CPopQA: Ranking Cultural Concept Popularity by LLMs

Anonymous ACL submission

Abstract

Prior work has demonstrated large language models' (LLMs) potential to discern statistical tendencies within their pre-training corpora. Despite that, many examinations of LLMs' knowledge capacity focus on knowledge explicitly appearing in the training data or implicitly inferable from similar contexts. How well an LLM captures the corpus-level statistical trends of concepts for reasoning, especially long-tail ones, is still underexplored. In this study, we introduce a novel few-shot question-answering task (CPopQA) that examines LLMs' statistical ranking abilities for long-tail cultural concepts (e.g., holidays), with a specific focus on these concepts' popularity in the United States and the United Kingdom, respectively. We curate a dataset containing 459 holidays across 58 countries, generating a total of 9,000 QA testing pairs. Experiments on four strong LLMs show that large models are capable of ranking long-tail cultural concepts regarding their statistical tendency. Notably, GPT-3.5 displayed superior performance and exhibited its potential to identify geo-cultural proximity across continents.

1 Introduction

Large language models (LLMs) have shown their potential to capture multiple facets of the world, benefiting a variety of downstream applications such as constructing knowledge bases with reduced reliance on human intervention (Bosselut et al., 2019; Wei et al., 2023). Despite that, the capacity of knowledge embedded in LLMs is still an open question, causing uncertainty in practical model deployment. To address this concern, researchers have been actively examining LLMs on various knowledge-intensive tasks, from word- or sentence-level linguistic phenomena(Hewitt and Manning, 2019; Conneau et al., 2018) to language's statistical tendencies (Meister and Cotterell, 2021; Takahashi and Tanaka-Ishii, 2017, 2019), and from commonsense (Li et al., 2022), factual knowledge (Petroni et al., 2019) to basic science (Horawalavithana et al., 2022; Imani et al., 2023). Given the generative nature of LLMs, these tasks can be easily conducted in a question-answering (QA) format. Early studies in this thread emphasize LLMs' memory capacity and discover that LLMs possess a remarkable ability to embed the aforementioned knowledge. Recently, advances in prompting like chain-of-thought (Wei et al., 2022; Wang et al., 2022) and self-reflection (Shinn et al., 2023) have enabled LLMs to elicit complex multi-hop reasoning from in-context examples. Beyond common knowledge, the latest studies further explore LLMs' ability to process long-tail knowledge (Kandpal et al., 2022), particularly centering geo-diverse knowledge in sociocultural contexts (Yin et al., 2022; Kabra et al., 2023; Arora et al., 2023a).

Despite insightful findings, existing examinations largely focus on the capability of LLMs to grasp knowledge explicitly appearing in the training data or implicitly inferable from similar contexts. There is scant research exploring the capacity of LLMs to *capture the broad statistical patterns of concepts within extensive datasets for in-depth comparisons, especially long-tail concepts spanning significantly diverse sociocultural contexts*. This alternative perspective focuses on models' potential to embed macro-level phenomena derived from widely scattered knowledge points in the training corpus, which can broaden LLMs' benefits as exploratory tools in support of corpus-centered computational analysis (Roberts, 2020), such as helping digital humanists and social scientists to gain new insights into historical, cultural, and social problems (Card et al., 2022; Underwood and So, 2021).

In this study, we attempt to explore the statistical ranking ability of LLMs, with a specific focus on a research question: *Can large language models compare cultural concepts, especially long-tail ones, regarding their popularity?* To examine this question, we design a ranking-based statistical

QA task that compares cultural concept popularity across countries (called CPopQA). This new task extends the study of LLMs' cultural awareness into their capacity to encode broader cross-cultural social visibility patterns, which may provide insights into the potential utility of LLMs for tracking cross-cultural evolution tendencies. To support this study, we curate a benchmark dataset of 9,000 QA testing pairs, covering 459 holidays across 58 countries. Note that, our dataset construction is flexible and scalable, allowing for the easy generation of diverse testing instances. Experiments on four popular LLMs show that the large models are capable of ranking holidays regarding their statistical tendency. In particular, GPT-3.5 outperformed other models and showed a potential to identify geo-cultural proximity across continents.

2 CPopQA

In this section, we introduce our CPopQA by describing the task formulation, the process of dataset construction, and a prompt-based LLM approach.

Tasks Considering the geo-association between holidays and countries, we propose two levels of ranking-based statistical QA tasks: (1) fine-level holiday ranking (see below) and (2) coarse-level country ranking (see Appendix F).

- **Task 1. Holiday ranking:** Given a set of holidays $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$ from a query country c_q, the goal is to sort $\mathcal{H}$ in a descending order based on their popularity in a target country c_t.

Holiday List Curation To create the QA dataset, we started by curating global holidays from Wikipedia's list of public holidays by country[1], considering the following factors:

1. **Diversity and inclusivity across geo-cultures:** we considered the holidays from both under-represented and well-represented countries. Specifically, we referred to the population statistics and the number of holidays on each country's wiki page, and collected holidays from the top five and bottom five countries regarding population statistics in each continent.
2. **Valid wiki page:** We required the selected holidays to have valid wiki pages to guarantee the authenticity of collected items. Meanwhile, we extracted the first paragraph from each holiday's wiki page as its description. This enables the future development of methods using text descriptions of these holidays.
3. **Date variation** We curated the holiday list by adding their countries and dates because many common holidays across countries are celebrated on different dates. For example, *Labor Day* is celebrated on *May 1st in China*, whereas it falls on *September 5th in the United States*.

