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

まとめと考察：長いコンテキスト

- 時間とメモリ両方で二次関数的な計算コストがかかるため、コンテキスト長は制限される
- 外挿
  - トレーニングコーパスの最大長より長いテキストをエンコードする能力を外挿能力と呼ぶ
  - いくつかのPE手法は一定の外挿能力を持つことが知られ、ALiBi を用いたモデルは学習時より10倍長い配列でも比較的安定したperplexityを維持する
- 効率
  - メモリ消費量のスケールをほぼ線型できる効率的な計算方法が設計されている
  - FleshAttentionはシステムレベル(GPUのメモリI/O)で効率を改善する
  - Transformer以外のアーキテクチャ (状態空間モデルやRWKVなど) を考案する研究もある