Conceptual Cultural Index: A Metric for Cultural Specificity via Relative Generality

Takumi Ohashi
Hosei University, Tokyo, Japan
takumi.ohashi.4g@gmail.com

Hitoshi Iyatomi
Hosei University, Tokyo, Japan
iyatomi@hosei.ac.jp

Abstract

Large language models (LLMs) are increasingly deployed in multicultural settings; however, systematic evaluation of cultural specificity at the sentence level remains underexplored. We propose the Conceptual Cultural Index (CCI), which estimates cultural specificity at the sentence level. CCI is defined as the difference between the generality estimate within the target culture and the average generality estimate across other cultures. This formulation enables users to operationally control the scope of culture via comparison settings and provides interpretability, since the score derives from the underlying generality estimates. We validate CCI on 400 sentences (200 culture-specific and 200 general), and the resulting score distribution exhibits the anticipated pattern: higher for culture-specific sentences and lower for general ones. For binary separability, CCI outperforms direct LLM scoring, yielding more than a 10-point improvement in AUC for models specialized to the target culture. Our code is available at https://github.com/IyatomiLab/CCI.

1 Introduction

Large language models (LLMs) exhibit broad multilingual competence and are increasingly used for tasks such as search, summarization, and dialogue (Brown et al., 2020; Achiam et al., 2023; Grattafiori et al., 2024; Yang et al., 2024). However, as applications expand, a key challenge remains in determining whether it is possible to ensure consistent and culturally aware responses across different regions. Everyday knowledge, such as dietary practices, greeting conventions, linguistic expressions, and seasonal events, varies systematically across cultures and regions, and how models handle this knowledge has direct implications for fairness, safety, and reliability (Cao et al., 2023; Naous et al., 2024; Shen et al., 2024; Rao et al., 2025). Consequently, there is a need to develop models that can appropriately accommodate diverse cultural characteristics and differences, as well as models specialized for a specific culture.

Benchmarks of LLM capabilities have progressed beyond general-knowledge evaluations toward frameworks that focus on cultural knowledge, which advance the visualization of regional differences and biases (Myung et al., 2024; Chiu et al., 2025). Many benchmarks use QA formats with overall accuracy as the main metric, consistently showing that culturally specific questions are harder than culture-agnostic ones (Shen et al., 2024; Arora et al., 2025). However, cultural knowledge spans a continuum—from phenomena shared across regions to those unique to a specific locale—and existing benchmarks fail to capture this aspect, making it difficult to conduct error analysis and formulate targeted improvement strategies.

On the data side, large-scale corpora of cultural knowledge have been proposed (Nguyen et al., 2023; Shi et al., 2024), but they lack annotations indicating the degree of cultural specificity of each sentence. Thus, both evaluation and data resource development require a framework for quantitatively assessing sentence-level cultural specificity, yet no such framework currently exists. Manual annotation of sentence-level cultural specificity is labor-intensive, requires domain expertise and contextual understanding, and often yields low inter-annotator agreement, highlighting the need for automation. However, culture is a multifaceted, high-level construct, and prior work has rarely provided an explicit definition (Adilazuarda et al., 2024).

In this paper, we propose the Conceptual Cultural Index (CCI), a sentence-level metric for quantifying cultural specificity, to address this challenge. CCI uses an LLM to estimate a sentence's generality across multiple cultures and, based on these scores, quantifies the target culture's specificity relative to others. This formulation allows users to control the scope of "culture" by adjusting the set of non-target cultures used for comparison.

Proceedings of the First Workshop on Multilingual Multicultural Evaluation, pages 67–75

The contributions of this study are as follows:

- We introduce CCI, a new sentence-level metric for quantifying cultural specificity.
- Compared with direct LLM-based scoring of cultural specificity, CCI yields clearer separability between culture-specific and general sentences and offers greater interpretability.
- We present a practical use case of CCI by assigning item-level CCI scores to existing benchmarks and showing that model performance varies with the level of cultural specificity, enabling culture-aware error analysis.