Due to the editions of different crowd workers and the unique characteristics of holidays, we conducted a series of data cleaning (Appendix B).

Holiday Popularity Collection We estimate the overall engagement per holiday among people based on the average frequency of each holiday's name over ∼220 years (1800–2019), counted by Google Books Ngram Viewer[2] (GBNV). This tool has been widely used to analyze user-selected n-gram frequency. The corpus of GBNV consists of digitized books, which, to our best knowledge, is the largest public collections of books across a wide spectrum of domains and time periods [3], making it affordable with representative n-gram statistics. Moreover, GBNV provides several sub-corpora categorized by the books' publication country, making it an ideal resource to collect worldwide holiday popularity within a specific country. Since English is the most accessible language denoting various worldwide holidays, especially for holidays from non-English speaking countries, we estimate holiday popularity based on two English corpora (i.e., American and British English corpus), respectively.

To validate our estimation strategy on holiday popularity, we conducted a human evaluation with 6 annotators (details in Appendix C). Our results show that the GBNV statistics achieved ∼60% consistency with human judgments on average. In total, we collected information on 459 holidays in 58 countries on 5 continents. Each holiday is annotated with its country, date, description, and frequency in American/British corpora. The details of the data statistics are in Appendix D.

QA Pair Construction To investigate the influence of ranking complexity on model performance, we constructed questions to rank n items for both tasks, where $n \in \{2, 3, 5\}$. For example in Task 1, we sampled a holiday set $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$ in a query country c_q from our complete holiday

[1] https://en.wikipedia.org/wiki/Category:Lists_of_public_holidays_by_country

[2] https://books.google.com/ngrams/

[3] Over 6% of all books ever published in the 2012 version (Lin et al., 2012), and we use the updated 2020 version.

Setting	#Countries	#Holidays
2-item ranking	58	255
3-item ranking	57	265
5-item ranking	55	271

Table 1: Holiday Diversity in Testing Data.

list, and sorted them by their popularity in a target country c_t (e.g., US or UK) to get a ranked list $[h'_1, h'_2, \ldots, h'_n]$. We followed prior work (Yin et al., 2022) by using either the country names (e.g., holidays in Nigeria) or their corresponding modifiers (e.g., Nigerian holidays) to denote the query countries in the questions. We then selected the optimal question template with the ranked holiday list as the answer for further analysis:

> Question = "Can you provide a descending order for the following c_q holidays by their popularity in the c_t: $h_1, h_2, \ldots, h_n$"
> Answer = "1.h'_1, 2.h'_2, ..., $n.h'_n$"

For each $n \in \{2, 3, 5\}$ in either holiday or country ranking, we created 200 pairs for ranking n items. To examine the variation of results, in each setting, we conducted five rounds of ranking set sampling for QA pair generation, and then we repeated all the experiments. Thus, our QA dataset consists of 9,000 QA pairs in total. Table 1 shows the holiday and country statistics in each ranking setting. Note that, we considered both US and UK as the target countries in this holiday ranking task.

Prompting Following Kandpal et al. (2022), we used a simple prompt template: "Q: [Question] \n A:[Answer]" and randomly selected 3 in-context examples [4] to form a prompt. Feeding the prompt to an LLM, we generated ranks by greedy decoding and we compared them with the ground truth.

3 Experimental Setting

LLMs and Baselines We chose 4 popular LLMs for evaluation. The first LLM is **GPT-3.5** (i.e., text-davinci-003, 175B parameters). Through fine-tuning GPT-3 by reinforcement learning from human preferences (Ouyang et al., 2022), GPT-3.5 shows a higher quality in handling complex instructions compared to prior GPT-based models. We next chose LLaMA, with 7B (**LLaMA-7B**) and 13B (**LLaMA-13B**) parameters, pre-trained on the English-dominated corpora covering diverse domains (Touvron et al., 2023). The final model is **BLOOM-7b1**, a multilingual LLM with ~7B parameters (Scao et al., 2023). We selected this BLOOM variant because of its comparable model size with LLaMA-7B. We employed 3 baselines, including random guess and statistical simulation by Google Trends and Wikipedia article length, respectively (see details in Appendix E).

Setting	Model	P@1 (%)	Acc. (%)	Diff.
2-item ranking	random guess	50.00 ± 0.00	50.00 ± 0.00	-
	google stat	57.00 ± 0.03	57.00 ± 0.03	-
	wiki len	**59.20** ± 0.04	**59.20** ± 0.04	-
	bloom-7b1	42.90 ± 0.04	42.90 ± 0.04	-
	llama-7b	48.20 ± 0.03	48.20 ± 0.03	-
	llama-13b	51.10 ± 0.05	51.10 ± 0.05	-
	gpt-3.5	54.80 ± 0.03	54.80 ± 0.03	-
3-item ranking	random guess	33.33 ± 0.00	16.67 ± 0.00	0.500 ± 0.00
	google stat	37.00 ± 0.03	21.40 ± 0.01	0.441 ± 0.01
	wiki len	53.5 ± 0.03	28.10 ± 0.03	0.378 ± 0.02
	bloom-7b1	32.20 ± 0.04	16.00 ± 0.03	0.472 ± 0.03
	llama-7b	43.10 ± 0.03	19.90 ± 0.04	0.427 ± 0.03
	llama-13b	36.30 ± 0.04	17.70 ± 0.02	0.460 ± 0.03
	gpt-3.5	**59.30** ± 0.04	**34.60** ± 0.02	**0.305** ± 0.03
5-item ranking	random guess	20.00 ± 0.00	0.83 ± 0.00	0.500 ± 0.00
	google stat	27.30 ± 0.01	3.30 ± 0.01	0.419 ± 0.02
	wiki len	46.70 ± 0.04	2.30 ± 0.01	0.337 ± 0.01
	bloom-7b1	17.90 ± 0.02	0.50 ± 0.01	0.473 ± 0.01
	llama-7b	27.80 ± 0.02	1.10 ± 0.01	0.461 ± 0.01
	llama-13b	24.90 ± 0.03	1.60 ± 0.01	0.446 ± 0.02
	gpt-3.5	**62.00** ± 0.03	**6.60** ± 0.00	**0.267** ± 0.02

