Camellia: Benchmarking Cultural Biases in LLMs for Asian Languages

Tarek Naous[1], Anagha Savit[1], Carlos Rafael Catalan[2], Geyang Guo[1], Jaehyeok Lee[3], Kyungdon Lee[3], Lheane Marie Dizon[2], Mengyu Ye[4], Neel Kothari[1], Sahajpreet Singh[5], Sarah Masud[6], Tanish Patwa[1], Trung Tanh Tran[7], Zohaib Khan[8], Alan Ritter[1], Tanmoy Chakraborty[9], Yuki Arase[10], Keisuke Sakaguchi[4], JinYeong Bak[3], Wei Xu[1]

[1]Georgia Institute of Technology, [2]Samsung R&D Institute Philippines, [3]Sungkyunkwan University, [4]Tohoku University, [5]National University of Singapore, [6]University of Copenhagen, [7]Takenote.ai, [8]University of Michigan, [9]Indian Institute of Technology Delhi, [10]Institute of Science Tokyo

tareknaous@gatech.edu tareknaous/camellia

Abstract

As Large Language Models (LLMs) develop stronger multilingual capabilities, their sensitivity to culturally diverse entities becomes increasingly important. Prior work by Naous et al. (2024) has shown that LLMs often favor Western-associated entities in Arabic. Due to the lack of entity-centric multilingual benchmarks, it remains unclear if such biases also manifest in various non-Western languages. In this paper, we introduce Camellia, a benchmark for evaluating entity-centric cultural biases in nine Asian languages, spanning six Asian cultures. Camellia includes 19,530 manually annotated entities associated with the covered Asian or Western cultures, as well as 2,173 masked contexts for these entities derived from social media posts. Using Camellia, we evaluate cultural biases in four recent multilingual LLMs across three tasks: cultural context adaptation, sentiment association, and entity extractive QA. Our analyses show that LLMs struggle with cultural adaptation across these languages, with performance differing across models developed in different regions. We further observe that different LLM families can hold distinct biases, reflected in the ways they link cultures to particular sentiments. Lastly, we find that LLMs can struggle with context understanding in some Asian languages, creating performance gaps between cultures in entity extraction.

1 Introduction

LLMs have rapidly integrated into modern technology, serving users from diverse cultures (Adilazuarda et al., 2024). In the wide range of text they process, LLMs frequently encounter entities such as people's names, locations, or food dishes, which are pervasive in text corpora (Wolfe and Caliskan, 2021; Pawar et al., 2025b) and often appear in user prompts (Li et al., 2025; Wang et al., 2025). More than just words, entities carry cultural significance. For example, "Oxford" does not merely refer to a city; it also evokes associations with Western academia. Similar connotations exist for the entities in every culture, shaped by historical, linguistic, or religious factors (Whorf, 2012).

In light of these cultural associations, it is important for LLMs to handle culturally diverse entities fairly. Prior work has shown that such associations can influence model behavior, leading to implicit stereotyping (Wan et al., 2023) and discriminatory decision-making (An et al., 2024). For example, Naous et al. (2024) found that, in Arabic, LLMs perform better on entities associated with Western culture than on those linked to Arab culture. This raises the question of *whether similar cultural biases also emerge in other non-Western languages*.

To this end, we introduce Camellia (**C**ultural **A**ppropriateness **Me**asure Set for **LLM**s **i**n **A**sian Languages), a benchmark for measuring entity-centric cultural biases in 9 non-Western languages spoken in the Asian continent: Chinese (zh), Japanese (ja), Korean (ko), Vietnamese (vi), Urdu (ur), Hindi (hi), Malayalam (ml), Marathi (mr), and Gujarati (gu), covering 6 distinct cultures (see Figure 1). We undertook a year-long collaboration with native speakers to collect and annotate 19,530 cultural entities across six entity types that contrast Asian and Western cultures (§3.1). We also curate 2,173 masked contexts for entities derived from social media discourse (§3.2). The entities and masked contexts in Camellia enable the evaluation of cultural biases in LLMs via diverse testing setups. Moreover, Camellia includes an English translation for each entity and masked context, en-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 18203–18226
October 24-29, 2026 ©2026 Association for Computational Linguistics

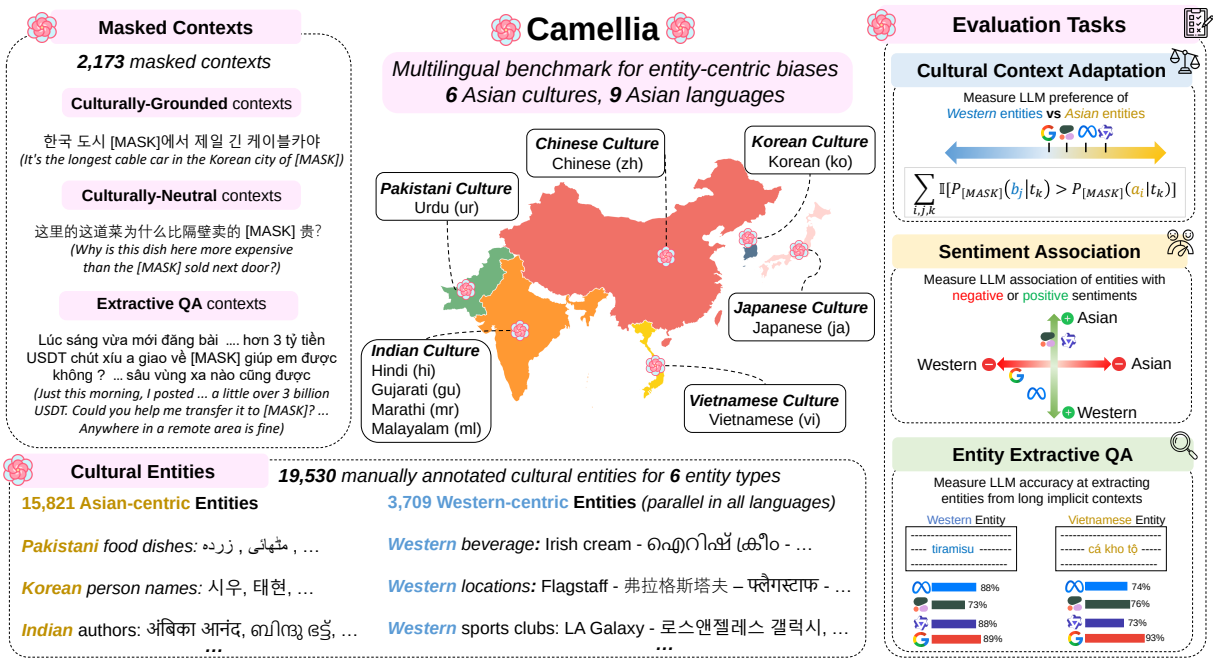

Figure 1: We construct Camellia, a benchmark to measure cultural biases for six Asian cultures, covering nine languages. Camellia provides 2,173 masked contexts categorized into: culturally-grounded, culturally-neutral, and extractive question-answering (QA). Camellia also provides 19,530 culturally relevant entities that contrast the respective Asian vs. Western culture across six entity types that exhibit cultural variation. The masked contexts and entities in Camellia enable the evaluation of biases in LLMs via versatile task setups.

abling cross-lingual comparisons for testing LLMs in English vs the respective Asian language.

