WikiOriginQA: 知識の起源を組み込んだ文化的バイアス分析用 QA データセットの自動構築

羽根田 賢和 [1,2]　岸波 洋介 [1]　藤井 諒 [1]　坂口 慶祐 [2]　森下 睦 [1]
[1] フューチャー株式会社　[2] 東北大学
team-nlp-research@future.co.jp

概要

本研究では，Wikipedia の各言語版の履歴情報をもとに知識の起源を客観的に表現することで抽出した，各文化圏由来の知識に基づく QA データセットを提案し，大規模言語モデル (LLM) における文化的バイアスの検証を行った．既存の文化的知識データセットは構築に各文化圏に精通したアノテータによる検証が必要であり，自動化や多言語への拡張が困難であった．本手法は構築プロセス全体が自動化されているため，人的コストを削減しつつ高い拡張性を有する．検証実験の結果，LLM の性能が英語圏文化に偏重する傾向が確認され，本データセットの文化的バイアス検証における有効性が確認された．

1　はじめに

大規模言語モデル (LLM) が多くの言語で活用されている一方で，英語とそれ以外の言語における性能差や，西洋文化に偏重した文化的バイアスの存在が指摘されている [1, 2]．LLM の持つ知識や能力が英語圏に対応するためのものへ偏重していくことは，英語圏以外の文化における知識が自然言語処理技術から取りこぼされていく格差を引き起こしうると考えられ，重要な課題となる．

多言語 LLM における文化的バイアスの問題の把握には，さまざまな言語・文化圏を対象とした評価手法の確立が求められる．しかしながら既存の評価手法の多くは，各文化圏に精通した評価者によって，対象となる知識がどの文化圏に属するものであるかを人手で判断する必要があり [3, 4, 5]，多言語への大規模な拡張が困難であった．また，知識がどの言語圏に由来するものであるかを客観的に判定することは容易ではなく，文化と知識の関係性を十分に捉えられていないという課題も存在する．そこで本研究では，客観的に評価した知識の起源を考慮した QA データセットを構築し，LLM が有する知識の文化的偏りを調査することを試みる．

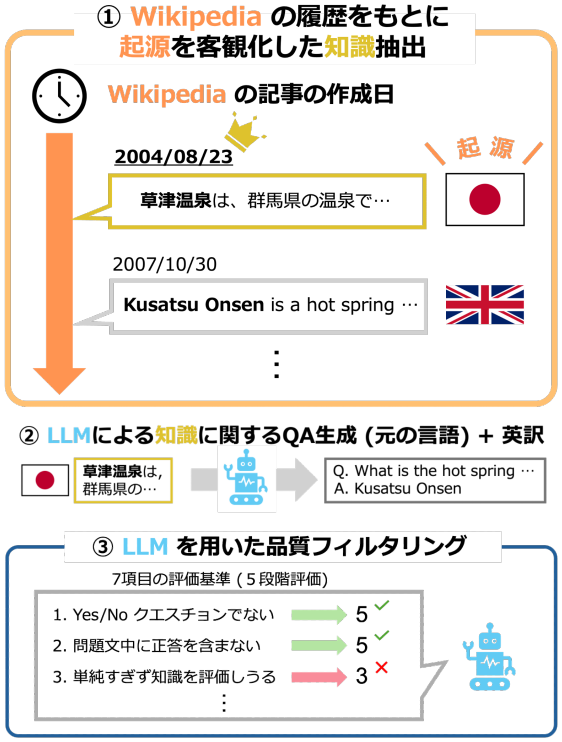

図 1　本研究における QA データ自動作成手法の概要．

本研究では，各言語に特有な知識はほかの言語に先駆けてテキスト資源が作成されるという仮説のもと，知識の起源の定量的な表現のために，Wikipedia の各言語版の作成日時を参照した．図 1 に本手法の概要を示す．例えば，「草津温泉」のように日本や日本語文化圏において一般的な知識であれば，Wikipedia の記事は他の言語に先駆けて日本語版が最も早く作成されると考えられる．このようにして各言語圏由来の知識を抽出し，これ関して述べた Wikipedia の記事を用いて QA データセット “WikiOriginQA” を構築した．本手法は，データセット構築に人手を必要とせず，様々な言語に対して容易に拡張可能であるという特徴をもつ．本データ