Table 2: Results of ranking worldwide holiday popularity in the US (mean ± standard deviation).

Metrics We used three evaluation metrics, including **Accuracy (Acc.)** measures the degree of the exact match; **Precision@1 (P@1)** calculates the precision of the first ranked item; **Average difference (Diff.** $= \frac{1}{N}\sum_{j=1}^{N}\frac{1-\rho_j}{2}$) measures the overall ranking difference, where ρ_j is the Spearman correlation coefficient between the model prediction and the ground-truth ranking on the j-th QA example.

4 Results and Analysis

Can LLMs elicit long-tail cultural statistics and rankings? Table 2 shows the ranking results of LLMs regarding holiday popularity in the US (see UK results in Appendix H, country ranking in Appendix F). In general, GPT-3.5 and LLaMa (7B and 13B) significantly outperform the random baseline, while BLOOM-7b1 tends to underperform on rankings. Both statistical baselines outperformed the random guess with a high margin. Notably, GPT-3.5 shows an obvious enhancement in most cases, except for pairwise holiday comparisons. Interestingly, the wiki baseline shows the highest accuracy in pairwise ranking and beats LLAMA variants in all ranking cases. Our observations demonstrate that GPT-3.5 and LLaMa exhibit the potential to capture the popularity tendencies of long-tail cul-

[4] we tried different sizes of in-context examples (e.g., 2, 3, 5) and observed similar trends regarding model performance.

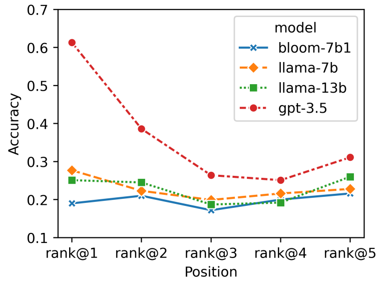

Figure 1: LLM results at each position on holiday ranking in the US

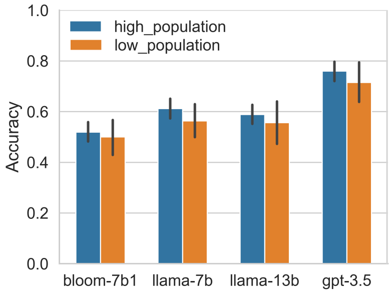

Figure 2: Pairwise ranking accuracy in the US regarding geo-cultural representativeness

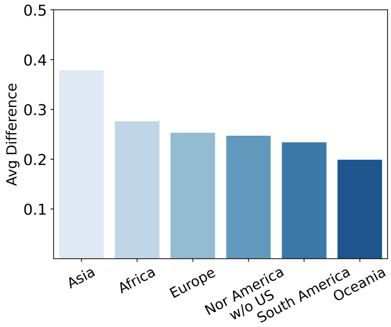

Figure 3: The results of GPT-3.5 on holiday ranking in the US across continents.

tural concepts for ranking. A detailed holiday description on Wikipedia shows a positive signal related to the holiday popularity.

What ranking-based factors challenge the prediction? Looking into the ranking setting, we find that LLMs show a noticeable drop in P@1 (∼5%-10%)and Acc (∼20%-30%) when adding ranked items (see Table 2), suggesting that LLMs are sensitive to the ranking complexity. With further exploration of model performance at each ranking position (see Figure 1 on 5-item ranking results in the US, Appendix H on results in the UK), we find that holidays ordered at two ends are usually easier to be predicted than those in between. Items ordered at the third and fourth positions are more prone to confuse LLMs than others.

Influence of geo-cultural representativeness? To examine how LLMs respond to geo-cultural representativeness, we conducted an analysis of pairwise ranking accuracy by comparing the most commonly shared holidays (at least 10 countries) with country-specific holidays in high- versus low-population countries. As shown in Figure 2, we observe that models tend to exhibit a higher accuracy when dealing with unique holidays from high-population countries as opposed to low-population ones. This outcome suggests that LLMs face challenges in capturing statistical trends related to under-represented geo-cultural concepts.

Possibility of LLMs embedding geo-cultural proximity? As countries with similar cultures tend to share common holidays, the cultural disparity between the query and target countries can influence LLM predictions. To investigate this, we group QA pairs by query locations and analyze model performance across distinct geo-groups. Given the dispersed distribution of QA pairs at the country level, we concentrate on continent-based comparisons using the optimal model, GPT-3.5.

