# Wikipedia リダイレクト情報を活用したエンティティベース質問応答データセットの構築

西田 悠人 [1,2] 志子田 直輝 [1] 岸波 洋介 [2] 藤井 諒 [2] 森下 睦 [2] 上垣外 英剛 [1] 渡辺 太郎 [1]
[1] 奈良先端科学技術大学院大学 [2] フューチャー株式会社
{nishida.yuto.nu8, shikoda.naoki.sm1, kamigaito.h, taro.watanabe}@naist.ac.jp
{y.kishinami.rh, r.fujii.6d, m.morishita.pi}@future.co.jp

## 概要

大規模言語モデル (LLM) がどのような知識を記憶しているかを調べるために，しばしばエンティティに基づく質問応答データセットが利用される．しかし，そのような既存のデータセットはエンティティの単一の表層のみに依存しており，LLM がエンティティを記憶しているのか，特定の表層を記憶しているのかを弁別できない．そのため，本研究では Wikipedia のリダイレクト情報を活用し，複数の表層を考慮可能なエンティティベースの質問応答データセット RedirectQA を構築する．本データセットはエンティティの複数の表層と，それらに対応するカテゴリが付与されており，LLM がどのような表層を記憶しているかの調査に適している．

## 1 はじめに

大規模言語モデル (LLM) はその内部に多種多様な知識を蓄えており [1, 2]，自然言語生成に関する広範な応用が期待されている．しかし，LLM はその記憶に存在しない知識について問われるとハルシネーションや性能低下を引き起こすことがある [3]．そのため，LLM がどのような知識を記憶しているかを測ることは，LLM の適正な利用や性能の改善にとって重要である．

これまでに，訓練データ中における出現頻度の低い知識は LLM に記憶されにくいこと [4, 5] や有名ではない知識は記憶されにくいこと [5] などが明らかとなっている．これらの研究では，主にエンティティに関する事実知識を対象として，その記憶を測るためにエンティティベースの（質問と回答にそれぞれエンティティを含む）質問応答データセット [6, 7, 5] が用いられてきており，質問に対する正答率に基づいて記憶が測られてきた．しかし，既存のエンティティベースのデータセットは，エンティティの単一の表層のみに依存している．エンティティはしばしば複数の表層を有するため，このようなデータセットによる評価では LLM がエンティティを記憶しているのか，特定の表層を記憶しているのかを弁別できない．実際に，3.2 節で詳述する質問応答タスクを対象とした事前実験では，表 1 に示すように，同一のエンティティについての質問であっても，代表的な表層に基づく質問とそうでない質問で LLM の生成結果の正答/誤答が異なるケースが 2 割程度存在した．このことは，LLM の記憶がエンティティではなく表層に紐づいていることを示唆するものであり，LLM の知識の記憶を深く理解するためには複数の表層を考慮した評価が必要である．

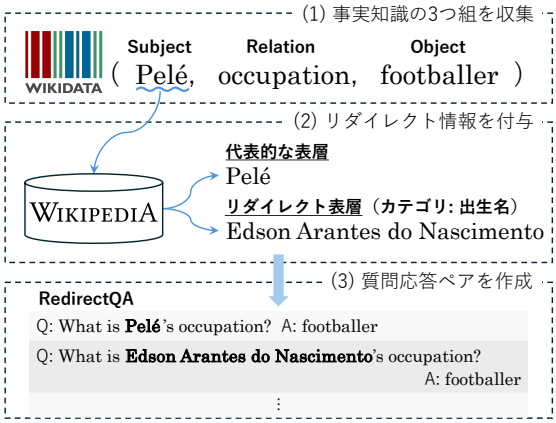


図 1: RedirectQA の構築

上記を踏まえ，本稿では，Wikipedia のリダイレクト情報を活用し，1 つのエンティティに対して複数の表層を考慮できるエンティティベースの質問応答データセット RedirectQA[1)] を構築する（図 1）．RedirectQA には，あるエンティティに対して複数の表層に基づく質問が含まれ，それぞれの表層の性質

1) https://huggingface.co/datasets/naist-nlp/RedirectQA

表 1: PopQA のリダイレクト拡張に対する評価結果

| | | 代表的な表層 | |
|---|---|---|---|
| | | 正答 | 誤答 |
| リダイレクト表層 | 正答 | 2,691 | 341 |
| | 誤答 | 1,269 | 3,628 |