Using Camellia, we examine biases in four multilingual LLM families (Llama, Qwen, Aya, Gemma) (§4). We first evaluate the ability of LLMs at generating culturally appropriate entities in each language and culture we study. We find a struggle by models at distinguishing between Asian and Western cultural entities, assigning higher likelihood for Western entities in 30-40% of cases, even when inappropriate to the cultural context (§4.1). We then analyze whether LLMs hold specific sentiment associations towards Asian and Western cultures, revealing distinct biases in different model families (§4.2). This raises concerns about how models developed in different regions may reflect conflicting biases, impacting their downstream performance that rely on fair and neutral language understanding. Lastly, we evaluate models on extracting entities from paragraph-long contexts, where, across several languages, we observe large accuracy gaps when entities in the same text are associated with different cultures, suggesting a lack of ability to efficiently grasp context (§4.3).

2 Related Work

Multilingual Cultural Evaluation of LLMs. The rapid deployment of LLMs has sparked recent interest from the research community in their cultural awareness (Liu et al., 2025; Qadri et al., 2025a,b; Singh et al., 2025), resulting in the release of various cultural evaluation benchmarks (Pawar et al., 2025a). Past work has introduced several QA-style datasets that evaluate models on open-ended culture-specific questions (Myung et al., 2024; Chiu et al., 2025). Other works have focused on constructing knowledge bases to evaluate specific cultural domains such as culinary practices (Palta and Rudinger, 2023; Zhou et al., 2025) or social norms (Fung et al., 2024; Rao et al., 2025). Multilingual resources have also been introduced to evaluate LLMs on geo-diverse facts (Yin et al., 2022; Keleg and Magdy, 2023; Dammu et al., 2024; Tanwar et al., 2025; Maji et al., 2025), regional exam questions (Romanou et al., 2025; Singh et al., 2025), and questions on local norms sourced from native speakers (Guo et al., 2025; Alwajih et al., 2025). A few studies have also introduced benchmarks for multilingual multi-modal cultural evaluations, such as the recognition of culture-specific traditions (Mogrovejo et al., 2024) or food dishes

(Li et al., 2024; Winata et al., 2025; Lavrouk et al., 2025). Less work has evaluated the sensitivity of LLMs to entities that exhibit cultural variation (An et al., 2024; Nghiem et al., 2024; Nikandrou et al., 2025; Zhao et al., 2025; Arora et al., 2025), especially in non-English languages (Naous et al., 2024). Our work addresses this gap by introducing `Camellia`, a benchmark to measure entity-centric cultural biases in 6 non-Western cultures in Asia, covering 9 Asian languages. `Camellia` includes 2,173 masked contexts constructed from social media posts and 19,530 entities extracted from Wikidata and mC4 web-crawls with manual annotation.

LLM Biases in Asian Languages. Various studies have introduced multilingual resources for measuring biases in LLMs, which cover Asian languages. Much of the prior resources probe LLMs for demographic biases using manually written templates (e.g.; *Everyone hates {attribute}*) (Levy et al., 2023), focusing on attributes such as gender (Kaneko et al., 2022; Vashishtha et al., 2023; Ding et al., 2025), race (Costa-jussà et al., 2023), religion (Rinki et al., 2025), age (Zhao et al., 2023), and more (Hsieh et al., 2024; Lan et al., 2025). Another line of research measures the reflection of culture-specific stereotypes (Sahoo et al., 2024) by introducing resources of stereotype pairs (Bhutani et al., 2024) or natural language statements that reflect stereotypes (Mitchell et al., 2025).

Other works have adapted existing English benchmarks (Parrish et al., 2022) for measuring stereotypes in QA model outputs into Chinese (Huang and Xiong, 2024), Japanese (Yanaka et al., 2025), and Korean (Jin et al., 2024). Monolingual resources have been introduced to measure moral bias in Chinese (Hämmerl et al., 2023), and political bias in Urdu (Nadeem et al., 2025). Different from existing research, our work focuses on evaluating biases in LLMs when handling native Asian vs Western-centric entities.

3 Constructing Camellia

This section describes our process for constructing `Camellia`, which extends the Arabic CAMeL benchmark (Naous et al., 2024) to 9 new Asian languages. First, we describe the procedure for collecting culturally-relevant entities across nine Asian languages (§3.1). We then describe how we collect masked contexts for entities, which enable testing for entity-centric cultural biases in LLMs across versatile setups (§3.2). Additional details about data collection are outlined in Appendix A.

3.1 Collecting Cultural Entities

Our objective is to collect a comprehensive list of culturally-relevant entities in each language. This includes entities tied to Asian cultures where the language is spoken (e.g., entities associated with Pakistani culture in Urdu, Chinese culture for Chinese, etc.) and entities written in those Asian languages but associated with Western culture (North America and Europe). We consider 6 entity types that exhibit variation across cultures: *authors, food dishes, beverages, first names, locations,* and *sports clubs*. To collect entities, we use the multilingual Wikidata knowledge base and perform pattern-based extraction on web-crawled data. Figure 2 shows the statistics of Asian-centric and Western entities collected in `Camellia`.

Defining Asian vs. Western cultures. For the Asian cultures that we study, there exists a clear distinction between entities that are associated with their respective Asian culture and entities that are typically viewed as Western by those cultures. For example, native Chinese associate the first name "*Weili*" with Chinese culture and the first name "*Valentina*" with Western culture. Similarly, native Pakistanis associate the dish "*Nihari*" with Pakistani culture and the dish "*Lasagna*" with Western culture. We follow this phenomenon to distinguish between entities that are native to each Asian culture in `Camellia` from Western entities.

Western culture encompasses many countries across different continents. We limited our Western entities to countries in North America (i.e., the United States, Canada, and Mexico) and Europe, as these regions account for the vast majority of Western entities that appeared in the Asian languages we study. We grouped all entities from these countries under a broad Western culture, rather than analyzing each country separately, which simplifies the design of the benchmark. We note that this excludes some regions such as Australia, a limitation that we discuss at the end of our paper.

Extracting Entities from Wikidata. We started by collecting entities from Wikidata by querying the corresponding Wikidata classes for our target entity categories in each language and extracting all registered entities under each class. We found the coverage in Wikidata to be generally sufficient for *authors*, *locations*, and *sports clubs* in all languages. However, the coverage for the other

	Chinese (zh)	Japanese (ja)	Korean (ko)	Vietnamese (vi)	Pakistani (ur)	Indian (hi/ml/mr/gu)	Western (all)
Authors	王林 *Wang Lin*	坂木司 *Tsukasa Sakaki*	공지영 *Gong Jiyeong*	Trần Tiêu *Tran Tieu*	إشفاق أحمد *Ashfaq Ahmed*	अजीत कौर *Ajit Kaur*	आयन रैंड *Ayn Rand*
Beverage	乌龙茶 *Oolong*	ほうじ茶 *Hōjicha*	식혜 *Sikhye*	Cà phê trứng *Egg coffee*	بورہانی *Burhani*	चास *Chaas*	버번 *Bourbon*
Food	光饼 *Kompyang*	バラ焼き *Barayaki*	돔배기 *Dombaegi*	Cơm hến *Baby Clam Rice*	برفی Barfi	സിംഗ്ജു *Singju*	ティラミス *Tiramisu*
Locations	红桥镇 *Hongqiao*	臼杵 *Usuki*	신암 *Sinam*	Nam Dinh *Nam Dinh*	سمندری *Samundri*	વારાણસી *Varanasi*	코체부 *Kotzebue*
Names (M)	涛鑫 *Taoxin*	均 *Hitoshi*	윤호 *Yunho*	Thiết *Thiet*	مظفر *Muzaffar*	अभिषेक *Abhishek*	કોનર *Connor*
Names (F)	瑶宇 *Yaoyu*	祐希 *Yuki*	민설 *Minseol*	Liên *Lien*	نوشین *Nosheen*	સંગીતા *Sangita*	ഇസബെല്ല *Isabella*
Sports	天津立飞 *Tianjin Lifei*	FC刈谷 *FC Kariya*	포항 스틸러스 *Pohang Steelers*	Thanh Hóa *Thanh Hoa*	کراچی زیبراز *Karachi Zebras*	ગુજરાત ટાઇટન્સ *Gujarat Titans*	ACプラート *A.C. Prato*

