No Culture Left Behind: Massively Multi-Cultural LM Knowledge Acquisition Benchmarking on 1000+ Sub-Country Regions and 2000+ Ethnolinguistic Groups

Anonymous ACL submission

Abstract

Pretrained large language models have revolutionized many applications but still face challenges related to cultural bias and a lack of cultural commonsense knowledge crucial for guiding cross-culture communication and interactions. Recognizing the shortcomings of existing methods in capturing the diverse and rich cultures across the world, this paper introduces a novel approach for massively multicultural knowledge acquisition. Specifically, our method strategically navigates from densely informative Wikipedia documents on cultural topics to an extensive network of linked pages. Leveraging this valuable source of data collection, we construct the **CultureAtlas** dataset, which covers a wide range of sub-country level geographical regions and ethnolinguistic groups, with data cleaning and preprocessing to ensure textual assertion sentence self-containment, as well as fine-grained cultural profile information extraction. Our dataset not only facilitates the evaluation of language model performance in culturally diverse contexts but also serves as a foundational tool for the development of culturally sensitive and aware language models. Our work marks an important step towards deeper understanding and bridging the gaps of cultural disparities in AI, to promote a more inclusive and balanced representation of global cultures in the digital domain.[1]

1 Introduction

Pretrained large language models (LMs) are increasingly used in applications across diverse domains, ranging from question answering (Brown et al., 2020; Gangi Reddy et al., 2022) and chatbots (Lin et al., 2020; Ouyang et al., 2022) to content recommendation (Wu et al., 2020) and norm violation detection (Fung et al., 2023). However, as their usage proliferates, an important concern emerges

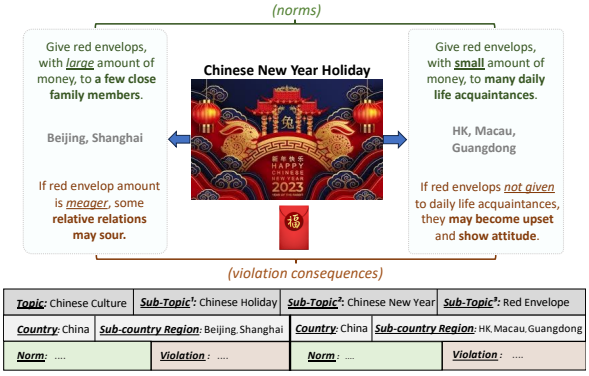

Figure 1: An example illustration of different cultural practices, with regards to Chinese New Year red envelopes across different geographical regions such as *Beijing/Shanghai* versus *HK/Macau/Guangdong*.

– the potential for cultural bias and misappropriation (Hershcovich et al., 2022; Palta and Rudinger, 2023; Li et al., 2023b). LMs, when not equipped with geo-diverse knowledge and cultural sensitivity, can exhibit significant disparities in performance when faced with textual data from different regions. This imbalance disadvantages certain users groups and exacerbates existing biases in NLP applications, perpetuating the presently predominant Western-centric perspectives and knowledge bases. Hence, tackling this issue is paramount not only for promoting fairness and inclusivity, but also for fostering a more culturally aware and harmonious digital landscape.

Previous approaches in benchmarking and improving the cross-cultural knowledge of language models tend to either 1) focus on a predefined, narrow set of coarse-grained cultures and cultural topics (Yin et al., 2022), or 2) discover cultural knowledge from large noisy corpora (Nguyen et al., 2023) in which important cultural elements often get filtered out in the data processing stage or lost as cultural differences get intermingled, leading to a scenario where LMs fail to learn information specific to individual subregions. Our goal is to investigate and set the grounds for empowering language

[1] Our code and resource will be released for publication.

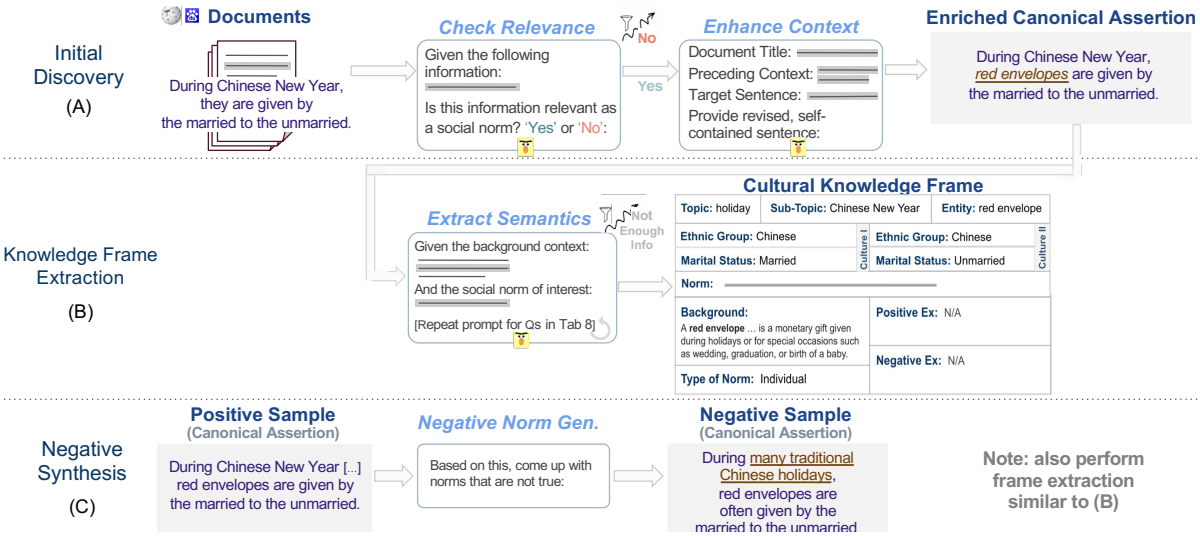

Figure 2: An overarching view of our **CultureAtlas** benchmark construction process.

model with reasoning capability on finer-grained cultural nuances that pertain to different cultural subgroups and deeper topic coverage. For example, as depicted in Figure 1, Chinese New Year red envelop gifting practices vary by geographical subregions within a same country, which may potentially lead to norm violations for people newly settling into an area who are unaware of the common cultural practice that differ by region or recipient group.

