

大型语言模型的最大上下文长度

最近，我使用了 DeepSeek API 来生成提交信息，正如AI 驱动的 Git 提交信息一文所述。

当一次提交涉及大量更改文件时，DeepSeek API 报告输入超出了其 65,535 个令牌 ($2^{16} - 1$) 的上下文长度限制。

以下是其他一些模型的上下文窗口大小：

- **Claude 3 系列：** 2024 年 3 月推出，这些模型的上下文窗口从 200,000 个令牌起。
- **GPT-4：** 标准版本支持 8,192 个令牌，而扩展版本（GPT-4-32k）支持 32,768 个令牌。
- **Meta 的 LLaMA 2：** 标准版本支持 4,096 个令牌，但经过微调的版本可以处理多达 16,384 个令牌。
- **Mistral 7B：** 支持多达 8,000 个令牌。