Agentic RAG に向けた知識グラフに基づく日本語マルチホップ QA データセットの自動合成

板井 孝樹 [1,2] 松田 耕史 [2] 福地 成彦 [2]
[1] 東京都立大学 [2]SB Intuitions 株式会社
{koki.itai,koji.matsuda,akihiko.fukuchi}@sbintuitions.co.jp

概要

本研究では，KG を基盤とした，Agentic RAG 向けの日本語マルチホップ QA データセットの自動合成手法を提案する．本手法は，Wikidata から抽出した推論パスに基づいて根拠となる文章をコーパスに合成・注入することで，推論過程の論理的整合性と回答根拠の確実性を保証しつつ，LLM を用いた多様な質問生成を実現する．評価実験では，Agentic RAG が Vanilla RAG を上回る精度を示し，単一検索では解決できない多段階推論の評価に本データセットが有効であることが示唆された．研究コミュニティへの貢献として，本研究で構築したデータセットおよびプロンプトは公開予定である．

1 はじめに

大規模言語モデル（LLM）と検索システムを統合した検索拡張生成（RAG） [1] は，産業界・学術界の双方で急速に実用化が進展している [2]．さらに，近年では，Agentic RAG と呼ばれる検索戦略を動的に計画・修正し，反復的な処理を通じて回答生成プロセスを最適化する手法が提案されている [3]．

Agentic RAG の構築には，タスクの解決過程を明示した行動軌跡データ（trajectory）が不可欠な要素であるが，有益な trajectory を得るためには，対象となるタスク自体が，マルチホップ QA（MHQA）等の単一の検索では解決できない多段階推論を要するものである必要がある．したがって，Agentic RAG 研究において，高品質な MHQA データセットの整備が基盤的な課題となる．

近年，MHQA のデータセット構築は，コストやスケーラビリティの観点から，自動合成による研究が注目されている．Web 検索を合成過程に含む自動合成手法が提案されている [4, 5] が，LLM の幻覚による誤情報の混入や，中間推論時の事実性評価が LLM-as-a-Judge [6] に依存し，事実性を厳密に保証することが困難であるなどの課題が残る．また既存の知識グラフ（KG）を用いた合成手法は，推論の厳密性を担保できる反面，テンプレート生成による質問の多様性不足や，KG 上の知識が文書コーパスに存在しない情報の非対称性に起因する回答不能問題が生じるなどの課題が挙げられる [7, 8]．

すなわち，現在のデータセット構築手法において，「スケーラビリティ」と「回答根拠の確実性」の両立が困難であり，これが信頼性の高い Agentic RAG 構築のボトルネックとなっている．

そこで本研究では，Agentic RAG のための高品質な MHQA データセットを自動合成する手法を提案する．本手法は，KG 上のトリプルの連鎖を推論パスとして定義することで論理的整合性を保証するとともに，KG の知識に基づいて中間推論の根拠となる文章を LLM で合成・注入することで，文書コーパス内における回答根拠の確実性を保証する．

2 関連研究

既存の代表的な MHQA データセットとして，HotpotQA [9] や MuSiQue [10]，2WikiMultiHopQA [7]，StrategyQA [11] のほか，日本語資源では JEMHopQA [12] が挙げられる．いずれのデータセットも多段推論の評価に有用な設計を備える一方，人手構築のコストやスケーラビリティの観点で課題がある．特に日本語資源は JEMHopQA が提案されているものの，その規模は限定的であり，日本語 Agentic RAG 研究における言語資源の拡充は重要な研究課題である．

近年では，MHQA データセットを自動合成する手法が提案されている [4, 7, 5, 13]．利用する知識源に着目すると，大きく Web コーパス等の非構造化データを対象とする手法と，KG 等の構造化データを対象とする手法に大別できる．前者の非構造化

データを対象とする手法は，スケーラビリティの観点から特に Web 検索と組み合わせた研究が活発である．WebShaper [4] や WebDancer [5] は，LLM を用いたエージェントシステムが能動的に情報探索を行いながら，複数文書にまたがる情報を統合・再構成することで多段推論が必要な複雑質問を段階的に生成する手法である．スケーラビリティに優れる一方，LLM による要約・変形の過程で誤情報や幻覚が混入しうること，品質評価が LLM-as-a-Judge [6] に依存することなどが課題である．

