiento**: si el texto excede el límite, la parte inicial se corta y el modelo “olvida” esa información.
* **Degradación de calidad**: al acercarse al máximo, puede responder más lento o con menor precisión contextual.
* **Estrategias para mitigar**:  
  + **RAG (Retrieval-Augmented Generation)**: en vez de mandar todo, se busca solo lo relevante en una base vectorial.
  + **Resúmenes**: condensar partes largas antes de pasarlas al modelo.
  + **Chunking**: dividir textos extensos en bloques más pequeños procesados por separado.

### **3. Comparación de capacidad (qué cabe en distintos contextos)**

Regla práctica: **700–800 palabras ≈ 1 000 tokens**

* **4K tokens (~3.000 palabras, 6–7 páginas de texto)** Caben documentos cortos (artículos, capítulos pequeños).
* **128K tokens (~100.000 palabras, un libro de ~350 páginas)** Caben manuales técnicos completos o novelas enteras.
* **200K tokens (~160.000 palabras, más de 500 páginas)** Caben colecciones de documentos extensos (varios libros o bases de conocimiento grandes).

👉 En resumen:

* El límite de tokens es como la “memoria de corto plazo” del modelo.
* Superarlo implica pérdida de contexto, por lo que conviene **usar estrategias como RAG, resúmenes y chunking**.
* Modelos modernos (Claude 3.5, GPT-4o) ya manejan decenas o cientos de miles de tokens, permitiendo trabajar con textos de tamaño libro.

**INFORMACIÓN SOBRE COPILOT**

1. **COSTOS**

### **💬 Modelo de cobro: por mensajes**

Copilot Studio mide el uso en **mensajes**, no en tokens. Cada interacción del usuario con el agente consume una cantidad determinada de mensajes, según el tipo de acción:

| **Tipo de acción en el agente** | **Costo en mensajes** |
| --- | --- |
| Respuesta clásica (predefinida) | 1 mensaje |
| Respuesta generativa (IA) | 2 mensajes |
| Acción de agente (flujo, decisión, llamada API) | 5 mensajes |
| Herramientas de IA generativa (premium) | 100 mensajes por cada 10 respuestas |

### **💳 Opciones de licenciamiento**

#### **🔹 Licencia fija mensual**

* **Precio**: $200 USD/año (≈ $16.67 USD/mes)
* **Incluye**: 25.000 mensajes al mes
* **Ventajas**: acceso completo a IA generativa, flujos, conectores premium, publicación en canales externos (web, WhatsApp, etc.)
* **Ideal para**: proyectos con uso constante o empresas con alto volumen de usuarios

#### **🔹 Pago por uso (Pay-as-you-go)**

* **Precio**: $0.01 USD por mensaje
* **Ventajas**: misma funcionalidad que el plan fijo, pero sin compromiso mensual
* **Ideal para**: freelancers, pilotos, MVPs, proyectos educativos en validación

#### **🔹 Incluido con Microsoft 365 Copilot**

* **Precio**: $30 USD por usuario/mes
* **Mensajes**: incluidos sin costo adicional, pero **solo en entornos internos** (Word, Excel, Teams)
* **Limitaciones**: no permite publicar agentes en sitios web ni canales externos

### **📊 Ejemplo de estimación mensual**

Supongamos que desarrollas un agente educativo en tu sitio web y recibes 500 visitantes al mes. Cada uno intercambia 6 mensajes (3 respuestas clásicas + 3 generativas):

* Clásicas: 3 × 1 = 3 mensajes
* Generativas: 3 × 2 = 6 mensajes
* Total por usuario: 9 mensajes
* Total mensual: 500 × 9 = **4.500 mensajes**

**Costo mensual:**

* Con licencia fija: incluido (dentro de los 25.000)
* Con pago por uso: 4.500 × $0.01 = **$45 USD**

### **📐 Fórmula general para estimar el costo en Copilot Studio**

plaintext

Costo mensual = (Mensajes por usuario × Número de usuarios × Costo por mensaje)

Donde:

* **Mensajes por usuario** = promedio de mensajes que cada usuario intercambia con el agente
* **Número de usuarios** = cantidad de personas que usan el agente al mes
* **Costo por mensaje** = depende del plan:
  + $0.01 USD por mensaje (pago por uso)
  + $0 USD si estás dentro del límite de 25.000 mensajes del plan fijo

### **Embeddings y fine-tuning**

* **Embeddings (para RAG)**: no se cobran por separado, están incluidos en el uso de herramientas de IA generativa.
* **Fine-tuning**: Copilot Studio no permite entrenamiento personalizado del modelo. La personalización se hace mediante flujos, reglas y conectores.
* **Herramientas premium**: si usas funciones avanzadas como razonamiento profundo o generación enriquecida, el costo puede subir a 100 mensajes por cada 10 respuestas.

### **🧮 Ejemplo práctico**

Supongamos que tu agente educativo recibe 300 usuarios al mes, y cada uno intercambia 8 mensajes (5 clásicos + 3 generativos):

* Clásicos: 5 × 1 = 5 mensajes
* Generativos: 3 × 2 = 6 mensajes
* Total por usuario: 11 mensajes

**Costo mensual (pago por uso):**

plaintext

11 × 300 × $0.01 = $33 USD

**Costo mensual (plan fijo):**

* Si estás dentro de los 25.000 mensajes: **$0 adicional**
* Si superas ese límite, los mensajes extra se cobran a $0.01 cada uno

### **🧠 Fórmula extendida (si usas herramientas premium)**

Si usas funciones como generación enriquecida o razonamiento profundo, puedes ajustar así:

plaintext

Costo mensual = [(Mensajes estándar × Costo estándar) + (Mensajes premium × Costo premium)]

Ejemplo: si usas 1.000 mensajes premium al mes (100 mensajes por cada 10 respuestas generativas avanzadas):

plaintext

1.000 × $0.01 = $10 USD adicionales.