In this work, we propose a culture knowledge acquisition process for constructing a novel benchmark for assessing language models' massively multicultural reasoning capabilities. In particular, our culture knowledge acquisition process seeks to combine the best of both worlds between *bottom-up discovery* of culture knowledge discovery from the open web documents (relatively noisy but large-scale data) and *top-down discovery* of culture knowledge from targeted topic guidance (relatively clean but limited data). We start from Wikipedia documents as our source of data, chosen for their clean nature as their contents inside are subject to public audits and back-and-forth information edits to strip away controversies until common ground is reached. Specifically, we include the documents for each country that revolves around an initial set of cultural topics, including education, dating/marriage, and holiday customs, among others. We then continue to expand on the relevant document sets based on linked topic pages within (*e.g.,* "Chinese culture"->"Chinese Holidays"->"Chinese New Year Holiday"->"Red Envelope"), and consider the sentences in the documents, in which a pretrained LM categorizes it as a generalizable social or cultural norm rather than instance-specific history or fact, to be the positive samples of cultural knowledge in our dataset. We also propose an effective method to construct negative (*i.e.,* non-factual) cultural knowledge samples, cross-validated through web search, for the purpose of probing language model cultural reasoning robustness. Furthermore, we perform information extraction on these positive and negative cultural knowledge samples, to derive fine-grained cultural profile fields, including sub-country geographical regions, ethno-linguistic identity, demographics, etc., for enabling deeper analysis on situationalized socio-cultural context frames.

Our contributions can be summarized as follows:

- We present a novel data collection process to construct a massively multicultural knowledge benchmarking dataset, **CultureAtlas**, spanning significantly greater diversity in coverage of sub-country level geographical regions, ethno-linguistic groups, etc., compared to prior benchmarks in the multi-cultural NLP domain (Yin et al., 2022; Nguyen et al., 2023; Ziems et al., 2023).
- This dataset comprises of high-quality positive data samples and negative data samples, with a $90^{+}\%$ pass rate in data quality check via human assessment.
- Finally, we evaluate the performance of state-of-the-art foundation language models on **CultureAtlas**, and demonstrate this new dataset as

a useful resource for identifying rooms for improvement in LM cultural awareness and debiasing.

2 Benchmark Construction

2.1 Data Collection and Data Preprocessing

Constructing a benchmark for assessing the fine-grained cultural knowledge of language models is the first and most important step in enabling training language models to incorporate better cultural knowledge downstream via finetuning. To achieve this goal, we construct a novel dataset, **CultureAtlas**, by collecting positive and negative samples of cultural knowledge assertions that span diverse geographical subregions and ethnolinguistic groups, with subsequent data processing as illustrated in Figure 2.

Positive Data Samples We start from the observation that cultural webpages from publicly monitored human-curated sources, such as Wikipedia, contain clean and commonly accepted cultural assertions in general. These cultural sources are dense in information, and, while not yet entirely comprehensive, serves as a valuable initial source of information and seedling for further expansion. First, we consider the set of documents from all countries worldwide. So we systematically explore culturally relevant topics (e.g., culture, holidays, dining etiquette, dating and marriage, education, honorifics, etc.) for each country, using the Wikipedia API[2] tool to match and download corresponding Wikipedia pages. We expand on this target set of documents by further including the hyperlinked document pages up to two hops down. In addition to incorporating the default English documents of these pages, we also include their document versions in the main language corresponding to the culture in discussion, and translate the text content into English. This ensures a more well-rounded understanding of cultural nuances and perspectives, as complementary information exists across language versions for the same document topic. Next, we process the corpus sentence by sentence, first filtering out sentences that focus on very specific, socio-culturally non-generalizable events or instances (see Appendix A). We refine these sentences into self-contained cultural knowledge assertions by eliminating ambiguous pronoun or phrase references and enriching each sentence with any necessary information from the preceding context in the same paragraph. Together, these steps constitute the **initial discovery** of cultural knowledge assertions in **CultureAtlas**.

Because each unique combination of population dimension – ethnicity, language, location, demographic background, etc. – plays a key role in shaping distinct cultural practices, we proceed to better discern subtle situational differences in norms across various cultures by **extracting cultural knowledge frames** for each remaining sentence, with a fine-grained profiling approach covering the following fields of information element:

- *country location*
- *sub-country regional location* – cities, states, and provinces under the GeoNames[3] knowledge base.
- *ethnicity* – ethno-linguistic groups from the ISO 639-3 code table[4].
- *religion* – all religious groups and denominations with a population of 1 million followers or more
- *age* – {infant, young children, teenager, young adult, adult, elderly}
- *gender* – {male, female, other}
- *marital status* – {single, engaged, married, divorced, widowed}
- *occupation* – open-domain fill in the blank

with *age*/*gender*/*marital status*/*occupation* pertaining to any person entities involved in the norms. As depicted in Fig 2(A), language model prompting is utilized to extract the values for these fields automatically, via a directed question such as "[norm] \n *Which gender group is mentioned or implied in the sentence (male, female, transgender, other, or N/A):*", with further details shown in Appendix B. Note that information unknown or not mentioned is regarded as "*N/A*".