Figure 3 presents 5-item ranking results for holiday popularity in the US, while Appendix H includes results for the UK. In Figure 3, Oceania shows the lowest ranking differences, whereas Asia exhibits the highest. The predictability of GPT-3.5 aligns with geo-cultural proximity across continents, as seen in the Inglehart–Welzel cultural map Inglehart and Welzel (2010). The Eastern culture dominant in Asia is more distant from the Western culture shared by Europe, Oceania, and the Americas. According to the cultural map, major Oceania countries (Australia and New Zealand) share a cultural group with the US. Despite the US being in North America, sampled non-US data in this continent mainly comes from low-population countries (e.g., Belize and Greenland), posing a potential challenge for GPT-3.5 predictions.

5 Conclusion

We introduce a novel QA task, CPopQA, to assess LLMs in ranking holiday-centered cultural concepts based on their popularity in the US and UK. Our results show that LLaMA and GPT-3.5 tend to capture implicit statistical tendencies of holiday popularity. Comparatively, GPT-3.5 displays superior ranking abilities. The model predictions are highly sensitive to the number of ranked items, and they encounter more difficulties in capturing statistical trends for under-represented geo-cultural concepts than well-represented ones. Further exploring the optimal LLM (GPT-3.5), we observe its potential to show geo-cultural proximity across continents. By examining LLMs' statistical ranking ability on long-tail cultural knowledge, this preliminary work benefits incentivizing future work on sociocultural tendency exploration by LLMs.

6 Limitations

With a systematic review of our study, we summarize a list of limitations as follows.

First, regarding the holiday list, since we curated the holiday list based on Wikipedia, the potential data biases in Wikipedia such as missing holidays and countries, and misrepresentation of communities may cause issues of data representativeness in our dataset. Moreover, despite the diverse countries considered in this study, we focused on a sample of countries based on the accessible data from Wikipedia. The limited coverage of geo-political regions may also lead to unwanted data biases.

Second, with respect to the holiday popularity collection, there may exist two concerns with the employment of Google Ngram Viewer to estimate holiday popularity. One is about the OCR quality of machine-digitized books, which may influence the n-gram statistical results. However, the tool developers have carefully considered this issue when building the tool (Michel et al., 2011) and the later version further updated the OCR technology to improve the corpus quality (Lin et al., 2012). Considering the corpus in Google Ngram Viewer mainly consists of Google books, the other concern is about the domain shift issue. We will extend our study to consider diverse web resources for n-gram statistics in the future.

Third, in this preliminary study, we mainly focus on the use case of holiday popularity to investigate LLMs' potential on ranking-based statistical analysis questions. Moreover, the prompting template is simple as our study emphasizes the fundamental ability of LLMs in CPopQA. In the future, we will consider more diverse cultural concepts and a variety of prompting strategies for model evaluation.

References

Anurag Acharya, Kartik Talamadupula, and Mark A Finlayson. 2020. Towards an atlas of cultural commonsense for machine reasoning. *arXiv preprint arXiv:2009.05664*.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023a. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023b. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Manvi Breja and Sanjay Kumar Jain. 2022. Analyzing linguistic features for answer re-ranking of why-questions. *Journal of Cases on Information Technology (JCIT)*, 24(3):1–16.

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683, Osaka, Japan. The COLING 2016 Organizing Committee.

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 (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Sameera Horawalavithana, Ellyn Ayton, Shivam Sharma, Scott Howland, Megha Subramanian, Scott

Vasquez, Robin Cosbey, Maria Glenski, and Svitlana Volkova. 2022. Foundation models of scientific knowledge for chemistry: Opportunities, challenges and lessons learned. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 160–172, virtual+Dublin. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Ronald Inglehart and Chris Welzel. 2010. The wvs cultural map of the world. *World Values Survey*.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Indra Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. *arXiv preprint arXiv:2305.16171*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. Large language models struggle to learn long-tail knowledge. *arXiv preprint arXiv:2211.08411*.

Bernhard Kratzwald, Anna Eigenmann, and Stefan Feuerriegel. 2019. RankQA: Neural question answering with answer re-ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6076–6085, Florence, Italy. Association for Computational Linguistics.

Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2022. A systematic investigation of commonsense knowledge in large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11838–11855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books NGram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174, Jeju Island, Korea. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339, Online. Association for Computational Linguistics.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48, Vancouver, Canada. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. 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 (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Parker Riley, Timothy Dozat, Jan A Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2022. Frmt: A benchmark for few-shot region-aware machine translation. *arXiv preprint arXiv:2210.00193*.

Dor Ringel, Gal Lavee, Ido Guy, and Kira Radinsky. 2019. Cross-cultural transfer learning for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3873–3883, Hong Kong, China. Association for Computational Linguistics.

Carl W Roberts. 2020. *Text analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts*. Routledge.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, and et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R. Mortensen. 2021. Cross-cultural similarity features for cross-lingual transfer learning of pragmatically motivated tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2403–2414, Online. Association for Computational Linguistics.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. 2017. Do neural nets learn statistical laws behind natural language? *PloS one*, 12(12):e0189326.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. 2019. Evaluating computational language models with scaling properties of natural language. *Computational Linguistics*, 45(3):481–513.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ted Underwood and Richard Jean So. 2021. Can we map culture? *Journal of Cultural Analytics*, 6(3).