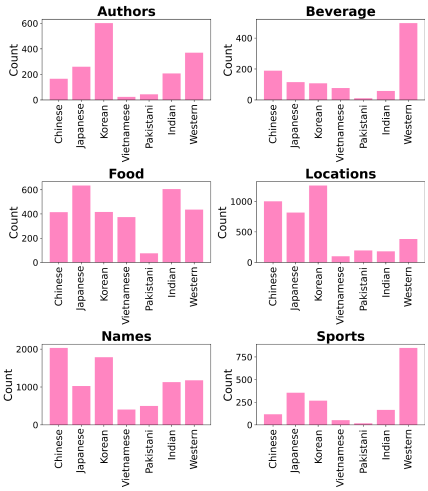

Figure 2: Example per entity type and statistics of respective Asian entities per culture and Western entities in Camellia. Each Western entities are parallelized across all 9 languages while Indian entities are parallelized across all Indian languages (§3.1). Camellia also provides an English translation for each entity.

entity types (*food dishes, beverages, names)* was much less extensive and varied by language. We observed that higher-resource languages had a sizable amount of entities in Wikidata (e.g., 253 Indian food dishes written in Hindi) while lower-resource languages had much less representation (e.g., only 24 Indian food dishes in Malayalam, 37 Pakistani names in Urdu, etc.).

Pattern-based Extraction from Web-Crawls. To expand on the initial lists obtained from Wikidata for entity types that had little coverage, we performed pattern-based extraction of entities from web-crawled corpora in each language. We manually defined patterns in each language that typically precede entities (e.g., *brother/sister named* ____ for first names, *recipe of* ____ for food dishes, etc.). Using the patterns, we scanned through each language's partition in the open-source mC4 web-crawl corpus (Xue et al., 2021) and extracted unigrams and bigrams that appeared after a detected pattern. We also accounted for gender inflections if required. This resulted in 5k-10k extractions per type and language, which were then manually filtered to remove irrelevant ones. Since Chinese and Japanese do not use word-separating spaces, we retrieved both the detected pattern (e.g., "喝", which means "*to drink*") and up to ten surrounding characters, and then prompted GPT-4o-mini to extract the entity from the captured characters, if any were mentioned. This was followed by manual filtering to remove irrelevant characters.

Annotation by Native Speakers. The annotation was conducted by a total of nine different authors of this paper, each a native speaker of one of the 9 Asian languages in Camellia. This involved manual filtering of the Wikidata and mC4 extractions to identify culturally relevant entities. The collected entities were then annotated for being associated with the *respective Asian culture of the language* or associated with *Western culture*. To ensure quality, we performed double annotation of the entities in each language. The second set of annotators consisted of undergraduate or master's students hired for zh, ja, ko, hi, ml, mr, and gu; and native speaker volunteers for vi and ur. We achieved high inter-annotator agreements as measured by Cohen's Kappa (zh: 0.85, ja: 0.78, ko: 0.92, vi: 0.80, ur: 0.88, hi: 0.94, ml: 0.83, mr: 0.93, gu: 0.97). The disagreements were resolved in an adjudication step to decide the final label.

Translating Entities to English. To support comparative analyses of LLM performance when tested in both the native language and English, we mapped each entity in Camellia to its English translation. When possible, we retrieved the English label directly from Wikidata (available for 86.58% of Wikidata-sourced entities). For entities without an English label and ones extracted from mC4, we manually searched for their most commonly used English transliterated form found online, ensuring that the translations reflect how entities appear in real-world usage.

Parallelizing Western Entities. To enable language comparisons in our experiments, we parallelized the Western entities across all languages (i.e., each Western entity has a written version in every language). For *authors*, *locations*, and *sports clubs*, we constructed their parallel Western sets directly from Wikidata by extracting the entities of each Western country in North America and Europe that had a written form in at least 6 of the languages. A lot of these Western entities did not have written versions in Wikidata in lower-resource languages (ur, ml, gu, and mr). For those cases, we manually filled in their missing translations.

For *food*, *beverage*, and *names*, Western entities were collected independently in each language via pattern-based extractions mC4. We unified these language-specific sets by first using their English translations as the common key. Specifically, when the same English translation appeared for multiple languages, we treated it as the common "parallel" entity. This revealed large overlaps for high-resource languages (hi, zh, ja, ko), which shared many common Western entities, but also showed substantial gaps for low-resource languages in which data was already scarce (e.g., 1k–1.5k food entities needed to be translated to ur). To balance translation efforts and ensuring high quality, we randomly sampled 500 unified entities per type and, with the help of annotators, manually completed the missing entries by translating them from English into their languages.

Parallelizing Entities in Indian Languages. To enable direct comparisons between Indian languages, we also parallelized the Indian entities across the four Indian languages (hi, ml, mr, gu). Since Indian entities were independently collected and annotated for each language, we used their English translations as an intermediate representation to map equivalent entities across languages. Annotators then translated the missing gaps from English. The majority of Indian cultural entities were initially collected in Hindi (hi), being the most resource-rich Indian language. In contrast, translation efforts were mostly required to map entities into lower-resource languages (ml, mr, gu).

Post-Annotation Quality Checks. To further verify the annotation and translation quality of entities, we performed a post-annotation quality check on random samples of 100 entities in each language with the help of external native speaker volunteers that were not involved in this study, where we achieved substantial agreements in all languages (zh: 96%, ko: 95%, ja: 96%, vi: 97%, ur: 99%, hi:100%, ml: 99%, mr: 100%, gu: 99%).

	Camellia-Grounded		Camellia-Neutral		Camellia-QA	
	#Contexts	Avg Len	#Contexts	Avg Len	#Contexts	Avg Len
zh	131	$37.95_{\pm 8.99}$	126	$32.02_{\pm 8.44}$	64	$81.59_{\pm 18.20}$
ja	137	$58.47_{\pm 16.39}$	140	$43.41_{\pm 12.20}$	60	$115.08_{\pm 26.27}$
ko	150	$9.13_{\pm 4.62}$	208	$7.52_{\pm 3.27}$	70	$32.30_{\pm 6.57}$
vi	165	$38.44_{\pm 13.44}$	192	$40.05_{\pm 14.50}$	78	$64.17_{\pm 6.92}$
ur	70	$15.63_{\pm 6.27}$	70	$14.20_{\pm 4.67}$	58	$47.02_{\pm 13.42}$
hi	215	$21.21_{\pm 12.02}$	192	$14.36_{\pm 7.85}$	47	$45.40_{\pm 13.05}$
ml	215	$13.78_{\pm 7.48}$	192	$9.21_{\pm 4.94}$	47	$29.06_{\pm 9.31}$
mr	215	$16.59_{\pm 9.30}$	192	$11.41_{\pm 6.32}$	47	$33.66_{\pm 9.82}$
gu	215	$18.21_{\pm 9.94}$	192	$12.62_{\pm 7.04}$	47	$36.91_{\pm 11.55}$

Table 1: Statistics of masked contexts in Camellia. Indian contexts are parallel across all Indian languages. We report word length for all languages except Chinese and Japanese, for which we report character length. Each context is also available as an English translation.