Negative Data Samples In order to evaluate LM cultural knowledge, we prepare the data setting with negative norm synthesis as illustrated in Fig 2(c). Basically, we take a pristine original norm assertion and manipulate it through LLM prompting for adversarial knowledge via the template of: *"[orig. norm] \n Based on this topic, come up with norms that are not true:"*. To ensure the negative norm generation is indeed a non-factual fabrication, we perform automatic verification, which is easy

[2] https://pypi.org/project/Wikipedia-API/

[3] https://www.geonames.org/

[4] 1https://iso639-3.sil.org/code/oci

	CANDLE (Nguyen et al., 2023)	GeoMLAMA (Yin et al., 2022)	NormsKB (Fung et al., 2023)	NormBank (Ziems et al., 2023)	CultureAtlas (Ours)
# countries	176	5	5	160	**193**
# local regions					
(state/province-level)	298	0	12	102	**1089**
(city-level)	1,376	0	15	493	**10,436**
# religion	14	0	3	30	**42**
# ethnolinguistic groups	298	5	10	551	**2,557**
fine-grained norm framing	x	x	✓	✓	✓
multi-ling. data source	x	x	x	x	✓

Table 1: Our data collection of cultural norms contains greater coverage in *local regions* and *ethno-linguistic groups*, compared to previous work. It also involves data from multi-lingual sources as well as fine-grained cultural knowledge frame extraction.

Original Cultural Knowledge	Negative Cultural Knowledge Generated
During the Chinese New Year, in **Southern China**, **red envelopes** are typically given by the married to the unmarried, most of whom are children.	In **China**, it is customary for students to present their teachers with **red envelopes** containing handwritten notes of gratitude at the end of each school term, symbolizing respect and appreciation for their guidance.
In **Bhutan culture** for special occasions and festivals, colourfully patterned silk kira and, more rarely, gho may be **worn**.	In **Bhutan**, there is a unique tradition of **wearing** "Khyenkhor Robes", woven with threads infused with blessings from Buddhist monks, during special ceremonies and festivals.

Table 2: Visualization of the original positive data samples and negative data synthesized counterpart, in our **CultureAtlas** benchmark construction.

to scale. Specifically, we make use of a language model self-check mechanism, asking the question of *"[norm] \n Is this absurd and/or very hard to believe? 'Yes' or 'No':"* to GPT4, to filter out negative sample candidates that are obviously absurd to believe. In particular, our motivation in leveraging GPT4 is that it stands as the most advanced LM backbone, with notable performance gap for open-sourced LMs or other propriety LMs to bridge in and thereby deeming our benchmark negative samples especially valuable. Subsequently, we follow with a web-check mechanism – web retrieving relevant context and ensuring no entailment is found. Examples of negative data samples, along with its original positive form, are visualized in Table 2.

2.2 Quality Check on Data

To ensure data quality, we perform manual assessment on the **CultureAtlas** dataset construction. Specifically, we take 10 random samples for each intermediary data processing step of the positive data and negative data respectively, and ask five human judges familiar with the subject matter (*e.g.*, self-identifying with geographical regions and cultural subgroups across US, China, Korea, India) to determine whether the data samples that is indeed either a correct cultural knowledge assertion when their ground truth label (based on our expectation from the Sec 2.1 procedure) is "TRUE" or an incorrect cultural knowledge assertion when their label is "FALSE". In our quality check assessment guidelines, we clarify examples of poor positive samples, such as having ambiguous pronoun references or lacking culture-specificity in non-universal norms, as well as examples of poor negative samples, such as contradicting known norms. Our study has been approved by an ethics review. Finally, as we can see in the qualitative results of our dataset construction reported in Table 3, the final post-processed positive and negative samples are high-quality, achieving $90^{+}\%$ pass rate. The interannotator agreement is 0.79.

Approach	Pass Rate (%)
Pos. Data	
- *Orig Sent.*	49.5
- *Post Proc. Sent.*	**93.2**
Neg. Data	
- *Direct Gen.*	81.1
- *Direct Gen. w/ self-check*	90.1
- *Direct Gen. w/ self-check & web check*	**92.0**

Table 3: The average pass rate for **CultureAtlas** dataset samples, at each processing step, based on human validation of post-processed cultural knowledge assertions.

2.3 Descriptive Stats

As shown in Table 1, our dataset covers over $1,089$ state or province level regions, $10,436$ city level regions, and $2,557$ ethnolinguistic groups, significantly exceeding prior work (Nguyen et al., 2023;

Yin et al., 2022; Fung et al., 2023) in the cultural knowledge for NLP domain. We provide detailed information on the # of subregion specific cultural knowledge frames per country in Table 9 of Appendix B, and the # of cultural knowledge frames per ethnolinguistic group in Table 10 of Appendix B, due to page restrictions. In particular, as an example, all 56 official ethnic groups of China (*e.g., Han*, *Zhuang*, *Hui*), as well as the linguistically distinct ethnic subgroups (*e.g., Yue* and *Hakka* of the Chinese *Han* population), are included in **CultureAtlas**.

Training-Testing Data Split Table 4 summarizes the number of culturally related document pages scraped and sentences parsed, as well as post norm-relevance filtering cultural knowledge assertion sentences and frame extractions. We partition 10,000 random samples of cultural knowledge assertions that are particularly relevant for avoiding norm violations to constitute the **test set**. All other data are provided for future language model training and development purpose.

# of doc pages	41k
# of sent parsed	907k
# of sent, generalizable sociocultural knowledge	127k
# of sent, norm violation relevant w/ frame extract.	21k

Table 4: Size and scale of our collected dataset, where 'k' represents the kilo unit of a thousand data.

3 Experiments

3.1 Task Setting

We evaluated the cultural knowledge and reasoning capability of state-of-the-art pretrained large language models (LLMs) on the canonical norm descriptions in our constructed benchmark, through a true-or-false binary classification setup. As a reminder, the derivation of ground truth labels for "correct "cultural knowledge assertion samples and "incorrect" cultural knowledge assertion samples have been detailed under Sec 2.1 ("Positive Data Samples" and "Negative Data Samples" paragraphs).

