マルチホップ QA データセットの多様な観点からの自動分析と LLM の解答分析への応用

大崎稜司 [1]　三輪誠 [1,2]
[1] 豊田工業大学　[2] 産業技術総合研究所人工知能研究センター
{sd22019,makoto-miwa}@toyota-ti.ac.jp

概要

近年，大規模言語モデル (Large Language Models; LLM) の推論能力向上に伴い，マルチホップ質問応答 (QA) が注目され，様々なデータセットが提案されている．既存研究では，データセットや LLM の推論過程の構造に対する分析は行われているものの，推論タイプや知識分野といった多様な観点からの分析は十分とは言えない．本研究では，LLM を用いたデータセットの自動分析と，LLM の解答結果の分析への応用を提案する．データセットの自動分析では，質問から正解への推論過程と，多様な観点に基づく分類基準を自動生成し，生成した基準に基づいて推論過程を自動分類する．解答結果の分析では分類結果を用いた決定木分析を行う．実験により，LLM にとって解答困難な問題の特徴を LLM によって部分的に明らかにできることを示した．

1　はじめに

マルチホップ質問応答（QA）[1] は，複数の情報源を用いた複雑な推論を要求するタスクであり，様々なデータセットが提案されている．QA データセットを言語モデルの評価に用いる際には，QA データセット自体の特徴を分析し，理解した上で，LLM の解答結果を分析する必要がある．しかし，QA データセットに含まれる質問や解答，参照文書のデータには，各 QA に対して期待される具体的な推論過程は明記されていない．このような推論過程が明記されていない大規模な QA データセットに対し，人手による網羅的な特徴の分析は容易ではない．また，分析の明確な指針がなく，人手では分析の再現性と一貫性に欠ける懸念がある．

こうした背景から，QA データセットや解答の自動分析の研究が進められている．Liu ら [2] は，マルチホップ QA データセットの質問と解答を元に，正解の推論過程を LLM を用いて事後的に生成することで，マルチホップ QA データセットの推論構造を明示した．また，Jiang ら [3] は，LLM の解答の自動分析手法として，LLM に出力させた推論過程を木構造に変換し，木の深さや分岐の数などの構造を分析した．結果として，モデルや QA データセットごとの特徴的な推論の構造を明らかにした．

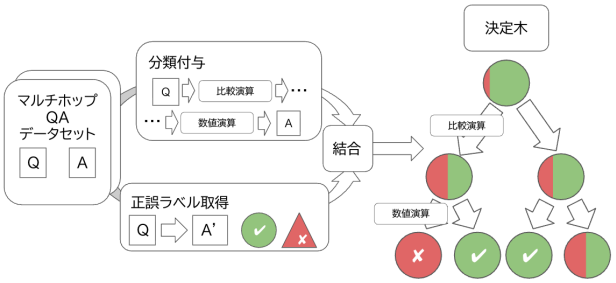

図 1　データセットの自動分析と決定木による LLM の解答分析

これらの手法は推論の構造のみを分析の対象としており，推論過程に含まれる，比較・橋渡しなどの異なる推論タイプや，異なる分野の知識，時間や因果関係などの意味的な推論など，様々な推論の特徴を考慮できていない．こうした推論の特徴を捉えるには，1 つの推論過程に対して，構造のみでなく，推論タイプや知識の分野などの多様な観点からの分析が求められる．

そこで本研究では，マルチホップ QA データセットの複数の観点からの多様な自動分析とその LLM の解答結果の分析への応用を提案する．具体的に，データセットの自動分析では，質問から解答への推論過程を自動生成し，多様な観点から推論過程の分類の基準を自動生成し，生成した基準に基づいて推論過程を自動分類する．また，解答結果の分析では，各分類を特徴量とした決定木を用いて，分類の組み合わせと LLM の解答時の正答率を分析する．この分析を通じて，LLM にとって困難な QA の特徴

を明らかにし，LLM の改善や今後のマルチホップ QA データセット作成のための指針の獲得を目指す．

本研究における貢献は以下の 2 点である．

- 既存のマルチホップ QA データセットを対象に，LLM を用いて多様な観点に基づいて推論過程を分類する自動分析手法を提案し，その分類が妥当であることを示した．
- LLM による分析結果を元に，LLM の正答・誤答の予測を行う決定木を用いた解答分析手法を提案し，どのような分類を持つ問題が LLM にとって困難であるかを部分的に明らかにした．