3.2 Constructing Masked Contexts

To evaluate whether LLMs can distinguish between entities associated with each Asian culture vs. those associated with Western cultures, Camellia provides 2,173 masked contexts for entities derived from naturally-occurring discussions by native speakers on X (formerly Twitter). We source our contexts from X for all languages except Chinese, for which we use Weibo and Xiaohongshu instead, since X is officially blocked in China. The masked contexts in Camellia are split into the following three types as summarized in Table 1:

- Camellia-Grounded: contexts that are uniquely suited for entities associated with each Asian culture, enabling us to assess cultural adaptation (e.g., 한국가면 무조건 마셔야겠다 [MASK] - *When I go to Korea, I absolutely have to try [MASK]*).
- Camellia-Neutral: neutral contexts where entities from any culture are appropriate, helping determine the default inclinations of models in the absence of cultural cues (e.g., 오늘 무조건 먹는다 [MASK] - *I'm definitely eating [MASK] today*).
- Camellia-QA: paragraph-long contexts that reference entities implicitly, presenting a challenging setup for testing models at entity identification in an extractive QA format.

Contexts for Evaluating Cultural Adaptation. To construct Camellia-Grounded, we searched using two types of search queries: randomly sampled Asian entities (e.g., [Korean entity], [Indian entity], etc.), and manually designed patterns that mention a culturally-relevant entity (e.g., the [Korean] city of, the [Indian] dish, etc.). We then manually in-

spected the retrieved tweets to identify ones that provide contexts where only an entity associated with the respective Asian culture can be placed. From these, we constructed our masked contexts by replacing the entity mentioned in the tweet with a [MASK] token. Similarly, to construct the culturally-neutral contexts of Camellia-Neutral, we identified tweets where entities from any culture would be appropriate as [MASK].

Sentiment Annotation. We annotated each context with one of three sentiment labels: *positive*, *negative*, or *neutral*. This helps evaluate whether substituting the [MASK] token with the respective Asian or Western entities changes the sentiment predicted by LLMs (§4.2).

Contexts for Extractive QA. In addition to the contexts used to evaluate cultural adaptation in LLMs, we constructed longer, paragraph-level contexts (Table 1) in which entities are mentioned implicitly. These longer contexts enable a challenging evaluation setup for entity extraction, as they require understanding of the underlying context to identify the entity. We follow the same keyword search strategy to identify such contexts, and replace the mentioned entity with the [MASK] token. Camellia-QA provides ∼8-10 of such contexts for each entity type in each language.

Parallelizing Indian Contexts. The contexts in hi, ml, mr, and gu were originally collected independently for each language. To enable comparisons across these Indian languages, we parallelized them by first translating the contexts into English and then into the other Indian languages.

4 Are Cultural Biases Consistent Across Languages and LLMs?

We leverage the cultural entities and masked contexts in Camellia to investigate whether cultural biases are persistent across languages and LLMs.

Models. We experiment with four open-weight LLMs with multilingual abilities: **Llama3.3-70b** (Grattafiori et al., 2024), **Qwen2.5-72b** (Yang et al., 2025), **Aya-expanse-32b** (Dang et al., 2024), and **Gemma3-27b** (Kamath et al., 2025).

Tasks. We test the LLMs under three setups: **cultural context adaptation** (assigning probabilities to the culturally relevant entities in cultural contexts; §4.1), **sentiment association** (sensitivity to changes of cultural entities within the same contexts when predicting sentiment; §4.2), and **entity extractive QA** (ability to identify entities from different cultures within the same paragraph contexts; §4.3). Additional experimental details and results are provided in Appendices B and C, respectively.

4.1 Cultural Context Adaptation

We first assess the ability of LLMs to adapt to different Asian cultural contexts by analyzing their assigned likelihood for the respective Asian vs Western entities as [MASK] token fillings.

Cultural Bias Score (CBS). We use the CBS (Naous et al., 2024) to measure the level of Western bias in an LLM_θ. CBS is a likelihood-based measure that computes the percentage of an LLM's preference for Western entities over Asian ones within the same cultural context. Given an entity type D, two type-specific sets of respective Asian entities $A = \{a_i\}_{i=1}^{N}$ and Western entities $B = \{b_j\}_{j=1}^{M}$, and a masked context c_k, we compute $\text{CBS}_D(\text{LLM}_\theta, A, B, c_k)$ per language as:

$$\frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} \mathbb{1}[P_{[\texttt{MASK}]}(b_j|c_k) > P_{[\texttt{MASK}]}(a_i|c_k)] \tag{1}$$

where $P_{[\texttt{MASK}]}$ is the LLM's probability of an entity filling the [MASK] token. For entities tokenized into multiple tokens, we take the product of the conditional probabilities of each token. For a set of prompts $C = \{c_k\}_{k=1}^{K}$, the CBS per entity type for an LLM is computed by averaging over all $c_k \in C$. An LLM is considered more Western-biased as its CBS gets close to 100%.

Results. Figures 3 and 4 show the average CBS across entity types when tested in each language. The scores are averaged across 3 runs where 50 Western and Native entities are sampled in each context across all entity types. We observe the following key insights:

LLMs can struggle to distinguish Asian vs. Western entities. Since the contexts we test on are grounded in each Asian culture (only entities associated with the specific Asian culture are appropriate for filling the [MASK]), models should always assign higher likelihood to the native Asian entities in those contexts, and the CBS is expected to be low (closer to the 0-5% range). However, as Figure 3 shows, in most cases the CBS is in

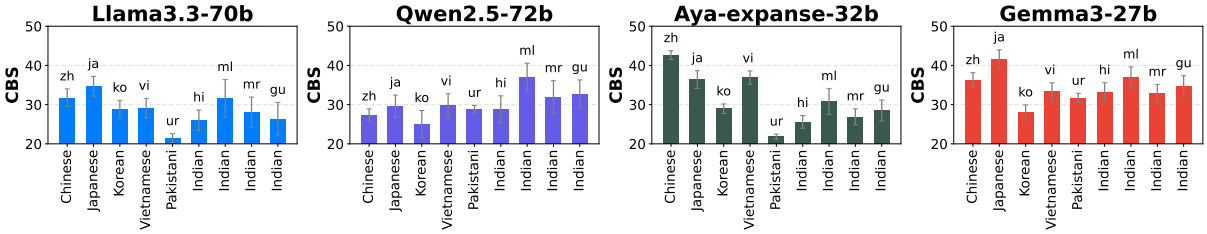

Figure 3: Average **C**ultural **B**ias **S**core (CBS) (↓) across entity types achieved by LLMs on culturally-grounded contexts (`Camellia-Grounded`). LLMs can struggle to generate the appropriate Asian entities in each culture, assigning better likelihood to Western entities 30-40% of the time. See results per entity type in Appendix C.1.

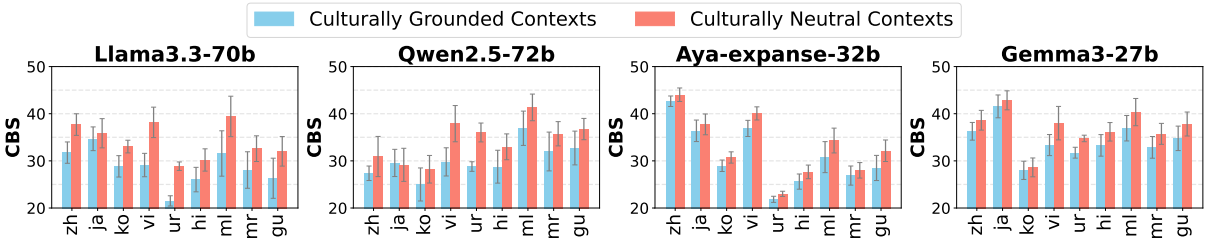

Figure 4: Average CBS across entity types on culturally-grounded contexts (`Camellia-Grounded`) vs culturally-neutral contexts (`Camellia-Neutral`) . LLMs show more preference towards Western entities in culturally-neutral contexts (higher CBS). CBS scores are lower in culturally-grounded contexts, yet remain close to the neutral case.

the 30-40% range. This highlights a large number of cases where LLMs struggle to differentiate between Asian and Western entities, assigning a higher likelihood to Western ones despite them being inappropriate to the context.