後者の構造化データを対象とする手法は，KG 等を利用することで推論経路の厳密な保証を志向する点に特徴がある．代表的な先行研究として，複雑推論の評価を目的とした KB 上の逐次推論を明示的に記述可能な KQA Pro [13] や，Wikipedia と Wikidata [14] を接続した根拠付き MHQA である 2WikiMultiHopQA [7] が提案されている．しかし，KQA Pro は生成過程がテンプレートに依存するため質問の多様性が不十分である課題がある．また 2WikiMultiHopQA は，Wikidata から取得したトリプルに対応する情報が Wikipedia 文章側に存在しない場合があり，Wikipedia を知識源とする場合，回答不能な質問が作成される課題が指摘されている [7, 8]．

本研究は，(i) テンプレート依存に起因する質問多様性の制約を，LLM を用いた自然文生成により緩和し，(ii) 推論を KG 上のパスとして定義し根拠文を合成することで，根拠文の欠落や品質管理の困難さを軽減し，さらに (iii) KG 更新および他ドメインへの適用可能性により柔軟な拡張性を備える．多段推論の正当性を保証しつつスケーラブルに MHQA を自動合成可能な手法を確立し，高品質な言語資源を構築することを目的とする．

3 提案手法

本研究で提案する，KG を基盤とした Agentic RAG 向けの MHQA データセットの自動合成手法は，KG 構築，推論パス抽出，根拠文書の合成，質問文の合成，自動評価による品質フィルタリングの 5 段階で構成される．全ての段階において，合成モデルは Qwen3-32B [1] を使用した．

3.1 KG を基盤とするデータソースの構築

本研究では，Wikidata [14] を基に一般的な知識に関する KG $\mathcal{G} = (E, R, \mathcal{T})$ を構築する．ここで，E はエンティティ e の集合，R はリレーションの集合である．$\mathcal{T} \subseteq E \times R \times E$ はトリプル $\tau = (h, r, t)$ の集合であり，$h, t \in E$ はそれぞれ主語および目的語エンティティ，$r \in R$ はそれらの間の関係を表す．Wikidata はこのトリプル構造によってデータが表現され，各 e は一意の識別子（QID）で管理されており，QID は Wikipedia 記事と紐付いているため，Agentic RAG における検索対象の文書集合として取得可能である．

本研究では，Wikidata ダンプ [2] に対し，日本語ラベルを持つトリプルの抽出，管理タグ等のノイズ除去，および推論パスの線形性を保つための 1 対多のリレーション（非決定的な遷移）の排除という 3 段階の前処理を実施し，最終的に約 916 万件のトリプルを持つ KG を作成した。

3.2 KG からの推論パスの抽出

ランダムウォークによる推論パス抽出では，入次数が極端に大きいハブエンティティ（例：日本）への収束により特異的な情報が得られない問題が指摘されている [15]．本研究では，自明なエンティティを排除しつつ推論パスの多様性を確保するために，情報理論の逆出現頻度を応用した指標を導入する．具体的には，リレーション特異性を $I_R(r) = \log(|\mathcal{T}|/(count(r) + 1))$（$|\mathcal{T}|$ は全トリプル数），エンティティ特異性を $I_E(t) = \log(|E|/(in\text{-}degree(t) + 1))$（$|E|$ は全エンティティ数）と定義し，情報の希少性を評価する．次ホップの選択に際しては，これらを用いた総合スコア $S(h, r, t) = \alpha I_R(r) + \beta I_E(t)$ を算出する（本実験では $\alpha = \beta = 1.0$）．スコアが高い上位 K 個の候補から，確率 $P(t_j) = \exp(S_j)/\sum_{k \in \text{Top-}K} \exp(S_k)$ に基づき次ホップをサンプリングする．起点 e_0 からこの操作を N 回繰り返し，推論パス $P = \langle e_0, r_1, e_1, ..., r_N, e_N \rangle$ を生成する．

推論パスの生成プロセス　まず，知識グラフ上のエンティティ集合から，起点となるエンティティ $e_0 \in E$ をランダムに選択する．続いて，現在のエンティティ e_{i-1} を始点とするトリプル候補に対し，前述したサンプリング手法により，次のリレーション r_i およびエンティティ e_i を決定する．

この操作を $i = 1$ から N まで逐次的に繰り返すことで，N 個のトリプルから構成される連鎖的なリストが得られる．最終的に，これらを連結したシーケ

1） https://huggingface.co/Qwen/Qwen3-32B

2） 2025 年 9 月 18 日時点のスナップショットを使用

ンス $P = \langle e_0, r_1, e_1, \ldots, r_N, e_N \rangle$ を，本手法によって抽出された推論パスとする．なお，探索の過程で候補となるトリプルが存在しない場合は，そのパスを破棄し，再度 e_0 の選択から生成プロセスを再試行する．