3.2 Model Setup

For the choice of language model in our experiments, we consider the present day commonly-used open-source language model backbones:

- **Llama-2** (Touvron et al., 2023b) - an open source foundation model from Meta that is trained on 40% more data compared to its predecessor, LLaMA (Touvron et al., 2023a). For each backbone size, we further include its variant of **-chat**, denoting finetuned with dialogue data, and **-hf**, denoting trained with human preference alignment data.
- **Vicuna** (Chiang et al., 2023) - an open source foundation model trained by fine-tuning LlamA-2 on 700k samples of user conversations with ChatGPT collected from the ShareGPT website [5]. This model series is optimized for efficiency through gradient checkpointing and flash attention.

We also consider the propriety closed-source language model backbones, which may generally tend to have higher performance compared to publicly available open-source model checkpoints but make research development and transparency challenging:

- **ChatGPT** (Ouyang et al., 2022) - a closed-source foundation model from OpenAI, trained *with alignment* data.

Our LM experiments are conducted in a zero temperature setting to remove randomness factor.

3.3 Results

Table 5 shows the results of our LM benchmarking. In particular, we notice several interesting observations. First, we find that out of the open-source models, Vicuna consistently perform better than Llama2 in cultural knowledge when comparing across the same model backbone sizes. This indicates that the training approach (*e.g.,* choice of training data, optimization objective, etc.) plays a key role in the cultural knowledge reasoning capability of these LLMs. Secondly, we find that pre-existing explicit human feedback (HF) alignment approaches does not necessary improve model performance in massively multicultural fine-grained reasoning domains, potentially due to non-desirable domain shift and catastrophic forgetting. This reaffirms the values in our new benchmark proposal, for continuously measuring cultural awareness progress in future language model development. We also observe a general positive correlation between model performance in cultural-aware inference and model parameter size, as expected.

In addition, we investigate LLM performance in culture reasoning patterns across resource availabil-

[5] https://sharegpt.com/

			All Culture			High Resource			Mid Resource			Low Resource		
			P	R	F	P	R	F	P	R	F	P	R	F
Llama2	**7B**	*chat*	84.2	42.1	56.1	86.8	45.6	59.8	83.3	42.9	56.6	87.0	20.7	33.5
		chat-HF	75.1	28.2	41.0	76.9	26.9	39.9	83.3	28.0	41.9	78.9	26.2	39.2
	13B	*chat*	63.6	77.1	69.7	56.1	80.9	66.3	64.1	75.5	69.3	53.3	20.5	29.6
		chat-HF	89.9	20.0	32.7	91.8	91.8	28.3	91.8	20.6	33.6	92.2	19.3	31.9
Vicuna	**7B**	*chat-HF*	79.6	56.8	66.3	77.3	47.2	58.6	79.4	57.9	67.0	81.3	55.7	66.1
	13B	*chat-HF*	67.4	81.2	**73.7**	68.9	81.0	**74.5**	69.4	82.4	**75.3**	67.8	82.3	**74.3**
ChatGPT	**20B**	*chat-HF*	95.8	90.6	93.1	95.9	91.4	93.6	94.9	92.1	93.5	94.1	90.1	92.1

Table 5: Experimental results on benchmarking state-of-the-art foundation large language model performance on the new **CultureAtlas** cultural knowledge assessment benchmark. Precision (P), recall (R), F-score (F) scores are reported.

ity and topic domains. As shown in Table 5, we investigated LM awareness in the cultural knowledge pertaining to country-level 'high-resource' (*e.g.,* US/China/France/Spain/Japan), 'mid-resource' (*e.g.,* Turkiye/Egypt/Iran/Malaysia/Argentina), and 'low-resource' (*e.g.,* Lao/Bhutan/Congo/Serbia) culture groups, as categorized by societal-wide economic development, which in turn affects the linguistic resources availability for constituting LM training data.

Moreover, LM performance tends to differ across cultural topics, demonstrating higher performance for example in "education" and "holiday" practices over "clothing" and "cuisine" practices, as shown in Fig 3. This may potentially be due to diverse finer-grained domain-specific and region-specific information elements typically involved in "clothing" and "cuisine" topic discussions, whereas "education" and "holiday" practices may tend to be more universal. Finally, while our research community lacks specific training details of the closed-source models (*e.g.,* ChatGPT, GPT4), we believe that by including them in our benchmark comparison, we can help shed light on the performance gap between open-source pretrained LLMs and these closed-source models, to better bridge this performance difference in future work.

3.3.1 Ablation on the Challenging Setting brought by Fine-Grained Cultural Knowledge Frame Profiling

In this subsection, we further investigate the potential limitations of existing pretrained Large Language Models (LLMs) in understanding nuanced cultural nuances within different situational contexts, along our massively multicultural task domain. Our natural intuition is that pretrained LLMs may general tend to lack finer-grained knowledge

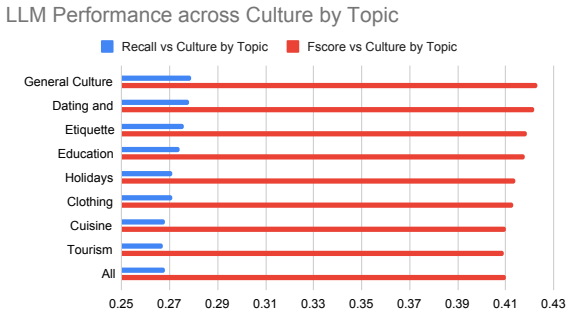

Figure 3: LLM performance by Topic (result from a single LM backbone ad-hoc, llama2-7b).

on the cultural nuances pertaining to subtle situational differences with respect to cultural frame profiling. To verify this hypothesis through empirical study, for each culture frame dimension, such as sub-country geo-region, ethnicity, age, gender, etc., we first isolate a subset of the **CultureAtlas** evaluation data with cultural knowledge assertions that generally applies across this dimension, which we refer to as "Gen", and isolate another subset of the **CultureAtlas** evaluation data with cultural knowledge assertions that applies specifically to a certain bucket/criteria across this dimension, which we refer to as "Spec". Then, we perform cross-comparison on LLM performance patterns, under data scenarios that are general ("Gen") versus specific ("Spec") in condition/critieria along each of these cultural frame profiling dimensions. Indeed, we find that lack fine-grained cultural commonsense knowledge is an area where there remains interesting rooms for improvement for LLM models. As shown in Table 7 (the result from a single ad-hoc LM backbone, llama2-7b), the zero-shot true/false inference capability of LLM significantly drops as we probe finer-grained cultural information.