2 Related Work

Cultural evaluation benchmarks for language models include broad, multi-region datasets such as GeoMLAMA (Yin et al., 2022), BLEnD (Myung et al., 2024), CDEval (Wang et al., 2024), and CulturalBench (Chiu et al., 2025), as well as country- or region-specific benchmarks such as CLIcK (Kim et al., 2024), IndoCulture (Koto et al., 2024), and CHARM (Sun et al., 2024). These resources support comparisons of cultural knowledge across models, but most adopt a QA format and rely on overall accuracy, which conflates culture-specific difficulty with general knowledge errors.

Text-based resources for collecting cultural knowledge, including StereoKG (Deshpande et al., 2022), CANDLE (Nguyen et al., 2023), CultureAtlas (Fung et al., 2024), CultureBank (Shi et al., 2024), and MANGO (Nguyen et al., 2024), extract and organize cultural assertions at scale. However, they do not provide sentence-level annotations with continuous scores of cultural specificity, leaving a gap that our sentence-level metric CCI aims to fill.

3 CCI

We propose the Conceptual Cultural Index (CCI), a sentence-level index of cultural specificity. As illustrated in Figure 1, given a target culture and a set of comparison cultures, we use an LLM to estimate how common a sentence is in each culture and derive a specificity score for the target culture.

3.1 Obtaining Generality Scores

Given an input sentence x, a set of cultures C, and a target culture $t \in C$, we use an LLM[1] to estimate, for each $c \in C$, how common x is in the culture c, yielding a continuous generality score $p_c(x) \in [0, 1]$. In practice, we query all cultures in C within a single prompt and parse the scores from a JSON-formatted response. The prompt used for the generality score is provided in Appendix A.

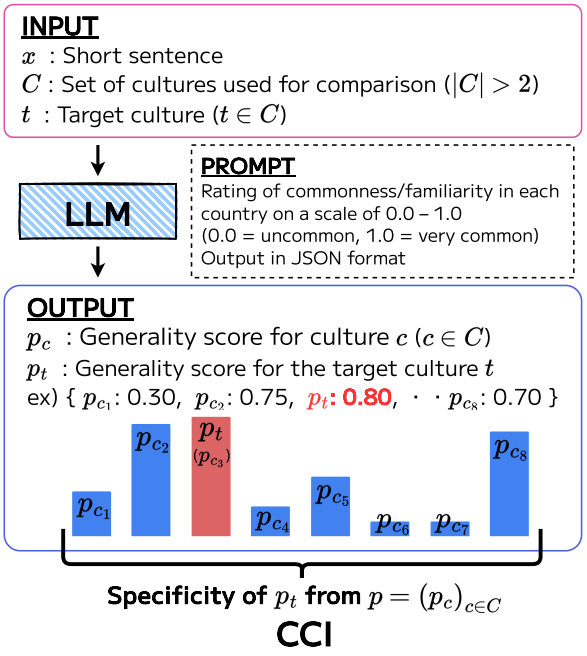

Figure 1: Overview of CCI.

To mitigate run-to-run variability in LLM outputs, we average results over N independent runs (in this paper, $N = 3$):

$$\bar{p}_c(x) = \frac{1}{N} \sum_{n=1}^{N} f_{\mathrm{LLM}}^{(n)}(x; C)[c], \quad c \in C. \tag{1}$$

Here, $f_{\mathrm{LLM}}^{(n)}(x; C)[c]$ denotes the score for culture c returned by the n-th run.

3.2 Definition of CCI

For a target culture $t \in C$, we define CCI as the difference between the generality score in the target culture and the average generality score across the other cultures:

$$CCI(x; t, C) = \bar{p}_t(x) - \frac{1}{|C| - 1} \sum_{c \in C \setminus \{t\}} \bar{p}_c(x). \tag{2}$$

CCI takes values in $[-1, 1]$: values near 0 indicate that x is cross-culturally general, values near 1 indicate that x is specific to the target culture, and values near -1 indicate that x is specific to non-target cultures.

For Eq. (2), we also examined a sharpness-based formulation that weights the target culture

[1]Model selection is discussed in the experiments section.

by its log-softmax-normalized generality[2], but it produced nearly constant scores across input sentences and compressed the score range, especially for larger $|C|$. In contrast, the simple difference is less sensitive to $|C|$ and directly measures the gap on the original $[0, 1]$ scale, so we adopt this definition.