に関するカテゴリが付与されている．本データセットを活用することで，LLM が多様な表層を汎化してエンティティの記憶を獲得しているか，あるいはどのような表層を記憶しやすいかについての調査が可能となる．実際に，RedirectQA を活用して LLM の表層の記憶について調査し，LLM は表記揺れのような表層の差異が小さいケースについては記憶を紐付けやすいことが示唆された．

## 2 RedirectQA

RedirectQA は，エンティティの複数の表層を用いたオープンドメイン [8] の質問応答データセットである．対象とする知識は Wikidata から収集した (subject, relationship, object) の 3 つ組からなる事実知識であり，複数の表層を考慮するために Wikipedia のリダイレクト情報を活用した．

### 2.1 Wikipedia リダイレクト

ある事柄について Wikipedia 記事が作成される際には，記事のタイトルはガイドライン[2]に基づいて，その事柄を表現する様々な表層のなかから認知度の高さや検索性などの観点により 1 つに決定される．このとき，異なる表記で検索された場合でも当該記事に辿り着けるように，記事タイトルとは異なる表記を用いたリダイレクト（転送）ページが作られることがある．たとえば，略称である “APEC” というページにアクセスした際には，正式名称である “Asia-Pacific Economic Cooperation” のページに自動的に転送される．

リダイレクトページにはそのページの表層と元の記事タイトルの表層の関係を分類するリダイレクトカテゴリが付与されている[3]．たとえば “APEC” のページには “Redirects from acronyms” というリダイレクトカテゴリ[4]が付与されており，“APEC” が “Asia-Pacific Economic Cooperation” の acronym（頭字語の一種）であることがわかる．

Wikipedia リダイレクトは，典型的には元の記事タイトルと異なる表層の情報として利用できるが，一部に元の表層の言い換えではないリダイレクトも存在することに注意が必要である．たとえば，“from books” というカテゴリでは，ある本のタイトルがその著者の記事にリンクされていることがある．

そのため，我々は頻度の高いリダイレクトカテゴリのなかから，元の表層と異なる表層として利用できるカテゴリを人手で 33 件選択し，これらのカテゴリに属するリダイレクト情報のみを利用した．これらのカテゴリは，その性質から 3 つの類型に大別できる．1 つ目は，**別名・略称**であり，上述の “APEC” などが該当する．2 つ目は，**表記揺れ**であり，別のスペルなどが該当する．3 つ目は，**典型的な誤り**であり，スペルミスなどが該当する．選択したカテゴリとその類型は付録 A の表 4 に示した．

### 2.2 複数の表層を用いた質問応答ペア

RedirectQA は，1 つの事実知識について，subject エンティティの表層が相異なる複数の質問応答ペアをもつ．ここで，複数の表層は，1 つの代表的な表層（当該エンティティに紐づく Wikipedia 記事のタイトル）と 1 つ以上のリダイレクト表層（当該 Wikipedia ページに張られたリダイレクトページのタイトル）からなり，リダイレクト表層にはリダイレクトカテゴリが付与されている．

RedirectQA を用いることで，言語モデルの事実知識に関する記憶を，表層の違いやその種類の観点から分析することが可能となる．

### 2.3 データセットの構築

本稿では，10,401 件の事実知識の 3 つ組を対象として，22,869 個の質問応答ペアを作成した．収集したリダイレクト表層のうち，別名・略称の類型に属するものは 4,731 件，表記揺れの類型に属するものは 7,104 件，典型的な誤りの類型に属するものは 705 件であった[5]．本データセットは，図 1 に示すように構築した．以下に詳述する．

**3 つ組の収集**　エンティティベースの質問応答データセットの 1 つである PopQA [7] の設定に倣い，15 種類の relation[6]のみを対象として，Wikidata

2） https://en.wikipedia.org/wiki/Wikipedia:Article_titles
3） 付与されるカテゴリは 0 個や 2 個以上のことがある．
4） 以降，カテゴリ名の “Redirects” を省略する．
5） これらの類型は必ずしも互いに排反ではなく，2 つ以上のリダイレクトカテゴリをもつリダイレクト表層が複数の類型に属する場合があるため，それぞれの件数の和はリダイレクト表層の総数（12,468 件）と一致しない．
6） PopQA で対象となっている 16 件の relation から，religion を除外した 15 件を対象とした．

