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

なぜ次の単語を予測することが有効なのか

- デコーダのみのアーキテクチャの本質は、事前学習データを再構築するために次の単語を正確に予測することである
- 現在のところ、他のアーキテクチャに対する優位性を理論的に示した研究はない
- 多くの単語が存在するが、それらの単語をより良く予測することで、文章の理解度は改善する
- GPT-4は次の単語をよりよい精度で予測する