マルチホップ QA の根拠情報を用いた LLM の “偽” 正解の分析

石井愛 [1] 井之上直也 [2,1] 鈴木久美 [1] 関根聡 [1]
[1] 理化学研究所 [2] 北陸先端科学技術大学院大学
ai.ishii@riken.jp naoya-i@jaist.ac.jp hisami.suzuki@a.riken.jp
satoshi.sekine@riken.jp

概要

LLM がどのような知識に基づいて推論しているのか，そのきめ細かい調査の一つとして，回答の根拠となる導出情報が付与されたマルチホップ QA データセットを用いて GPT-4 の出力を分析する．分析の結果，回答が正解しているにも関わらず，根拠となる導出に誤りが含まれる “偽” 正解といえる現象が発生し，“偽” 正解を除くと，正解率は 60%から 40%に低下することを示す．さらに，外部知識を組み合わせた場合のカバー率の調査結果から，LLM を既存の外部知識と適切に組み合わせることによる改善の可能性を示す．

1 はじめに

大規模言語モデル（LLM）が様々なタスクで高い精度を達成する一方で，LLM が事実と矛盾する内容を生成することが示され [1, 2]，LLM の内部知識の評価や LLM が事実に基づいて回答をできるかといった事実性に関する検証結果が複数報告されている [3, 4, 5, 6, 7]．

QA 問題の解決において，LLM がどのような知識に基づいて推論して回答しているのかは，回答の導出に至った導出過程を出力させ，その出力を精査する必要がある．Zheng ら [7] はマルチホップ QA データセットを用いてエラー内容を人手で分類し，エラーの半数以上が事実性に関するエラーであり，知識を文章や段落等の粒度で与えることによる結果の改善を示している．ただし，このような詳細な調査においても，問題に正解した場合も含めてどの程度導出が正確に実行されていたかは明らかになっていない．

我々は LLM の導出過程のきめ細かい調査の一つとして，導出情報が付与されているマルチホップ QA データセットである JEMHopQA[1)] [8] を用いて GPT-4 が出力する質問への回答と，導出を調査した．調査結果から，回答は正解しているにも関わらず，根拠となる導出に誤りがある “偽” 正解といえる現象が発生することがわかった（図 1）．

JEMHopQA の正解セット
質問: 長嶋茂雄と小林旭，どちらが年上ですか？
導出: (長嶋茂雄, 生年月日, 1936 年 2 月 20 日);
(小林旭, 生年月日, 1938 年 11 月 3 日)
回答: 長嶋茂雄

GPT-4 の出力
導出: (長嶋茂雄, 生年月日, 1936 年 1 月 20 日);
(小林旭, 生年月日, 1939 年 4 月 13 日)
回答: 長嶋茂雄

図 1: “偽” 正解の例（赤字：導出のエラー部分）

本稿では，JEMHopQA を用いた GPT-4 の出力の評価結果および，手作業で分析した結果から “偽” 正解の発生頻度やその具体例を示す．さらに，導出の誤りを，既存の構造化知識ベース (KB) がどの程度カバーできるかを調査し，LLM と KB の適切な組み合わせによる改善の可能性を示す．

2 分析方法

2.1 データセット

本分析では，回答の導出情報が付与されている Wikipedia をベースとした日本語のマルチホップ QA データセットである JEMHopQA を用いる．JEMHopQA の導出情報は，図 1 の”(長嶋茂雄, 生年月日, 1936 年 2 月 20 日)”のように，主語エンティティ（長嶋茂雄）と目的語エンティティ（1936 年 2 月 20 日）間の半構造化された関係（生年月日）を表すトリプルの形式である．質問はマルチホップ推論が必要な質問であり，各質問と回答のペアには 2 つ以上の導出ステップが付与されている．

1) https://github.com/aiishii/JEMHopQA

JEMHopQA を用いて評価するタスクは，質問 Q が与えられたとき，タスクは (i) 答え A を予測し，(ii)A の根拠となる導出 D を生成するタスクである．

2.2 評価指標

導出はトリプルの形で半構造化されているが，エンティティも関係も類義語のバリエーションや言い換えの影響を受けるため，厳密な文字列マッチによる自動評価は困難である．そこで，井之上ら [9] を参考に，導出の評価において類似度スコアを用いる．具体的には，トークン化[2] された単語に対する正規化レーベンシュタイン距離を使用し，前処理として類義語辞書[3] を用いる[4]．類義語辞書は，データセットで観測された類似する単語群を追加して用いる．

