WikiOriginQA: Automated Construction of a QA Dataset with Knowledge Origins for Analyzing Cultural Bias

Kento Haneda[1,2], **Yosuke Kishinami**[1], **Ryo Fujii**[1], **Keisuke Sakaguchi**[2], **Makoto Morishita**[1]
[1]Future Corporation, [2]Tohoku University
team-nlp-research@future.co.jp

Abstract

Existing datasets on cultural knowledge have been typically created by human annotators familiar with each culture. However, the process requires considerable human effort, presenting a significant challenge to scalability. To address this challenge, we introduce **WikiOriginQA**, a QA dataset automatically constructed by objectively representing the origins of knowledge through the creation dates of different Wikipedia language editions. Experimental results with the created dataset confirm that LLM performance tends to be biased toward Anglophone cultures. Furthermore, the results suggest the importance of knowledge propagation into the Anglosphere. These findings demonstrate the validity of our automated approach, enabling cultural bias evaluations with superior scalability across diverse languages without extensive manual annotation.

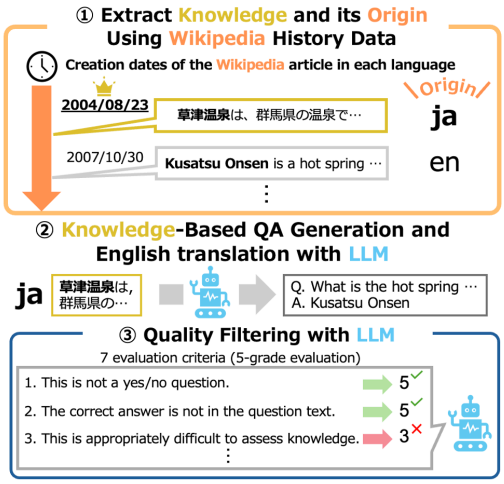


Figure 1: The concept of WikiOriginQA. Culture-specific knowledge is expected to appear first in the corresponding Wikipedia language edition.

1 Introduction

While Large Language Models (LLMs) are being used in many languages, previous studies have pointed out the existence of Anglocentric biases (Wang et al., 2024; Hershcovich et al., 2022), potentially leading to the exclusion of knowledge from non-English-speaking cultures in NLP technologies. To understand the extent of cultural bias in multilingual LLMs, it is essential to establish evaluation methods that cover a diverse range of languages and cultural spheres. However, many existing evaluation methods require human experts familiar with each culture to manually determine which culture a specific piece of knowledge belongs to (Arora et al., 2025; Myung et al., 2024; Romanou et al., 2024), making large-scale expansion to multiple cultures difficult.

In this study, we construct a QA dataset for validating cultural bias in a completely automated manner by identifying cultural spheres based on objective information that indicates the origin of knowledge. Specifically, we refer to the article creation dates across various Wikipedia language editions, based on the hypothesis that knowledge specific to a particular cultural sphere is documented earlier in its associated language than in other languages. For instance, common knowledge within the Japanese cultural sphere, such as “草津温泉” (Kusatsu Onsen, a famous hot spring in Japan), is expected to have its Wikipedia article created in the Japanese version earlier than in any other language (Figure 1). By using the knowledge extracted in this manner, we constructed **WikiOriginQA**,[1] an automatically generated QA dataset for analyzing cultural biases of LLMs. Our approach achieves high precision in extracting culture-specific knowledge while maintaining high multilingual scalability due to the complete elimination of human intervention. Experiments using this dataset confirmed that LLMs tend to be biased toward Anglosphere cultures, demonstrating the effectiveness of fully automated datasets for validating cultural bias.

[1]https://huggingface.co/datasets/future-architect/WikiOriginQA

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 69–77
November 6-10, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Several benchmarks and datasets have been proposed to evaluate cultural bias in LLMs. Global MMLU (Singh et al., 2025) has confirmed that performance on culture-dependent questions deteriorates in low-resource languages. Shen et al. (2024) constructed a benchmark for cultural common sense and demonstrated a decline in model performance regarding non-Western cultures. Other datasets focusing on cultural knowledge include CaLMQA (Arora et al., 2025) for long-form QA, BLEnD (Myung et al., 2024) for everyday knowledge, and INCLUDE (Romanou et al., 2024) for licensing exams. Furthermore, CANDLE (Nguyen et al., 2023) efficiently collects cultural knowledge from the English corpora focusing on categories such as food and clothing. These studies have reported Anglocentric biases in LLMs.

