メモリ効率の良いモデル適応

パラメータ効率の良いチューニング方法

- LLMは膨大なパラメータを持つため、推論に必要なメモリ量が莫大
  - ➡ モデルを圧縮することでメモリ使用を削減できる
- 圧縮の方法として一般的には「量子化」が行われる
  - ➡ 量子化とは、浮動小数店からビット数の小さい整数にモデルのパラメータやアクティベーション (中間層) の値を変換して圧縮すること
  - ➡ 量子化を考慮した学習 (QAT) と学習後の量子化 (PTQ) があるが、後者の方が計算コストがはるかに低い
  - ➡ LLMは小さい言語モデルよりも量子化の影響を受けにくく、同じメモリコストであれば小さなモデルよりも量子化した大きなモデルを使用する方が良い