Conceptual Cultural Index: 相対的一般性に基づく文化特有性の尺度

大橋 巧 [1]　彌冨 仁 [1]
[1] 法政大学大学院 理工学研究科 応用情報工学専攻
takumi.ohashi.4g@stu.hosei.ac.jp　iyatomi@hosei.ac.jp

概要

大規模言語モデルの多文化利用が進み，モデル面とデータ面の双方での整備が求められる一方で，文単位で文化特有性を定量化する指標は未整備である．本研究では，任意の文に対して文化特有性を推定する尺度 Conceptual Cultural Index（CCI）を提案する．CCI は，評価文の対象文化圏における一般性推定値と，他文化圏における一般性推定値の平均との差として定義される．この定式化により，文化圏集合を調整することで評価する文化スコープを制御でき，背後の一般性推定に基づくため解釈性も高い．文化依存文と一般文からなる計 400 文の評価セットで検証した結果，文化依存文では CCI が高く，一般文では低いという望ましいスコア分布が得られた．

1　はじめに

大規模言語モデル（Large Language Models; LLM）は，多言語にわたる知識と高い言語流暢性を備え，検索・要約・対話など幅広いタスクにおいて実用段階に達しつつある [1, 2, 3, 4]．一方で，応用現場への展開が進むほど，文化圏間の違いを考慮した応答を担保できるかという課題が残されている．例えば，食習慣や挨拶規範，言語表現，季節行事など日常に根ざした知識は文化や地域ごとに体系的な差異をもち，LLM がこれらをどのように扱うかはモデルの公平性や安全性にも直結する [5, 6, 7, 8]．そのため，様々な文化の特性や違いに適切に対応できるモデル，あるいはある一文化に特化したモデルを構築するためのデータや評価基盤が求められている．

LLM の能力を測るベンチマークは，一般的な知識タスク [9, 10] に加えて文化知識に焦点を当てた枠組みが提案され，地域差や偏りの可視化が進んでいる [11, 12, 13, 14]．多くのベンチマークは QA 形式で全体の正解率を主指標としており，文化依存的な

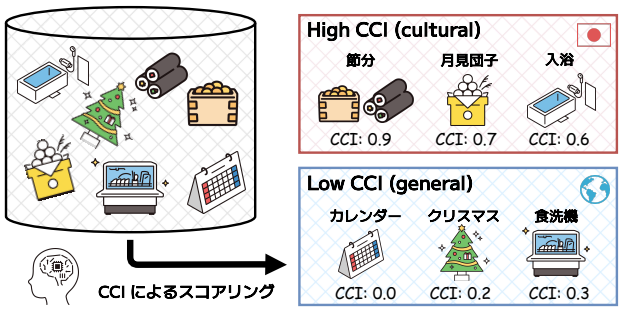

図 1　CCI によるスコアリング例．

質問は文化非依存な質問に比べて難しいことが報告されている [7, 15]．しかし，文化知識の中でも多地域で観察されるものから特定地域に固有のものまで幅があり，現行ベンチマークではその区別が十分に整理されておらず，誤答分析や改善方針の立案に直結しにくいと考えられる．またデータ資源の面でも，文化知識を大規模に収集したコーパスが提案されているが [16, 17]，各文がどの文化にどの程度特有かを示す注釈は付与されていない．

LLM の評価や資源整備の双方において，文単位で文化特有性を定量評価する枠組みが求められているが，現状は未整備である．文単位の文化特有性の人手注釈は，専門知識と文脈理解を要して工数が大きく，注釈者間合意の確保も難しいため自動化の必要性が高い．その実現には「文化」を数値化可能な形で定義する尺度の導入が不可欠である．しかし，文化は多面的かつ高次の概念であり，既存研究でも明示的な定義がほとんど与えられていない [18]．

本研究はこの課題に対処するため，文単位で文化特有性を定量化する新たな尺度 CCI（Conceptual Cultural Index）を提案する．CCI は LLM を用いて，複数の文化圏における文の一般性を推定し，対象文化圏の特異性を測定する尺度である．ある文化圏では一般的だが他文化圏では一般的でない文は高スコアに，逆に多くの文化圏で広く共有される文は低スコアとする．この設計により，比較対象の文化圏を

