Maximum Context Length of Large Language Models

I recently used the DeepSeek API to generate a commit message, as described in AI-Powered Git Commit Messages.

When a commit involves many changed files, the DeepSeek API reported that the input exceeded its context length limit of 65,535 tokens (2^16 - 1).

Here are the context window sizes of some other models:

- **Claude 3 Family:** Introduced in March 2024, these models have context windows starting at 200,000 tokens.
- **GPT-4:** The standard version supports 8,192 tokens, while the extended version (GPT-4-32k) supports 32,768 tokens.
- **Meta's LLaMA 2:** The standard version supports 4,096 tokens, but fine-tuned versions can handle up to 16,384 tokens.
- **Mistral 7B:** Supports up to 8,000 tokens.