However, constructing these datasets typically requires considerable human effort, such as collection and annotation of questions by native speakers, presenting a significant challenge to scalability especially for low-resource languages. Furthermore, manual judgment raises concerns regarding subjectivity.

To address this, Jiang et al. (2025) recently proposed KoLasSimpleQA, an objective approach that extracts culture-specific knowledge by identifying items existing only in a specific Wikipedia language edition. However, this method fails to capture propagated knowledge that has been subsequently added to other language editions. Therefore, there is a risk that the dataset becomes biased toward excessively niche information. While our method is also based on Wikipedia, it incorporates temporal information to offer the advantage of being robust against knowledge propagation.

Furthermore, previous studies have reported a close relationship between language and culture regarding the development of Wikipedia articles, and from this perspective, extracting cultural knowledge using article creation timestamps is a rational approach. For example, articles on topics of high interest to a specific cultural group tend to exist only in the language edition corresponding to that group (Roy et al., 2020), individuals closely associated with a particular language are more likely to be documented in that language edition (Eom et al., 2015), and the cultural background of the editors community influences article contents (Miquel-Ribé and Laniado, 2018).

Language	# Questions	% En	% Other
Afrikaans (af)	81	29.6	34.6
Catalan (ca)	98	16.3	32.7
Estonian (et)	43	23.3	39.5
Japanese (ja)	108	17.6	30.6
Slovak (sk)	45	31.1	60.0
Georgian (ka)	61	13.1	31.1
Pashto (ps)	26	11.5	15.4
Ukrainian (uk)	41	7.3	26.8
English (en)	120	-	64.2

Table 1: The composition of WikiOriginQA. % En and % Other indicate the percentages of questions whose source items have an article in English and in any other language other than the target language, respectively.

Question: Who is the Japanese Western-style painter who was born in Tokushima City, Tokushima Prefecture (Chuo-dori), withdrew together with Asai Chu from the Kobu Art School in 1878 to form the “Juuichikai,” became an art teacher at Osaka Specialized School in 1879, later oversaw the art department at Doshisha from 1895 to 1917, and additionally published all ten volumes of Taisei Futsu Gagakuhon from Garyukan in 1882–1883?

Answer: [“Isana Morizumi”, “守住勇魚”]

Table 2: Example of an automatically generated QA pair for Japanese culture in WikiOriginQA.

3 WikiOriginQA

In this section, we describe the construction process of WikiOriginQA, a short-answer QA dataset designed to evaluate cultural biases in LLMs.

3.1 Extraction of Knowledge for QA Data

To reference the creation dates of Wikipedia articles across different languages representing the same concept, we used Wikipedia dump files and Wikidata. Wikidata is a database of items representing topics, concepts, and objects, where a unique entity ID is linked to corresponding Wikipedia articles in each language. This ID enables referencing creation dates for all associated Wikipedia language editions. By comparing creation dates across 342 languages from the List of Wikipedias[2], we inferred each item’s originating language sphere to serve as the basis for QA data.

3.2 Generation of QA Data

For each extracted entity, we used an LLM to generate a question from its Wikipedia article, with the

[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias

Language	Randomly selected examples of extracted source knowledge
Afrikaans	Swartvlei (South African saltwater lake), Andrew Motjuoadi (South African painter), Tienie Holloway (South African writer)
Japanese	守住勇魚 (Japanese painter), 新党さきがけ (Japanese political party), 十六夜日記 (medieval Japanese travel diary)
Pashto	خیشکي (Pashtun tribes), لوګزامبورګ (Luxembourg), زلی هیوادمل (Afghan writer)
Ukrainian	Пригородок (City in Ukraine), SLC2A8 (name of a protein), Автошлях Т 2601 (road in Ukraine)