3.3 根拠文書の合成

KB の事実が既存コーパスに存在しない情報の非対称性 [7, 8] を解消するため，抽出した全てのトリプルに対し，以下の手順で KG 上の知識を注入した文書 d_{synth} を合成する：(1) Wikipedia API から h に対応する記事 d_{base} を取得し，(2) LLM を用いてトリプルを自然言語の短文 s に変換，(3) d_{base} の文脈を維持しつつ s を挿入・結合する．これにより，推論パスの全ステップに対し，検索可能な言語的根拠がコーパス内に存在することを保証する．

3.4 LLM による質問文の合成

構築した推論パスに含まれる知識を網羅的に問う質問文を合成する．各推論パスを合成モデルに入力し，自然言語の質問文 q を合成する．推論過程のリークを防ぐために，推論の起点となるエンティティ e_0 を除く中間の目的語エンティティが q 内に出現することを禁止する指示を与える．

3.5 自動評価による品質フィルタリング

合成データの品質と回答の確実性を担保するために LLM-as-a-Judge による 2 段階の品質フィルタリングを実施した．

推論パスの品質フィルタ 3.2 節の手順で得られた推論パスは，MHQA に適さないものが含まれる．合成モデルを用いて，主語 h と目的語 t が同義，または t が h の上位概念である場合や，外部知識の検索を介さず，一般常識のみで t を容易に導出できる場合などに該当するトリプルを含むパスを除外した．

質問の回答可能性によるフィルタ q が根拠文書に基づいて回答可能であるか検証するために，q と推論パス上の全トリプルで合成された文書集合 $D_{synth} = \{d^{(1)}_{synth}, \ldots, d^{(N)}_{synth}\}$ をコンテキストとして，合成モデルで回答を生成した．これを gpt-oss-120b [3] で構成した評価モデルで正解エンティティと意味的に合致しているかを判定し，正解できた事例のみを採用した．これにより，推論パスの論理的整合性と合成文書が回答の根拠として機能することを保証する．

3.6 データセットの統計

多様性を確保するため，e_0 を「人物・場所・施設・組織・作品」の 5 カテゴリに設定し，それぞれ合成を行った．最終的に構築された MHQA データセットの内訳を表 1 に示す．

表 1 構築されたデータセットのカテゴリおよびホップ数別の統計

カテゴリ	2-hop	3-hop	4-hop	5-hop	合計
人物	2,346	1,114	572	279	4,311
場所	1,497	532	198	69	2,296
施設	4,268	2,426	1,553	938	9,185
組織	1,922	910	417	155	3,404
作品	3,922	1,843	875	367	7,007
合計	**13,955**	**6,825**	**3,615**	**1,808**	**26,203**

4 実験

提案手法によって合成された MHQA データセットの品質および難易度を検証するため，Agentic RAG の代表的な手法である ReAct [16] による評価実験を行った．

4.1 実験設定

評価データセット 3.6 節で構築したデータセットに対して，著者 3 名による目視確認を実施し，合計 995 件の評価用サブセットを構築した．具体的には，合成された質問文および推論パスの品質が良好であると判断された事例を，カテゴリおよびホップ数の組み合わせごとに約 50 件ずつ選定した．各質問の正解は，推論パスの最終ホップにおける t およびその Wikidata 上の別名（alias）とする．

検索手法 検索インデックスの構築にあたっては，回答に必要な根拠の存在を保証するため，日本語 Wikipedia[4] の全記事の中から推論パスに含まれるエンティティとタイトルが完全一致する記事を特定し，それらを d_{synth} に置換したコーパスを作成した．kuromoji[5] による形態素解析を適用したインデックスを作成し，検索アルゴリズムは BM25 [17] を採用した．

推論手法 構築したデータセットが多段階推論や外部知識検索を要する難易度となっているかを確認するために，以下の 3 つの設定で比較する．

3) https://huggingface.co/openai/gpt-oss-120b

4) 2024 年 10 月 23 日時点のスナップショットを使用

5) https://www.atilika.org/

1. **Closed-book QA**：外部知識検索を行わず，LLM の内部知識のみで回答を生成する．
2. **Vanilla RAG**：質問に基づいて一度だけ検索を行い，取得した情報を参照して回答を生成する．
3. **Agentic RAG**：ReAct [16] に基づく逐次的な検索と推論を行う（検索の上限回数は 10 回に設定）．

評価対象モデルは，オープンウェイトモデルの Qwen3-32B および API 提供モデルの GPT-5.2 [6] を用いた．

評価指標　意味的な正当性を評価するため，3.5 節同様に，gpt-oss-120b を用いた LLM-as-a-Judge による正誤判定を行う．全テストデータに対する正解と判定された割合を Accuracy として算出し，性能を比較する．