表 2: RedirectQA の評価結果．代表的な表層に対する正誤を所与とした場合のリダイレクト表層に対する正誤の割合を括弧内に付記した．

(a) 全体

| Redirect 表層 | 代表的な表層 正答 | 誤答 |
|---|---|---|
| 正答 | 3,433 (79.8%) | 471 ( 5.8%) |
| 誤答 | 870 (20.2%) | 7,694 (94.2%) |

(b) 別名・略称

| Redirect 表層 | 代表的な表層 正答 | 誤答 |
|---|---|---|
| 正答 | 766 (68.6%) | 215 ( 5.9%) |
| 誤答 | 351 (31.4%) | 3,399 (94.1%) |

(c) 表記揺れ

| Redirect 表層 | 代表的な表層 正答 | 誤答 |
|---|---|---|
| 正答 | 2,550 (84.2%) | 235 ( 5.8%) |
| 誤答 | 478 (15.8%) | 3,841 (94.2%) |

(d) 典型的な誤り

| Redirect 表層 | 代表的な表層 正答 | 誤答 |
|---|---|---|
| 正答 | 129 (71.3%) | 21 ( 4.0%) |
| 誤答 | 52 (28.7%) | 503 (96.0%) |

(e) Redirects from short names

| Redirect 表層 | 代表的な表層 正答 | 誤答 |
|---|---|---|
| 正答 | 130 (87.2%) | 21 ( 7.4%) |
| 誤答 | 19 (12.8%) | 262 (92.6%) |

(f) Redirects from long names

| Redirect 表層 | 代表的な表層 正答 | 誤答 |
|---|---|---|
| 正答 | 124 (63.3%) | 41 ( 4.7%) |
| 誤答 | 72 (36.7%) | 827 (95.3%) |

のダンプデータから事実知識の 3 つ組をランダムにサンプリングした．図 1 の (1) に対応する．

**リダイレクト情報の付与**　Wikipedia のダンプデータを用いて，それぞれの subject エンティティにリダイレクト表層とそのカテゴリを付与した．なお，これらの情報が存在しない 3 つ組はこの段階で除外した．図 1 の (2) に対応する．

**質問応答ペアの作成**　PopQA と同一のテンプレートをもとに各表層における質問文を生成し，質問応答ペアを作成した．図 1 の (3) に対応する．

# 3　実験

## 3.1　実験設定

LLM として Pythia [9] の 12B モデルを利用し，推論時には 8bit 量子化を施した．PopQA と同一のプロンプトテンプレート “`Q: <question> A:`” を用いて，14-shot の設定で推論を行った．生成結果の評価は完全一致によって行った．

## 3.2　事前実験

事前実験として，PopQA に対して 2.3 節と同様の手順でリダイレクト情報を付与し，7,929 件の質問応答ペアを作成した．

表 1 に，評価結果を示す．代表的な表層とそうでない表層で LLM の正誤が異なるケースが全体の 20.3%を占めた．

## 3.3　RedirectQA を活用した分析

表 2a に RedirectQA 全体に対する評価結果を示す．LLM が代表的な表層を記憶しているとき，多くの場合はリダイレクト表層も記憶しているが，一部では代表的な表層のみを記憶しているケースがあることが示された．表 3 の事例 1 では，代表的な表層（本名）の “Theodore Hook” では正答できるが，リダイレクト表層（ペンネームの 1 つ）の “Alfred Allendale” では誤答している．また，LLM が代表的な表層を記憶していないときに，リダイレクト表層のみを記憶しているケースについても少ないながらも存在することが示された．エンティティの代表的な表層はその認知度の高さなどから選ばれていることを踏まえると，このことは人間にとっての認知度の高さと LLM の記憶しやすさは必ずしも一致しないことを示唆している．表 3 の事例 2 において，“George Clanton” と “Mirror Kisses” はどちらも同一人物の（歌手としての）活動名である．前者の活動名におけるディスコグラフィには，後者に比べて多くの独立した Wikipedia 記事が存在し，認知度も高いと思われるが，LLM は後者の表層に関連する知識のみを記憶していた．