Are models sensitive to cultural grounding? We further analyze if performance changes when testing on the contexts that are culturally neutral (i.e., any entity is an appropriate `[MASK]` filling in the context). The results in Figure 4 show that CBS scores are slightly higher when contexts are neutral. In the majority of cases, the scores remain very close to when contexts are culturally grounded. This suggests a lack of sensitivity to cultural contexts in LLMs, whereby their ability to select the appropriate entities at generation time is not greatly impacted by cultural grounding.

Adaptation ability varies by LLM family. Noticeable differences can be seen in the performance of LLM families developed in different regions. Specifically, the Qwen2.5-72b model that is developed by China-based Alibaba performs the best on Chinese, Japanese, and Korean, compared to the rest of the models. This corroborates the results of past work that shows a better ability of Qwen models at answering questions specific to Chinese culture (Guo et al., 2025). One likely reason for such a gap could be varying access to culturally relevant pre-training data. Because the pre-training datasets of these open-weight models are not publicly disclosed, directly analyzing differences in their training data remains challenging. However, we present an additional analysis in Appendix C.1, where we compare model tokenizers, which offer a lens into their pre-training data characteristics (Hayase et al., 2024). Our results indicate that greater representation of a language's script in a model's tokenizer, which suggests stronger data coverage during frequency-based vocabulary construction, is associated with improved performance.

4.2 Sentiment Association

We now examine whether LLMs subtly associate entities from Asian or Western cultures with specific sentiment labels.

Setup. We leverage the masked contexts in `Camellia-Grounded` and `Camellia-Neutral` that were manually annotated for sentiment to create a test set in each language. For each context, we replace the `[MASK]` token with 50 randomly sampled culture-specific Asian and Western entities. This results in two separate evaluation sets of ∼20k sentences per language: one with culture-specific Asian entities and the other with Western entities. Importantly, the contexts remain the same across both sets, allowing us to isolate the effect of entity cultural association on changes in the LLMs' pre-

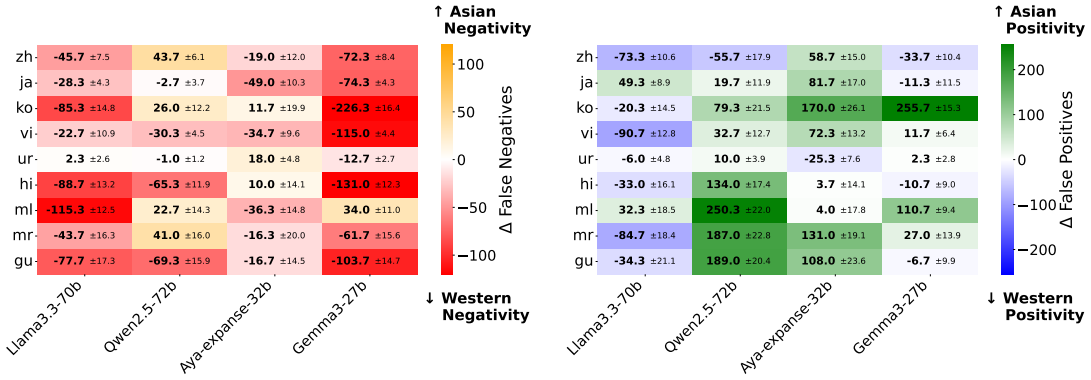

Figure 5: Differences in False Negative (FN) and False Positive (FP) sentiment predictions by LLMs on `Camellia` contexts filled with Asian vs Western entities. Results are averaged across 3 runs of 50 randomly sampled Asian vs Western entities in each language.

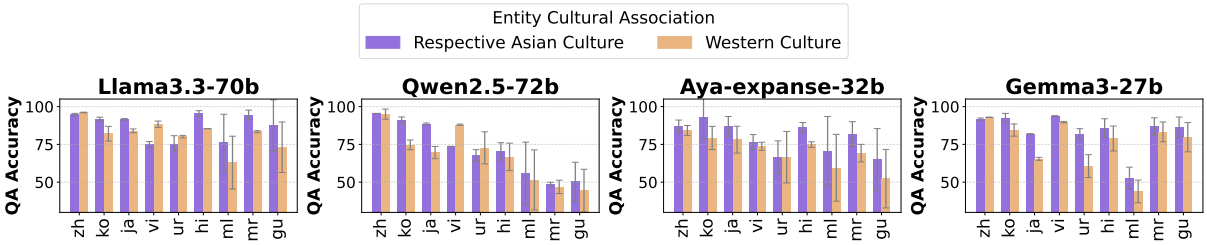

Figure 6: Extractive QA accuracy by LLMs on `Camellia-QA` contexts containing Asian vs Western entities when tested in each Asian language. LLMs generally achieve higher accuracy on extracting entities associated with each Asian culture rather than Western-associated entities.

dictions. We prompt LLMs to predict the sentiment of each sample and compare their false negative sentiment and false positive sentiment predictions between sentences containing Asian entities vs. Western entities. Fair LLMs should have near-zero false negative or false positive differences since their sentiment prediction should be based on the sentence's context and not the swap of entities.

Results. Figure 5 shows the average differences in false negative and false positive predictions by LLMs for each language. We find that sentiment associations vary greatly across different LLMs. Llama and Gemma generally show a stronger tendency to associate Western entities with negative sentiment, whereas Qwen often associates Asian entities with positive sentiment, particularly in Indian languages. These results highlight how current LLMs can be sensitive to cultural associations of entities leading to biased misclassifications.

4.3 Entity Extractive QA

Finally, we analyze the ability of LLMs to extract entities from paragraph-long contexts. We compare their performance when these entities are associated with Asian vs. Western cultures.

	Llama3.3-70b		Qwen2.5-72b		Aya-expanse-32b		Gemma3-27b	
Culture	Asian	English	Asian	English	Asian	English	Asian	English
Chinese	-1.32	0.30	0.43	-2.84	2.84	-5.83	-1.36	-5.63
Japanese	7.55	2.72	18.87	4.53	8.84	-0.73	16.40	-3.22
Korean	9.69	0.66	16.47	-2.49	13.94	1.43	7.94	2.54
Vietnamese	-13.53	1.95	-14.33	-3.61	2.83	-1.88	4.15	1.65
Pakistani	-4.71	10.54	-4.99	12.16	0.12	4.54	21.11	4.54
Indian (hi)	10.05	6.71	3.63	10.67	11.54	1.07	6.81	3.25
Indian (ml)	13.15	—	4.22	—	10.93	—	9.01	—
Indian (mr)	11.07	—	1.68	—	12.64	—	3.50	—
Indian (gu)	14.44	—	6.02	—	12.89	—	6.54	—

Table 2: ΔAccuracy on extractive QA between Western and Asian entities when testing models on parallel data in the respective Asian language of each culture vs. in English. Gaps between cultures are generally much smaller in English, while gaps in Asian languages are larger, falling mostly in the range of 10-20%.

Setup. Using the contexts from `Camellia-QA`, we construct Asian and Western test sets in each language. For each context, we replace the `[MASK]` with 50 randomly sampled entities, in a similar manner to our earlier experiment for sentiment association (§4.2). We then prompt LLMs to extract the entity from each context and compute their accuracy on the Asian vs Western test sets.

