Maximale Kontextlänge von Large Language Models

Ich habe kürzlich die DeepSeek API verwendet, um eine Commit-Nachricht zu generieren, wie in AI-Powered Git Commit Messages beschrieben.

Wenn ein Commit viele geänderte Dateien umfasst, meldete die DeepSeek API, dass die Eingabe das Kontextlängenlimit von 65.535 Tokens (2^16 - 1) überschritten hat.

Hier sind die Kontextfenstergrößen einiger anderer Modelle:

- **Claude-3-Familie:** Diese Modelle, die im März 2024 eingeführt wurden, verfügen über Kontextfenster, die bei 200.000 Tokens beginnen.
- **GPT-4:** Die Standardversion unterstützt 8.192 Tokens, während die erweiterte Version (GPT-4-32k) 32.768 Tokens unterstützt.
- **Meta’s LLaMA 2:** Die Standardversion unterstützt 4.096 Tokens, aber feinabgestimmte Versionen können bis zu 16.384 Tokens verarbeiten.
- **Mistral 7B:** Unterstützt bis zu 8.000 Tokens.