導出　回答である導出 G と，システムが出力する導出 D が与えられたときのスコアを $c(D;G)$ とし，そのスコアに基づく，精度，再現率，f_1 を次のように定義する：

$$\mathrm{pr}(D) = \frac{c(D;G)}{|D|}, \mathrm{rc}(D) = \frac{c(D;G)}{|G|}$$
$$\mathrm{f}_1(D) = \frac{2 \cdot \mathrm{pr}(D;G) \cdot \mathrm{rc}(D;G)}{\mathrm{pr}(D;G) + \mathrm{rc}(D;G)}$$

ここで，$|G|$ と $|D|$ は与えられた質問に対するトリプルの数である．$c(D;G)$ は，G と D に含まれるトリプル d_i，g_j について，d_i と g_j の類似度スコア $a(d_i, g_j) \in [0, 1]$ を計算し，スコアの合計が最大となる組み合わせのスコアとする．

また，導出の正確性を評価するため，f_1^{ent}，f_1^{rel}，および f_1^{full} の 3 つのスコアを用いる．f_1^{ent} は主語と目的語エンティティの $a(d_i, g_j)$ の平均，f_1^{rel} は d_i と g_j の関係についての $a(d_i, g_j)$ の平均，f_1^{full} は主語と目的語エンティティ，および関係についての $a(d_i, g_j)$ の平均である．

回答　回答の評価には Exact Match(EM) と Similarity Match(SM) スコアを使用する．SM は上述の類似性計算を使用する．

2.3 実験設定

評価には，47 の構成問題と 73 の比較問題から構成される JEMHopQA 開発セットを対象に，OpenAI API[5] 経由で `gpt-4-0613` モデルを用い，以下の 5 種類のプロンプトで実験を行った：

1. **Zero-shot**: 質問のみを与える
   > 次の質問に，簡潔な名詞句または YES/NO で答えてください:
   > 杉咲花の父の職業は？
   > =>
2. **5-shot**: Train セットから 5 つのランダムなサンプルを Few-shot の例として含める
   > 次の質問に，簡潔な名詞句または YES/NO で答えてください:
   > ルーヴル美術館が所在する都市の市長の名前は？ => アンヌ・イダルゴ
   > 奈良市とドバイではどちらが人口が多いですか？ => ドバイ
   > (... 他 3 例)
   > 杉咲花の父の職業は？
   > =>
3. **Chain-of-Thought (CoT) 5-shot**: CoT[10] を提供する命令を，5 つのサンプルとともに追加する
   > 次の質問に，根拠を提示しながら，簡潔な名詞句または YES か NO で答えてください:
   > ルーヴル美術館が所在する都市の市長の名前は？ => (ルーヴル美術館, 所在地, パリ); (パリ, 市長, アンヌ・イダルゴ) => アンヌ・イダルゴ
   > (... 他 4 例)
   > 杉咲花の父の職業は？
   > =>
4. **Gold D**: 正解の導出 D を与える
   > 次の質問に，簡潔な名詞句または YES か NO で答えてください:
   > 杉咲花の父の職業は？ => (杉咲花, 父, 木暮武彦); (木暮武彦, 職業, ギタリスト)
   > =>
5. **Gold D 5-shot**: 正解の導出 D を，5 つのサンプルとともに追加する
   > 次の質問に，簡潔な名詞句または YES か NO で答えてください:
   > ルーヴル美術館が所在する都市の市長の名前は？ => (パリ, 市長, アンヌ・イダルゴ); (ルーヴル美術館, 所在地, パリ) => アンヌ・イダルゴ
   > (... 他 4 例)
   > 杉咲花の父の職業は？ => (杉咲花, 父, 木暮武彦); (木暮武彦, 職業, ギタリスト)
   > =>

パラメータは，`temperature` は 1.0，`max_tokens` は 32，根拠も生成する際は 256 に設定する．各実験は 3 回実行した結果の平均を報告する．

2) トークン化には Sudachi (https://github.com/WorksApplications/SudachiPy) を用いた
3) https://github.com/WorksApplications/SudachiDict/blob/develop/docs/synonyms.md
4) https://github.com/WorksApplications/chikkar
5) https://platform.openai.com/

