

大規模言語モデルの最大コンテキスト長

最近、AI-Powered Git Commit Messagesで説明されているように、DeepSeek API を使用してコミットメッセージを生成しました。

コミットに多くの変更ファイルが含まれている場合、DeepSeek API は入力が 65,535 トークン ($2^{16} - 1$) のコンテキスト長制限を超えていると報告しました。

以下は、他のいくつかのモデルのコンテキストウィンドウサイズです：

- **Claude 3 Family:** 2024 年 3 月に導入され、これらのモデルは 200,000 トークンから始まるコンテキストウィンドウを持っています。
- **GPT-4:** 標準版は 8,192 トークンをサポートし、拡張版 (GPT-4-32k) は 32,768 トークンをサポートします。
- **Meta の LLaMA 2:** 標準版は 4,096 トークンをサポートしますが、微調整されたバージョンは最大 16,384 トークンを処理できます。
- **Mistral 7B:** 最大 8,000 トークンをサポートします。