Longueur de Contexte Maximale des Modèles de Langue à Grande Échelle

J’ai récemment utilisé l’API DeepSeek pour générer un message de commit, comme décrit dans Messages de Commit Git Pilotés par l’IA.

Lorsqu’un commit implique de nombreux fichiers modifiés, l’API DeepSeek a signalé que l’entrée dépassait sa limite de longueur de contexte de 65 535 tokens (2^16 - 1).

Voici les tailles de fenêtres de contexte de quelques autres modèles :

- **Famille Claude 3 :** Introduits en mars 2024, ces modèles ont des fenêtres de contexte commençant à 200 000 tokens.
- **GPT-4 :** La version standard prend en charge 8 192 tokens, tandis que la version étendue (GPT-4-32k) prend en charge 32 768 tokens.
- **LLaMA 2 de Meta :** La version standard prend en charge 4 096 tokens, mais les versions affinées peuvent gérer jusqu’à 16 384 tokens.
- **Mistral 7B :** Prend en charge jusqu’à 8 000 tokens.