課題名（和文）LLM の信頼性向上のためのデータ管理基盤

課題名（英文）Data management system for improving the trustworthiness of LLMs

天笠俊之
筑波大学計算科学研究センター

１．研究目的

大規模言語モデル（LLM）はこの数年で急速な発展を遂げ、様々な応用において利用が進んでいる。その一方で、LLM の出力に基づいてクリティカルな意思決定を行おうとする際には、信頼性に関連して以下のような問題がある。1）モデル学習の際には、異なる教師データやプロンプトを用いて複数のモデルが作成される。あるモデルの出力に何らかの問題が見つかった場合、それに関連するモデル、学習に使用された教師データやプロンプトを特定可能にする必要がある。このために、LLM に関連するデータの来歴情報を管理する必要がある。 2) LLM の問題の一つに、誤りを含む情報を出力してしまう「ハルシネーション」がある。ハルシネーションが発生した際、それに関与した教師データを特定する技術が必要である。 これらを踏まえ、本研究では、LLM の信頼性向上を目的としたデータ管理基盤技術に関する研究開発をった。

２．研究成果の内容

知識グラフの特徴量を用いたエントロピー誘導型プローブによる LLM のハルシネーション予測

大規模言語モデル（LLM）は、質問応答を含む幅広いタスクにおいて優れた性能を示している。しかし、時として誤解を招く回答や不正確な回答を生成する「ハルシネーション（幻覚）」の傾向は、依然として重大な課題である。

本研究では、LLM が生成した回答が、知識グラフ（KG）に由来する構造化された関係知識とどのように整合しているかを調査した。具体的には、複数のオープンソース LLM を用いた質問応答において、KG から抽出された特徴量と回答精度の相関関係を分析し、両者の間に密接な関連があることを明らかにした。本調査は主に Wikidata に焦点を当てているが、他の知識グラフにも一般化が可能である。

これらの知見を活かし、LLM の限界を露呈させる可能性が高い「困難な質問」を特定する、特徴量ベースのプロービング戦略を提案する。我々は、相関の確認された特徴量を用いて、KG と LLM の整合性が低い脆弱な箇所を検出するための評価プロービングを実施した。実験の結果、KG の特徴量がプロービングにおいて効果的な指針となることが示された。本研究は、より信頼性が高く、信頼に足る生成 AI システ

ムを構築する上で、構造化された知識グラフを活用することの価値を浮き彫りにしている。

Boltz-2 の潜在表現操作によるタンパク質多状態構造予測の効率化

従来のタンパク質立体構造予測 AI は、単一の安定な構造を出力する傾向が強く、輸送体やアロステリックタンパク質のように複数の状態間を遷移する系の構造多様性を捉えきれない課題があった。従来は入力となる MSA（多重配列アラインメント）のサンプリングやマスキングといった入力データへの摂動によって別状態の誘導を試みてきたが、これらは試行錯誤的で計算コストも高い。本研究では、AI 創薬モデル「Boltz-2」の内部表現を直接操作することで、構造多様性を系統的に引き出す手法「Boltz-sample」を提案した。本手法の中核は、モデル内部の「潜在ペア表現（pair representation）」を単一のスカラー値βによって一様に再スケーリングする点にある。このβは残基間相互作用の有効強度を制御する変数として機能する。$¥beta$が正の値では構造制約が強まり優勢な状態に予測が集中する一方、負の値では制約が緩和されて代替状態への遷移が起こりやすくなる。この操作は追加の学習や複雑な前処理を必要とせず、単なるテンソル乗算のみで実装できるため、極めて低い計算コストで多状態サンプリングが可能となる。評価実験の結果、OC23 や TP16 といったベンチマークにおいて、提案手法は標準的な推論（vanilla）を大幅に上回る状態カバー率を達成し、高コストな MSA クラスタリング法に匹敵する性能を示した。また、βの正負を調整することで予測構造の分布を効率的に制御できることも確認された。本研究は、入力データの加工に頼っていた従来の探索手法を、モデル内部の潜在空間を直接かつ系統的に制御するアプローチへと転換した点に大きな意義がある。

３．学際共同利用プログラムが果たした役割と意義

近年の LLM や深層学習を用いた研究では、最新の GPU を用いた大規模な計算が必須であり、本研究は、学祭共同利用プログラムなしでは実現が難しかったと言える。

４．今後の展望

今後も継続して、LLM や深層学習を用いた研究を推進する予定である。

５．成果発表

(1) 学術論文

Ushtar Ali, Steven J. Lynden, Akiyoshi Matono, Toshiyuki Amagasa, "Entropy-Guided Probing for Predicting LLM Hallucinations with Knowledge Graph Features," Proc. DEXA 2025, Vol. 1, pp. 68-82, August 2025.

(2) 学会発表

(3) その他

Shosuke Suzuki, Toshiyuki Amagasa, "Steering Conformational Sampling in Boltz-2 via Pair Representation Scaling," bioRxiv, January 23, 2026. doi: https://doi.org/10.64898/2026.01.23.701250

使用計算機	使用計算機に○	配分リソース※		
		当初配分	移行*	一般利用による追加
Pegasus	○	3840		
Miyabi-G				
Miyabi-C				
	※配分リソースについてはノード時間積をご記入ください。 ＊バジェット移行を行った場合、「+2000」「-1000」のように記入			