変更することで，対象文化圏に特有とみなす範囲と一般的とみなす範囲を操作できる．

2 関連研究

言語モデルの文化的側面に焦点を当てた評価ベンチマークとして，GeoMLAMA [11]，BLEnD [12]，CDEval [13]，CulturalBench [14] など，多地域を対象としたデータセットに加え，CLIcK [19]，IndoCulture [20]，CHARM [21] など特定の国・地域に焦点を当てたベンチマークが提案されている．これらのリソースはモデル間の文化知識の比較を可能にする一方，多くは QA 形式の正解率を主指標として用いており，誤答の要因が文化依存の困難さなのか，文化非依存の一般知識不足なのかが混在したまま評価されるため，両者を区別した分析が難しい．

評価ベンチマークに加えて，様々な国・地域の文化知識を大規模に収集したデータ資源として，StereoKG [22]，CANDLE [16]，CultureAtlas [23]，CultureBank [17]，MANGO [24] などが構築されている．しかし，文単位の文化特有性を注釈した資源はこれまでに整備されていない．提案する CCI はこれらを補う，文単位で文化特有性を連続値として測定する手法として位置付けられる．

3 CCI

本研究では，任意の文に対して文化特有性を測る尺度 Conceptual Cultural Index（CCI）を提案する．図 2 のように，対象文化圏とその他の文化圏集合を与え，各文化圏において入力文がどの程度一般的であるかを LLM に推定させる．その推定値を用いて，対象文化圏における相対的な特有性を数値化する．

3.1 一般性推定スコアの取得

入力文 x，文化圏集合 C，および対象文化圏 $t \in C$ が与えられたとき，まず LLM を用いて，対象文化圏を含む各文化圏 $c \in C$ において x がどの程度一般的かを推定し，連続値の一般性スコア $p_c \in [0,1]$ を得る．実装上は，単一のプロンプト内で C に含まれる全ての文化を一括で問い合わせ，JSON 形式の応答からスコアを抽出する．一般性スコアの推定に用いるプロンプトは 付録 A に示す．LLM 出力の実行間のばらつきを抑えるため，本研究では N 回の実行結果を平均する（本論文では $N = 3$ とする）：

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

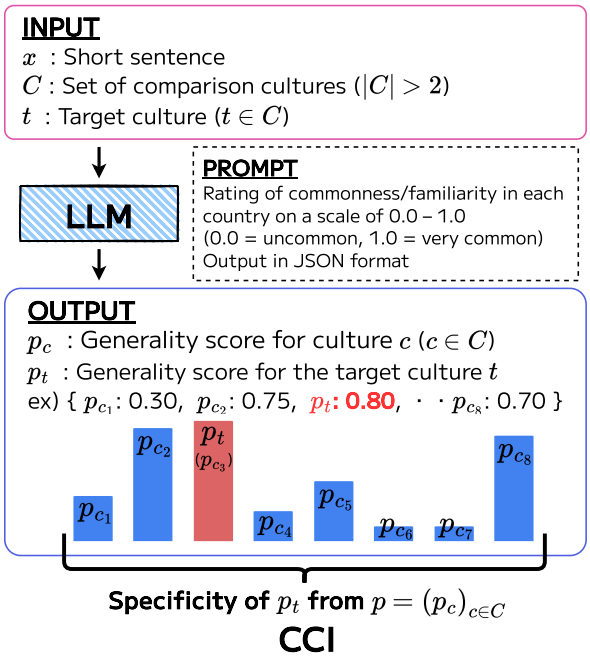

図 2 CCI の概要図.

ここで，$f_{\mathrm{LLM}}^{(n)}(x; C)[c]$ は，n 回目の実行において文化 c に対して返されるスコアを表す．

3.2 CCI の定義

我々は CCI を，対象文化圏における一般性スコアと，他の文化圏における一般性スコアの平均との差として次のように定義する：

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

CCI の値域は $[-1, 1]$ であり，値が 0 付近であれば x が文化横断的に一般的，1 付近であれば対象文化圏で特有，-1 付近であれば他の文化圏で特有であることを意味する．文化圏集合 C は目的に応じて柔軟に選択でき，どの文化圏を比較対象として含めるかによって，対象文化圏に特有とみなす範囲と一般的とみなす範囲（以下，文化スコープ）を調整できる．