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sheng Zhang, Xin Zhang, Weiming Zhang, and Anders Søgaard. 2021. Sociolectal analysis of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4588, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Related Work

Cultural-aware NLP Language and culture are intertwined (Hershcovich et al., 2022; Hovy and Yang, 2021). Overall, research on the interaction of language technologies and cultures can be divided into two groups. The first group focuses on improving specific language technologies inspired by cultural diversity (Sun et al., 2021; Jurgens et al., 2017; Riley et al., 2022). For example, Jurgens et al. (2017) proposed a new distance measure between languages based on linguistic proxies of culture, hoping to improve cross-lingual transfer learning. Riley et al. (2022) constructed a benchmark called FRMT to improve matching translation with an emphasis on geo-cultural diversity.

The second group concentrates on investigating the cultural awareness of language technologies (Arora et al., 2023b; Ringel et al., 2019; Garimella et al., 2016). Popular research topics in this thread include cross-cultural differences in word usage (Garimella et al., 2016), dialect-associated biases (Zhang et al., 2021), and geo-diverse commonsense (Liu et al., 2021; Acharya et al., 2020; Yin et al., 2022).

Ranking-based QA Existing work in the field of ranking-based QA primarily focus on answer re-ranking to identify the optimal one (Nakov et al., 2017; Breja and Jain, 2022; Kratzwald et al., 2019). Following Rogers et al. (2023), one of the major motivations behind this group of studies lies in the diversity in both the quality and quantity of questions and answers. Differing from prior studies that focus on developing ranking-based QA models to identify the best answer from a pool of candidates, our study specifically centers around a QA task that aims to generate a ranking of cultural concepts (holidays) based on their popularity.

B Data Cleaning

We conducted both rule-based cleaning and post-human edition to improve the data quality. Specifically, we filtered out holidays that lose the time description for further consideration. Regarding temporal diversity, we employed two human annotators to unify the holiday date following Gregorian Calendar. Considering the temporal dynamics of some holidays caused by calendar conversion, we further required annotators to assign the label "movable" to these holidays' dates. Moreover, the paraphrase phenomenon of some holidays may cause their popularity distribution to be dispersed. To avoid this issue, we examined each holiday concept and grouped its aliases. In addition, through the empirical examination of the holiday list, annotators also removed false positives (e.g., special events like the memory of an emperor) and improved holiday descriptions by manual edition.

C Human Evaluation of Holiday Popularity Collection

To further validate our strategy for deriving holiday popularity, we additionally conducted a human evaluation of holiday popularity rankings. Specifically, we randomly sampled 5 countries and selected the top 10 holidays per country based on their frequency in GBNV's American English corpus. For each country's holiday list, we asked 6 non-immigrant US citizens, who grew up in the US, to compare holidays regarding their popularity in the US and generated a rank based on annotators' average votes. Toward a correlation analysis of two ranked holiday lists per country, our results show that the statistics of Google Books Ngram Viewer achieved 60% consistency (i.e., Pearson p=63.34%, Spearman rho=58.65%) with human judgments on average.

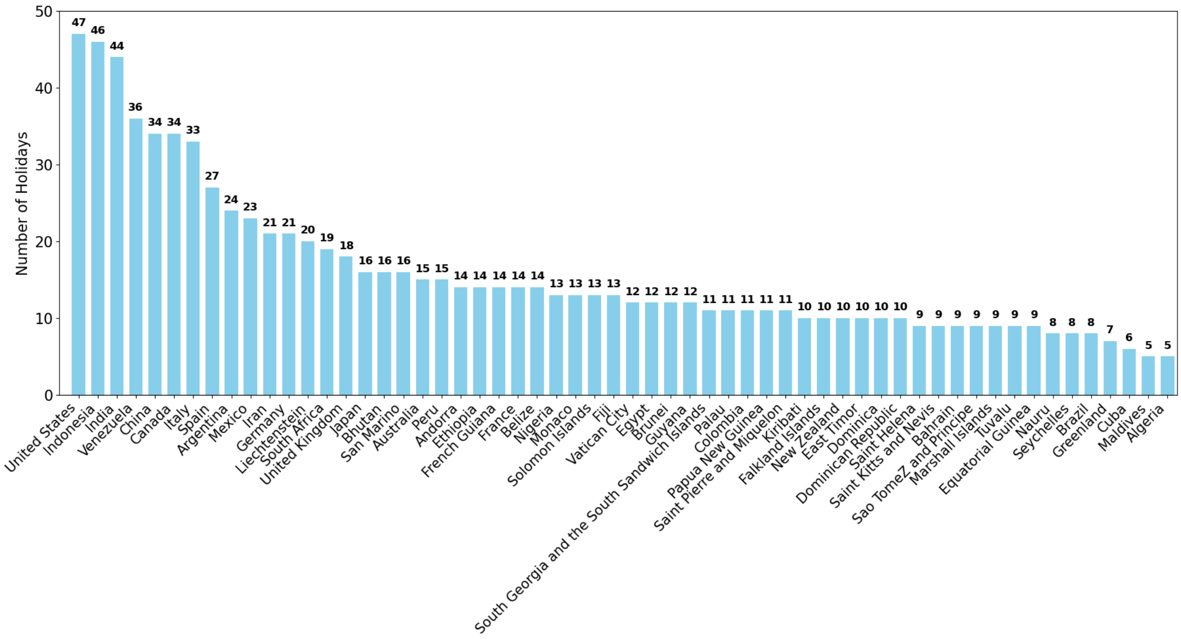

Figure 4: Distribution of holidays by country in descending order.