Table 3: Randomly selected examples. Red text indicates cases where non-cultural items were incorrectly extracted. Our approach effectively extracted cultural knowledge.

title as the correct answer. The prompts were provided in the same language as the corresponding articles. In addition, the generated questions were iteratively refined by the LLM to improve their quality. To minimize the influence of the model's linguistic ability on the results, we translated the QA data into English using an LLM. Furthermore, we applied an LLM-based filtering to ensure data quality, as illustrated in Figure 1. We designed a filter consisting of seven criteria, each evaluated on a 5-point scale, building upon previous studies (Jiang et al., 2025; He et al., 2024; Lelkes et al., 2021). We retained only questions that received maximum scores across all criteria.[3] We used GPT-5 for these processes with the reasoning effort set to low.

We set gold answers for each question to the original item name supplemented by aliases in Wikidata. In addition, to account for instances where a response in the original script is inherently expected (e.g., *kanji* idioms), we accepted answers in the source language.

3.3 Target Languages and Dataset Details

We selected languages whose speakers are concentrated in specific regions as candidates for dataset construction. Specifically, we selected 75 candidate languages that are designated as official languages[4] in exactly one of the 193 United Nations member states. Among these candidates, we selected 8 languages, listed in Table 1, considering their regionality and Wikipedia resource availability. For each language, we selected 1,000 items and applied the procedure described in Section 3.2 to construct the dataset. The composition of the dataset is in Table 1, and an example of questions remaining after filtering is shown in Table 2. For comparative evaluation with these languages, we also created the questions in English using the same methodology.

As shown in Table 3, knowledge specific to each language sphere was effectively extracted. Furthermore, Table 1 demonstrates ability to capture knowledge propagated to other languages. Note that there were a small number of cases, such as "Luxembourg" in Pashto, where items were extracted separately due to missing Wikidata links, despite existing in other languages. However, a manual case study of Japanese data confirmed that 96 out of 108 questions (88.9%) originated from Japan, indicating that such errors are exceptional and demonstrating the high validity of our method.

4 Experiment

4.1 Settings

To verify the cultural biases in multilingual LLMs, we prompted several models to generate answers for the questions in WikiOriginQA. The models used were GPT-5, llm-jp-3.1-13b-instruct4 (LLM-jp) (LLM-jp, 2024), and Qwen3 (Yang et al., 2025) in various parameter sizes (8B, 14B, 32B, and 235B). For GPT-5, we set the reasoning effort to low and kept all other parameters to default. Since the task is in short-answer form, we evaluated correctness not only through Exact Match accuracy (Acc) but also by using the character-level F1 score (chrF) (Popović, 2015).

4.2 Results

The accuracy of the models' responses is presented in Figure 2. Overall, the Acc scores remained at a low level across all languages. For most models, questions based on English articles achieved higher Acc than those based on other language editions, and this tendency was stronger for models with larger parameter sizes. For each model, questions whose source items have an English article tended to achieve higher Acc than the overall questions based on that language.[5] This indicates that the information propagation to the Anglosphere is essential for a model's knowledge acquisition.

Considering the impact of the translation, we also conducted additional experiments for certain

[3] See Appendix A for detailed filtering criteria.

[4] This includes de-facto official languages.

[5] Detailed results are presented in Appendix C.

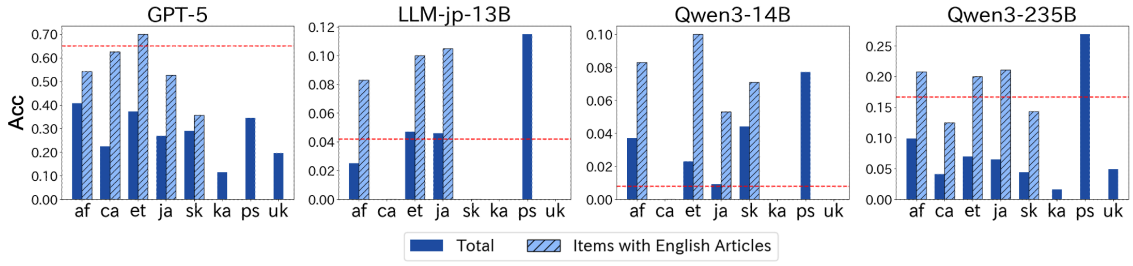


Figure 2: Accuracy of questions based on each source language. The red line indicates the accuracy of questions derived from English articles. Higher accuracy was observed for items with English articles. Since the number of questions with English articles was fewer than 10, this analysis was not conducted for ka, ps, and uk.