	Norm Violation Relevant Culture Knowledge Assertion Samples
True Positive	• In **Indian** culture, when eating rice, it is mixed with curry, picking up small quantities with the fingers and pushing it into the mouth with the thumb. • In **Bhutan** culture, for special occasions and festivals, colourfully patterned silk kira and, more rarely, gho may be worn.
False Positive	• The **American** flag protocol dictates that the flag should be flown on all buildings, both public and private, as a sign of respect and loyalty to the nation. • In **Rioplatense Spanish**, the pronoun "vos" is commonly used to address strangers and authority figures, while it is considered impolite to use it with friends, family members, and close acquaintances. • **South Korea** and **North Korea** have different vocabulary standards because they consider their languages to be completely distinct and unrelated to each other. • In **Chinese culture**, particularly in **Macao**, it is believed that giving money in amounts that include the number four brings good luck and prosperity. • The society in **Kuwait City** is highly strict about traditions in the Gulf Arab region.
False Negative	• In **Catalonia**, there is a strong social and political consensus on the *language policies that support the use of Catalan*. • In **Botswana**, it is customary to use the *salutation 'kgosi'* before the name of a chief, providing an important cultural and social distinction. • In **Barbados**, people drive on the left side of the road, similar to the driving habits in the United Kingdom due to their history as a former British colony. • The practice of *inserting Mandarin into conversations in* ***Shanghainese*** is very common, especially among young people. • *Personal pronouns in* ***Shanghainese*** do not distinguish gender or case. • In **Argentina** culture, hot but not boiling water is poured into the gourd, drunk, then the mate is refilled.
True Negative	• In **Indian** culture, when eating rice, people commonly mix it with chutney, a flavorful condiment made from a variety of ingredients such as fruits, vegetables, herbs, and spices. • In **Malay** culture, people commonly greet each other with the phrase "Khabar", which roughly translates to "what's up" or "how are you" in English.

Table 6: Qualitative error analysis on LM culture knowledge reasoning capability (of **Vicuna-13B**) in zero-shot true/false inference.

	Spec.	Gen
Sub-Country Location	22.1	35.0
Ethnicity	27.5	35.5
Religion	17.9	35.0
Marital Status	13.7	33.6
Occupation	27.3	35.0

Table 7: F-score performance comparison, in %, by general (*'gen'*) versus fine-grained cultural profile framing specific (*'spec'*) knowledge, such as between country-level and province-level cultural group.

3.3.2 Error Analysis

Table 6 visualizes results qualitatively, shedding light on the challenges for an off-the-shelf pretrained language model (LM) in accurately reasoning about cultural practices across different societies. While a LM may correctly grasp certain cultural practices, such as properly recognizing traditional ways of eating rice in Indian culture and the ceremonial dress in Bhutan for special occasions, as well as accurately recognizing misconceptions in negative samples on mixing rice with chutney in Indian culture or common greetings in Malay, it demonstrates a lack of cultural knowledge and reasoning robustness in other less well-represented topics and geographical regions. For example, the model also produced false positives, such as in regards to the exaggerated protocol around the American flag, suggesting misunderstandings of cultural norms. False negatives, such as the underappreciated practice of drinking mate in Argentina, point to the model's oversight of genuine cultural customs. Overall, the error analysis reveals the inconsistent performance of LMs in capturing the breadth and depth of the different cultural knowledge in the world around us, revealing a significant area for improvement in LM cultural commonsense reasoning.

4 Related Work

Importance of Cultural Knowledge in NLP Tasks While large language models have generally embedded large parametric knowledge from large text corpora during its pretraining stage (Petroni et al., 2019), these models are also typically imposed with normative bias due to imbalanced representation at the data source (Emelin and Sennrich, 2021; Arora et al., 2022). Cultural knowledge is an integral part to the success of large language model (LLM) reasoning in a wide array of downstream applications. For example, there has been recent explorations on the vital role cultural knowledge plans in

helping answer commonsense questions (Palta and Rudinger, 2023; Yin et al., 2022), understand societal moral conventions (Ramezani and Xu, 2023; Emelin et al., 2021), analyze and mitigate social biases (Sap et al., 2020; Yang et al., 2023), detect norm violations (Fung et al., 2023; Li et al., 2023a), correct conversational dialogues (Ziems et al., 2022), and ultimately tune LLMs to align with the helpful and harmless principles of constitutional AI (Bai et al., 2022). Of ongoing interest to the NLP community is scrutinizing LMs on social minority understanding (Sun et al., 2023), which turns out that PLMs can learn norms diverging from social majority only when they are fine-tuned accordingly due to a presence of normative bias (Kiehne et al., 2022).

Cultural Knowledge Acquisition for improved LLM Training Hershcovich et al. (2022) explains culture as a concept of identity that can be examined from the dimensions of *objectives and values*, *linguistic form and style* (e.g., honorific reference terms when addressing a person), and *common ground* (e.g, socio-cultural norms, shared event occurrences, etc.). Our cultural knowledge acquisition process follows this theory of cultural definition and covers the dimensions of culture outlined above. In terms of the genre of data source for cultural knowledge discovery in practice, cultural knowledge has been predominantly gleaned from conversational dialogues (Fung et al., 2023) or web sources such as Reddit/Zhihu discussion forums (Forbes et al., 2020; CH-Wang et al., 2023) and the Common Crawl (Nguyen et al., 2023) – both of which tends to be relatively sparse in culturally relevant information and also noisy – or through knowledge elicitation of LLM parametric knowledge through prompting (Ziems et al., 2023), but this may be limited in the scope of available information that can be extracted due to the stochastic parrot property observed in LLMs when a cultural topic falls out of a LLM's pretrained knowledge boundary.