4 Experiments

4.1 Experimental Setup

To assess whether CCI reflects cultural specificity, we use Japan as the target culture t and compute CCI for two classes: culture-specific sentences (positive) and general sentences (negative). We plot ROC curves for detecting Japanese cultural sentences and evaluate separability using the area under the ROC curve (AUC) and the difference in class medians. We also conduct a qualitative analysis of representative examples.

As a baseline, we use an LLM that directly outputs a $[0, 1]$ specificity score and compute AUC with the same protocol to compare its discriminative performance with CCI. As with CCI, 0 denotes cross-cultural generality and 1 denotes specificity to the target culture. The prompt used for the baseline is provided in Appendix A.

Data We first used GPT-5 (model as of August 7, 2025) to generate 300 short Japanese cultural sentences and 300 general sentences. All generated sentences were manually reviewed and filtered, with duplicates and clear misclassifications removed, although some borderline cases may remain because cultural boundaries are inherently ambiguous. The final evaluation set consists of 200 Japanese cultural sentences and 200 general sentences. The prompts used for data generation are provided in Appendix A.

Models To identify suitable LLMs for computing CCI, we compared CCI and the baseline under a common protocol across five LLMs: multilingual models (Llama 3.1 (Grattafiori et al., 2024), Qwen 2.5 (Yang et al., 2024), gpt-oss (Agarwal et al., 2025)) and Japanese-specialized models (Llama 3.1 Swallow (Fujii et al., 2024), llm-jp 3.1 (Aizawa et al., 2024)). The exact model identifiers and links are listed in Appendix B.

[2] $q_t = \frac{\exp(\bar{p}_t)}{\sum_{c \in C} \exp(\bar{p}_c)}$, $CCI_{\log} = \left(1 + \frac{\log q_t}{\log(|C|)}\right) \bar{p}_t$.

4.2 Varying the Set of Cultures C

CCI allows the set C of cultures to vary. To assess how controllable the cultural scope is, we conduct experiments under two modes.

Global mode C is fixed to the 19 G20 member countries, excluding the European Union and the African Union, as we restrict C to country names.

Custom mode C is configured to match the task objective. In this experiment, to test whether the inclusion of neighboring cultures can be controlled, we defined two conditions across four countries:

1. ***+Neighbor Culture***, which includes neighboring countries in C: ["China", "Republic of Korea", "United States of America", "Japan"];

2. ***–Neighbor Culture***, which excludes neighboring countries from C: ["Brazil", "France", "United States of America", "Japan"];

For the baseline, we also evaluate two prompting conditions: one prompt that explicitly includes the instruction *"If the practice is also common in neighboring or culturally adjacent countries, do not consider it specific to the target."* (***+Neighbor Culture***), and another that omits this instruction.

4.3 CCI-Based Benchmark Stratification

As a use case of CCI, we perform benchmark stratification by assigning CCI scores to each item and analyzing how task accuracy varies across CCI levels. We use two datasets that capture Japanese commonsense: JCommonsenseQA (JCQA) (Kurihara et al., 2022)[3] and JCommonsenseMorality (JCM) (Takeshita and Rzepka, 2025)[4], both of which include items that may reflect phenomena specific to the Japanese cultural sphere. JCQA is a five-way multiple-choice commonsense question answering task; we compute CCI using as input x the concatenation of the question text and the gold option text. JCM is a binary classification task that judges whether an action is morally acceptable; we compute CCI using as input x the target sentence together with the gold label. For scoring, we use CCI in Global mode computed with gpt-oss. We evaluate on the JCQA dev set (1,119 items) and the JCM test set (3,992 items) using predictions from Qwen 2.5, Llama 3.1, and llm-jp 3.1.