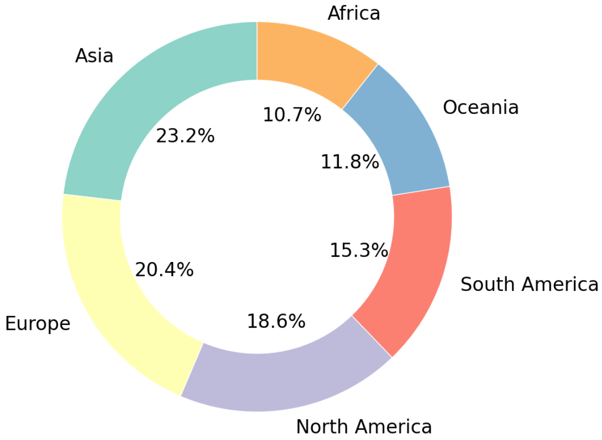

Figure 5: Distribution of holidays by continent.

D Holiday Statistics

Figure 4 displays the distribution of holidays across various countries. Our dataset comprises a total of 459 unique holidays in 58 countries. Notably, the United States, Indonesia, and India are the top three countries with the highest number of holidays. Conversely, Cuba, Maldives, and Algeria have the lowest number of holidays among the countries included in our dataset.

Figure 5 presents the distribution of holidays by continent. In comparison, Asia (23.2%) and Europe (20.4%) emerge as the top two continents with a higher number of holidays compared to others. On the other hand, Africa (10.7%) and Oceania (11.8%) have a relatively lower ratio of holidays in comparison to the other continents.

E Details of LLM Tuning and Baselines

All the experiments are built upon an RTX3090. We tuned LLMs with the optimal temperature values and the default values of the other hyperparameters. Specifically, we used a temperature of 1 for BLOOM, 0.7 for LLaMA, and 0.3 for GPT-3.5.

Regarding baselines, in addition to widely used random guess, we proposed two additional baselines. For the first baseline which we called "google stat", we quantified the general holiday popularity by querying each holiday in a search engine and estimating its cumulative search volume over time. Given that Google is one of the most popular search engines, we used Google Trends (https://trends.google.com/trends/) to access the search volume of each holiday query in Google Search across 20 years (2004-1-1 - 2023-8-1, the maximum accessible timeline in the tool) and sum up the statistics over the selected time span. The second baseline is called "wiki len". With the assumption that a well-known holiday tends to contain a more comprehensive and lengthy description compared to a lesser-known one, we calculated the word length of the description on each holiday's Wikipedia page to approximate the holiday's popularity.

F LLM Investigation on Country Ranking

In addition to fine-level holiday ranking, we further explore LLM performance on a high-level country ranking, which is defined as below:

- **Task 2. Country ranking:** Given a set of countries $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$, this task aims to sort $\mathcal{C}$ in a descending order based on each country's overall holiday popularity in a target country c_t.

Differing from Task 1 where we explore holiday ranking in both the US and the UK, we specifically concentrate on the country ranking in the US in Task 2, as we found that there is a high agreement on ranking countries by their overall holiday popularity in the US and UK.

Table 3 displays the overall performance of LLMs in this level of ranking.

Setting	Model	P@1 (%)	Acc. (%)	Diff.
2-item ranking	random guess	50.00 ± 0.00	50.00 ± 0.00	-
	google stat	80.10 ± 0.03	80.10 ± 0.03	-
	wiki len	**80.90** ± 0.02	**80.90** ± 0.02	-
	bloom-7b1	42.80 ± 0.02	42.80 ± 0.02	-
	llama-7b	53.20 ± 0.03	53.20 ± 0.03	-
	llama-13b	52.50 ± 0.02	52.50 ± 0.02	-
	gpt-3.5	60.90 ± 0.04	60.90 ± 0.04	-
3-item ranking	random guess	33.33 ± 0.00	16.67 ± 0.00	0.500 ± 0.00
	google stat	52.90 ± 0.03	42.20 ± 0.02	0.235 ± 0.02
	wiki len	**55.40** ± 0.03	**45.70** ± 0.04	**0.224** ± 0.03
	bloom-7b1	30.30 ± 0.03	15.20 ± 0.01	0.470 ± 0.02
	llama-7b	37.80 ± 0.02	18.40 ± 0.04	0.481 ± 0.02
	llama-13b	36.60 ± 0.04	17.70 ± 0.03	0.466 ± 0.02
	gpt-3.5	48.70 ± 0.01	25.10 ± 0.03	0.398 ± 0.01
5-item ranking	random guess	20.00 ± 0.00	0.83 ± 0.00	0.500 ± 0.00
	google stat	39.40 ± 0.02	9.90 ± 0.02	0.225 ± 0.01
	wiki len	**47.70** ± 0.04	**16.90**± 0.02	**0.221** ± 0.01
	bloom-7b1	14.10 ± 0.03	0.60 ± 0.01	0.514 ± 0.03
	llama-7b	27.10 ± 0.02	0.50 ± 0.00	0.490 ± 0.03
	llama-13b	27.40 ± 0.03	1.60 ± 0.01	0.450 ± 0.03
	gpt-3.5	42.10 ± 0.02	3.70 ± 0.01	0.353 ± 0.01

Table 3: Performance of LLMs on country rankings regarding their overall holiday popularity in the US (mean ± standard deviation).

G Consistency of Holiday Rank in US versus UK

Table 4 displays the level of ranking consistency between the popularity of holidays in American culture compared to British culture. In particular, the agreement metric measures the extent of the exact match between the rank in the US and the rank in the UK. We also compute the average ranking difference across queries.