	GPT-5	Qwen3-235B
True Negative	63 / 79	93 / 101
Random guess	44 / 63	69 / 93
Use info in the question	19 / 63	24 / 93
False Negative	14 / 79	6 / 101
Spelling variation (English)	7 / 14	2 / 6
Spelling variation (Other)	3 / 14	0 / 6
Other (e.g., another name)	4 /14	4 / 6
Other (e.g., unable to answer)	2 / 79	0 / 101

Table 4: Error analysis of GPT-5 and Qwen3-235B. The results demonstrate the validity of automatic evaluation through high true negative rates.

languages using source-language QA pairs without English translation. As a result, similar to the translated data, higher performance was observed for items that have corresponding English articles. Furthermore, to address potential extraction errors, we restricted the Japanese dataset to cultural QA pairs that were manually verified, and found that the results mirrored the findings from the overall dataset, which was generated automatically.[6]

5 Analysis

5.1 Result Analysis

Experimental results suggest a cultural bias toward the Anglosphere in LLMs, as they achieved higher Acc on questions based on English articles or knowledge propagated into English. Afrikaans questions were answered relatively accurately; this is likely because English is also an official language in South Africa. Consequently, knowledge in Afrikaans is more prone to translation into English compared to other languages, making it easier for the LLM to answer. However, the intrinsic difficulty of the questions may vary between entities with and without a corresponding English article. Therefore, further investigation is necessary to distinguish whether the performance gap stems from the availability of English resources for the entity or from the intrinsic difficulty.

5.2 Error Analysis

We analyzed the tendencies of the responses that were judged incorrect in the answer of GPT-5 and Qwen3-235B. Among the 108 questions based on Japanese, 79 for GPT-5 and 101 for Qwen3-235B were judged incorrect by exact match. Table 4 shows the manual verification for these errors. In most cases, there was no discrepancy between the automatic and manual evaluation, and the responses could reasonably be considered incorrect.

On the other hand, in some cases, responses that should have been judged correct were classified as incorrect. In many of these cases, as shown in Table 4, the responses were judged incorrect due to factors related to the English translation and the conversion of Japanese, particularly *kanji* into romanized forms.[7] Additionally, some other responses were observed, including GPT-5's refusal to answer. Further refinement of question construction and answer evaluation remains future work.

6 Conclusion

We proposed a fully automated methodology to construct **WikiOriginQA**, a highly scalable QA dataset for evaluating cultural biases in LLMs. By identifying knowledge origins using Wikipedia article creation dates, our approach overcomes the scalability and subjectivity issues of traditional human-annotated datasets. Experimental results across various LLMs revealed a significant performance skew toward the Anglosphere, highlighting the persistent risk of excluding non-English cultures. We

[6]Detailed results are in Appendix D and E, respectively.

[7]A specific example is provided in Appendix H.

expect our approach and dataset will help develop more culturally inclusive language models.

Limitations

Target Languages The experiments conducted in this study primarily aim to demonstrate the practical viability of our proposed dataset construction method, which is inherently scalable to a much broader range of languages. As described in Section 3.3, there are 75 languages designated as official languages in exactly one country. From this pool, we selected eight languages, taking regional diversity and other factors into account. This selection serves merely as an illustrative example and does not limit the overall scalability of our approach.

However, because our method relies on the intrinsic connection between a language and a specific region, it is difficult to simply apply it to widely spoken languages across multiple regions, such as French or Spanish. While identifying a knowledge's originating region using the IP addresses of Wikipedia editors is a potential alternative approach, we did not adopt it due to concerns regarding long-term sustainability and scalability. Notably, several versions of Wikipedia, including the English edition, have begun phasing out the public display of IP addresses[8].