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

セットを用いた実験を通して，LLM は英語圏文化に偏重する傾向が改めて確認され，LLM の多文化対応における課題を明らかにした．

2 関連研究

LLM の有する知識の文化的バイアスの有無の検証を目的としたベンチマークやデータセットはこれまでいくつか提案されている．

マルチタスク評価用のベンチマークである MMLU [6] を多言語に拡張した Global MMLU [7] では，文化に依存した問題に関して，低リソース言語での性能の悪化が確認されている．また，Shen ら [8] は，LLM の有する文化的常識に関するベンチマークを構築し，非西洋的な文化に対するモデルの性能低下を示している．また，長文 QA に着目した CaLMQA [3] や，より日常的な文化知識に着目した BLEND [4]，各地域のライセンス試験をベースとした INCLUDE [5] など，文化的知識に関するデータセットが提案されており，いずれも，英語圏や西洋文化へのバイアスが報告されている．一方で，このようなデータセットを構築する際には，人的コストが大きな課題となる．CaLMQA では，低リソース言語に関する問題の収集に母語話者の協力を得ており，BLEND や INCLUDE では，収集した問題の品質の確保において，人手によるアノテーションを実施している．データセット構築に人手が介在することは，拡張性の観点から大きな制約となり，話者人口の少ない言語や文化的知識を対象とする際，その問題は顕著となる．これに対し，本研究の提案するデータセットは，人手によるアノテーションを必要とせず，拡張が容易という利点を持つ．

また，人手による判断は主観性の課題も抱えている．客観的な軸を用いて知識と文化圏の関係を評価する研究として，Jiang ら [9] は，Wikipedia のある記事が何言語に展開されているか，という数的情報をもとに，特定の言語版のみに存在する項目を文化特有の知識として扱う文化的 QA データセット KoLasSimpleQA を提案した．しかしながら，この手法はデータ全体が過度にニッチな情報に偏るおそれがある点や，時間の経過に伴い他言語版の記事が追加されることで，伝播した知識の取りこぼしが生じる点に課題がある．本手法も同じく Wikipedia を基盤とするが，記事の作成日時という時間情報を組み込むことで，起源を明確にしながら，知識の伝播に対しても頑健であるという利点を持つ．

3 WikiOriginQA

本研究では，知識の起源となる言語圏を Wikipedia の記事作成日時という客観的情報に基づいて特定し，LLM が有する知識の文化的偏りを検証するための自由回答式 QA データセットを構築する．

3.1 問題のベースとなる知識の抽出

同じ概念を表す Wikipedia の各言語版の作成日時を参照するため，Wikipedia のダンプファイルと Wikidata の項目 ID を利用した．Wikidata はトピック，概念，オブジェクトなどを表す項目に焦点を当てたデータベースであり，各項目に固有の項目 ID (以降 QID) が各言語版の Wikipedia に紐づいている．例えば，“地球” と “Earth” はそれぞれ日本語版と英語版の Wikipedia では別の記事であり，内容もそれぞれが独自に書かれている．しかし，これらが指す概念は同一であるため，それぞれの Wikipedia のページには，Wikidata における “Earth” を示す同一の QID が紐づけられている．この QID をもとにそこに紐づくすべての言語版の Wikipedia の記事を参照し作成日時を比較することで，各項目の起源となる言語圏を評価することが可能である．本研究では，Wikipedia の言語版リスト[1)]中の言語のうち，閉鎖されておらず，ダンプファイルが公開されている 342 言語を対象に編集履歴を取得し，Wikipedia 記事群の中からある項目に関して最も早く作成された言語を特定した．このようにして取得した，各言語圏に起源を持つ項目の集合を，その言語における QA データのベースとした．表 1 に示すように，全体として各言語圏に特有の知識が抽出されていることが確認できる．一方で，パシュトー語のルクセンブルクのように，既に他言語版が存在しているのにも関わらず，Wikipedia の整備状況により，独立した別項目として抽出されるケースも少量存在した．