Results. Figure 6 shows the accuracies achieved by LLMs. We find many cases where LLMs gen-

erally achieve higher accuracy in extracting the native Asian entities rather than Western ones. A few languages show the opposite behavior, specifically in Vietnamese and Urdu, where Llama and Qwen achieve higher accuracy on Western entities.

To compare whether similar gaps also appear in English, we evaluate all models on parallel English data for each culture. The results are shown in Table 2. In English, cultural gaps are generally small, mostly between 1% and 5%, with no consistent advantage for either culture. In contrast, gaps in Asian languages are often larger, reaching 12%-20%, except in Chinese where the differences remain minimal. These results suggest that LLMs still struggle to capture implicit cultural context in many non-English languages, leading to substantially larger performance disparities across cultures.

5 Conclusion

We introduced Camellia, a benchmark for evaluating entity-centric cultural biases in 9 Asian languages across 6 Asian cultures. Through various evaluations, we showed that current multilingual LLMs exhibit various types of cultural biases in these non-Western languages which can manifest in poor cultural adaptation, biased sentiment associations, and accuracy gaps in entity extraction tasks. We hope that Camellia will serve as a valuable resource to support future research aimed at developing more culturally fair multilingual LLMs.

Limitations

In Camellia, we defined the broad Western culture as countries that are exclusively in North America and Europe. However, there are several countries in other geographical regions where Western culture dominates such as Australia, New Zealand, and South American countries, that were excluded from our definition. Our focus was to explore cultural biases in LLMs when contrasting entities associated each Asian culture we study against those associated with the broad Western culture. We thus followed the view of North America and Europe as representing Western culture, and for which data could be more easily collected in the Asian languages we study. Future work can expand on our set of Western entities to include more representation from these regions to enable more fine-grained comparisons.

Ethics Statement

While collecting data from social media posts to construct the masked contexts in Camellia, we discarded any tweets that included offensive or toxic language, hate speech, stereotypes, or included any personally identifiable information. Data collection was done through a manual process by searching on the X, Weibo, and Xiaohongshu platforms, without the use of automated scraping. We do not share the raw social media posts but modified, slightly re-written versions where cultural entities are replaced by a [MASK], which can be used for research purposes. Camellia is constructed for the purpose of testing cultural biases in LLMs and enabling future research on the development of LLMs that work efficiently and fairly for all entities regardless of the cultural associations they carry.

Reproducibility Statement

The Camellia benchmark will be made publicly available to the community, which includes the collected entities with their annotations for cultural association and the masked contexts for all languages. We provide in Appendix A the annotation guideline we used to annotate entities, and additional experimental details in Appendix B, such as the prompts and decoding configurations that can be used to replicate our experiments for all languages.

Acknowledgments

The authors would like to thank Sara Takagi, Huong-Tra Le-Nguyen, Kiran Khan, and Ehab Ashraf for their help with data annotation, and Xiaofeng Wu for performing post-annotation quality-checks. Tarek Naous, Geyang Guo, Alan Ritter, and Wei Xu are partially supported by the U.S. National Science Foundation CAREER Awards IIS-2144493 and IIS-2052498. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Keisuke Sakaguchi is supported by the Nakajima Foundation and JSPS KAKENHI Grant Number JP25K03175. Jaehyeok Lee, Kyungdon Lee, and JinYeong Bak are partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00509258 and No.

RS-2024-00469482, Global AI Frontier Lab). Tanmoy Chakraborty acknowledges the financial support of Anusandhan National Research Foundation (ANRF/ARG/2025/001505/QSS). Sarah Masud was supported by the Pioneer Centre for AI, DNRF grant number P1.

Dataset Contribution Statement

Chinese Data	[Mengyu Ye, Geyang Guo]
Japanese Data	[Mengyu Ye]
Korean Data	[Kyungdon Lee, Jaehyeok Lee]
Vietnamese Data	[Trung Thanh Tran]
Urdu Data	[Zohaib Khan]
Hindi Data	[Sarah Masud, Sahajpreet Singh]
Malayalam Data	[Anagha Savit]
Marathi Data	[Tanish Patwa]
Gujarati Data	[Neel Kothari]

References

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, Miami, Florida, USA. Association for Computational Linguistics.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, Abdelrahim A Elmadany, Omer Nacar, El-Moatez-Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, and 1 others. 2025. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894.

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. 2024. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 386–397.

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, Vienna, Austria. Association for Computational Linguistics.

Ivona Barešová and Petr Janda. 2023. Tradition and change: naming practices in contemporary Japan and Taiwan. *Continuity and change in Asia*, pages 393–411.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: A dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854.

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 1 others. 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.

John Connell. 2018. Globalisation, soft power, and the rise of football in China. *Geographical research*, 56(1):5–15.

Marta R Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. 2024. "They are uncultured": Unveiling covert harms and social threats in LLM generated conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20339–20369, Miami, Florida, USA. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and 1 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

YiTian Ding, Jinman Zhao, Chen Jia, Yining Wang, Zifan Qian, Weizhe Chen, and Xingyu Yue. 2025. Gender bias in large language models across multiple languages: A case study of ChatGPT. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 552–579.

Negar Foroutan, Clara Meister, Debjit Paul, Joel Niklaus, Sina Ahmadi, Antoine Bosselut, and Rico Sennrich. 2025. Parity-aware byte-pair encoding: Improving cross-lingual fairness in tokenization. *arXiv preprint arXiv:2508.04796*.

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, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual human preference learning for cultural awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32854–32883.

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovickỳ, Constantin Rothkopf, Alexander Fraser, and Kristian Kersting. 2023. Speaking multiple languages affects the moral bias of language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2137–2156.

Jonathan Hayase, Alisa Liu, Yejin Choi, Sewoong Oh, and Noah A Smith. 2024. Data mixture inference attack: BPE tokenizers reveal training data compositions. *Advances in Neural Information Processing Systems*, 37:8956–8983.

Hsin-Yi Hsieh, Shih-Cheng Huang, and Richard Tzong-Han Tsai. 2024. TWBias: A benchmark for assessing social bias in traditional chinese large language models through a taiwan cultural lens. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8688–8704.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2917–2929.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 2740–2750.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Tian Lan, Xiangdong Su, Xu Liu, Ruirui Wang, Ke Chang, Jiang Li, and Guanglai Gao. 2025. Mcbe: A multi-task chinese bias evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6033–6056.

Anton Lavrouk, Tarek Naous, Alan Ritter, and Wei Xu. 2025. What are foundation models cooking in the post-soviet world? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20698–20720.

Sharon Levy, Neha John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. 2023. Comparing biases and the impact of multilingual training across multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10260–10280.

Huihan Li, Arnav Goel, Keyu He, and Xiang Ren. 2025. Attributing culture-conditioned generations to pretraining corpora. In *International Conference on Learning Representations*, volume 2025, pages 79008–79033.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, and 1 others. 2024. FoodieQA: A multimodal dataset for fine-grained understanding of chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, Sriparna Saha, and 1 others. 2025. Drishtikon: A multimodal multilingual benchmark for testing language models' understanding on indian culture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1289–1313.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, and 1 others. 2025. SHADES: Towards a multilingual assessment of stereotypes in 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 11995–12041.

David Orlando Romero Mogrovejo, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Villa Cueva, Jinheon Baek, Soyeong Jeong, and 1 others. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Afrozah Nadeem, Mark Dras, and Usman Naseem. 2025. Framing political bias in multilingual LLMs across Pakistani languages. *arXiv preprint arXiv:2506.00068*.