Coverage in Knowledge Extraction Particularly for low-resource languages or those with underdeveloped Wikipedia editions, it is often the case that cultural items native to that language only exist as English articles or are documented in other major global languages. In such cases, our method cannot extract knowledge concerning these items, which presents a limitation in achieving high recall when extracting cultural knowledge. On the other hand, an entity that is documented first in a specific language ahead of all other global languages strongly implies a close connection to that language's cultural sphere, leading to the high-precision extraction of culture-specific knowledge.

Dataset Scale The dataset constructed in this study is relatively small in scale owing to financial constraints. Nevertheless, since Wikipedia contains tens of thousands of extractable entities per language, there is substantial room for future expansion. Specifically, there are over 800,000 items in the Japanese edition, and over 300,000 in the Ukrainian edition, that were created earlier than in any other language version. In this work, we sampled only a subset of these candidates to build the dataset.

Reliability of Automated Construction Because our method focuses on fully automated dataset construction, it is difficult to guarantee the absolute reliability typically associated with human-annotated approaches. However, as demonstrated in Section 3.3, 88.9% of the entities in the Japanese subset were confirmed to originate from Japan, indicating that our entity extraction process was successful. Admittedly, this extraction is not perfectly accurate and leaves room for improvement, however, as demonstrated in Section 4.2, it does not have a critical impact on evaluating the model's cultural bias, at least regarding the Japanese questions. Furthermore, regarding the automatically generated QA pairs, although we have not exhaustively verified the entire dataset, the authors manually reviewed a sampled subset and confirmed its high quality. Utilizing our method as a foundation while incorporating human-assisted quality control could potentially enable the construction of an even higher-quality dataset; we leave this exploration for future work.

[8] https://en.wikipedia.org/wiki/Wikipedia:Temporary_accounts

References

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. 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)*, pages 11772–11817.

Young-Ho Eom, Pablo Aragón, David Laniado, Andreas Kaltenbrunner, Sebastiano Vigna, and Dima L. Shepelyansky. 2015. Interactions of cultures and top people of wikipedia from ranking of 24 language editions. *PLOS ONE*, 10(3):1–27.

Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, Zhuoran Lin, Xuepeng Liu, Dekai Sun, Shirong Lin, Zhicheng Zheng, Xiaoyong Zhu, Wenbo Su, and Bo Zheng. 2024. Chinese SimpleQA: A chinese factuality evaluation for large language models. In *arXiv:2411.07140*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza

Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6997–7013.

Bowen Jiang, Runchuan Zhu, Jiang Wu, Zinco Jiang, Yifan He, Junyuan Gao, Jia Yu, Rui Min, Yinfan Wang, Haote Yang, Songyang Zhang, Dahua Lin, Lijun Wu, and Conghui He. 2025. Evaluating large language model with knowledge oriented language specific simple question answering. In *arXiv:2505.16591*.

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

LLM-jp. 2024. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. In *arXiv:2407.03963*.

Marc Miquel-Ribé and David Laniado. 2018. Wikipedia culture gap: Quantifying content imbalances across 40 language editions. *Frontiers in Physics*, Volume 6 - 2018.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. 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)*, pages 78104–78146.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1907–1917, New York, NY, USA. Association for Computing Machinery.

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

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, and 40 others. 2024. Include: Evaluating multilingual language understanding with regional knowledge. In *arXiv:2411.19799*.

Dwaipayan Roy, Sumit Bhatia, and Prateek Jain. 2020. A topic-aligned multilingual corpus of Wikipedia articles for studying information asymmetry in low resource languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2373–2380, Marseille, France. European Language Resources Association.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. 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)*, pages 5668–5680.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. 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)*, pages 18761–18799.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. 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)*, pages 6349–6384.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. In *arXiv:2505.09388*.

A Details of LLM-based Filtering

For data filtering of WikiOriginQA, we employed an LLM to evaluate the questions according to the following criteria:

1. The question is not a 'yes'/'no' question.
2. It is not a question that can only be answered with full sentences like 'how' or 'why.'
3. The answer does not include unnecessary parts and does not restate parts of the problem.
4. Questions can be answered using information up until December 31, 2023.
5. These are questions whose answers are unlikely to change over time.
6. The questions are not overly simplistic and adequately assess the depth of knowledge of the respondent.
7. The question stands alone and provides enough context to understand what is being asked without any additional information.