[3] https://github.com/yahoojapan/JGLUE

[4] https://github.com/Language-Media-Lab/jethics

Models	Baseline		CCI (Global)	CCI (Custom)	
		+Neighbor		+Neighbor	-Neighbor
	($C_{\text{median}}\uparrow$, $G_{\text{median}}\downarrow$)	($C_{\text{median}}\uparrow$, $G_{\text{median}}\downarrow$)	($C_{\text{median}}\uparrow$, $G_{\text{median}}\downarrow$)	($C_{\text{median}}\uparrow$, $G_{\text{median}}\downarrow$)	($C_{\text{median}}\uparrow$, $G_{\text{median}}\downarrow$)
Qwen2.5-7B	(0.815, 0.800)	(0.980, 0.800)	(0.800, 0.505)	(0.633, 0.267)	(0.833, 0.467)
Llama-3.1-8B	(0.870, 0.800)	(0.870, 0.800)	(0.778, 0.648)	(0.664, 0.283)	(0.980, 0.711)
Llama-3.1-Swallow-8B	(0.950, 0.850)	(0.950, 0.850)	(0.761, 0.324)	(0.331, 0.117)	(0.933, 0.300)
llm-jp-3.1-13b	(0.800, 0.785)	(0.800, 0.700)	(0.869, 0.568)	(0.792, 0.467)	(0.897, 0.593)
gpt-oss-20b	(0.880, 0.100)	(0.775, 0.100)	(0.836, 0.063)	(0.697, 0.111)	(0.817, 0.104)

Table 1: Class-wise medians of specificity scores for CCI and the baseline. C_{median} denotes the median score for culture-specific sentences, and G_{median} denotes the median score for general sentences. While C_{median} should ideally be close to 1 and G_{median} close to 0, it is not necessary for every instance to reach these extremes; it suffices that the overall trend is observed. The cultural specificity score aims to assign context-appropriate values to each sentence.

Models	Baseline AUC / Δ	CCI (Global) AUC / Δ
Qwen2.5-7B	0.816 / 0.015	**0.884 / 0.295**
Llama-3.1-8B	**0.803** / 0.070	0.796 / **0.130**
Llama-3.1-Swallow-8B	0.842 / 0.100	**0.945 / 0.437**
llm-jp-3.1-13b	0.768 / 0.015	**0.908 / 0.301**
gpt-oss-20b	**0.963 / 0.780**	0.956 / 0.773

Table 2: Separability between culture-specific and general sentences, reported in terms of the AUC and the median gap $\Delta = C_{\text{median}} - G_{\text{median}}$.

5 Results and Discussion

5.1 Separability between Culture-Specific and General Sentences

Table 1 shows, for each LLM, the class-wise median specificity scores for CCI and the direct-estimation baseline: C_{median} for culture-specific sentences and G_{median} for general sentences. Table 2 shows separability, measured by AUC, and the gap between the medians ($\Delta = C_{median} - G_{median}$). We report median gaps rather than mean gaps because the baseline scores are bounded in $[0, 1]$, whereas CCI scores lie in $[-1, 1]$. Given these different ranges, mean differences could inadvertently favor CCI; using medians avoids this issue in our experiments.

CCI vs. Baseline CCI achieves AUC comparable to or higher than the baseline and yields clearer separation, with higher scores for culture-specific sentences and lower scores for general ones. The baseline, in contrast, tends to assign relatively high scores to many sentences, and for some models the class medians are nearly identical. This may be because directly quantifying "culture" as a single scalar is inherently difficult, whereas CCI decomposes the task into per-culture generality estimates, thereby stabilizing the inference process.

Model suitability Regarding which LLMs are suitable for computing cultural specificity scores, gpt-oss achieves near-ideal separation under both CCI and the baseline. This appears to reflect not only model size but also its reasoning-oriented architecture, which captures cultural differences through step-by-step reasoning. In addition, Japanese-specialized models show better separation than multilingual models. Overall, models that combine strong reasoning capabilities and a deep understanding of the target culture, while also possessing knowledge of other cultures, are most suitable for computing cultural specificity scores.

5.2 Controllability of Cultural Scope

CCI vs. Baseline From Table 1, we observe that under CCI's Custom mode (+*Neighbor*), the median score for the culture-specific class is lower than in the Global mode. This suggests that the cultural scope can be adjusted to avoid overestimating practices common in neighboring cultures. By contrast, the baseline appears to be sensitive to prompt wording and input language, indicating that cultural scope is difficult to control solely through textual instructions. Additionally, CCI provides a numerical assessment of cultural specificity together with per-culture generality scores; even when baseline accuracy is high, CCI offers greater interpretability by indicating in which cultures a sentence is considered common or uncommon.