4 実験

4.1 実験方法

CCI が文化特有性を適切に反映するかの検証のため，日本を対象文化圏 t とし，文化に依存する文（正例）と，特定の文化に依存しない一般文（負例）の 2 クラスに対して CCI を算出する．得られた CCI から日本文化の検出に関する ROC 曲線を描き，ROC 曲線下面積（area under the ROC curve; AUC）と，両クラスの中央値差により分離性能を評価する．あわ

表 1　文化依存文と一般文の CCI 中央値．C_{median} は文化依存文の中央値，G_{median} は一般文の中央値を表す．C_{median} は 1 に近く，G_{median} は 0 に近いことが望ましいが，全ての文が必ずしもこれらの極値に到達する必要はなく，全体としてこの傾向が観測されれば十分である．

Models	Baseline		CCI	CCI (Custom mode)	
		+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)

表 2　文化依存文と一般文の分離性能．

Models	Baseline AUC / Δ	CCI 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

せて，代表事例について定性的分析を行う．

Baseline として，LLM を用いて 文化特有性スコアを $[0, 1]$ の連続値として直接推定し，CCI と同一の手順で AUC を算出し，分類性能を比較する．CCI の場合と同様に，0 は文化横断的な一般性を，1 は対象文化に対する特有性を表す．

4.2　文化圏集合 C の選択

文化圏集合 C の選択が CCI に与える影響を評価する．本実験では，まず C を G20 に加盟している 19 か国[1)]で固定した設定を用いるとともに，文化スコープの操作可能性を検証するため，C を任意に選択した Custom mode でも実験を行う．

Custom mode　C をタスク目的に合わせて設定する．近隣文化を対象文化に含めるか否かを操作可能か検証するため，日本とアメリカを固定した状態で，次の 2 条件を定義する：

1. **+Neighbor Culture**：近隣諸国を C に含める [China, Republic of Korea, United States of America, Japan];
2. **−Neighbor Culture**：近隣諸国を C から除外する [Brazil, France, United States of America, Japan];

Baseline については，**+Neighbor Culture** に対応して，“If the practice is also common in neighboring or culturally adjacent countries, do not consider it specific to the target.” という明示的な指示文をプロンプトに含める条件と，含めない条件の 2 条件で比較した．

1) ここでは C を国名のみに制限するため，欧州連合およびアフリカ連合は含めない．

4.3　実験設定

Data　評価用コーパスは，GPT-5（2025 年 8 月 7 日時点のモデル）により文化依存文・一般文それぞれ 300 件の短文を生成し，重複や明らかな誤分類を人手で除外して選別した．なお，文化の境界は本質的に曖昧であるため，解釈の揺らぎを含む事例が混在し得る．最終的な評価セットは，文化依存文 200 文と一般文 200 文から構成される．

Models　CCI の算出に適した LLM を選定するため，我々は 5 つの LLM を対象に，共通のプロトコルのもとで CCI と Baseline を比較した．対象モデルは，多言語モデル（Llama 3.1 [3], Qwen 2.5 [4], gpt-oss [25]）と，日本語特化モデル（Llama 3.1 Swallow [26], LLM-jp 3.1 [27]）である．各モデルの正確な識別子は 付録 B に記載する．

5　結果と考察

5.1　文化依存文と一般文の分離性

表 1 は，各 LLM について，CCI および Baseline による特有性スコアに基づく文化依存文 (C_{median}) と一般文 (G_{median}) の中央値を示す．また 表 2 は，AUC によって測定した分離性能と，中央値の差 ($\Delta = C_{median} - G_{median}$) を示す．CCI は，Baseline と同等以上の AUC を達成するとともに，文化依存文にはより高いスコアを，一般文にはより低いスコアを割り当てることで，明確なクラス分離を実現した．対照的に Baseline では，多くの文に対して高いスコアを付与する傾向があり，モデルによっては両クラスの中央値が接近した．CCI が Baseline より適