3 GPT-4 の評価結果と考察

3.1 マルチホップ QA にどの程度答えられるのか

表 1 に Zero-shot，5-shot，Cot 5-shot の設定の回答の Exact Match (EM) と Similarity Match (SM) 結果を示す．5-shot の設定が EM を 48.9%から 55.6%に向上させ，さらに，CoT 設定により約 4%向上した．JEMHopQA の開発セットにおいて，Cot 5-shot の設定が最も精度が高く約 60%正しく回答できることがわかった．

	Answer EM	Answer SM
Zero-shot	0.489	0.507
5-shot	0.556	0.571
CoT 5-shot	**0.597**	**0.629**

表 1: GPT-4 による評価結果

	Answer EM / SM	Derivation $f_1^{\text{ent}}/f_1^{\text{rel}}/f_1^{\text{full}}$
構成	0.305/0.385	0.552/0.720/0.606
比較	0.785/0.785	0.724/0.707/0.718
ALL	0.597/0.629	0.656/0.712/0.674

表 2: GPT-4 CoT 5-shot 設定の結果詳細

3.2 GPT-4 の誤答の要因はなにか

GPT-4 の誤答の要因を調査するため，CoT 5-shot の導出トリプルの評価結果を合わせた詳細を表 2 に示す．導出のエンティティの正しさの尺度である f_1^{ent} は，構成問題で 0.552，比較問題で 0.724 であり，約 17%の相違があった．これは，質問内で 2 つの主語エンティティが明示的に言及される比較問題に対し，構成問題ではブリッジエンティティが暗黙的であり，識別が困難なためである．

導出トリプルの誤りをさらに調査するため，CoT 5-shot の誤答のケースを手動で分類した[6]．誤答 48 問のうち，推論でのエラーは 1 問のみであり，残りの 47 問は出力した導出トリプルの誤りに起因することがわかった．構成問題における導出トリプルの誤りは 31 問で，そのうち 58%がブリッジエンティティの識別エラー，残りはブリッジエンティティを主語とする目的語エンティティのエラーであった．その他として，正解では導出トリプルが 4 つの問題で 2 つのみが出力されたエラーが 1 問，エンティティのエラーと同時に発生していた．トリプルの関係の表現に起因する誤答は発生していなかった．

6) 導出トリプルの個数や関係表現が正解と異なる場合，質問に回答するために十分な情報と判断できれば正解とした．また，人口は時期により変動するため，近い数値であれば正解とした．

3.3 導出 D を与えられれば正しく推論できるのか

誤答の要因の大部分はエンティティの識別に起因することが示されたことから，正しい導出トリプルを入力として与えられれば，GPT-4 は導出トリプルを用いて正しく推論できると考えられる．導出トリプルを用いて推論する能力を確認するため，正解の導出 D を入力として与えたケース Gold D および，Gold D 5-shot の設定での検証を実施した．各設定においては，GPT-4 が導出 D の順序から回答を導き出す可能性を考え，そのままの順序およびランダムな順序 (RND) として実施するケースをそれぞれ実施した．評価結果を表 3 に示す．

	Answer EM	Answer SM
Gold D	0.906	0.924
Gold D (RND)	0.914	0.939
Gold D 5-shot	0.956	0.976
Gold D 5-shot (RND)	0.953	0.969

表 3: 導出 D を入力として与えた評価結果

導出 D を与えなかった結果 Zero-shot，5-shot(表 1) と比較すると，Gold D，Gold D 5-shot はそれぞれ EM が約 40%向上し，Gold D 5-shot の設定の結果は約 95%であった．これにより，JEMHopQA を解く難しさは導出 D の識別にあり，導出 D を与えられればほぼすべてのケースで正しく推論できることがわかった．また，導出 D の順序をランダムに設定したケースとの結果の差異が小さいことから，導出 D の順序情報を用いているわけではなく，正しく推論に活用していると考えられる．

Gold D 5-shot の 3 回の試行で回答の EM が不正解となった問題数は 4〜6 問であった．そのうち，部分正解（大学を聞いているのに対し，学部まで答えた等）が 3〜4 問であり，その他は数値比較の問題であった．3 回の試行ですべて誤答となった数値比較の問題をその分析とともに付録図 2 に示す．

3.4 “偽” 正解はどの程度発生するのか

CoT 5-shot の設定での GPT-4 が出力した回答と導出トリプルについて，それぞれ正誤を手作業で分

類した結果を表 4 に示す．回答が正解であった 72 問のうち，32%（23/72）に導出トリプルのエラーが含まれており，導出にエラーが含まれているにも関わらず問題には “偽” 正解していることがわかった．回答の正解率は 60%(72/120) であるが，“偽” 正解を除いた回答と導出両方が正しい場合の正解率は 40%(49/120) であった．