2 関連研究

2.1 マルチホップ QA の推論構造

自然言語による推論過程を明示したマルチホップ QA データセットは著者の知る限り存在しない．PER [2] では，マルチホップ QA の質問・解答・参照文書を元に正解の推論過程を LLM を用いて事後的に生成し，実際の LLM の推論過程と比較することで，推論過程を考慮した正誤評価を試みた．PER により，マルチホップ QA の推論構造を LLM によって明示できるが，その推論構造は評価にのみ用いられており，分析の対象とはしていない．

2.2 推論構造の分析

LLM の Chain of Thought (CoT）における出力を分析することで，複雑な推論タスクにおける言語モデルの振る舞いや失敗要因を調べる手法が提案されている．LCoT2Tree [3] では，CoT 出力を推論木と呼ばれるグラフ構造に変換し，得られたグラフを Graph Neural Network (GNN) [4] に学習させ，CoT 出力の構造から LLM の推論過程における失敗要因を特定した．また，分析を通じて，モデルやタスクごとに特有の推論構造が存在することを明らかにした．

このような推論構造はモデルやタスクの推論過程の性質に関する重要な特徴ではあるが，分析の対象が限定的であり，LLM の実際の推論過程を十分に考慮できているとは言えない．

3 提案手法

提案手法はマルチホップ QA データセットの自動分析と分析結果を用いた LLM の解答結果の分析で構成される．提案手法の概要を図 1 に示す．

3.1 QA データセットの自動分析

本節では，データセットに含まれる質問応答について，LLM を用いて質問から解答への推論過程を意味的な特徴に自動分類する手法について述べる．手法は推論過程の生成，推論過程の多様な観点からの分類の基準の自動生成，推論過程の分類からなる．

3.1.1 推論過程の生成

推論過程の生成では，LLM にマルチホップ QA データセットの質問・正解・参照文書を与え，質問から正解に至るまでの推論過程を自動生成する．以下，ここで生成された問題の性質を表すための推論過程を単に「推論過程」と呼ぶ．1 つの質問に対して，複数の推論過程があることや正解に至らない推論過程が出力されることも考えられるが，本稿では，簡単のため，出力された推論過程が問題に対する正しい解法の一つであると仮定する．推論過程の生成は Zero-shot で行い，推論過程が複数のステップから構成されるように指示を与える．

3.1.2 多様な観点からの分類基準の生成

既存のデータセットに対する分類は，一般に人が予め定義した固定的な観点に限られることが多い．一方で，LLM を用いて観点を含めた分類を一括で作成することもできるが，作成のたびに結果が変わるため，安定性に欠ける．本手法では，これらの問題を回避するため，3.1.1 節で生成された推論過程に含まれる各ステップ集合を整理するための多様な観点に基づく分類基準を，実際に分類を行う前に生成する．具体的には，まず，教育工学，論理学，認知科学などの専門的な知識体系を LLM に提示させ，それらの観点に基づいて複数のプロンプトを作成する[1)]．これらのプロンプトは，論理的誤謬やボトムアップなどの基準を用いてデータから特性を抽出するよう設計されている．次に，推論の集合すべてを分析対象として，作成したプロンプトを LLM に与え，推論過程を分類するための具体的な項目とその定義を生成させる．これにより，推論過程の深さ，論理構造の複雑さ，あるいは推論過程において LLM の失敗を誘発する要因など，多様な観点からの分類基準を獲得する．この過程では，後述する分類タスクにおいて一貫した判定が可能となるよう，各項目

1) LLM の生成結果を人手で取捨選択し，半自動でプロンプトを作成した．生成された基準の詳細は 4.1 節を参照．

の明確な定義と判断基準を記述させる．分類基準の生成は各観点につき一回のみ行い，データセット全体で共通の基準を用いる．

3.1.3 推論過程の分類

最後の推論過程の分類では，3.1.2 節で構築された分類基準を用いて，個々の推論過程がどの分類項目に該当するかを判定する．ここでは，生成された推論過程と，前節で生成された分類基準 (分類項目およびその定義) を LLM に入力し，推論過程を分析させる．分類は無作為に選定した 5 件の事例を入力した Few-shot で行う．分類はマルチラベル分類であり，一つの推論過程が複数の分類基準に該当する．LLM は，個々の推論過程に対して，定義に基づき分類する．また，分類基準が推論過程中の特定の推論ステップに対応する場合は，どの推論ステップに対応するかという具体的な判定根拠を生成させ，分類結果の解釈性を高める．