	Agreement (%)	Diff.
2-item ranking	93.00	-
3-item ranking	75.80	0.073
5-item ranking	39.20	0.061

Table 4: Consistency of holiday popularity rank in US versus UK.

H Results of LLMs on Holiday Ranking in the UK

Setting	Model	P@1 (%)	Acc. (%)	Diff.
2-item ranking	random guess	50.00 ± 0.00	50.00 ± 0.00	-
	google stat	56.40 ± 0.03	56.40 ± 0.03	-
	wiki len	**61.30** ± 0.04	**61.30** ± 0.04	-
	bloom-7b1	39.40 ± 0.04	39.40 ± 0.04	-
	llama-7b	46.80 ± 0.02	46.80 ± 0.02	-
	llama-13b	49.50 ± 0.03	49.50 ± 0.03	-
	gpt-3.5	53.50 ± 0.04	53.50 ± 0.04	-
3-item ranking	random guess	33.33 ± 0.00	16.67 ± 0.00	0.500 ± 0.00
	google stat	36.30 ± 0.03	20.00 ± 0.02	0.448 ± 0.01
	wiki len	54.60 ± 0.03	29.80 ± 0.03	0.361 ± 0.02
	bloom-7b1	31.90 ± 0.01	16.60 ± 0.03	0.481 ± 0.01
	llama-7b	37.60 ± 0.03	17.00 ± 0.03	0.466 ± 0.01
	llama-13b	37.10 ± 0.03	19.80 ± 0.04	0.455 ± 0.02
	gpt-3.5	**62.60** ± 0.01	**38.80** ± 0.01	**0.278** ± 0.01
5-item ranking	random guess	20.00 ± 0.00	0.83 ± 0.00	0.500 ± 0.00
	google stat	23.80 ± 0.01	2.40 ± 0.01	0.431 ± 0.02
	wiki len	53.10 ± 0.04	2.90 ± 0.01	0.322 ± 0.01
	bloom-7b1	18.10 ± 0.03	0.80 ± 0.01	0.477 ± 0.02
	llama-7b	29.50 ± 0.03	1.50 ± 0.01	0.462 ± 0.01
	llama-13b	28.30 ± 0.03	2.20 ± 0.00	0.420 ± 0.03
	gpt-3.5	**60.60** ± 0.04	**7.60** ± 0.02	**0.258** ± 0.01

Table 5: Performance of LLMs on worldwide holiday popularity rankings in the UK (mean ± standard deviation).

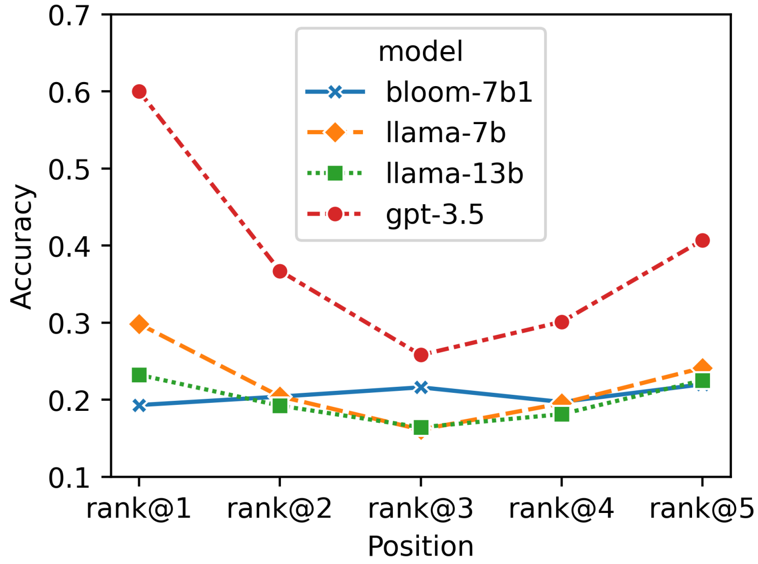

Figure 6: LLM results at each position on holiday ranking in the UK.

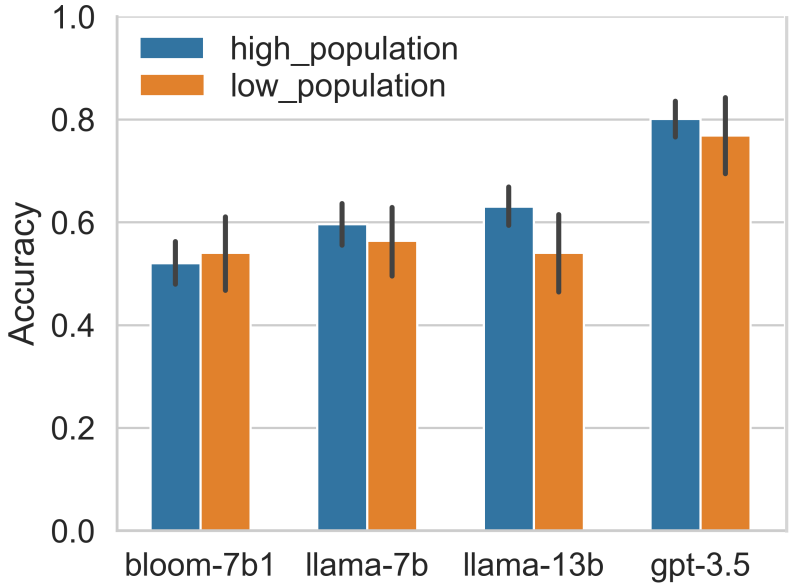

Figure 7: Pairwise ranking accuracy in the UK regarding geo-cultural representativeness