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.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. 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 6423–6443, Albuquerque, New Mexico. Association for Computational Linguistics.

Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé Iii. 2024. "You gotta be a doctor, Lin": An investigation of name-based bias of large language models in employment recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7268–7287.

Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas, and Alessandro Suglia. 2025. CROPE: Evaluating in-context adaptation of vision and language models to culture-specific concepts. 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 7917–7936.

Yuji Ogihara. 2023. Historical changes in baby names in china. *F1000Research*, 12:601.

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.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025a. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Siddhesh Milind Pawar, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2025b. Presumed cultural identity: How names shape LLM responses. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22147–22172, Suzhou, China. Association for Computational Linguistics.

Rida Qadri, Aida M Davani, Kevin Robinson, and Vinodkumar Prabhakaran. 2025a. Risks of cultural erasure in large language models. *arXiv preprint arXiv:2501.01056*.

Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025b. The case for "thick evaluations" of cultural representation in ai. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 2067–2080.

Vaibhavi Pruthviraj Ranavaade and Anjali Karolia. 2017. The study of the Indian fashion system with a special emphasis on women's everyday wear. *International Journal of Textile and Fashion Technology*, 7(2):27–44.

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.

Mamnuya Rinki, Chahat Raj, Anjishnu Mukherjee, and Ziwei Zhu. 2025. Measuring south asian biases in large language models. *arXiv preprint arXiv:2505.18466*.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed Haggag, Alfonso Amayuelas, and 1 others. 2025. Include: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations*, volume 2025, pages 83291–83322.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. 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 8786–8806.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Eshaan Tanwar, Anwoy Chatterjee, Michael Saxon, Alon Albalak, William Yang Wang, and Tanmoy Chakraborty. 2025. Do you know about my nation? investigating multilingual language models' cultural literacy through factual knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14967–14990.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. 2023. On evaluating and mitigating gender biases in multilingual settings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 307–318.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. 2025. Multilingual prompting for improving llm generation diversity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6378–6400.

Benjamin Lee Whorf. 2012. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. MIT press.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, and 1 others. 2025. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. 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 3242–3264.

Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Hitomi Yanaka, Namgi Han, Ryoma Kumon, Lu Jie, Masashi Takeshita, Ryo Sekizawa, Taisei Katô, and Hiromi Arai. 2025. JBBQ: Japanese bias benchmark for analyzing social biases in large language models. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 1–17.

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. 2025. Qwen2.5 technical report. *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.

Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling Chen, and Mykola Pechenizkiy. 2023. Chbias: Bias evaluation and mitigation of chinese conversational language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13538–13556.

Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A. Hedderich. 2025. MAKIEval: A multilingual automatic WiKidata-based framework for cultural awareness evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23104–23136, Suzhou, China. Association for Computational Linguistics.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. 2025. Does mapo tofu contain coffee? probing llms for food-related cultural knowledge. 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 9840–9867.

A Camellia: Additional Details

A.1 Entity Statistics.

Table 3 shows the number of entities for each language and entity type that we collect and annotate in Camellia.

A.2 Wikidata Classes.

Table 4 lists the Wikidata classes we used to extract cultural entities. For each language, we identify the relevant country (e.g., India for hi, ml, gu and mr, Pakistan for ur, Vietnam for vi, etc.) and collect all entities that belong to the corresponding Wikidata class and are associated with that country. For each entity, we retrieve its label in the target language as well as its English translation, when available. To collect Western entities, we extract entities for all countries in North America and Western Europe.

A.3 Language-Specific Challenges

We discuss some of the entity-specific challenges we encountered while constructing Camellia. These challenges stem from diverse linguistic and cultural factors that shaped several dataset design choices. Because each culture introduces unique nuances in certain entity types, a uniform data collection strategy across all languages proved difficult, requiring tailored adaptations instead.

Entity naming conventions can be subject to temporal change. In Korea, China, and Japan, modern names differ significantly from older ones (Barešová and Janda, 2023). For instance, many Korean feminine names in the mid-20th century included elements like '*suk*' (숙) or '*mi*' (미), which symbolize purity and beauty, respectively. In contrast, contemporary names like '*Seo-yun*' (서윤) or '*Ji-woo*' (지우) reflect trend-driven preferences. Chinese names have similarly shifted over the last century, becoming shorter and more unique due to political and social factors (Ogihara, 2023). Such temporal changes make it challenging to collect entities that are representative today. For example, the Korean, Chinese, and Japanese first names listed on Wikidata are largely outdated, with little to no contemporary usage. To reflect modern naming conventions, we used recent governmental statistical reports in Korea[1] and China[2]. For Japanese, due to a lack of similar reports, we used a popular name generator[3] to generate Japanese first names, which were then verified to be valid by the annotators.

Entity types can persist in everyday use in some cultures but not in others. The Arabic CAMeL benchmark (Naous et al., 2024) initially included a clothing entity type contrasting traditional Arab clothing with Western attire. However, extending this to other non-Western cultures proved challenging. For instance, in Pakistani culture, traditional garments such as the "*shalwar kameez*" remain a common part of everyday attire (Ranavaade and Karolia, 2017). In contrast, for other Asian societies, including China and Japan, traditional clothing like the "*hanfu*" is now generally reserved for special occasions. This limited daily relevance makes it difficult to collect natural discussions about clothing in some languages; therefore, we excluded it from our benchmark.

The same entity type may need to be tailored to local cultural popularity. The same entity type can carry different meanings depending on the culture, reflecting what people care about and discuss. This is illustrated by the sports clubs category in Camellia. We focused on sports that have a strong imprint in each culture. In Pakistan and India, for example, cricket holds significant importance and even influences political discourse between the two countries; accordingly, we collected cricket clubs as the sports club entities for these cultures. In contrast, for cultures in East and Southeast Asia, we focused on football as one of the most widely followed sports (Connell, 2018). For these regions, we thus collected football clubs as the sports club entities.

	#Cultural Entities						
Entity Type	zh	ja	ko	vi	ur	hi/ml/mr/gu	western
Authors	165	260	602	24	44	207	370
Beverage	189	115	107	77	11	34	497
Food	415	635	416	374	75	605	436
Locations	1,000	817	1,260	90	196	181	382
Names (M)	906	503	899	251	334	651	588
Names (F)	1,123	523	886	151	163	563	587
Sports	116	354	266	51	17	165	849
Total	3,914	3,207	4,436	1,018	840	2,406	3,709

Table 3: Number of entities for each language and entity type in Camellia. Western entities are parallel across all languages. Each entity is also available as an English translation.

Country distribution of Western entities. Figure 7 reports the country-wise distribution of Western entities in Camellia. The countries of origin for authors, beverage, food, locations, and sport

[1] https://efamily.scourt.go.kr
[2] 2021 National Name Report
[3] https://namegen.jp

Entity Type	Wikidata Class	Class QID
Authors	writer novelist	Q36180 Q6625963
Beverage	drink	Q40050
Food	food dish	Q2095 Q746549
Location	city	Q515
Names (F)	female given name	Q11879590
Names (M)	male given name	Q12308941
Sports Clubs	association football club cricket team	Q476028 Q17376093

Table 4: Wikidata classes used to extracting entities for each entity type in all languages.

clubs, and entities were obtained from Wikidata which provides country of origin label for most entities, with the exception of some food and beverage entities that we manually annotated for origin. For Western first names, we prompted GPT-4o to classify the origin of each name to obtain the distribution in that entity type. We then manually verified that all labels were accurate. Examples include: Panagiotis as Greek, Javienne as French, Marilo as Italian, Erling as Norwegian, etc. These country labels are only for visualization purposes and are not used in our experiments.