3.2 QA データの作成

各言語の QA のベースとなる項目に対し，項目名が答えとなるよう，各項目の Wikipedia の記事をもとに LLM による問題生成を行った．生成指示は対象の言語で行い，問題の質の向上のため，作成された問題を改善するよう，複数回モデルに指示を与えた．その後，モデルの言語理解能力が回答精度に影響することを防ぐため，作成した問題を LLM を用

1) https://meta.wikimedia.org/wiki/List_of_Wikipedias

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 1　QA のベースとなる知識の具体例.

言語	項目
アフリカーンス語	Swartvlei (南アフリカの塩水湖)，Andrew Motjuoadi (南アフリカ人の画家)
日本語	守住勇魚，新党さきがけ，道氏，泉谷祐勝，コシヒカリ BL
パシュトー語	خیشکي (パシュトゥーン人の部族)，لوکزامبورګ (ルクセンブルク)，زلی هېوادمل (アフガニスタンの作家)
ウクライナ語	Пригородок (ウクライナの街)，SLC2A8 (タンパク質名)

表 2　QA の具体例.（実際の問題は英語）

問題: 徳島県徳島市（中央通）に生まれ、1878 年に浅井忠らと工部美術学校を連袂退学して「十一会」を結成し、翌 1879 年に大阪専門学校の画学教員となり、さらに 1895 年から 1917 年まで同志社の画学科を担当、加えて 1882〜1883 年に臥龍館から『大成普通画学本』全 10 冊を刊行した、日本の洋画家は誰でしょうか？
正答: ["Isana Morizumi", "守住勇魚"]

いて英訳し，回答は英語で行う形式とした.

データの品質を確保するため，図 1 のように LLM によるフィルタリングを実施した．フィルタは既存研究 [9, 10, 11] を参考に，計 7 つの基準により構成し，各基準に対してそれぞれ 5 段階評価を行った．全ての基準において最高評価の 5 点を得た問題のみを，品質を満たした問題と判定した.[2] なお，問題作成，英訳，フィルタリングには，いずれも GPT-5 を用い，reasoning-effort は low とした.

各問題の正答は，元の項目名に，Wikidata にまとめられている別称を加えたものとした．問題は英訳されており回答も英語で行われるが，LLM の回答が元の言語の表記にならざるを得ない場合があることも考慮し，元となった言語での表記も正答とした.[3]

3.3　データセットの構成

本研究の手法は，文化や地域性が言語と密接に関連づいていることを前提としている．一方で，例えばスペイン語のように広域的に使用されている言語では，文化と言語を正確に紐づけることは難しい．そこで本研究では，大半の話者が特定の地域に集中していると考えられる言語を対象言語の候補として選定した．具体的には，Wikipedia の言語版リストに掲載されている 342 言語のうち，国連加盟国 193 か国の中で，1 か国でのみ公用語[4] とされている 75 言語を問題作成の候補とした．このうち，地域性や Wikipedia におけるリソース量を考慮し，本研究では表 3 記載の 8 言語を対象とした．これらの言語における問題のベースとなる項目，各 1000 件に対し，3.2 節の手順を適用し，データセットを構築した．フィルタリング実施後の実際の問題は表 2 に示した形であり，各言語の問題数は表 3 にまとめた通りである．日本語の問題 224 問のうち，90.2%にあたる 202 問に関しては日本由来の知識であることを目視によって確認した．またデータセット全体のうち，ベースとなった項目の記事が複数言語に展開されているものは 31.0%，10 言語以上の展開がなされているものは 5.42%であり，他言語圏に伝播した文化的知識に関してもカバーされていた.