Case analysis Figure 2 shows a subset of the evaluation instances along with the CCI scores produced by gpt-oss. In individual cases, actions that may be taboo in neighboring cultures (e.g., "Pick up the small bowl and bring it to your mouth.") should not receive excessively low scores even under +*Neighbor*. Conversely, practices widely observed across regions (e.g., "Taking milk out of the refrigerator.") should not receive high scores even under –*Neighbor*. Consistent with these ex-

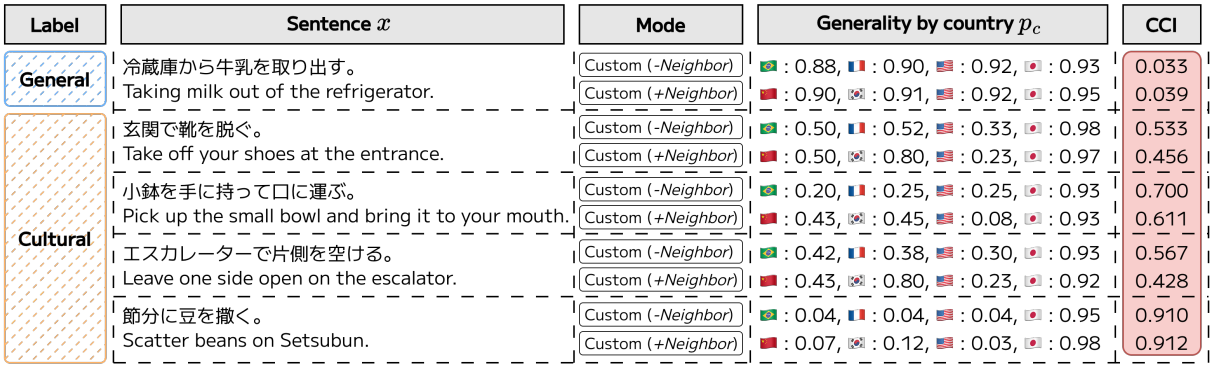

Figure 2: Example sentences and their corresponding CCI scores computed by gpt-oss.

Bin statistics		Accuracy		
Range	#Items	Qwen 2.5	Llama 3.1	llm-jp 3.1
CCI ≤ 0.1	583	0.940	0.899	0.967
$0.1 <$ CCI ≤ 0.2	75	0.893	0.880	0.987
$0.2 <$ CCI ≤ 0.3	69	0.826	0.870	0.971
$0.3 <$ CCI ≤ 0.4	40	0.900	0.775	0.925
$0.4 <$ CCI ≤ 0.5	19	0.789	0.789	0.842
$0.5 <$ CCI ≤ 0.6	63	0.873	0.825	0.905
$0.6 <$ CCI ≤ 0.7	36	0.889	0.806	0.972
$0.7 <$ CCI ≤ 0.8	45	0.867	0.800	0.978
$0.8 <$ CCI ≤ 0.9	173	0.873	0.821	0.942
$0.9 <$ CCI ≤ 1.0	16	0.750	0.750	0.938
Overall Accuracy		0.904	0.864	0.958

Table 3: Number of JCQA items and model-wise accuracy when CCI is binned in increments of 0.1.

Bin statistics		Accuracy		
Range	#Items	Qwen 2.5	Llama 3.1	llm-jp 3.1
CCI ≤ 0.1	3217	0.846	0.836	0.924
$0.1 <$ CCI ≤ 0.2	117	0.778	0.795	0.906
$0.2 <$ CCI ≤ 0.3	116	0.741	0.724	0.897
$0.3 <$ CCI ≤ 0.4	72	0.708	0.694	0.931
$0.4 <$ CCI ≤ 0.5	60	0.733	0.767	0.883
$0.5 <$ CCI ≤ 0.6	75	0.680	0.773	0.880
$0.6 <$ CCI ≤ 0.7	94	0.755	0.691	0.894
$0.7 <$ CCI ≤ 0.8	138	0.761	0.681	0.891
$0.8 <$ CCI ≤ 0.9	102	0.745	0.657	0.892
$0.9 <$ CCI ≤ 1.0	1	1.000	1.000	1.000
Overall Accuracy		0.826	0.814	0.918

Table 4: Number of JCM items and model-wise accuracy when CCI is binned in increments of 0.1.

pectations, the presence or absence of similar practices in neighboring cultures yields systematic differences between the *+Neighbor* and *–Neighbor* conditions, indicating that CCI effectively operationalizes cultural scope control as intended.

