**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

**⚠ ️ IMPORTANTE:** Estos son valores aproximados para referencia. Los precios reales pueden ser muy diferentes.

**Debes consultar:**

* https://docs.anthropic.com para precios exactos
* https://support.anthropic.com para dudas específicas

## **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