4　実験

4.1　実験設定

多言語 LLM に潜む文化的バイアスを検証するために，作成したデータセットを用いて，LLM の質問応答精度を計測した．本研究では，reasoning-effort を low に設定した GPT-5 を用い，問題に対して自由回答をさせる形で質問応答を行った．評価は LLM の提示した回答と，データセットに収録した各問題の正答の照合により行った．自由回答形式であるため，正誤判定は LLM の回答と正答との完全一致で評価するとともに，表記揺れなどを考慮し，正答に含まれる文字列のうち，回答とのレーベンシュタイン距離が最も小さいものとの文字レベルの F1 スコア [12] (以降 chrF) での評価も行った.[5]

4.2　実験結果

各言語のデータに対する回答結果を表 3 に示す．全体として，いずれの言語においても，正解率や chrF は低い水準にとどまっていた．うち，最も高い精度での回答がなされたのはアフリカーンス語であり，次いでスロバキア語であった．一方，どの指標においても，エストニア語，ウクライナ語，グルジア語といった東欧諸国の言語に対する性能は低い傾

2)　フィルタの詳細は付録 A に記載.
3)　具体例は付録 D に掲載.
4)　実質的な公用語とされている言語を含む.
5)　大文字小文字の統一，および NFKC 正規化を適用した.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 3 データセットの構成と検証実験の結果．() 中の数字は英語版記事が存在する項目の問題数．

言語	問題数	完全一致正解率 (↑) 全体	英語版有	英語版無	平均 chrF (↑) 全体	英語版有	英語版無
日本語 (東アジア)	224 (35)	0.223	0.600	0.153	0.374	0.725	0.309
パシュトー語 (中東)	78 (9)	0.262	0.889	0.161	0.412	0.973	0.321
アフリカーンス語 (アフリカ)	210 (43)	0.352	0.605	0.287	0.554	0.766	0.500
カタルーニャ語 (西欧)	120 (17)	0.283	0.824	0.194	0.478	0.899	0.408
スロバキア語 (中欧)	68 (20)	0.309	0.700	0.146	0.498	0.789	0.377
エストニア語 (北・東欧)	74 (9)	0.189	0.333	0.169	0.357	0.386	0.352
ウクライナ語 (中・東欧)	58 (4)	0.138	0.500	0.111	0.340	0.824	0.304
グルジア語 (東欧)	145 (20)	0.103	0.450	0.048	0.307	0.642	0.253

表 4 GPT-5 の誤答の例．

問題: 幼名を鶴之助とし、通称を主計・右近とした盛岡藩 3 代藩主・**南部**重信の子で … (中略) … 三田**南部**家の祖となった人物は誰でしょうか？

GPT の回答: nanbu nobuoki
正答: ["Katsunobu Nambu", "南部勝信"]

表 5 意図にそぐわない誤答判定の例．

問題: 兵庫県神戸市兵庫区にある … (中略) … この岬の名称は何でしょうか？

GPT の回答: wadamisaki
正答: ["Cape Wada", "和田岬"]

向が見られた．また，英語版記事が存在する項目の問題か否かで正解率や chrF に差が見られた．英語版記事が存在する項目に関しては言語を問わず高い回答精度を示しており，一部の言語では正解率が 8 割を超えるなど，英語版記事が存在しない項目との性能差が顕著にみられた．

以上より，各言語圏に特有な文化や知識に関する問題は，LLM における正解率が低く，反対に英語圏に伝播した知識に関しては正解率が高いことから，知識の文化的バイアスが発生していると考えられる．また，南アフリカではアフリカーンス語のほかに英語も公用語であるため，他の言語と比べ文化特有の知識が英語化されやすい環境にあると予想され，英語版記事の有無による性能差の結果と併せて考えると，このことが他の言語と比して高い性能を示した一因であると考えられる．なお，英語版が存在する項目の問題とそうでない問題では，知識がどの程度一般的であるかという要因以外の観点で問題の難易度が異なる可能性もあるため，さらなる検証が必要であると考える．