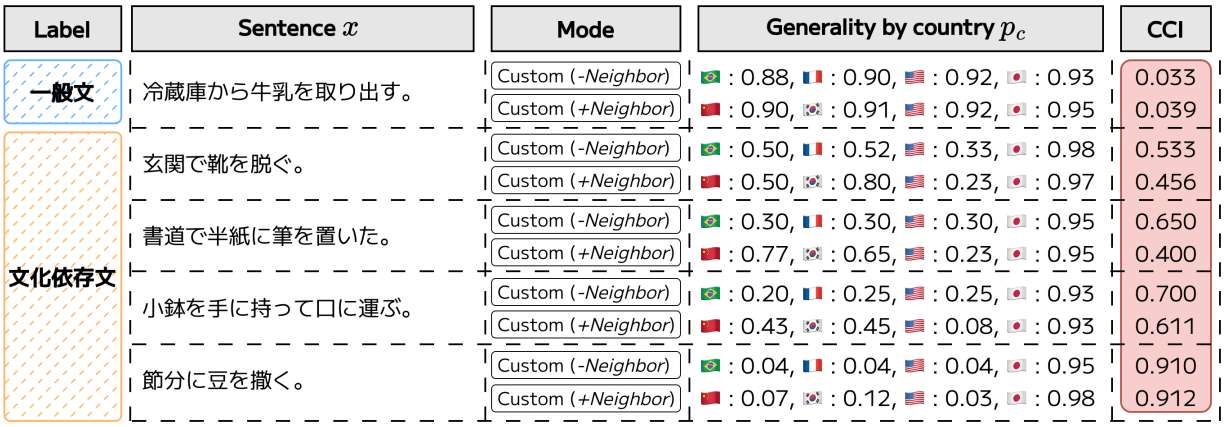

図 3　評価に用いた事例と gpt-oss で算出した CCI.

切にスコアリングできた要因として，「文化」を単一のスカラーとして直接定量化することが本質的に難しい一方で，CCI はタスクを文化圏ごとの一般性推定に分解する設計により，推論過程の安定化に寄与した点が挙げられる．

また，文化特有性スコアの算出に適した LLM について検討する．gpt-oss は CCI と Baseline のどちらでも，ほぼ理想的なクラス分離を達成した．この結果は，モデルサイズだけでなく，reasoning 指向の特性により，文化間の差異を段階的な推論を通じて捉えられている可能性を示唆する．さらに，日本語特化モデルは，多言語モデルと比べてより良いクラス分離を示した．以上より，強力な推論能力と対象文化に対する深い理解を備えつつ，他文化への理解も十分に有するモデルが適していると示唆される．

5.2　文化スコープの制御性

表 1 から，CCI の Custom mode（**+Neighbor**）は通常モードと比べて，文化依存文の中央値が低下していることがわかる．これは，近隣文化圏にも共通する慣習を過大評価しないよう，文化スコープを調整できていることを示す．一方，Baseline は入力言語の影響を受けやすく，プロンプト記述のみで文化スコープを厳密に制御することは難しいことが示唆された．さらに，CCI は文化特有性スコアに加えて文化ごとの一般性スコアも数値として提供するため，仮に Baseline の精度が十分な場合でも，「どの文化で一般的／非一般的と判断されたか」を確認できる点で解釈性が高い．

図 3 は，評価に用いた一部事例と，gpt-oss により算出された CCI を示す．事例別に見ると，日本と近隣文化でよく見られる慣習（例：書道で半紙に筆を置いた。）は，**+Neighbor** 条件下で低く，**−Neighbor** 条件下で高いスコアを取る．近隣文化でもタブーとみなされうる行為（例：小鉢を手に持って口に運ぶ。）は，**+Neighbor** 条件下でも過度に低いスコアを取るべきではない．以上より，近隣文化における同様の慣習が存在するか否かに応じて，**+Neighbor** と **−Neighbor** の条件の間でスコアに差が生じ，CCI が想定通りに文化スコープ制御を反映できている．

5.3　CCI の制限事項

本研究には主に 3 つの限界がある．第一に，実験では日本を対象文化とする設定に焦点を当てており，他言語・他地域への一般化可能性は今後の検証課題である．第二に，CCI は LLM が推定した一般性スコアに依存するため，基盤モデルのバイアスやキャリブレーションの問題を継承する可能性がある．第三に，文化を国レベルで近似しているため，国内の地域差を十分に捉えきれていない．今後は，文化集団の粒度を洗練し，広範な多言語・多地域評価を行うことで，CCI をより堅牢で汎用的な評価フレームワークへ発展させる．

6　おわりに

本研究では，対象文化圏における一般性推定値と，他文化圏における一般性推定値の差分として，文単位の文化特有性を定量化する尺度 Conceptual Cultural Index（CCI）を提案した．CCI は，LLM による直接推定では安定にスコア付与しにくい設定でも有効に機能し，明示的な定義に基づく解釈可能な算出を提供する．また，目的に応じて比較対象とする文化圏を変更することで，文化スコープを制御できる．本手法は特定モデルに依存しない評価枠組みとして設計されており，基盤モデルが発展しても，同一のプロトコルで適用可能である．

参考文献

[1] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language Models are Few-Shot Learners. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, Vol. 33, pp. 1877–1901, 2020.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, et al. GPT-4 Technical Report. **arXiv preprint arXiv:2303.08774**, 2023.