3.2 LLM の解答結果の分析

本節では，3.1.3 節で得られた推論過程の分類を元にした決定木を用いて，LLM の解答結果を分析する手法について述べる．結果の解釈が容易な決定木の性質を活かし，LLM が誤答に至る推論に含まれる特徴的な分類基準の組み合わせを可視化し，解答困難な問題の要因の特定を目指す．

3.2.1 正誤ラベルの付与

分析の対象とするデータセットの各質問について，LLM が正答できる問題か否かの正誤ラベルを付与する．データセットの各質問を参照文書とともに LLM に与えて得た解答を，LLM を用いて正解と比較し，正誤ラベルを得る．

3.2.2 モデル構築と分析手法

3.1.3 節で各推論過程に対して付与された分類ラベル（該当する分類項目）の出現頻度を特徴量とし，正誤ラベルを目的変数とした決定木を学習する．

得られた決定木の構造に基づき，分類ラベルの組み合わせとして LLM にとって困難な推論を特定し，分析を行う．本手法では，得られた決定木の構造を 2 つの視点から分析する．一つは，木の根に近い節点で選択された分類ラベルである．これらは正答・誤答を分ける支配的な要因であり，LLM の推論能力に大きく影響する問題の特性を示唆していると考えられる．もう一つは，誤答の割合が著しく高い葉に至るまでの，特定の分類ラベルの有無や頻度による分岐条件の組み合わせである．この組み合わせを用いることで，どのような条件下で LLM が失敗しやすいかを明らかにする．

4 実験と考察

4.1 実験設定

実験では，検索拡張生成システムのマルチホップ質問応答能力を評価する QA データセット FRAMES[5] に含まれる 588 問を用いた[2)]．分析及び解答生成には Gemini 2.5-Flash を用いた[3)]．

分類ラベル生成には，ブルームの分類法 [6] による分類，ボトムアップな分類，論理的誤謬に基づいた分類，論理的推論に基づいた分類，3 つの専門家ペルソナによる分類，トップダウンな分類プロセスによる分類，データ間の遷移に着目した分類の計 7 種類のプロンプトを用いた．生成されたラベルに対し，出現頻度が 2 回以下のもの，およびラベルが付与された問題集合間の Jaccard 係数が 0.9 以上の重複するラベルを除外し，144 種類の分類ラベルを特徴量として使用した．決定木分析には scikit-learn の DecisionTreeClassifier を用い，葉節点に含まれる最小サンプル数は 10，最大深さは制限無しに設定した．

4.2 LLM による自動分類の妥当性

LLM による推論過程の自動分類の信頼性を検証するため，無作為に抽出した 100 問に付与された 385 件の分類ラベルを人手評価した．評価は「妥当」「部分的に妥当」「妥当でない」の 3 段階とした．

結果として，85%が「妥当」と評価され，「部分的に妥当」を合わせると 98%となり，LLM による自動分類は推論の内容を適切に分類できているとわかった．このため，後の決定木分析の特徴量として用いた．人手評価の詳細を付録 B に示す．

4.3 決定木による解答結果の分析

決定木分析の結果，全 144 の分類ラベルのうち 12 種類が分岐に使用された．実際に分岐に使用された

2) 入力されるデータのモダリティをテキストに統一するため，全 824 問から表形式のデータを用いる問題 236 問を除外した．

3) LLM を用いた評価では同一モデルによるバイアスの問題が指摘されているが，本研究の目的は，評価ではなく分析であり，解答モデルは正誤ラベルの付与のみに用いるため，分析の信頼性への影響は少ないと考えられる．

分類ラベルと，その分類がどの観点から生成されたかを表 1 に示す．対象の QA に対する正答率は約 93%となった．学習された決定木では頻度による分岐ではなく有無による分岐のみが現れたため，以降はラベルの有無に着目して議論する．主要な分類ラベルとそれに基づく知見を以下に述べる．