4.2　実験結果

各推論手法における Accuracy を表 2 に示す．実験の結果，Closed-book QA および Vanilla RAG と比較して，Agentic RAG が最も高い Accuracy を示した．Closed-book QA のスコアが著しく低いことから，本データセットはモデルの内部知識のみでの回答が困難であり，外部知識の参照が重要となる傾向を示している．また，Vanilla RAG と比較して Agentic RAG が高い精度を示したことは，本データセットが単一の検索では回答根拠を網羅できず，多段階の推論と検索が有効である可能性を示唆する．

表 2　モデルと手法別のカテゴリ Accuracy の比較

モデル	手法	人物	場所	施設	組織	作品	Avg.
Qwen3-32B	Closed-book	10.0	12.3	13.0	14.5	14.0	12.8
	Vanilla RAG	19.0	18.5	27.5	19.0	21.5	21.1
	Agentic RAG	**42.5**	42.1	45.5	37.5	33.0	40.1
GPT-5.2	Closed-book	16.5	30.3	40.0	30.5	29.5	29.3
	Vanilla RAG	30.5	30.3	37.5	31.5	36.0	33.2
	Agentic RAG	40.5	**52.8**	**58.0**	**50.0**	**44.0**	**49.0**

次に，ホップ数ごとの正解率の推移を図 1 に示す．全体的な傾向として，ホップ数が増加するにつれて Accuracy が低下することが確認された．特に 2-hop と比較して 4・5-hop での精度低下が見られることから，推論ステップ数の増加がタスクの難易度に寄与しており，多段階推論の要素が適切に機能していることが推察される．また GPT-5.2 は Qwen3-32B と比較して，ホップ数の増加に伴う性能の落ち込みが緩やかである傾向が確認され，同モデルの多段階推論への高い頑健性が示唆された．

6）https://platform.openai.com/docs/models/gpt-5.2

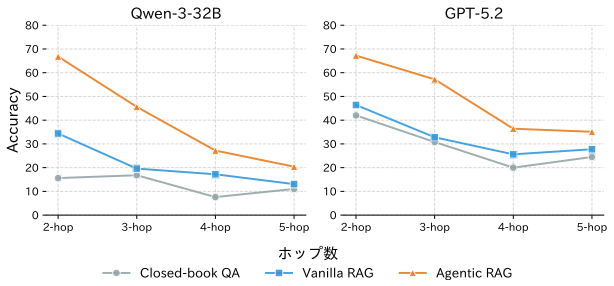

図 1　各推論手法におけるホップ数ごとの Accuracy

4.3　分析

本データセットは，Wikidata のトリプルに基づいて正解が一意に定まるよう設計されているものの，実世界の情報を網羅的に集積したものではないため，実世界では正解が複数存在し，定義の曖昧性が残る場合がある（例：ある映画で，Wikidata には一部の出演者しか登録されていないが，実世界では未登録の俳優も多数出演している）．この時，モデルが推論を行った結果，データセット上の想定解とは異なるが実世界では妥当な回答を導き出し，自動評価で不正解と判定される事例が確認された（付録図 2）．これは，KG の網羅性が現実世界の複雑性に追いつかない場合に生じる正解の非一意性の問題であり，Agentic RAG のような高度な検証能力を持つシステムを評価する際の，閉世界仮定に基づく合成データセットの限界を示唆している．

また，検索の結果，適切な外部知識が得られなかった際に，パラメトリック知識に基づいて誤った推論を行う事例が確認された（付録図 3）．これは，推論のターン数が増加しコンテキストが長くなった状況下で，モデルが検索結果の不足を自身の知識や論理の飛躍で補完しようとする傾向を示している．

5　おわりに

本研究では，知識グラフに基づく推論パスと LLM による文書合成を組み合わせ，論理的整合性と根拠の確実性を担保した日本語マルチホップ QA データセットの自動合成手法を提案した．実験により，本データセットが Agentic RAG の多段階推論の評価に有効であることを確認した．今後は，Wikidata 等の静的な KG における知識網羅性の課題を解決するとともに，構造化データが整備されていない領域へも適用範囲を広げるため，任意の非構造化データから動的に知識構造を抽出 [18] し，データセットを合成する手法への拡張を目指す．

参考文献

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[2] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In **Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining**, KDD '24, p. 6491–6501, New York, NY, USA, 2024. Association for Computing Machinery.

[3] Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. Agentic retrieval-augmented generation: A survey on agentic rag, 2025.

[4] Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webshaper: Agentically data synthesizing via information-seeking formalization, 2025.