Multilingual LM Reasoning and implicit multicultural knowledge Previous research (Jiang et al., 2020; Clark et al., 2020) on language models have demonstrated strong capability to perform reasoning in multilingual settings, which is an initial step towards overcoming cultural barriers. The progression of research includes extending LM reasoning to low-resource language setting, such as for name tagging & knowledge base linking (Pan et al., 2017; Wen et al., 2021) through annotation transferring, which lays a promising foundation for reasoning across linguistic groups. However, the existing language models struggle with the cultural bias, primarily due to the lack of awareness of implicit multicultural knowledge. Recent studies have highlighted these issues; for instance, Havaldar et al. (2023) has identified the models' underperformance in recognizing cultural variations in specific phenomena, such as emotion detection across different countries. Another category of recent work has focused on evaluating performance on underrepresented languages, where Deas et al. (2023) has revealed biases against African American languages, leading to overlooked race-related issues in speech recognition and toxicity detection tasks. These findings underline the necessity to develop a new framework capable of acquiring cultural knowledge, aimed at addressing the cultural imbalances present in existing datasets used for training language models. Instead of stylistic linguistic conveyance, our work focuses on cultural knowledge acquisition based on fine-grained semantic variations, sourcing from over 500+ geosubregions and 2000+ ethnolinguistic groups.

5 Conclusions and Future Work

In this work, we explored an exciting area of massively multicultural knowledge probing of language model backbones, which has important impacts to norm violation detection and mitigation in assisting human-human interaction behavior across the many different subregions and ethnolinguistic groups around the world. We proposed a novel method for large-scale data collection across curated data sources and web-retrieval enhancement, which we verified with quality checks. Leveraging this constructed dataset, we achieved benchmarking the fine-grained cultural knowledge of popular language models. We foresee the impact of our work to contribute a useful resource and LM evaluation setting for the NLP community to develop more culturally-inclusive foundation models upon in the future. Finally, for future research directions, we will also investigate the effect that multimedia settings (e.g., vision and language) and multilingual (e.g., low-resource) settings have on foundation cultural reasoning across different fine-grained sub-cultures.

Ethical Considerations & Broader Impact

Our work is dedicated to advancing language model (LM) development practices to more accurately reflect the diverse social practices and cultural customs present in human societies across the world. This effort serves to enable improvements in LM norm violation reasoning capabilities, and at a higher-level, re-emphasizes the importance of overarching principles to human-model alignment. In formulating the ethical framework for our research, we carefully consider the key dimensions inherent to our work, to ensure that the modeling outcomes are in line with ethical standards and societal expectations. We introduce a novel normative framework aimed at enhancing the detection of incorrect cultural knowledge in future models. Our experimental setup incorporates commonsense reasoning, grounding the model's cultural knowledge in alignment with socially-situated human preferences.

We acknowledge limitations in our current work. The ground truth labels (*i.e.,* "positive" and "negative" data samples) that we utilize for LM cultural knowledge benchmark evaluation are automatically derived, due to scalability advantages. Our constructed dataset has been manually assessed and determined to be high-quality with a $90^{+}\%$ pass rate in data sample quality check. In other words, cultural knowledge assertions labeled "positive" are indeed rated as a true sample by human assessors, while cultural knowledge assertions labeled "negative" are indeed rated as non-factual samples for the large majority. If time and manual labor cost were ample, it would be meaningful to prepare human-curated cultural knowledge assertions and investigate any potential differences as a future research direction. However, we also point that it is extremely challenging to find competent human annotators covering the different geographical subregions and ethno-linguistic groups at a massive scale. Many crowdsourcing platforms for data annotation are skewed towards certain demographic subgroups, such as primarily American-based or Western centric in the Amazon MTurk service. Hence, our work serves as an extremely valuable first step towards preparing a meaningful benchmark for massively multicultural fine-grained LM knowledge reasoning, based on publicly audited Wikipedia document sources and additional data preprocessing procedure.

An essential element of our ethical framework is ensuring balanced representations across cultural groups, with a special focus on including perspectives from low-resource settings. Due to current design and imbalanced training data, LLMs may lack the capability to fully understand social norms and including biases across different countries, suffering from normative bias (Emelin and Sennrich, 2021; Arora et al., 2022). We commit to the ongoing journey of cultural knowledge enhancement and normative correction. In particular, we aim to not only address issues of fairness and equity, but also enrich our NLP models by considering a broader range of experiences and cultural knowledge. By adhering to cultural definitions, we can build cultural knowledge systems that are inclusive and reflective of the global community. Our approach paves the path for mitigating the risks of subtle cultural bias in language model training, and fosters understanding across fine-grained cultural divides.

References

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. 2022. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 700–710, Dublin, Ireland. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. *arXiv preprint arXiv:2305.14492*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and

Jennimaria Palomaki. 2020. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *CoRR*, abs/2003.05002.

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Yi R. Fung, Kevin Small, and Heng Ji. 2022. A zero-shot claim detection framework using question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6927–6933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion.

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.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Niklas Kiehne, Hermann Kroll, and Wolf-Tilo Balke. 2022. Contextualizing language models for norms diverging from social majority. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4620–4633, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Oliver Li, Mallika Subramanian, Arkadiy Saakyan, Sky CH-Wang, and Smaranda Muresan. 2023a. NormDial: A comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15732–15744, Singapore. Association for Computational Linguistics.

