

ABCI 3.0開発加速利用（2024年度）成果概要（公開用）

課題名：LLM構築の効率化に資する内部挙動分析と計算コスト削減法の確立	実施時期：2025年1月20日から2025年3月30日 所属機関名：東北大学 代表者氏名：鈴木潤
成果概要： 本開発加速利用では、構築に時間やコストがかかる大規模言語モデル（LLM）の事前学習に対して、実験的検証に基づいて内部挙動を分析することで、計算コストの削減や性能向上を実現する研究を行っており、それらの実験を実施するための計算機利用に用いた。	
成果のポイント： 昨今のTransformerに基づく大規模言語モデル（LLM）の中間層は、自己注意機構とフィードフォワードネットワーク（FFN）という代表的な二つの構成要素の繰り返しより構成されている。ここでは、主に知識を記憶していると考えられているFFNに着目し、モデル内の層単位の配置場所に依存したFFNの重要度について検証する。具体的には、モデル全体のパラメータ数を維持したまま、一部の連続する層でFFNの中間次元を拡大し、残りの層からFFNを除去したモデル構成を考える。この層の位置に対して偏りのある構成と、従来の偏りのない標準的なモデル構成において、様々な言語モデルのベンチマークデータや知識の記憶に関するタスクの性能を比較する。複数のモデルサイズで評価を行った結果、FFN層の配置を偏らせることで、様々なタスク性能に大きな違いが生じることが明らかになった。これにより、Transformerの浅い層や深い層によって重要度や役割が異なることが示された。また、従来の偏りのない標準的な配置によるTransformerよりも、特定の位置にFFNを偏らせた場合のほうが高いタスク性能を示す傾向も確認され、偏りのない配置のモデル構成が必ずしも最良とは限らないことが実験的に示された。 上記成果以外にも、Transformerに基づくLLMのモデル構成が、入力と出力の埋め込み層および自己注意機構とFFNによる中間層（Transformerブロックと呼ばれる場合もある）の大きく2種類に分割できることに着目し、埋め込み層と中間層の二段階学習により、事前学習を効率的に進める方法や、事前学習済みモデルにおいて各層の予測分布の意味的収束性が示される一方で、隠れ状態ベクトルの数値的収束性が確認されていない点に注目した。これを踏まえ、事前学習において意味的収束性と数値的収束性の両方を満たす事前学習法を考案するための研究に関する実証実験を、試行錯誤も含めて多数実施した。	
成果についてより詳細な情報を提供しているWebページ、発表論文などの情報： ① 池田 航, 矢野 一樹, 高橋 良允, 李 宰成, 柴田 圭悟, 鈴木 潤. Transformer LLMにおける層単位のFFN層の重要度検証, 言語処理学会 第31回年次大会, 2025-03 [Link], に関する発展研究 [国際会議投稿済/査読中] ② 李 宰成, 矢野 一樹, 高橋 良允, 柴田 圭悟, 池田 航, 鈴木 潤. LLM事前学習の効率化と性質改善：埋め込み層および出力層のパラメータ固定による再利用, 言語処理学会 第31回年次大会, 2025-03 [Link], に関する発展研究 [国際会議投稿予定] ③ 柴田 圭悟, 高橋 良允, 矢野 一樹, 李 宰成, 池田 航, 鈴木 潤. Transformer LLMの内部挙動改善：隠れ状態ベクトルの数値的収束性の向上, 言語処理学会 第31回年次大会, 2025-03 [Link], に関する発展研究 [国際会議投予定]	