5.3 Task Accuracy Shifts across CCI Levels

Tables 3 and 4 show the results for JCQA and JCM, respectively, where items are binned by CCI in increments of 0.1. Across both datasets, the CCI distribution is skewed toward lower values, suggesting that JCQA and JCM contain many culturally non-specific commonsense questions.

Accuracy tends to decrease as CCI increases, and higher-CCI bins often fall below the overall dataset accuracy. This indicates that items with higher cultural specificity tend to be more challenging, and that current models may not sufficiently acquire culture-specific knowledge. In contrast, llm-jp maintains higher overall accuracy and exhibits a comparatively smaller drop in high-CCI bins. This suggests that models trained on Japanese data may have an advantage on items that can reflect commonsense knowledge in the Japanese cultural sphere. Overall, CCI-based stratification makes clear how performance varies with cultural specificity. This variation is difficult to observe from overall accuracy alone, and the results suggest that model gaps tend to widen in higher-CCI bins.

6 Conclusion

We introduced the Conceptual Cultural Index (CCI), a metric that quantifies sentence-level cultural specificity as the relative difference in generality between a target culture and other cultures. CCI remains effective even when the baseline struggles to assign stable scores and provides interpretable estimates grounded in an explicit definition. CCI supports practical culture-related workflows, including annotating and stratifying benchmarks for model evaluation, as well as filtering culture-specific knowledge data. Our method is presented as an evaluation framework that can be applied consistently as models evolve.

Limitations

This study has three limitations. First, because cultures were approximated primarily at the country level, intra-country heterogeneity, such as regional or generational differences, may not be fully captured. Second, our experiments focused on Japan as the target culture, and thus the generalizability to other languages and regions remains to be established. Third, CCI relies on LLM-generated generality scores and thus inherits the biases and calibration issues of the underlying models. With these points in mind, we plan to develop CCI into a more robust and general-purpose evaluation framework by refining the granularity of cultural groupings and conducting broader multilingual and cross-regional evaluations.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, and 261 others. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards Measuring and Modeling "Culture" in LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, and 106 others. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint arXiv:2508.10925*.

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, and 62 others. 2024. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. *arXiv preprint arXiv:2407.03963*.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11772–11817.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, pages 1877–1901.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP*, pages 53–67.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A Robust, Diverse and Challenging Benchmark for Measuring LMs' Cultural Knowledge Through Human-AI Red-Teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701.

Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. 2022. StereoKG: Data-Driven Knowledge Graph Construction For Cultural Knowledge and Stereotypes. In *Proceedings of the Sixth Workshop on Online Abuse and Harms*, pages 67–78.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In *Proceedings of the First Conference on Language Modeling*.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively Multi-Cultural Knowledge Acquisition & LM Benchmarking. *arXiv preprint arXiv:2402.09369*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics,*

Language Resources and Evaluation, pages 3335–3346.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring Geographically Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese General Language Understanding Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pages 78104–78146.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393.

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*, page 1907–1917.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. Cultural Commonsense Knowledge for Intercultural Dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, page 1774–1784.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A Framework for Measuring the Cultural Adaptability of Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the Capabilities and Limitations of Large Language Models for Cultural Commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025.

Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. 2024. Benchmarking Chinese Commonsense Reasoning of LLMs: From Chinese-Specifics to Reasoning-Memorization Correlations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11205–11228.

Masashi Takeshita and Rafal Rzepka. 2025. JETHICS: Japanese Ethics Understanding Evaluation Dataset. *arXiv preprint arXiv:2506.16187*.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024. CDEval: A Benchmark for Measuring the Cultural Dimensions of Large Language Models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

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.

