

Longitud Máxima de Contexto de los Modelos de Lenguaje de Gran Escala

Recientemente utilicé la API de DeepSeek para generar un mensaje de commit, como se describe en Mensajes de Commit en Git con IA.

Cuando un commit involucra muchos archivos modificados, la API de DeepSeek informó que la entrada excedía su límite de longitud de contexto de 65,535 tokens ($2^{16} - 1$).

Aquí están los tamaños de ventana de contexto de algunos otros modelos:

- **Familia Claude 3:** Introducidos en marzo de 2024, estos modelos tienen ventanas de contexto que comienzan en 200,000 tokens.
- **GPT-4:** La versión estándar soporta 8,192 tokens, mientras que la versión extendida (GPT-4-32k) soporta 32,768 tokens.
- **LLaMA 2 de Meta:** La versión estándar soporta 4,096 tokens, pero las versiones ajustadas pueden manejar hasta 16,384 tokens.
- **Mistral 7B:** Soporta hasta 8,000 tokens.