表 2b, 2c, 2d にリダイレクトカテゴリの類型ごとの評価結果を示す．表記揺れの類型は，代表的な表層に正答した場合のリダイレクト表層での正答率 (84.2%) が別名・略称の類型の場合 (68.6%) と比較して高いことがわかる．このことから，LLM の記憶は表記揺れに関して頑健であることが示唆される．別名・略称と表記揺れの事例をそれぞれ表 3 の事例 3,4 に示す．これらの事例に代表されるように，表記揺れの類型は別名・略称の類型に比べて代表的な表層とリダイレクト表層の差異が小さい傾向にあると考えられ，別名や略称のような表層が異なる度合いの大きいものについては記憶が紐づきにくいということが示唆された．また，典型的な誤りの類型の事例を表 3 の事例 5 に示す．この事例に代表されるように，典型的な誤りの類型に該当するリダイレクト表層は（表記揺れの類型と同様に）代表的な表層

表 3: RedirectQA に対する評価結果の事例分析

| | カテゴリ | 類型 | 質問 | LLM の回答 | 正誤 |
|---|---|---|---|---|---|
| 事例 1 | （代表的な表層） | - | *In what city was Theodore Hook born?* | *London* | ○ |
| | from pseudonyms | 別名・略称 | *In what city was Alfred Allendale born?* | *New York* | × |
| 事例 2 | （代表的な表層） | - | *What is George Clanton's occupation?* | *boxer* | × |
| | from alternative names | 別名・略称 | *What is Mirror Kisses's occupation?* | *singer* | ○ |
| 事例 3 | （代表的な表層） | - | *What genre is Neverwinter Nights 2: Mysteries of Westgate?* | *RPG* | ○ |
| | from abbreviations | 別名・略称 | *What genre is NWN2:MOW?* | *action* | × |
| 事例 4 | （代表的な表層） | - | *In what country is Beesley's Point Bridge?* | *Australia* | ○ |
| | from modifications | 表記揺れ | *In what country is Beesleys Point Bridge?* | *Australia* | ○ |
| 事例 5 | （代表的な表層） | - | *What sport does Carl Pavano play?* | *baseball* | ○ |
| | from misspellings | 典型的な誤り | *What sport does Karl Pavano play?* | *ice hockey* | × |
| 事例 6 | （代表的な表層） | - | *Who is the father of Howard Graham Buffett?* | *Warren Buffett* | ○ |
| | from short names | 別名・略称 | *Who is the father of Howard G. Buffett?* | *Howard Buffett* | × |
| 事例 7 | （代表的な表層） | - | *What is Kevin Kiermaier's occupation?* | *baseball player* | ○ |
| | from long names | 別名・略称 | *What is Kevin James Kiermaier's occupation?* | *baseball player* | ○ |

に対して軽微な差異しかないことが多いと考えられる．しかし，この類型における代表的な表層に正答した場合のリダイレクト表層での正答率 (71.3%) は表記揺れの場合 (84.2%) と比べて低く，典型的な誤りはあまり記憶していないということが伺える．

別名・略称の類型に対するさらなる分析として，“from short names” と “from long names” という 2 つのカテゴリに着目する．“from short names” は表 3 の事例 6 のように一部の省略などによって代表的な表層よりも短い表層をもつカテゴリであり，反対に “from long names” は同表の事例 7 のように代表的な表層よりも長い表層をもつカテゴリである．これらのカテゴリについての評価結果を表 2e, 2f に示す．結果から前者のカテゴリは後者のカテゴリに比べて代表的な表層に正答した場合のリダイレクト表層での正答率が低いことがわかる．このように，LLM は長い表層（典型的にはより多くの情報をもつ表層）をより記憶しやすいことが示唆された．

## 4 関連研究

LLM の記憶についての分析は，その対象によって 2 種類に大別される [4]．一方は，逐語的な（丸覚えの）記憶を対象とするもの [10, 11, 12] であり，主にプライバシーの観点から LLM による情報漏洩の防止などについて議論されてきた．もう一方は，非逐語的な（汎化した）記憶を対象にするもの [4, 7, 5] であり，主にエンティティベースの質問応答データセットを用いて，どのような特徴をもつ事実知識がより記憶されやすいかなどが議論されてきた．本研究は後者の分析のためのデータセット作成に位置付けられる．