A Prompts

This section provides the prompts used in our experiments. Table 5 presents the prompt for obtaining per-culture generality scores used to compute CCI; Table 6 shows the prompt for direct cultural specificity scoring used in the baseline comparison; and Table 7 and 8 provides the prompts used to generate the evaluation data (culture-specific and general sentences).

B Model Details

We use the following LLMs in our experiments:

- Llama 3.1 8B Instruct: `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

```
Task:
Rate how COMMON/FAMILIAR the following item is in each country (0.00 = not common, 1.00 = very common). Treat countries independently. Be language-agnostic: interpret the statement regardless of its language.

Statement: {sentence}
Countries: {cultures}

Rules:
- Use general knowledge; avoid stereotypes.
- If similarly common across many countries, use similar (even identical) scores.
- If unsure, use mid values (e.g., 0.50).
- Do NOT normalize across countries.

Output JSON ONLY (no prose):
Schema: {"scores": {"<country>": <float>}}
Constraints: use the country names exactly as provided; floats in [0.00, 1.00], rounded to two decimals.
```

Table 5: Prompt for obtaining per-culture generality scores used in computing CCI. The input consists of a sentence x and a set of cultures $\mathcal{C}$.

```
Task:
Decide whether the following item is culturally specific to the TARGET country, and output a numeric specificity score only
(0.00 = globally common; 1.00 = unique to the target).

Statement: {sentence}
Target country: {target_culture}

Rules:
- Use general knowledge; avoid stereotypes.
- Be language-agnostic; interpret the statement regardless of its language.
- If unsure, use mid values (e.g., 0.50).

Output JSON ONLY (no prose):
{"score": <float in [0.00,1.00] rounded to two decimals>}
```

Table 6: Prompt for directly predicting cultural specificity scores in the target culture by an LLM (Baseline). The input consists of a sentence x and a target culture t.

- Qwen 2.5 7B Instruct:
  https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
- gpt-oss-20B:
  https://huggingface.co/openai/gpt-oss-20b
- Llama 3.1 Swallow 8B Instruct v0.5:
  https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5
- LLM-jp 3.1 13B instruct4:
  https://huggingface.co/llm-jp/llm-jp-3.1-13b-instruct4

```
You are an assistant for creating a short-sentence corpus.
Strictly satisfy the requirements below.

Goal:
- Collect very ordinary events that could occur in any region of the world.

Output:
- Exactly one natural Japanese sentence, short in length (about 10-20 characters).
- Describe facts plainly without evaluations, impressions, or subjectivity.
- Avoid place names, specific store names, and personal names. Avoid excessive stereotypes.
- Vary expressions, vocabulary, and scenes so that the same sentence endings and the same constructions do not appear consecutively.

Strict requirements:
- The output must be a JSON array. Each element must have the form { "text": "<one sentence>" }.
- The number of items must be exactly 300.
- Do not include any additional explanations, labels, or numbering (do not output any strings other than JSON).

Example (format only):
[
  { "text": "Turn off the alarm in the morning." },
  { "text": "Wait for the train at the station." }
]
```

Table 7: Prompt for generating general sentences. We used a Japanese prompt in our experiments; the version shown here is the English translation.

```
You are an assistant for creating a short-sentence corpus.
Strictly satisfy the requirements below.

Goal:
- Broadly collect, in short sentences, Japan-specific customs, daily life culture, annual events, food culture, public manners, etc.

Output:
- Exactly one natural Japanese sentence, short in length (about 10-20 characters).
- Describe events plainly without evaluations or impressions.
- Avoid place names, specific store names, and personal names. Avoid excessive stereotypes.
- Vary expressions, vocabulary, and scenes so that the same sentence endings and the same constructions do not appear consecutively.

Strict requirements:
- The output must be a JSON array. Each element must have the form { "text": "<one sentence>" }.
- The number of items must be exactly 300.
- Do not include any additional explanations, labels, or numbering (do not output any strings other than JSON).

Example (format only):
[
  { "text": "Take off your shoes at the entrance." },
  { "text": "Use the purification basin at a shrine." }
]
```

Table 8: Prompt for generating Japan-specific (culture-specific) sentences. We used a Japanese prompt in our experiments; the version shown here is the English translation.