TSUBAME 共同利用　令和7年度　学術利用　成果報告書

利用課題名　大規模言語モデルにおけるスケーラブルな外部知識表現
英文：Scalable External Knowledge Representation of Large Language Models for Generative Information Retrieval

利用課題責任者　杜　キン
Xin DU

早稲田大学　理工学術院　基幹理工学部
Waseda University, School of Fundamental Science and Engineering
URL　https://www.ml-waseda.jp

邦文抄録（300 字程度）

本研究は、「大規模言語モデルにおけるスケーラブルな外部知識表現」を課題とし、LLM が外部知識を効率的に表現・活用するための数理的基盤と計算論モデルを構築することを目的とする。近年、LLM の高性能化に伴い外部知識の活用が重要となるが、そのスケーラビリティと非線形な構造の解析が課題となっている。本研究では、フラクタル幾何学に基づく相関次元を活用し、自己回帰型 LLM の認識論的複雑性を定量化し、学習過程における 3 つの明確な段階を発見した。さらに、相関次元が LLM の幻覚傾向や生成テキストの崩壊を検出する有効な指標となることを実証し、新たな生成的検索・クラスタリング手法を開発した。成果は NeurIPS 2025 を含む複数の国際学会・雑誌に発表した。今後は、相関次元の形成メカニズムを解明し、法律・金融分野への応用を推進する。

英文抄録（100 words 程度）

This study focuses on "Scalable External Knowledge Representation of Large Language Models for Generative Information Retrieval" and aims to establish a mathematical and computational framework for LLMs to efficiently represent and utilize external knowledge. We use correlation dimension based on fractal geometry to quantify the epistemological complexity of autoregressive LLMs, identifying three distinct phases during pretraining. Our results show that correlation dimension effectively detects LLM hallucinations and text degeneration. We also develop new generative retrieval and clustering methods. Our findings are published in NeurIPS 2025 and other international venues. Future work will clarify the formation mechanism of correlation dimension and promote applications in law and finance.

Keywords: 大規模言語モデル, 外部知識表現, 相関次元, 非線形性, 生成モデル

背景と目的

大規模言語モデル（LLM）の急速な発展により、自然言語処理技術は飛躍的に進歩し、外部知識を活用した高度なタスク遂行が可能となっている。しかし、LLM における外部知識の表現はスケーラビリティに課題があり、また、従来の評価指標（困惑度など）では、LLM の生成テキストに見られる反復や非一貫性といった問題を捉えられず、外部知識表現の本質的な非線形性と階層構造を解析する有効な手法が不足している。

本研究では、「大規模言語モデルにおけるスケーラブルな外部知識表現」を課題とし、フラクタル幾何学の相関次元を活用して LLM の外部知識表現の非線形性を定量化し、そのスケーラビリティを向上させる計算論モデルを構築することを目的とする。また、このモデルを基に、外部知識を効率的に活用する生成的検索・クラスタリング手法を開発し、法律・金融分野のテキスト処理に応用することで、実用的な情報処理技術の革新を目指す。

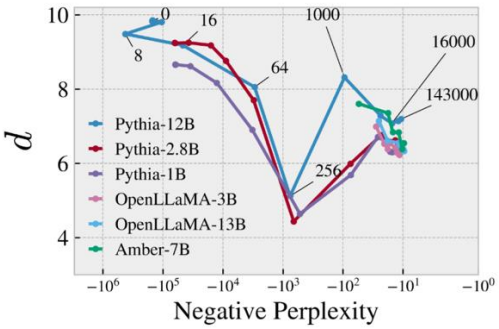
概要

LLM の高性能化に伴い、外部知識の表現と活用が自然言語処理の核心課題となっている。しかし、LLM が外部知識をどの程度適切に表現し、その非線形な構造を学習できているかについては、未だ十分な解析が行われていない。従来の自己回帰型 LLM は、局所的な予測精度に優れるものの、外部知識のグローバルな構造や階層的特性を捉えることが難しく、スケーラブルな表現が困難である。