B Details of QA Data Generation

Given a source such as the Japanese Wikipedia entry for "草津温泉" (Kusatsu Onsen), we provide the article text ("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...") to the LLM and instruct it to generate a question with "草津温泉" as the answer. The correct answers initially consist of the entity's title in the original language and its English translation. If a corresponding English Wikipedia article exists, its title (in this case, "Kusatsu Onsen") is added to the answer set. Furthermore, we include aliases listed under Wikidata's *Also known as* property as valid alternative answers. For example, the entity "Earth" has Wikidata aliases ("Planet Earth" and "The Blue Planet"). We equally treat these as correct answers.

C Details of Results

Table 5 shows the detailed QA performance for each source language.

D Results of Original Language

Table 6 shows the results of the GPT-5's responses to each source-language QA prior to English translation.

E Results of Manually Verified Data

Table 7 presents GPT-5's performance on manually verified Japanese QA.

F Examples of Source Knowledge

Table 8 summarizes randomly sampled examples of the knowledge that served as the basis for the QA data across several languages.

G Example of Incorrect Answer by GPT-5

Table 9 shows examples of incorrect responses where the model answered by inferring from the information in the question text.

H Examples Erroneously Judged as Incorrect

Table 10 shows examples of responses erroneously judged as incorrect.

I Target Languages for Retrieving Revision History

In this study, we retrieved the edit histories of 342 active Wikipedia language editions with publicly available data dumps (https://meta.wikimedia.org/wiki/List_of_Wikipedias) to identify the language of earliest creation for each item.

J Filtering Extracted Knowledge Using LLM

As described in Section 3.3, knowledge extraction using our method is not perfect. To improve extraction accuracy, we used GPT-5 to determine whether the extracted knowledge was cultural and filtered out knowledge that was judged not to be cultural. When this filtering was applied to the Japanese data, the extraction accuracy improved from 89% to 94%, confirming that LLM-based filtering can contribute to improving extraction accuracy to some extent.

K Dataset License

The WikiOriginQA dataset proposed in this paper is released under the Apache License 2.0 (https://www.apache.org/licenses/LICENSE-2.0).

Language	# Questions	GPT-5		LLM-jp-13B		Qwen3-8B		Qwen3-14B		Qwen3-32B		Qwen3-235B	
		Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF
Afrikaans (af)	81	.407	.588	.025	.148	.037	.191	.037	.182	.099	.228	.099	.226
Catalan (ca)	98	.224	.460	.000	.148	.010	.165	.000	.170	.061	.272	.041	.261
Estonian (et)	43	.372	.507	.047	.181	.000	.149	.023	.170	.070	.244	.070	.231
Japanese (ja)	108	.269	.403	.046	.131	.009	.155	.009	.167	.056	.224	.065	.229
Slovak (sk)	45	.289	.464	.000	.161	.044	.197	.044	.189	.044	.244	.044	.214
Georgian (ka)	61	.115	.336	.000	.123	.000	.156	.000	.169	.016	.188	.016	.185
Pashto (ps)	26	.346	.581	.115	.255	.154	.272	.077	.240	.269	.452	.269	.438
Ukrainian (uk)	41	.195	.405	.000	.179	.000	.187	.000	.160	.049	.234	.049	.220
English (en)	120	.650	.762	.042	.130	.008	.135	.042	.159	.183	.319	.167	.305

Table 5: Detailed results of the experiment. The table presents the overall evaluation for questions derived from each language. Overall performance was low for both Acc and chrF. However, English-based questions yielded relatively better performance than those from other languages, a tendency that became more prominent as model size increased.

	Translated		Original	
	Acc	chrF	Acc	chrF
af (total)	0.41	0.59	0.43	0.61
af (w/ en articles)	0.54	0.68	0.46	0.68
ja (total)	0.27	0.40	0.37	0.34
ja (w/ en articles)	0.52	0.66	0.68	0.55

Table 6: Results using the QA prior to English translation. Trends generally similar to those after English translation were observed.

