ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名：	実施時期：2025年4月- 2026年3月
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	所属機関名：国立大学法人東北大学 代表者氏名：鈴木　潤

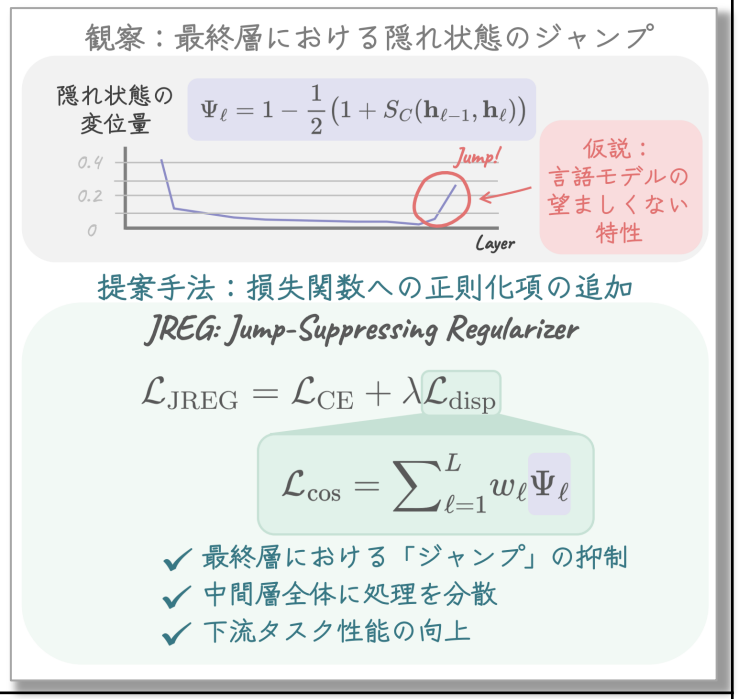
成果概要：本開発加速利用では、構築に時間やコストがかかる大規模言語モデル（LLM）の事前学習に対して、実験的検証に基づいて内部挙動を分析するとともに，そこで得られた知見に基づいて計算コストの削減や性能向上を実現する研究を行った。

成果のポイント：

ABCIの計算資源を活用して得た2025年度の代表的な研究成果として，大規模言語モデルの事前学習における性能向上に資する新たな学習手法を考案した。

具体的には，Transformer型言語モデルの内部表現に着目し，最終層付近で隠れ状態が急激に変化する現象を「ジャンプ」と名付け，その強さを定量化する指標を提案した。さらに，この現象が多様な既存モデルにおいて普遍的に観測されることを示した。また，事前学習の進行に伴ってジャンプが増幅される傾向を明らかにし，中間層の表現能力が十分に活用されていない可能性を示唆した。これを踏まえ，最終層付近の変化量を抑制する正則化手法「JREG」を提案した。JREGは，事前学習時にジャンプを罰則化することで，各層における情報処理の均等化を促す。Llamaベースの異なる規模のモデルを用いた評価の結果，モデル構造を変更することなく，ジャンプの低減と性能向上の両立を確認した。加えて，中間層の寄与が増大し，内部表現の遷移がより滑らかになることも確認した。以上より，本研究は，Transformer型言語モデルにおける学習効率と表現活用を改善する新たな正則化手法を示した。

さらに，ABCIを活用して得られたその他の研究成果として，NeurIPSなどのAI関連分野の難関国際会議を含む査読あり国際会議・ワークショップに7件の論文が採択された。また，国内学会でも9件の発表を行った。



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 Keigo Shibata, Kazuki Yano, Ryosuke Takahashi, Jaesung Lee, Wataru Ikeda, Jun Suzuki. “Suppressing Final Layer Hidden State Jumps in Transformer Pretraining” Findings of the Association for Computational Linguistics: EACL 2026. ② Mengyu Ye, Jun Suzuki, Tatsuro Inaba, Tatsuki Kuribayashi. ”Transformer Key-Value Memories Are Nearly as Interpretable as Sparse Autoencoders” Advances in Neural Information Processing Systems: NeurIPS 2025. ③ Wataru Ikeda, Kazuki Yano, Ryosuke Takahashi, Jaesung Lee, Keigo Shibata, Jun Suzuki. “Layerwise Importance Analysis of Feed-Forward Networks in Transformer-based Language Models” In Proceedings of the Second Conference on Language Modeling: COLM 2025.

他：査読あり国際会議/ワークショップ4件．国内学会発表9件．