4.3.1 正答率の高い分類の分析

118 問のインスタンスが属する決定木の最大の葉節点は，「単純数値計算」「フィルタリング」「関係性抽出」等を含まない問題群であり，これらは正答率 100%であった．また，「多段演繹」や「情報の統合」と分類された問題群においても正答率は極めて高かった．これは，LLM にとって，情報の検索や統合，あるいは形式的な多段推論そのものは，必ずしも困難なタスクではないことを示している．

4.3.2 誤答要因の分析

次のような分類ラベルの有無の組み合わせにおいて，正答率が低下していることがわかった．

(1)「単純数値計算」・「エンティティ抽出」 「単純数値計算」が含まれる問題において，正答率は「エンティティ抽出」ラベルの有無によって大きく分かれた (あり: 100%，なし: 56%) ．

「エンティティ抽出」は，テキストから固有表現や数値を明示的に切り出す処理を指す．この結果は，計算対象が明確に抽出可能な場合は LLM は正確に計算できるが，文脈から暗示的に数値を特定しなければならない（エンティティ抽出が明示的でない）場合，計算の前提となる数値の特定に失敗し，誤答に至ることを示唆している．

(2)「単純数値計算」・「フィルタリング/絞り込み」・「意図の把握」 「単純数値計算」を含まず，「フィルタリング/絞り込み」と「意図の把握」が含まれるノードにおいて，正答率は 66%となった．質問者の意図を推論した上で情報を絞り込む質問において，LLM の正答率が低下すると明らかになった．

4.4 考察

表 1 で示した分類の多くは，複数の観点で類似した分類として現れた．例えば，ブルームの分類における「単純数値計算」は，論理的推論による分類に「演繹的数値計算」として現れている．これらの分類は，データセット内の質問が共通して持つ性質であるために複数の観点で現れたと考えられる．このような分類は，データセットの基礎的な性質を表すため，決定木の分岐に現れたと考えられる．

表 1 生成されたうち，使用された分類ラベルと分類の観点

分類ラベル	観点
単純数値計算	ブルームの分類
エンティティ抽出	ペルソナ (データ科学)
フィルタリング/絞り込み	データ遷移
意図の把握	ペルソナ (認知心理学)
多段演繹	ペルソナ (形式論理学)
複雑な情報統合と多段階推論	ボトムアップ
関係性抽出	ペルソナ (データ科学)
情報の統合・合成	ブルームの分類
情報の関連付け	ブルームの分類
時系列順序演繹	ペルソナ (形式論理学)
多義語/曖昧な表現の誤解	論理的誤謬
演繹的ルール適用	論理的推論

一方で「意図の把握」のような観点に固有の分類も決定木の分岐として現れており，複数の観点からの分類が有効であることを示している．これ以外にも，決定木には現れなかったものの，ボトムアップにおける「地理と行政区画の階層的推論」のように複数の固有の分類が生成された．このような分類は，基礎的な分類に比べて事例が少ないものの，これらの詳細な分析を進めることで，マルチホップ QA の性質に対する理解が進むと期待される．

最後に，今回の分析ではマルチホップ QA において特徴的と考えられる推論ステップ数を決定木の特徴量としなかった．これは，推論ステップ数が 3 ステップから 6 ステップの範囲で正答率が 90%前後で安定しており，正答率とステップ数に顕著な相関はなかったことによる．この結果は，推論の長さよりも推論の内容が誤答の主因であることを示唆する．

5 おわりに

本研究では，マルチホップ QA データセットの分析・理解に向けて，QA の推論過程に対する多様な観点からの自動分析と，それに基づく決定木を用いた LLM の解答分析手法を提案した．FRAMES データセットを対象とした実験では，LLM の苦手とする推論パターンを可視化し，いくつかの誤答要因を特定した．今後の課題として，分析で明らかとなった困難な分類を持つ質問の LLM による自動生成や，決定木に現れなかった少数の分類に対する階層的な分類器による分析，LLM の QA の際の推論過程を利用したエラー解析，LLM のモデル変更やその組み合わせによる分析結果の変化の調査，および他のデータセットに対する分析が挙げられる．

謝辞

この成果の一部は，NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託業務（JPNP25006）の結果得られたものです．本研究に関して有益なご議論とご助言をいただきました産業技術総合研究所の浅田真生氏に深く感謝いたします．

参考文献

[1] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[2] Qichuan Liu, Chentao Zhang, Chenfeng Zheng, Guosheng Hu, Xiaodong Li, and Zhihong Zhang. Beyond the answer: Advancing multi-hop QA with fine-grained graph reasoning and evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 23433–23456, Vienna, Austria, July 2025. Association for Computational Linguistics.