Sha Li, Chi Han, Pengfei Yu, Carl Edwards, Manling Li, Xingyao Wang, Yi Fung, Charles Yu, Joel Tetreault, Eduard Hovy, and Heng Ji. 2023b. Defining a new NLP playground. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11932–11951, Singapore. Association for Computational Linguistics.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13622–13623.

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.

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.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

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.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Chenkai Sun, Jinning Li, Yi Fung, Hou Chan, Tarek Abdelzaher, ChengXiang Zhai, and Heng Ji. 2023. Decoding the silent majority: Inducing belief augmented social graph with large language model for response forecasting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 43–57, Singapore. Association for Computational Linguistics.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhailong Wang, Yi Fung, Piyush Mishra, Qing Lyu, Dídac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 133–143, Online. Association for Computational Linguistics.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online. Association for Computational Linguistics.

Ke Yang, Charles Yu, Yi R Fung, Manling Li, and Heng Ji. 2023. Adept: A debiasing prompt framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10780–10788.

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.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

A Characterizing Cultural Knowledge Specificity for Data Filtering

In our **CultureAtlas** dataset construction process, we want to focus on socioculturally relevant knowledge assertions that are not too event or instance specific. For example, *"In 2020, China tops the QS Asia University Rankings list with over 120 universities including in the ranking, and five Chinese universities appear in the Asia Top 10, which is more than any country."* would be culturally relevant but too event-specific. To filter out such instances, we utilize the FACEBOOK/BART-LARGE-MNLI model to perform classification on each candidate sentence, between the classes of "general assertion" and "specific fact or instance". Through this approach, we found that approximately 52% of the

original pristine sentences from culturally-relevant Wikipedia pages fall under the "general assertion" category, which we retain in the dataset.

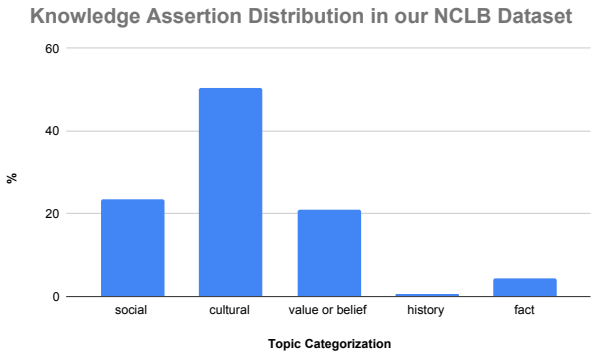

Figure 4: Bar graph on the data distribution of the original positive samples, in our **CultureAtlas** dataset construction.

B Culture Profile Extraction Performance

In this section, we expand on low-level details on the culture profile extraction process methodology and quality check results, leveraging prompting with various state-of-the-art pretrained large language model (LLM) backbones. Specifically, Table 8 details the prompt template details.

Culture Profile Field	Directed Question Answering
interaction nature categorization	Is this an individual human behavioral norm or human-human behavioral norm?
topic distribution modeling	Is this a social norm, cultural norm, belief or ritual, history, politics, or fact?
country-level extraction	Which country is mentioned or implied in the sentence? Answer N/A if unknown.
sub-country level extraction	Which state/province/city/subcountry region is mentioned or implied in the sentence (or answer N/A):
ethnicity extraction	Which ethnic group is mentioned or implied in the sentence? Answer N/A if unspecified?
ethnic subgroup extraction	Which ethnolinguistic subgroup is mentioned or implied in the sentence? Answer N/A if unspecified.
age extraction	Which age group is mentioned or implied in the sentence?
gender extraction	Which gender group is mentioned or implied in the sentence (male, female, transgender, or N/A)
marital status	Which marital status is mentioned or implied in the sentence.
religion belief extraction	Which religious group is mentioned or implied in the sentence (or answer N/A):
occupation extraction	Which occupation is mentioned or implied in the sentence? Answer N/A if unspecified

Table 8: Details on the prompt template for cultural profile extraction

Afghanistan	29	0.2k	Georgia	35	0.2k	Afghanistan	29	0.2k
Albania	0	0.0k	Germany	301	6.2k	Zimbabwe	26	0.3k
Algeria	38	0.4k	Ghana	30	0.3k	Zambia	9	0.2k
Andorra	4	0.0k	Greece	98	1.1k	Yemen	8	0.0k
Angola	22	0.2k	Grenada	0	0.0k	Viet Nam	43	0.8k
Antigua and Barbuda	1	0.0k	Guatemala	0	0.0k	Venezuela	9	0.1k
Argentina	222	1.4k	Guinea	0	0.0k	Vanuatu	1	0.0k
Armenia	118	0.6k	Guinea-Bissau	17	0.4k	Uzbekistan	34	0.2k
Australia	190	2.8k	Guyana	0	0.0k	Uruguay	40	0.4k
Austria	78	0.9k	Haiti	0	0.0k	United States	1452	42.4k
Azerbaijan	49	0.4k	Honduras	13	0.1k	Tanzania	118	0.6k
Bahamas	0	0.0k	Hungary	37	0.3k	United Kingdom	229	7.2k
Bahrain	6	0.1k	Iceland	0	0.0k	United Arab Emirates	38	0.6k
Bangladesh	105	0.8k	India	847	18.4k	Ukraine	87	1.3k
Barbados	4	0.1k	Indonesia	364	6.4k	Uganda	61	0.2k
Belarus	16	0.2k	Iran	22	0.1k	Tuvalu	2	0.0k
Belgium	75	1.4k	Iraq	34	0.3k	Turkmenistan	12	0.1k
Belize	2	0.0k	Ireland	99	1.2k	Türkiye	42	0.7k
Benin	14	0.1k	Israel	71	2.2k	Tunisia	94	1.8k
Bhutan	53	0.3k	Italy	560	8.5k	Trinidad and Tobago	9	0.2k
Bolivia	18	0.1k	Jamaica	0	0.0k	Tonga	0	0.0k
Bosnia and Herzegovina	40	0.2k	Japan	388	5.7k	Togo	16	0.0k
Botswana	61	0.2k	Jordan	15	0.1k	Timor-Leste	2	0.0k
Brazil	226	2.2k	Kazakhstan	23	0.1k	Thailand	223	2.9k
Brunei Darussalam	0	0.0k	Kenya	61	0.6k	Tajikistan	4	0.0k
Bulgaria	129	1.7k	Kiribati	9	0.1k	Syrian Arab Republic	0	0.0k
Burkina Faso	6	0.1k	Kuwait	0	0.0k	Switzerland	140	1.9k
Burundi	13	0.0k	Kyrgyzstan	6	0.1k	Sweden	124	1.9k
Cabo Verde	0	0.0k	Laos	80	0.2k	Suriname	3	0.0k
Cambodia	73	0.6k	Latvia	25	0.1k	Sudan	20	0.2k
Cameroon	29	0.1k	Lebanon	33	0.5k	Sri Lanka	12	0.2k
Canada	198	2.8k	Lesotho	22	0.3k	Spain	162	3.5k
Central African Republic	4	0.0k	Liberia	9	0.0k	South Sudan	13	0.3k
Chad	5	0.0k	Libya	10	0.1k	South Africa	79	1.0k
Chile	47	0.3k	Liechtenstein	6	0.0k	Somalia	23	0.1k
China	409	8.1k	Lithuania	25	0.4k	Solomon Islands	5	0.0k