[3] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 Herd of Models. **arXiv preprint arXiv:2407.21783**, 2024.

[4] An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 Technical Report. **arXiv preprint arXiv:2412.15115**, 2024.

[5] Yong Cao, Li Zhou, Seolhwa Lee, et al. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In **Proceedings of the First Workshop on Cross-Cultural Considerations in NLP**, pp. 53–67, 2023.

[6] Tarek Naous, Michael J Ryan, Alan Ritter, et al. 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)**, pp. 16366–16393, 2024.

[7] Siqi Shen, Lajanugen Logeswaran, Moontae Lee, et al. 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)**, pp. 5668–5680, 2024.

[8] Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, et al. 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)**, pp. 2373–2403, 2025.

[9] Alex Wang, Amanpreet Singh, Julian Michael, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In **Proceedings of the 7th International Conference on Learning Representations**, 2019.

[10] Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring Massive Multitask Language Understanding. In **Proceedings of the 9th International Conference on Learning Representations**, 2021.

[11] Da Yin, Hritik Bansal, Masoud Monajatipoor, et al. GeoMLAMA: Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 2039–2055, 2022.

[12] Junho Myung, Nayeon Lee, Yi Zhou, et al. 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**, Vol. 37, pp. 78104–78146, 2024.

[13] Yuhang Wang, Yanxu Zhu, Chao Kong, et al. CDEval: A Benchmark for Measuring the Cultural Dimensions of Large Language Models. In **Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP**, pp. 1–16, 2024.

[14] Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, et al. 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)**, pp. 25663–25701, 2025.

[15] Shane Arora, Marzena Karpinska, Hung-Ting Chen, et al. 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)**, pp. 11772–11817, 2025.

[16] Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, et al. Extracting Cultural Commonsense Knowledge at Scale. In **Proceedings of the ACM Web Conference 2023**, p. 1907–1917, 2023.

[17] Weiyan Shi, Ryan Li, Yutong Zhang, et al. CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. In **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 4996–5025, 2024.

[18] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, et al. Towards Measuring and Modeling "Culture" in LLMs: A Survey. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 15763–15784, 2024.

[19] Eunsu Kim, Juyoung Suk, Philhoon Oh, et al. 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**, pp. 3335–3346, 2024.

[20] Fajri Koto, Rahmad Mahendra, Nurul Aisyah, et al. IndoCulture: Exploring Geographically Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 1703–1719, 2024.

[21] Jiaxing Sun, Weiquan Huang, Jiang Wu, et al. 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)**, pp. 11205–11228, 2024.

[22] Awantee Deshpande, Dana Ruiter, Marius Mosbach, et al. StereoKG: Data-Driven Knowledge Graph Construction For Cultural Knowledge and Stereotypes. In **Proceedings of the Sixth Workshop on Online Abuse and Harms**, pp. 67–78, 2022.

[23] Yi Fung, Ruining Zhao, Jae Doo, et al. Massively Multi-Cultural Knowledge Acquisition & LM Benchmarking. **arXiv preprint arXiv:2402.09369**, 2024.

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

[25] Sandhini Agarwal, Lama Ahmad, Jason Ai, et al. gpt-oss-120b & gpt-oss-20b Model Card. **arXiv preprint arXiv:2508.10925**, 2025.

[26] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, et al. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In **Proceedings of the First Conference on Language Modeling**, 2024.

[27] Akiko Aizawa, Eiji Aramaki, Bowen Chen, et al. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. **arXiv preprint arXiv:2407.03963**, 2024.

付録

A　使用したプロンプト

LLM に文化圏ごとの一般性スコアを出力させるためのプロンプトを，表 3 に示す．

表 3　CCI の算出に用いる文化圏ごとの一般性スコアを推定させるプロンプト．入力は文 x と文化圏集合 C である．

```
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.
```

B　使用したモデル

本実験で使用した 5 つの LLM を以下に示す．

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