Transformerの構成要素：位置埋め込み

- 入力トークンの順番の情報を保持するために、位置埋め込み (Position Embeddings) が採用される
  - ❖ 絶対位置埋め込み
    - 正弦波を利用するなど
  - ❖ 相対位置埋め込み
    - キーとクエリの間のオフセットにしたがって生成する
  - ❖ 回転位置埋め込み
    - 各トークンの絶対位置に基づいて特定の回転行列を生成する
  - ❖ ALiBi
    - キーとクエリ間の距離に基づくペナルティでアテンションスコアにバイアスをかける

Transformerの構成要素：Attentionとバイアス

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