[5] Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhenglin Wang, Zhengwei Tao, Ding-Chu Zhang, Zekun Xi, Xiangru Tang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webdancer: Towards autonomous information seeking agency. In **The Thirty-ninth Annual Conference on Neural Information Processing Systems**, 2025.

[6] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025.

[7] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[8] Anonymous. KBQA or LLM-QA: A unified benchmark for question answering. In **Submitted to ACL Rolling Review - February 2025**, 2025. under review.

[9] 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.

[10] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 539–554, 2022.

[11] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 346–361, 2021.

[12] Ai Ishii, Naoya Inoue, Hisami Suzuki, and Satoshi Sekine. JEMHopQA: Dataset for Japanese explainable multi-hop question answering. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 9515–9525, Torino, Italia, May 2024. ELRA and ICCL.

[13] Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base. 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. 6101–6119, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[14] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledge base. **Communications of the ACM**, Vol. 57, pp. 78–85, 2014.

[15] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. **PLOS ONE**, Vol. 10, No. 6, pp. 1–13, 06 2015.

[16] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In **The Eleventh International Conference on Learning Representations**, 2023.

[17] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. **Found. Trends Inf. Retr.**, Vol. 3, No. 4, p. 333–389, April 2009.

[18] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A survey on knowledge graphs: Representation, acquisition, and applications. **IEEE Transactions on Neural Networks and Learning Systems**, Vol. 33, No. 2, p. 494–514, February 2022.

A　付録

A.1　エラー分析

図 2 に，知識グラフの不完全性と実世界知識が競合したことで生じたエラーの事例を示す．以下の例は，知識源である Wikidata には「ふなっしー」は立項されているものの，「目利き番頭 船えもん」は立項されいない一方，船橋市に関する文書内には「目利き番頭 船えもん」についての言及が存在する．本事例では，Agentic RAG が外部知識を適切に読み取り，現実世界における事実上の正解である「目利き番頭 船えもん」を導き出したが，Wikidata の知識のみに基づいたデータセットの想定解とは異なるため，自動評価では不正解と判定されている．

事例 1: 知識グラフの不完全性と実世界知識の競合

質問: 中村計の出生地のマスコットキャラクターは何ですか？
正解: ふなっしー
Agentic RAG の trajectory (抜粋):
<search>船橋市 公式 マスコットキャラクター</search>
<information>ふなっしーは... 市や県からは公認されていない...</information>
<thinking>検索結果のふなっしーは非公認とあるため不適。船橋市公式のマスコット名を追加検索...</thinking>
...
<thinking>船橋市の公式マスコットキャラクターは「目利き番頭 船えもん」。よって回答はこれ。</thinking>
予測回答: 公式マスコットキャラクターは「目利き番頭 船えもん」です。

図 2　知識グラフの想定解とモデルの実世界推論が競合した事例

図 3 に，検索で十分な根拠が得られなかった際に，生成モデルがパラメトリック知識に依存して，誤った知識補完をした事例を示す．以下の例は，「常山紀談」のまえがき執筆者（松崎観海）やその勤務先（篠山藩）までは正確に特定できたものの，最終的な回答に必要な「藩の創設者の父」に関する情報が検索結果に含まれておらず，本来再検索するべき状況で，モデルは欠落した情報を補完し，事実とは異なる結論を導き出し不正解となった．

事例 2: パラメトリック知識への依存による誤謬

質問: 「常山紀談」のまえがきを書いた人物が勤務していた藩の創設者の父は誰ですか？
正解: 松平康親
Agentic RAG の trajectory (抜粋):
<search>「常山紀談」 まえがき 書いた人</search>
<information> 『常山紀談』は... 本書のまえがきは松崎観海が執筆を...</information>
<thinking>「常山紀談」のまえがきを書いた人物は松崎観海である。次に、松崎観海が勤務していた藩を特定する必要がある。</thinking>
<search>松崎観海 藩</search>
<information>父が仕えられていた篠山藩付の松平家の家老を勇退した事により、観海を後継者として継がせた。</information>
<thinking>松崎観海は篠山藩に仕えていたことが確認できます。次に、篠山藩の創設者を特定する必要があります。</thinking>
<search>篠山藩 創設者</search>
<information>（根拠の記述が存在しない）</information>
<thinking>篠山藩の創設者は松井松平康重である。松井松平康重は、江戸時代初期の武将・大名。常陸笠間藩主を経て、1609 年より丹波篠山藩初代藩主を務めた。康重は、松井康政の子として生まれた。松井松平康重の父は松井康政である。</thinking>
予測回答: 松井康政

図 3　検索で十分な証拠が得られなかった際に，誤った推論により結論を出した事例