“偽” 正解のうち，91%（21/23）は比較問題であり，数値の大小や日付の前後関係を問う 14 問および，エンティティが同じかどうかを問う 7 問であった．数値や日付の前後関係を問う問題では，トリプルの数値が正解と異なっていても，相対的な大小・前後関係は正解と変わらず正答できたケースであった．エンティティが同じかどうかを問う問題では，正解が NO のとき，エンティティが正解と異なっていても，もう一方のエンティティと同じではないことで正解となっていた．残りの 9%(2/23) は構成問題で，ブリッジエンティティが Wikipedia に存在しない組織や個人であったが，出力された答えは正しいケースであった．具体的なエラーの例を付録表 6 に示す．

		導出トリプル	
		正 (構成/比較)	誤 (構成/比較)
回答	正	49(14/35)	**23(2/21)**
	誤	1(0/1)	47(31/16)

表 4: 回答と導出トリプルの正誤分類

3.5 導出トリプルのエラーは既存の KB によって改善できるのか

GPT-4 が出力した導出トリプルのエラーについて，既存の KB を利用することで改善できる可能性を探るため，JEMHopQA の導出トリプルのうち KB におけるカバー率を調査した．KB として，Wikipedia を対象とした 2 つの日本語の KB，Wikidata および森羅[7)][11] を使用した．森羅は Wikipedia の記事から属性と値のペアを抽出し，ENE カテゴリ [12] に従って構造化された KB データである．どちらの KB の知識表現も，JEMHopQA の導出トリプルと互換性がある．

表 5 において，最初の 3 列は，質問に必要な導出トリプルが Wikidata，森羅，または両方の結合のいずれで見つかったかの開発セット 120 問にする KB のカバー率を示している．各質問には 2 つ以上の導出トリプルが設定されているため，そのうちの一部正解した場合のカバー率を合わせて示す．Wikidata と森羅がすべての導出トリプルをカバーできたケースは，それぞれ 30%と 50%であり，両方の KB を合わせると 63%であった．両方の KB でカバーされていないトリプルの例としては，KB 内の関係より細かい関係（例えば，「兄弟」ではなく「妹」）や，対応のない関係（例えば，「ピアノを習いはじめた年」）が含まれる．

最後の 2 つの列は，GPT-4 と GPT-4 を両方の KB と組み合わせたカバー率を示している．GPT-4 のカバー率は 2 つの KB の組み合わせよりも低いが，すべて組み合わせるとカバー率が 77%となった．GPT-4 ではエンティティに誤りが含まれるエラーが 6 割程度発生するものの，上記のような KB 内の関係より細かい関係が必要とされるケースにおいて GPT-4 は適切な関係を出力できていたことから，互いに補完し合う結果となった．このことから，LLM を既存の KB と適切に組み合わせることができれば，本タスクのさらなる改善が可能であることが示された．

	Wikidata (W)	**森羅 (S)**	**W+S**	**GPT-4**	**GPT-4 +W+S**
完全	30.0%	50.0%	63.3%	40.8%	77.5%
一部	27.5%	29.2%	22.5%	22.5%	16.7%
無し	42.5%	20.8%	14.2%	36.7%	5.8%

表 5: 既存 KB および GPT-4 による導出のカバー率

4 おわりに

本稿では，根拠情報として導出がトリプルの形式で付与された日本語のマルチホップ QA 用のデータセットである JEMHopQA を用いて，GPT-4 の推論および出力した導出トリプルの評価結果を示した．分析結果から，GPT-4 の誤答はほぼすべて導出トリプルのエラーに起因し，正答した場合も導出トリプルにはエラーが含まれている “偽” 正解のケースがあることを示した．回答の正解率は 60%であったが，“偽” 正解を除いた回答と導出トリプル両方の正解率は 40%であることがわかった．さらに，根拠情報について既存の構造化された知識（KB）のカバー率および，GPT-4 と組み合わせた場合のカバー率を示し，既存 KB と組み合わせることでより良い性能を達成できる可能性を示した．そのような，KB と LLM の統合に向けた検討は今後の課題である．

7) http://shinra-project.info/

謝辞

本研究はJSPS科研費JP20269633, および19K20332の助成を受けたものです.