[3] Gangwei Jiang, Yahui Liu, Zhaoyi Li, Wei Bi, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. What makes a good reasoning chain? uncovering structural patterns in long chain-of-thought reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 6501–6525, Suzhou, China, November 2025. Association for Computational Linguistics.

[4] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications, 2021.

[5] Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 4745–4759, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[6] Benjamin Samuel Bloom. **Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain**. Longman, London, 1956.

A　作成された決定木

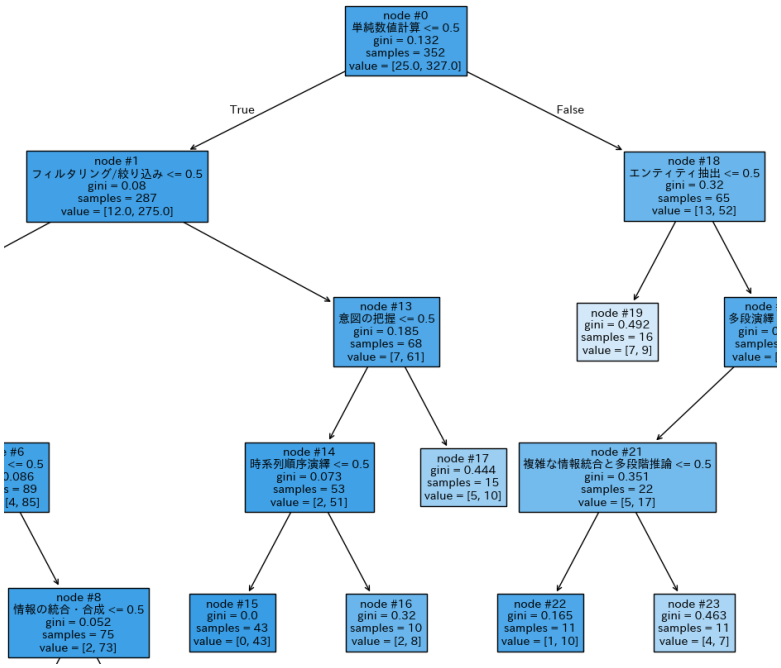

図 2　決定木モデルによる分析結果の一部．左の分岐は節点に示された分類を持たないことを，右の分岐はその分類を持つことを示す．

本稿の実験で実際に作成された決定木の図を図 2 に示す．

B　人手評価の詳細

人手による評価は，著者を含む 2 人の注釈者によって行った．2 人の注釈者間での一致率は 86%であった．注釈者間で結果が異なる場合は，より低い評価を採用した．注釈者ごとの評価結果を表 2 に示す．

表 2　人手による分類の妥当性評価 (100 問，ラベル 385 件)

評価	件数	割合
妥当	326	85%
部分的に妥当	51	13%
妥当でない	8	2%

C　使用したプロンプト

図 3 に分類生成の際に使用したプロンプトの例を示す．紙面の都合上，プロンプトの一部を省略した．

図 3　データサイエンティストのペルソナに基づいた分類生成のプロンプト

指示
(前略)
ペルソナ 3：データサイエンティスト
あなたの役割と分析視点：
あなたは、データフローと情報変換の観点からシステムを分析するデータサイエンティストです。あなたの目的は、推論過程を情報処理のパイプラインとして捉え、各ステップで情報がどのように抽出、変換、集約、フィルタリングされるかを明らかにすることです。入力情報（コンテキスト）、中間生成物、最終出力の関係性に注目してください。
生成すべきアウトプット：
- **分類軸の名称:** 情報変換プロセスに基づく推論パイプライン分類
- **分類カテゴリと定義:**（$n\_category$ つ程度）
- 各カテゴリの基準は明確かつ網羅的に記述してください。
- 分類名は，日本語名でのみ記載し，英語などの別名は表記しないでください．
- 各問題が複数のカテゴリに属する可能性もあります．