5 エラー分析

データセット中の日本語 QA において，完全一致で不正解とされた問題とその回答の傾向を分析した．日本語の問題計 224 問のうち，不正解は 174 問であった．174 問中，133 問に関しては，目視での確認においても，間違いと判定して差支えない回答がなされていた．ほとんどのケース (88 問) では，問題文中の情報と一切関係のない回答がなされており，モデルが全く知識を持たないと判断しうるものであった．残りの 45 問では，問題文から類推した情報をもとに回答がなされており，知識を有していない場合でも，推論能力により正解しうる可能性が示唆された．具体的な例は表 4 に示す．

一方，174 問中，34 問に関しては正解と判定されうる回答でありながら，不正解の判定がなされていた．そのほとんど (25 問) は，表 5 に示したような英訳に伴う表記揺れによるものであり，漢字が影響したケースも見られた．また，問題の不備や GPT-5 のフィルタリングによる回答拒否が計 2 問あるなど，問題作成プロセスの洗練が今後の課題として挙げられた．

6 おわりに

本研究では，LLM における文化的バイアスを検証するため，知識とその起源となる文化圏の情報を含む自由回答式 QA データセットを構築した．本手法は，知識と文化圏との関係を Wikipedia の作成日時という客観的情報に基づいて表現することから，人手による介入を必要とする既存研究とは異なり低コスト，かつ高い拡張性を有する．本データセットを用いた検証実験の結果，既存研究と同様に，LLM の性能における英語圏文化への偏重傾向が確認された．今後は，多文化への拡張やより高品質な QA の作成，日時情報を用いた知識の伝播への精緻な分析に取り組み，本課題の解決を目指す．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究の一部は JSPS 科研費 JP25K03175 の助成を受けたものです.

参考文献

[1] Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 6349–6384, 2024.

[2] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, et al. Challenges and strategies in cross-cultural NLP. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 6997–7013, 2022.

[3] Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 11772–11817, 2025.

[4] Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, et al. BLEnD: A benchmark for llms on everyday knowledge in diverse cultures and languages. In **Proceedings of the 38th Conference on Advances in Neural Information Processing Systems (NeurIPS)**, pp. 78104–78146, 2024.

[5] Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, et al. Include: Evaluating multilingual language understanding with regional knowledge. In **arXiv:2411.19799**, 2024.

[6] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In **arXiv:2009.03300**, 2021.

[7] Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, et al. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 18761–18799, 2025.

[8] Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. Understanding the capabilities and limitations of large language models for cultural commonsense. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)**, pp. 5668–5680, 2024.

[9] Bowen Jiang, Runchuan Zhu, Jiang Wu, Zinco Jiang, et al. Evaluating large language model with knowledge oriented language specific simple question answering. In **arXiv:2505.16591**, 2025.

[10] Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, et al. Chinese SimpleQA: A chinese factuality evaluation for large language models. In **arXiv:2411.07140**, 2024.

[11] Adam D. Lelkes, Vinh Q. Tran, and Cong Yu. Quiz-style question generation for news stories. In **arXiv:2102.09094**, 2021.

[12] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In **Proceedings of the Tenth Workshop on Statistical Machine Translation**, pp. 392–395, 2015.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 6　問題のベースとなった知識の例.

