パラメータ効率の良いチューニング方法

- プロンプト・チューニング
  - 入力層に学習可能なプロンプトのベクトルを埋め込む
  - 入力プロンプトを補強して、特定のタスクに特化させられる
- Low-Rank Adaptation (LoRA)
  - モデルのパラメータの一部をより小さな行列に近似し、この行列に対して更新を行うことで次元数を減らして効率的に学習する
  - 出力にかかる時間が増加したり入力テキストの長さが制限されない

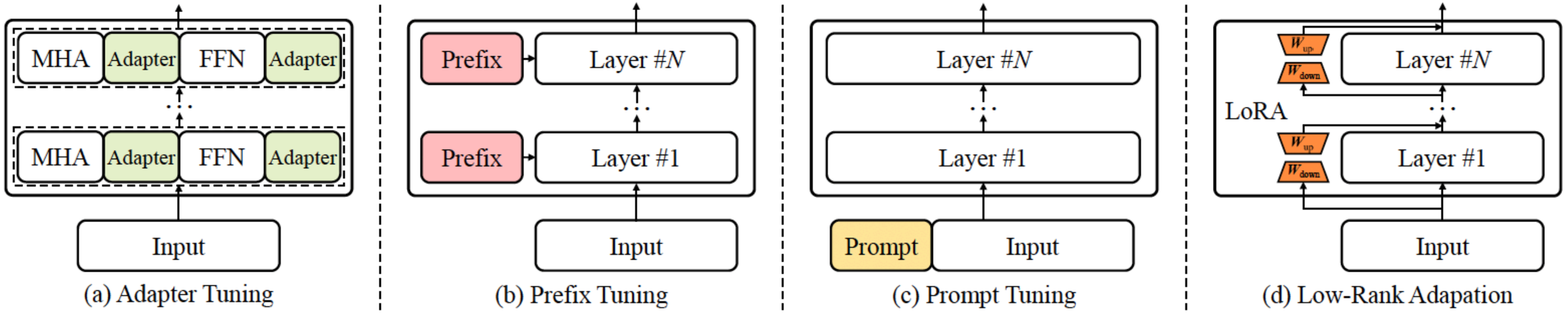

Fig. 10: An illustration of four different parameter-efficient fine-tuning methods. MHA and FFN denote the multi-head attention and feed-forward networks in the Transformer layer, respectively.