LLM の非逐語的な記憶はいくつかの観点から分析されてきた．Kandpal ら [4] は，オープンドメイン質問応答データセットの NaturalQuestions [13] および TriviaQA [14] から質問と回答にそれぞれエンティティを含むペアを抽出した評価データを用いて，LLM の訓練データ中における出現頻度の低い事実知識は記憶されづらいことを示した．Mallen ら [7] は，エンティティベースの質問応答データセットの PopQA を構築し，同じくエンティティベースの EntityQuestions [6] も用いた評価によって，LLM は有名度（Wikipedia 記事の閲覧数）の低いエンティティを記憶しにくいことを示した．Maekawa ら [5] は，エンティティベースの質問応答データセットの WitQA を構築し，relation の頻度も事実知識の記憶に影響することを示した．これらの研究で用いられる質問応答データセットは，単一の表層に依存したものであり，表層の記憶の分析には適さない．

## 5 おわりに

本稿では，エンティティの表層についての LLM の記憶を測るために，複数の表層を考慮できるエンティティベースの質問応答データセット RedirectQA を構築した．また，本データセットによって，LLM のエンティティの表層の種類によって記憶のされやすさが異なることを示した．表層の記憶の要因についての更なる分析は今後の課題であり，表層の頻度が記憶に与える影響の調査などが考えられる．

## 謝辞

本研究は JST 次世代研究者挑戦的研究プログラム JPMJSP2140 の支援を受けたものです.

## 参考文献

[1] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, pp. 2463–2473, 2019.

[2] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. Generate rather than retrieve: Large language models are strong context generators. In **Proceedings of the Eleventh International Conference on Learning Representations**, 2023.

[3] Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. Distinguishing ignorance from error in LLM hallucinations. **arXiv preprint arXiv:2410.22071**, 2024.

[4] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In **Proceedings of the 40th International Conference on Machine Learning**, pp. 15696–15707, 2023.

[5] Seiji Maekawa, Hayate Iso, Sairam Gurajada, and Nikita Bhutani. Retrieval helps or hurts? a deeper dive into the efficacy of retrieval augmentation to language models. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 5506–5521, 2024.

[6] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6138–6148, 2021.

[7] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 9802–9822, 2023.

[8] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 5418–5426, 2020.

[9] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: a suite for analyzing large language models across training and scaling. In **Proceedings of the 40th International Conference on Machine Learning**, pp. 2397–2430, 2023.

[10] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In **Proceedings of the 30th USENIX Security Symposium**, pp. 2633–2650, 2021.

[11] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In **Proceedings of the Eleventh International Conference on Learning Representations**, 2023.

[12] Bowen Chen, Namgi Han, and Yusuke Miyao. A multi-perspective analysis of memorization in large language models. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 11190–11209, 2024.

[13] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 452–466, 2019.

[14] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1601–1611, 2017.

表 4: リダイレクトカテゴリ

| 類型 | リダイレクトカテゴリ |
|---|---|
| 表記揺れ | Redirects_from_titles_without_diacritics |
| | Redirects_from_modifications |
| | Redirects_from_other_capitalisations |
| | Redirects_from_alternative_spellings |
| | Redirects_from_titles_with_diacritics |
| | Redirects_from_ASCII-only_titles |
| | Redirects_from_stylizations |
| | Redirects_from_titles_without_ligatures |
| | Redirects_from_numerals |
| | Redirects_to_ASCII-only_titles |
| 別名・略称 | Redirects_from_alternative_names |
| | Redirects_from_long_names |
| | Redirects_from_surnames |
| | Redirects_from_short_names |
| | Redirects_from_abbreviations |
| | Redirects_from_former_names |
| | Redirects_from_birth_names |
| | Redirects_from_initialisms |
| | Redirects_from_given_names |
| | Redirects_from_pseudonyms |
| | Redirects_from_personal_names |
| | Redirects_from_plurals |
| | Redirects_from_married_names |
| | Redirects_from_letter–word_combinations |
| | Redirects_from_technical_names |
| | Redirects_to_plurals |
| | Redirects_from_acronyms |
| | Redirects_from_synonyms |
| | Redirects_to_initialisms |
| | Redirects_to_acronyms |
| ミス | Redirects_from_misspellings |
| | Redirects_from_miscapitalisations |
| | Redirects_from_incorrect_names |

# A リダイレクトカテゴリ

RedirectQA で利用したリダイレクトカテゴリの一覧を表 4 に示す.