言語	例
カタルーニャ語	**Carlos Santiago Cela Pereira** → ガリシア人 (スペイン北西部ガリシア州に住む人) のサッカー選手 / **Neoptòlem de Milet** → ミレトス (かつてギリシャにあった町) に住んでいた作家 / **Vicenç Buron i Llorens** → カタルーニャ州のロマネスクが専門の歴史学者 / **Agustí Blanch** → バルセロナにある教会の合唱団の長 / **Rectoria de Salitja** → カタルーニャ州の Salitja という街の建物 / **José Antonio Bermúdez** → スペイン人の映画音響エンジニア / **Na Guardis** → スペインのマヨルカ海岸の小島 / **Rick Burgett** → アメリカ人の元プロ・モトクロス選手 / **Futur de l'Indicatiu (català)** → カタルーニャ語の未来を表す時制 / **Puig d'en Ponç (Santa Cristina d'Aro)** → カタルーニャ地方の山
エストニア語	**Ēķini mõis** → リヴォニア地域 (現ラトビア東北部・エストニア南部) にあった騎士の領地 / **Rahvusvaheline Sotsiaalteaduste Rakenduslik Kõrgkool LEX** → エストニアの首都タリンにかつてあった高等教育機関 / **Gertrud Üppis** → エストニア人の歌手 / **Koigi (Haljala)** → エストニア北部にかつてあった街 / **Ohvriaed** → エストニア南部の農園にあった宗教的な意味を持つ庭 / **Johann Christoph Schmidt** → エストニア、リヴォニア地域の聖職者 / **Tõnis Jõgiaas** → エストニア人の土木技師、活動家 / **Reiu jõgi** → エストニア、ラドビアを流れる川 / **Heinrich Steding** → リヴォニア騎士団の最後から 2 番目の団長 / **Voldemar Horst** → エストニア人の船乗り
スロバキア語	**Bílí Tygři Liberec** → チェコのアイスホッケーリーグ (エクストラリーグ) に属するチーム / **Dražkovce** → スロバキア中央部にある街 / **Xenombrotos z Kosu** → 紀元前 5 世紀のオリンピックの馬術チャンピオン / **Olympos (sochár)** → 紀元前 1 世紀頃のギリシャ人の彫刻家 / **Ján Longauer** → スロバキア人のカトリック司祭、教育者 / **Devínska hradná s ala** → スロバキアの城にあるにある天然記念物の崖 / **Božský manžel** → オーストリア人が書いた歌劇 / **Hronské Kosihy** → スロバキア南西部にある街 / **1. československá partizánska brigáda J. V. Stalina** → スロバキア民衆蜂起の頃に構成された軍隊 / **Rosinka** → スロバキア北部にある小川

A　フィルタ詳細

データフィルタリングの際には以下の基準を LLM により判断した.

1. 質問は「はい」／「いいえ」で回答される質問ではない.
2. 「どのように」や 「なぜ」のようなフルセンテンスでしか答えられないような問題ではない.
3. 答えは不必要な部分を含まず，問題の一部を再述していない.
4. 質問は 2023 年 12 月 31 日までの情報で回答可能である.
5. 時間が経過しても回答が変化する可能性のない質問である.
6. 質問は過度に単純ではなく，回答者の知識の深さを十分に評価するものである.
7. 問題は単独で成立しており，追加の情報がなくとも問われていることを理解するのに十分な背景を提供している.

B　QA 作成の詳細

日本語版 Wikipedia の “草津温泉” がベースとなる場合，「草津温泉は，群馬県吾妻郡草津町草津界隈に所在する温泉で ···」という記事の文章を LLM に与え，“草津温泉” が答えとなるような問題を生成させる．このようにして生成した問題に対する正答は元の言語版の項目名，元の言語版の項目名の英訳をベースとした．さらに，同じ項目の英語版が存在した場合の記事名 (この場合 “Kusatsu Onsen”) を加えた．また Wikidata に Also known as としてまとめられている表記を別解として用意した．例えば “Earth” は，Wikidata において “Planet Earth” や “The Blue Planet” などが Also known as とされており，これらの表記も正答とした.

表 7　元の言語での回答が想定される QA の例.

問題: 『阿弥陀経』に由来し、浄土真宗で念仏の信仰者が命終直後に極楽浄土へ生まれ、仏・菩薩や先に往生した先祖と同じ場所で再会できることを意味し、墓碑に刻まれることもある四字熟語は何でしょうか？
正答: ["Together we shall meet in the same place.", "倶会一処"]

C　問題のベースとなった知識の例

表 6 に，いくつかの言語についてランダムにサンプルした，問題のベースとなった知識の例をまとめる.

D　問題の例

表 7 に LLM の回答が元の言語の表記にならざるを得ない例を示す.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).