Table 9: The # of documents and cultural knowledge assertion sentences, per culture by country, that are specific to sub-country level geographical regions.

Medumba	145	Awabakal	35	Kalapalo, Kuikúro-Kalapálo	17
Germanic	117	Eleme	35	Bengkala	16
Central Dusun, Kadazan Dusun	115	Ahtena	33	Bangala	16
Notre	112	Ayoreo	32	Chilcotin	16
Doga	107	Bakumpai	32	Northern Dagara	16
Hopi	107	San Blas Kuna	32	Gbagyi	16
Mescalero-Chiricahua Apache	99	Kabardian	32	Kayardild	16
Kashubian	96	Hiberno-Scottish Gaelic	31	Jah Hut	16
Ainu (Japan)	94	Jahanka	31	Kanuri	16
Assiniboine	91	Kambaata	31	Kankanaey	16
Arapaho	89	Batek	30	Barai	15
Jicarilla Apache	88	Hunsrik	30	Djabugay, Dyaabugay	15
Batak languages	86	Bwa	29	Emilian	15
Izere	81	Gadang	29	Macushi	15
Kodava	75	Cowlitz	28	Azha	14
Cappadocian Greek	74	Dolpo	28	Igede	14
Efik	73	Kuku-Yalanji	28	Khorasani Turkish	14
Algonquin	72	Korak	27	Bundeli	13
Hittite	70	Harari	26	Columbia-Wenatchi	13
Etruscan	69	Humla	26	Gooniyandi	13
Huichol	67	Shambala	26	Kaska	13
Hajong	66	Dhanggatti, Dyangadi	25	Amahuaca	12
Coast Miwok	65	Gumatj	25	Atsugewi	12
Mycenaean Greek	62	Reel	24	Awetí	12
Jju	62	Banggarla	24	Southern Luri	12
Pacific Gulf Yupik	60	Hidatsa	24	Adhola	11
Heiltsuk	60	Khanty	24	Qimant	11
Jukun Takum	60	Gros Ventre	23	Bidyogo	11
Khasi	59	Baga Sitemu	23	Gambera	11
Javanese	58	Koasati	23	Guro	11
Garre	57	Igala	23	Kurdish	11
Gujarati	57	Yaka (Congo)	23	Hermit	11
Chickasaw	56	Adnyamathanha	22	Molale	11
Gunditjmara	55	Baras	22	Pemon	10
Esselen	54	Burarra	22	Sari	10
Yanomamö	54	Ejagham	22	Bhunjia	10
Bhojpuri	53	Ngadjuri	21	Laba	10
Beothuk	52	Kalenjin	21	Lasi	10
Goan Konkani	52	Gheg Albanian	20	Arbore	9
Idoma	52	Badaga	20	Kaingang	9
Atakapa	51	Chuvash	20	Col	9
Chipewyan, Dene Suline	50	Cocopa	20	Lozi	9
Daai Chin	50	Dimasa	20	Leti (Indonesia)	9
Colonia Tovar German	50	Eyak	20	Akuntsu	8
Jakun	50	Hyam	20	Djauan, Jawoyn	8
Bodo (India)	49	Greenlandic, Kalaallisut	20	Adiwasi Garasia	8
Shor	49	Mixed Great Andamanese	19	Holikachuk	8
Lushai	49	Karadjeri, Karajarri	19	Jiru	8
Con	48	Gilaki	19	Kurichiya	8
Kutenai	47	Ikulu	19	Korwa	8
Djawi	46	Worimi	19	Hijazi Arabic	7
Angor	44	Lobi	19	Bugun	7
Chitimacha	42	Ingrian	18	Cuban	7
Gitxsan	42	Keliko	18	Cheq Wong, Chewong	7
Kickapoo	41	Keiga	18	Gata’	7
Sudanese Arabic	40	Ladin	18	Gureng Gureng	7
Darlong	38	Guerrero Amuzgo	17	Gunwinggu	7
Haisla	38	Bora	17	Koba	7
Igbo	38	Chamacoco	17	Lanoh	7
Kalapuya	38	Mro-Khimi Chin	17	Apatani	6
Upper Kuskokwim	37	Isoko	17	Tuki	6
Atikamekw	36	Krymchak	17	Bote-Majhi	6
Esperanto	36	Kota (India)	17	Bwile	6

Table 10: The # of documents and cultural knowledge assertion sentences per culture by ethnolinguistic group