	Total		Verified	
	Acc	chrF	Acc	chrF
ja (total)	0.27	0.40	0.26	0.39
ja (w/ en articles)	0.52	0.66	0.50	0.62

Table 7: Results of GPT-5's responses on the manually verified data. Trends generally similar to those of the entire dataset were observed.

Language	Knowledge
Catalan	**Carlos Santiago Cela Pereira** → A Galician soccer player / **Neoptòlem de Milet** → A writer who lived in Miletus (a town that once existed in Greece) / **Vicenç Buron i Llorens** → A historian specializing in the Romanesque of Catalonia / **Agustí Blanch** → The director of a choir at a church in Barcelona / **Rectoria de Salitja** → A building in a town called *Salitja* in Catalonia / **José Antonio Bermúdez**→ A Spanish film sound engineer / **Na Guardis** → A small island off the coast of Mallorca, Spain / **Rick Burgett** → An American former professional motocross rider / **Futur de l'Indicatiu (català)** → A tense expressing the future in Catalan / **Puig d'en Ponç (Santa Cristina d'Aro)** → A mountain in the Catalonia region
Estonian	**Ēķini mõis** → A knight's estate in the Livonian region (present-day northeastern Latvia and southern Estonia) / **Rahvusvaheline Sotsiaalteaduste Rakenduslik Kõrgkool LEX** → A former higher education institution in Tallinn, the capital of Estonia / **Gertrud Üppis** → An Estonian singer / **Koigi (Haljala)** → A former town in northern Estonia / **Ohvriaed** → A garden with religious significance on a farm in southern Estonia / **Johann Christoph Schmidt** → A clergyman in Estonia and the Livonian region / **Tõnis Jõgiaas** → An Estonian civil engineer and activist / **Reiu jõgi** → A river flowing through Estonia and Latvia / **Heinrich Steding** → The penultimate Master of the Livonian Order / **Voldemar Horst** → An Estonian sailor
Slovak	**Bílí Tygři Liberec** → A team in the Czech ice hockey league (the Extralig) / **Dražkovce** → A town in central Slovakia / **Xenombrotos z Kosu** → An Olympic equestrian champion from the 5th century BC / **Olympos (sochár)** → A Greek sculptor from around the 1st century BC / **Ján Longauer** → A Slovak Catholic priest and educator / **Devínska hradná s ala** → A natural monument cliff, located at a castle in Slovakia / **Božský manžel** → An opera written by an Austrian / **Hronské Kosihy** → A town in southwestern Slovakia / **1. československá partizánska brigáda J. V. Stalina** → A military force formed during the Slovak National Uprising / **Rosinka** → A stream in northern Slovakia
English	**Victoria Vesna** → An American artist / **Huastec people** → The Indigenous people of Mexico / **Prosaposin** → The name of a protein / **Canal Street station (New York City Subway)** → A subway station in New York City / **Liberty Square Blueprint** → An Occupy Wall Street manifesto / **Infinity (Lee Morgan album)** → An American jazz trumpeter Lee Morgan's album / **Marianne Francis** → An English evangelical / **Silas Mix** → An American lawyer / **Mikey Chung** → A Jamaican musician / **Butch Gautreaux** → An American politician

Table 8: Examples of source knowledge for the QA data across several languages.

Question: Who was the person whose childhood name was Tsurunosuke and [· · ·] becoming the founder of the Mita **Nanbu** family?
GPT-5's Answer: Nanbu Yukishige **Correct Answers:** ["Katsunobu Nambu", "南部勝信"]

Table 9: Incorrect answer by GPT-5. The model answers by inferring from the family name **Nanbu** present in the question text.

Question: What is the name of the Class A river in the Arakawa river system [· · ·] referred to as "Nenfutorikawa" in the Edo-period Shinpen Musashi Fudokikō?
GPT-5's Answer: Furo River **Correct Answers:** ["Ageless River", "不老川"]

Table 10: An example of unintentional incorrect answers. "不老" is pronounced "Furo" and "川" means "river".