# LLM出力の因果・推論関係可視化に関する調査報告

## 研究の新規性と有用性

本研究は，日本語で生成された LLM の回答について，質問・概念・事実・プロセス・回答の 5 種類のノードを抽出し，それらの関係をグラフとして可視化する点に新規性がある。図は「イギリスの首都はどこか」という質問に対する LLM 回答の説明グラフの例であり，色分けされたノード（質問は橙，回答は青，概念・事実・プロセスは緑系）と矢印・線によって因果的・論理的つながりが示されている。従来，LLM は内部で暗黙的に知識を保持し推論を行うため，プロセスがブラックボックスであり透明性に欠けると指摘されてきた [1]。本手法では，LLM の出力内容から重要な概念や推論ステップを正規表現で抽出・分類し，NetworkX などで視覚化することで，モデルの推論過程を人間が追跡できる形にしている。この説明グラフはモデルの思考経路を構造的に示すため，ユーザが回答の根拠や因果関係を直感的に理解できる。結果として，LLM の動作に対する解釈可能性と透明性が向上し，回答の信頼性判断や誤り発見にも寄与する。

## 関連研究

### LLM の出力解釈・説明可能性

LLM の説明可能性とは，モデルの予測結果や動作原理を人間が理解できる形で示すことを指す。近年，LLM の推論過程を明示させる試みとして，逐次的な思考の鎖（Chain-of-Thought, CoT）を生成させるプロンプト手法が注目されている。Wei ら[1]は，段階的な推論過程を出力させることで複雑な問題の性能と可読性を同時に向上させた。また，モデル自身に根拠を出力させる自己説明の研究も進んでいるが，これらが本当に内部推論を反映しているかを検証する課題に対し，Bronzini[3] らは隠れ層のトークン表現を解析し，内部にエンコードされた知識を動的知識グラフとして抽出する枠組みを提案した [3]。因果関係や推論構造の可視化は，誤った推論や知識の欠落（いわゆる幻覚）を検知する手段としても有望である。

出力の可視化技術（グラフベース・因果構造の可視化）

テキスト生成結果をグラフ構造で表現・可視化する技術は，知識グラフや説明グラフの文脈で盛んに研究されている。Saha[3] らはスタンス推定タスク向けに ExplaGraphs データセットを提案し，推論に必要な概念ノードと因果関係エッジを含むグラフ生成課題を設定した。構造予測モデルは構造的完全性と意味的完全性の両面で従来手法を大きく上回った。さらに，LLM の推論プロセスをグラフ構造で探索・生成するプロンプト手法も登場している。Wen[4] らの MindMap は外部知識グラフをプロンプトに組み込み，モデルからマインドマップ相当のグラフを引き出して推論経路を可視化し，医学質問応答で精度と透明性を高めた。Hu[5] らの GRAG は Knowledge Graph を検索に活用し，文脈を保持したまま回答を生成する Graph RAG を提案している。本学生の研究は，正規表現による軽量実装で既存 LLM の出力を再利用し，説明可能性を付加できる点で実用上の意義が大きい。

### 今後の展開に向けた実験設計の提案

本アプローチをさらに発展させ有効性を客観的に評価するために、いくつかの観点での実験が考えられる。

1. 定量的な評価指標の設定と検証: 生成した説明グラフの品質を測る指標を設ける。例えば、グラフ中のノード・エッジがどれだけ正確に重要事実や因果関係を捉えているかを、人手で作成したゴールド標準のグラフや既知の事実と比較する評価が考えられる。ExplaGraphsの研究[2]ではグラフ構造の正しさ（構造的完全性）や意味的整合性を自動評価する手法が提案されており 、それにならって本手法のグラフについても、不要なノード・誤った関係の割合、重要情報の網羅率などを測定できる。また、類似研究ではBERTScore[6]によるテキスト類似度評価やGPT-4を用いた主観評価の自動化も試みられており 、大規模言語モデル自体にグラフ説明の妥当性を評価させる方法も検討に値します。
2. ユーザビリティ評価（主観評価）: 実際のユーザがこの説明グラフを見たときにどの程度理解が促進されるか、信頼感が向上するかを評価する。具体的には、エンドユーザや開発者に対し、(a)グラフ付きの回答とグラフなしの通常回答を提示して回答内容の理解度や納得感をアンケート調査、(b)グラフを用いてモデルの誤りを発見できるかをテスト、といったユーザ実験が有用である。前者ではLikert尺度[7]で「説明が分かりやすい」「根拠が明確」と感じた度合いを計測し、後者では意図的に含めた誤った事実をユーザが指摘できる率を比較する。
3. 他手法との比較検証: 提案手法の有効性を客観的に示すため、既存の説明手法やベースラインとの比較実験を行う。例えば、テキストによるChain-of-Thought説明や箇条書きの根拠列挙と、本研究のグラフ表示とでユーザの理解度・好みを比較したり、回答の事実誤り検出率を比較したりする。グラフは直感的視覚化が利点だが、一方でテキスト説明は読み慣れた形式でもあるため、それぞれの長所短所を定量的に評価する。また、自動評価ベンチマーク上で、グラフ化によって質問応答の正答率や一貫性が向上するか検証することも考えられる。例えば、因果推論を含む質問集合に対してグラフ説明の有無でモデル回答の妥当性を第三者評価させ、グラフが推論の齟齬を減らせているかを見る、といった比較が有効か。
4. 汎用性と拡張の評価: 異なるドメインやタスクへの適用実験も今後の展開として有意義です。本研究が地理的事実の質問に適用されているなら、医学や法律など専門領域の複雑な質問に対しても同様のグラフ可視化が通用するか試す。