参考文献

[1] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. **arXiv preprint arXiv:2302.04023**, 2023.

[2] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. **ACM Comput. Surv.**, Vol. 55, No. 12, mar 2023.

[3] Cunxiang Wang, Sirui Cheng, Zhikun Xu, Bowen Ding, Yidong Wang, and Yue Zhang. Evaluating open question answering evaluation. **CoRR**, Vol. abs/2305.12421, , 2023.

[4] Pouya Pezeshkpour. Measuring and modifying factual knowledge in large language models. **arXiv preprint arXiv:2306.06264**, 2023.

[5] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics.

[6] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[7] Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. Why does chatgpt fall short in answering questions faithfully? **arXiv preprint arXiv:2304.10513**, 2023.

[8] 石井愛, 井之上直也, 鈴木久美, 関根聡. JEMHopQA: 日本語マルチホップ QA データセットの改良. 言語処理学会第 30 回年次大会発表論文集 (NLP2024), 2024.

[9] Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6740–6750, Online, July 2020. Association for Computational Linguistics.

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 24824–24837, 2022.

[11] Satoshi Sekine, Akio Kobayashi, and Kouta Nakayama. Shinra: Structuring wikipedia by collaborative contribution. In **Conference on Automated Knowledge Base Construction**, 2019.

[12] 関根聡, 安藤まや, 小林暁雄, 隅田飛鳥. 拡張固有表現定義の更新と日本語 Wikipedia 分類データ 2019. 言語処理学会第 26 回年次大会発表論文集 (NLP2020), pp. 1221–1224, 2020.

付録

JEMHopQA の正解セット
質問: 漫画『テセウスの船』と『来世ではちゃんとします』,『来世ではちゃんとします』の方が先の掲載ですか？
導出: (テセウスの船 (漫画), 発表年月日, 2017 年 6 月 22 日); (来世ではちゃんとします, 発表年月日, 2018 年 2 月 7 日)
回答: NO

GPT-4 の出力
回答: YES

誤答の要因
導出トリプルの関係「発表年月日」と質問文の「連載」との結びつけの失敗が考えられる．「先」という文脈によって変化する意味の解釈の失敗も考えられるが，他のどちらが先かを問う他の問題 12 問はすべて正解しており，これら 12 問では関係と質問文の表現は同一または類義語にあたるものであった．

図 2: 導出 D を入力とした与える設定で不正解となった問題

種類	誤りの箇所	質問	正解	システムの予測例
比較問題	目的語エンティティ（数値）	長嶋茂雄と小林旭，どちらが年上ですか？	回答：長嶋茂雄 導出：(長嶋茂雄, 生年月日, 1936 年 2 月 20 日); (小林旭, 生年月日, 1938 年 11 月 3 日)	回答：長嶋茂雄 導出：(長嶋茂雄, 生年月日, 1936 年 1 月 20 日); (小林旭, 生年月日, 1939 年 4 月 13 日)
	目的語エンティティ（エンティティ）	安美錦竜児と千代鳳祐樹は二人とも九重部屋に所属していましたか？	回答：NO 導出：(安美錦竜児, 所属部屋, 伊勢ヶ濱部屋); (千代鳳祐樹, 所属部屋, 九重部屋)	回答：NO 導出：(安美錦竜児, 所属部屋, 宮城野部屋); (千代鳳祐樹, 所属部屋, 九重部屋)
構成問題	Wikipedia に存在しないブリッジエンティティおよび関係	東條英機が死没した施設は何という戦争の後に設置されましたか？	回答：第二次世界大戦 導出：(東條英機, 死没地, 巣鴨拘置所); (巣鴨拘置所, 設置されたきっかけとなった戦争, 第二次世界大戦)	回答：第二次世界大戦 導出：(東條英機, 死没地, 杉並区立舎人病院); (杉並区立舎人病院, 設立, 第二次世界大戦)
	Wikipedia に存在しないブリッジエンティティ	永山瑛太の実弟の職業は？	回答：俳優 導出： (永山瑛太, 実弟, 永山絢斗); (永山絢斗, 職業, 俳優)	回答：俳優 導出： (永山瑛太, 実弟, 永山俊輔); (永山俊輔, 職業, 俳優)

表 6: エンティティに誤りが含まれるが正解したケースの例（青字：正解となった回答，赤字：導出のエラー部分）