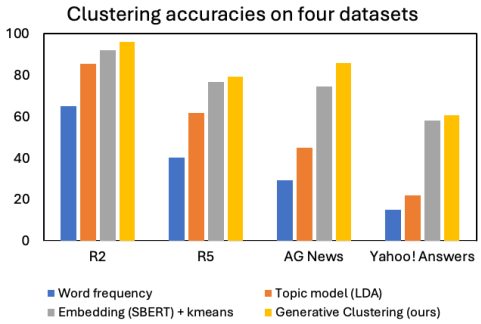
本研究では、TSUBAME の高性能計算環境を活用し、以下の 2 点を中心に研究を推進した。(1) 相関次元（フラクタル幾何学に基づく自己相似性の指標）を用いて、自己回帰型 LLM における外部知識表現の認識論的複雑性を定量化する手法を開発し、LLM の学習過程や生成特性を詳細に解析する。(2) 外部知識のスケーラブルな表現を活用し、情報論的アプローチに基づく生成的検索・文書クラスタリング手法を改良・開発し、従来手法の限界を突破する。TSUBAME の並列計算能力により、大規模データによるモデル学習や複雑な解析を効率化し、研究の進捗を大幅に加速させた。

結果および考察

本研究では、TSUBAME を活用して相関次元を用いた LLM の外部知識表現解析を行い、以下の重要な成果を得た。まず、自己回帰型 LLM（Transformer、Mamba など複数のアーキテクチャ）に相関次元を適用した結果、プレトレーニング過程に 3 つの明確な段階が存在することを発見した。この相関次元は、文脈依存的な複雑性を反映し、LLM の幻覚傾向や生成テキストの種々の崩壊形式（反復など）を確実に検出することができることを実証した。さらに、この手法は計算効率が高く、4 ビット量子化された LLM に対しても頑健で、複数の言語（自然言語 8 言語、プログラミング言語 3 言語）に適用可能であることを確認した。本成果は NeurIPS 2025 に発表され [5]、LLM の生成ダイナミクスに関する新たな知見を提供した。



また、外部知識のスケーラブルな表現を活用し、生成的検索モデルの改良と「生成クラスタリング」手法の最適化を行った。TSUBAME の計算資源を活用して大規模外部知識データによるモデル学習を行い、LLM の外部知識索引記憶限界を解明し、効率的な索引付与方式を提案した。その結果、従来のベクトル埋め込みに基づく手法と比較して、文書検索・クラスタリングの精度が大幅に向上し（図 2）、法律・金融分野のテキスト処理への応用可能性が高まった。本成果は AAAI 2025 [4]、ICML 2024 [3] に発表されたほか、相関次元に関する基礎研究成果は Physical Review Research [1] に掲載され、フラクタル分野の国際ワークショップ Geometry and Stochastics にて招待講演 [2] として発表された。



まとめ、今後の課題

本研究では、「大規模言語モデルにおけるスケーラブルな外部知識表現」を課題とし、TSUBAME の高性能計算環境を活用し、相関次元に基づく LLM 非線形性解析手法を開発・実証し、外部知識活用型の生成的検索・クラスタリング手法を改良し、NeurIPS 2025 を含む国際学会・雑誌に成果を発表した。

今後は、相関次元の形成メカニズムを解明し、その LLM 評価指標としての定着とモデル最適化技術との組み合わせを検討し、開発手法の法律・金融分野への実用化を推進し、多言語における外部知識表現のスケーラビリティ解析を行う。

参考文献

[1] Xin Du and Kumiko Tanaka-Ishii. Correlation dimension of natural language in a statistical manifold. Physical Review Research, 6 (2), L022028.
[2] Kumiko Tanaka-Ishii and Xin Du. Correlation dimension of large language models. In Proceedings of Fractal Geometry and Stochastics 7, 2024. Invited talk.
[3] Xin Du, Lixin Xiu, and Kumiko Tanaka-Ishii. Bottleneck-Minimal Indexing for Generative Document Retrieval. In Proceedings of ICML 2024.
[4] Xin Du and Kumiko Tanaka-Ishii. Information-Theoretic Generative Clustering of Documents. In Proceedings of AAAI 2025.
[5] Xin Du, & Tanaka-Ishii, K. Correlation Dimension of Autoregressive Large Language Models. In The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025).