Transformerの構成要素：Attentionとバイアス

- Full Attention
- Sparse Attention
- Multi-query Attention
- FlashAttention

まとめと考察：アーキテクチャの選択

- ほとんどのLLMは因果デコーダアーキテクチャーに基づき開発されており、他の選択肢に対する優位性についての理論的分析はまだ不十分
- 因果デコーダアーキテクチャは、優れたゼロショットと少数ショットの汎化能力を達成できる