IEEE Access への投稿適格性評価

1. スコープ適合性

IEEE Access は オープンアクセス型のマルチディシプリン総合誌 であり，機械学習・NLP・Explainable AI（XAI）を公式に扱う。本研究は

* LLM 出力に対する 新しい説明グラフ生成手法（因果・推論構造の可視化），
* 日本語という非英語圏言語への適用，
* 軽量実装（正規表現＋ NetworkX）という 実用性・再現性の高さ

を特色とするため，「XAI / NLP / 可視化技術」トピックに合致していると判断できる。

2. 新規性・学術的貢献

既存文献では英語を主対象とするグラフ説明が多い一方，本研究は

* 5 種ノード＋7 種関係 による汎用フレームワーク
* 日本語特有の構文パターンを利用した関係抽出
* 説明グラフの定量評価指標（構造的完全性・意味的完全性）とユーザビリティ評価案

を備える点で差別化できる。さらに，実証セクションで CoT・MindMap など代表的手法との比較を行い，精度・説明品質の優位性 を示せば，査読基準を満たす可能性が高い。

3. 追加で期待される実験と資料

1. ベンチマーク比較：ExplaGraphs タスクや日本語 QA コーパスで，F1／StCA／SeCA を測定し，既存グラフ生成モデル・テキスト CoT と比較。
2. ユーザ調査：グラフ表示有無で理解度・信頼度（Likert 5 段階），誤り検出率を測定。
3. 再現用コード／データ：GitHub で公開し，再現性を保証。
4. 動画サマリ：IEEE Access はビデオ提出を推奨している ￼ため，デモ動画（≤100 MB）を用意すると採択後の可視性が高まる。

4. 採択を高めるポイント

* 定量＋主観評価を併記 し，グラフ説明の有用性を実証する。
* 既存 XAI 可視化手法との比較 を明確にし，どの課題を解決したかを強調する。
* 応用シナリオ（教育支援，対話 AI デバッグなど）をケーススタディとして提示し，実用的インパクトを示す。
* 公開実装・データセット により再現性を担保し，IEEE Access の “High Reproducibility” 方針に沿う。

## 参考文献

[1] J. Wei et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” arXiv preprint arXiv:2201.11903, 2022.

[2] S. Saha et al., “ExplaGraphs: An Explanation Graph Generation Task for Structured Commonsense Reasoning,” in Proc. EMNLP 2021, pp. 7716–7740.

[3] M. Bronzini et al., “Unveiling LLMs: The Evolution of Latent Representations in a Dynamic Knowledge Graph,” arXiv preprint arXiv:2404.03623, 2024.

[4] Y. Wen, Z. Wang, and J. Sun, “MindMap: Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models,” in Proc. ACL 2024, pp. 10370–10388, doi: 10.18653/v1/2024.acl-long.558.

[5] Y. Hu et al., “GRAG: Graph Retrieval-Augmented Generation,” arXiv preprint arXiv:2405.16506, 2024.

[6] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” in Proc. 8th Int. Conf. Learn. Represent. (ICLR), Addis Ababa, Ethiopia, 2020. Online. Available: <https://openreview.net/forum?id=SkeHuCVFDr>.

[7] R. Likert, “A Technique for the Measurement of Attitudes,” Archives of Psychology, no. 140, pp. 1–55, 1932.