1. **VENTANA DE CONTEXTO**

## **¿Qué es la ventana de contexto en Copilot Studio?**

La **ventana de contexto** es el límite total de tokens que el modelo puede procesar en una sola interacción. Incluye:

* **Tokens de entrada**: instrucciones del usuario, historial de conversación, contenido de archivos, mensajes del sistema.
* **Tokens de salida**: la respuesta generada por el modelo.

En Copilot Studio, este límite depende del modelo que se seleccione dinámicamente (como GPT-4o o GPT-5). Aunque no se muestra explícitamente al usuario, los límites prácticos son:

| **Modelo usado por Copilot Studio** | **Límite de tokens (entrada + salida)** |
| --- | --- |
| GPT-4o | Hasta 128.000 tokens |
| GPT-5 (en entornos avanzados) | Hasta 200.000 tokens (estimado) |

## **⚠️ Impacto práctico del límite**

Cuando se acerca o supera el límite de tokens, ocurren varios efectos:

### **🔻 Truncamiento**

* El modelo elimina tokens antiguos o menos relevantes, especialmente del historial de conversación o documentos largos.

### **📉 Degradación de calidad**

* Las respuestas pueden volverse incoherentes, perder contexto o ser incompletas.

### **❌ Fallos de generación**

* Si el contenido excede el límite, el modelo puede fallar al generar una respuesta o devolver un error silencioso.

## **🛠️ Estrategias para optimizar el uso del contexto**

### **🔹 RAG (Retrieval-Augmented Generation)**

* En lugar de cargar todo el documento, se indexa y se recuperan solo los fragmentos relevantes durante la conversación. Copilot Studio lo hace automáticamente con fuentes como SharePoint o archivos cargados.

### **🔹 Resúmenes previos**

* Condensar el contenido antes de enviarlo al modelo. Puedes usar flujos para generar resúmenes intermedios.

### **🔹 Chunking**

* Dividir documentos largos en secciones pequeñas y procesarlas por separado. Ideal para agentes que trabajan con manuales, libros o bases de conocimiento extensas.

## **📏 Comparación de tamaños de ventana**

Usando la regla aproximada: **700–800 palabras ≈ 1.000 tokens**

| **Ventana de contexto** | **Equivalente en palabras** | **Qué cabe** |
| --- | --- | --- |
| 4K tokens | ~3.200 palabras | Un artículo largo o una conversación breve con historial |
| 128K tokens | ~102.000 palabras | Un libro completo, múltiples documentos + historial |
| 200K tokens | ~160.000 palabras | Enciclopedia temática, sesiones extendidas, corpus técnico |

En Copilot Studio, aunque puedes cargar archivos de hasta **512 MB**, solo los fragmentos utilizados en la respuesta ocupan tokens.

**INFORMACIÓN SOBRE CLAUDE**

**Tarifas aproximadas (por 1,000 tokens):**

**Claude Sonnet 4:**

* Tokens de entrada: ~$3.00
* Tokens de salida: ~$15.00

**Claude Opus 4:**

* Tokens de entrada: ~$15.00
* Tokens de salida: ~$75.00

**Ejemplo de cálculo:** Una consulta típica de 500 tokens de entrada + 300 tokens de salida con Sonnet:

* Entrada: (500/1000) × $3.00 = $1.50
* Salida: (300/1000) × $15.00 = $4.50
* **Total por llamada: ~$6.00**

**Estimación mensual (1,000 consultas similares):**

* 1,000 llamadas × $6.00 = ~$6,000/mes

**Otros costos aproximados:**

* **Embeddings**: ~$0.10-$0.50 por 1M tokens
* **Fine-tuning**: Varía significativamente según el modelo y datos
* **Herramientas especializadas**: Costos adicionales variables

## **Ventana de Contexto**

### **Límites actuales de Claude:**

* **Claude Sonnet 4 (API)**: Hasta 1 millón de tokens (equivalente a ~750,000 palabras o 75,000 líneas de código)
* **Claude planes pagos (web)**: 200K tokens (aproximadamente 500 páginas de texto)
* **Claude Enterprise**: 500K tokens para Claude Sonnet 4

### **Impacto práctico:**

**Problemas cerca del límite:**

* **Truncamiento**: El contenido más antiguo se elimina automáticamente
* **Degradación de calidad**: El modelo puede perder coherencia con conversaciones muy largas
* **Costos elevados**: Más tokens = mayor costo por consulta

**Estrategias de manejo:**

* **RAG (Retrieval-Augmented Generation)**: Buscar y recuperar solo información relevante
* **Resúmenes**: Comprimir conversaciones largas en puntos clave
* **Chunking**: Dividir documentos grandes en fragmentos procesables
* **Gestión inteligente**: Mantener solo el contexto esencial

### **Comparación práctica (usando ~750 palabras = 1,000 tokens):**

**4K tokens** (~3,000 palabras):

* Email largo + respuesta
* Artículo corto de blog
* 5-6 páginas de texto

**128K tokens** (~96,000 palabras):

* Libro completo de 200-300 páginas
* Base de código mediana
* Múltiples documentos técnicos

**200K tokens** (~150,000 palabras):

* Novela completa
* Documentación extensa de proyecto
* Análisis de datasets grandes

**1M tokens** (~750,000 palabras):

* Más que toda la trilogía de "El Señor de los Anillos"
* Repositorio completo de código
* Base de conocimiento empresarial completa