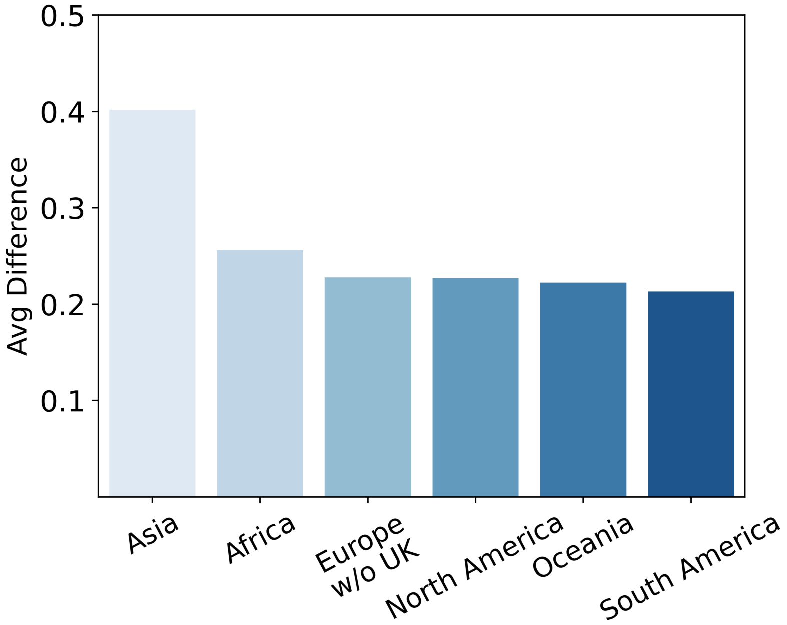

Figure 8: The results of GPT-3.5 on holiday ranking in the UK across continents.

I Detailed Examples of QA Pairs

In Table 6 and Table 7, we provide a few detailed examples of QA pairs for holiday ranking and country ranking, respectively. Note that, we don't use the popularity information or holiday descriptions in our prompts. However, such information can be served as valuable context for future studies.

I.1 Holiday Ranking

No.	Attribute	Content
1	Question	Can you provide a descending order for the following Chinese holidays by their popularity in the United States: Spring Festival, Children's Day, Lantern Festival, Martyrs' Day, Dragon Boat Festival?
	Answer	1. Spring Festival 2. Children's Day 3. Dragon Boat Festival 4. Lantern Festival 5. Martyrs' Day
	Holiday Popularity	The percentage of the frequency of the holiday Chinese New Year / Spring Festival mentioned in the United States corpus is 1.11e-07. The percentage of ...
	Description	Chinese New Year / Spring Festival: Chinese New Year is the festival that celebrates the beginning of a new year on the traditional lunisolar Chinese calendar. In Chinese, the festival is commonly referred to as the Spring Festival...
2	Question	Can you provide a descending order for the following Indian holidays by their popularity in the United States: Independence Day, Gandhi Jayanti, Bihu, Accession Day, Day of Ashura?
	Answer	1. Independence Day 2. Accession Day 3. Bihu 4. Day of Ashura 5. Gandhi Jayanti
	Holiday Popularity	The percentage of the frequency of the holiday Independence Day mentioned in the United States corpus is 4.07e-07. The percentage of ...
	Description	Independence Day: Independence Day is celebrated annually on 15 August as a public holiday in India commemorating the nation's independence from the United Kingdom...
3	Question	Can you provide a descending order for the following Fijian holidays by their popularity in the United States: New Year's Day, Ram Naumi, Fiji Day, National Youth Day, Palm Sunday?
	Answer	1. New Year's Day 2. Palm Sunday 3. National Youth Day 4. Fiji Day 5. Ram Naumi
	Holiday Popularity	The percentage of the frequency of the holiday New Year's Day mentioned in the United States corpus is 5.78e-07. The percentage of ...
	Description	New Year's Day: In the Gregorian calendar, New Year's Day is the first day of the year ...

Table 6: QA pair examples for 5-holiday ranking in the US.

I.2 Country Ranking

No.	Attribute	Content
1	Question	Can you provide a descending order for the following foreign countries by their holiday popularity in the United States: Brazil, Guyana, Egypt, New Zealand, China?
	Answer	1. Brazil 2. China 3. Guyana 4. New Zealand 5. Egypt
	Country Popularity	The percentage of the total frequency of holidays in Brazil mentioned in the United States corpus is 1.82e-05....
2	Question	Can you provide a descending order for the following foreign countries by their holiday popularity in the United States: Vatican City, Canada, Australia, Japan, Greenland?
	Answer	1. Australia 2. Canada 3. Greenland 4. Vatican City 5. Japan
	Country Popularity	The percentage of the total frequency of holidays in Vatican City mentioned in the United States corpus is 2.31e-06...
3	Question	Can you provide a descending order for the following foreign countries by their holiday popularity in the United States: Fiji, Greenland, India, Solomon Islands, Mexico?
	Answer	1. Mexico 2. Fiji 3. Greenland 4. Solomon Islands 5. India
	Country Popularity	The percentage of the total frequency of holidays in Fiji mentioned in the United States corpus is 6.4e-06...

Table 7: QA pair examples for 5-country ranking in the US.