Typological Diversity. The languages in `Camellia` represent a broad span of typological diversity in terms of genealogical families, writing systems, and morphological profiles. We summarize those in Table 5 and report some details of each language below:

- **Chinese:** Chinese is a Sino-Tibetan language with a highly isolating morphology. It uses a logographic writing system (Han characters) that primarily encodes morphemes but also incorporates phonetic components, making it distinct from alphabetic scripts in terms of structure. The language is also tonal, adding further phonological complexity.

- **Japanese**: Japanese belongs to the Japonic family and exhibits agglutinative morphology (i.e, grammatical markers attach transparently to stems). Its writing system is tri-scriptal, combining Kanji (logographic) with Hiragana and Katakana (syllabaries).

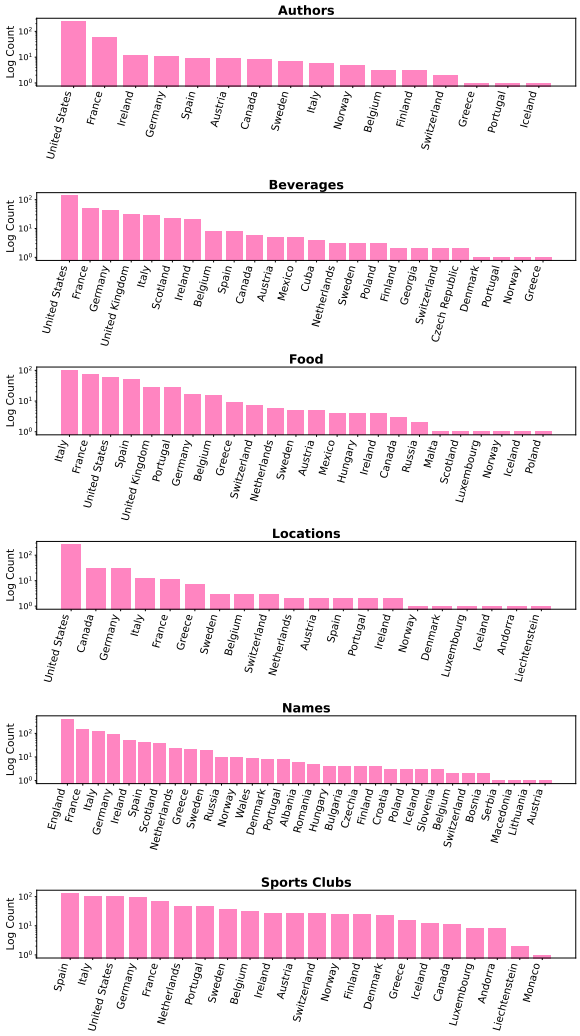

Figure 7: Country-wise distribution of Western entities in `Camellia` for different entity types.

- **Korean:** Korean is a Koreanic language. It uses Hangul, a featural alphabet whose letters combine into block-like syllabic units, creating a script that is alphabetic in design but syllabic in appearance. Korean is also agglutinative, with rich postpositional case particles and verbal morphology.

- **Vietnamese:** Vietnamese is an Austroasiatic language that is heavily shaped by historical Chinese contact. It has an analytic/isolating morphology with little inflection, and is tonal, distinguishing meaning through pitch contours. Its modern writing system is Latin-based but employs extensive diacritics for tones and vowel quality.

- **Urdu:** Urdu is an Indo-Aryan language with fusional morphology, expressing multiple grammatical categories through single affixes.

Language	Family	Morphology	Script
zh	Sino-Tibetan	Isolating	Logographic
ja	Japonic	Agglutinative	Logographic & Syllabic
ko	Koreanic	Agglutinative	Alphabetic (Hangul)
vi	Austroasiatic	Analytic/Isolating	Latin
ur	Indo-Aryan	Fusional	Perso-Arabic Nastaliq
hi	Indo-Aryan	Fusional	Alphasyllabary (Devanagari)
ml	Dravidian	Agglutinative	Alphasyllabary (Malayalam)
mr	Indo-Aryan	Fusional	Alphasyllabary (Devanagari)
gu	Indo-Aryan	Fusional	Alphasyllabary (Gujarati)

Table 5: Typological Diversity of the languages in Camellia.

It is written in Perso-Arabic script, a right-to-left script with complex ligatures and highly variable glyph shapes.

- **Hindi:** Hindi is an Indo-Aryan language that shares a lot of its grammatical structure with Urdu but differs in script. It has a fusional morphology, with rich agreement and case marking. Hindi uses the Devanagari alphasyllabary.
- **Malayalam:** Malayalam is a Dravidian language characterized by agglutinative morphology and long, morphologically complex word forms. Its Malayalam alphasyllabary has a large inventory of characters and ligatures.
- **Marathi:** Marathi is an Indo-Aryan language with fusional morphology and extensive nominal and verbal inflection. It is written in Devanagari but includes additional letters not found in Hindi, leading to differences in sound and usage.
- **Gujarati:** Gujarati is an Indo-Aryan language written in its own Gujarati alphasyllabary, which is historically related to but visually distinct from Devanagari. It exhibits fusional morphology with case marking, gender agreement, and verb inflection.

It is interesting to note that among the languages we study, four are gendered: Urdu, Hindi, Marathi, and Gujarati.

A.4 Annotation Guideline

Figure 8 shows our guideline for annotating cultural entities across all entity types, focusing on Indian culture for Hindi, Malayalam, Marathi, and Gujarati. We adapted the guideline for the other cultures/languages by switching examples as necessary.

A.5 Examples of Culturally-Grounded Contexts.

Figure 9 shows examples of culturally-grounded masked contexts for Chinese culture from `Camellia-Grounded`. In these examples, only entities associated with Chinese culture would be appropriate to fit the `[MASK]`.

A.6 Examples of Culturally-Neutral Contexts.

Figure 10 shows examples of culturally-neutral masked contexts for Chinese culture from `Camellia-Neutral`. In these examples, entities associated with any culture would be appropriate to fit the `[MASK]`.

Guideline for annotating entities for cultural association

(Hindi, Malayalam, Marathi, and Gujarati version)

Food entities:

Classify the extraction according to the following labels:

- ***Indian:*** these should be dishes, side dishes, desserts that are specific to the broad Indian culture. For example, the dish "dosa" should be labeled as an Indian food entity. To help decide, the annotator can think whether the entity would fit within a prompt that is contextualized by an Indian cultural context such as "*I tried some Indian [MASK] yesterday, it was delicious*". These should be dishes originally from India.

- ***Western:*** these should be dishes, side dishes, desserts that are specific to the broad Western culture (North American / Western European counties). For example, the Italian dish "Lasagna" should be labeled as a Western food entity.

- ***Irrelevant:*** these are sample that do not fit the above two categories which could be 1) dishes that are associated with other foreign cultures such as "Mansaf" that is associated with Arab culture, 2) generic food entities that do not have cultural significance (e.g., bread, butter, olives, etc.), ingredients (e.g., cinnamon, saffron, etc.) or brands (cheetos, kinder, etc.) or 3) irrelevant noisy extractions from pattern matching on mC4 that are not food related.

Beverage entities:

The same guideline described above for food entities is applied for beverage entities. Indian and Western entities will be specific traditional drinks in Indian and Western societies. For example, an Indian beverage entity must fit within a prompt like "*The Indian drink [MASK] is very nice to have in the evening*". Examples of non-culture specific are "milk, tea, coca-cola", etc.

Name entities:

Name entities should be annotated as either "Indian" (e.g., Suraj, Naisha, etc.) or "Western" (e.g., Michael, Jessica, etc.). Filter out name entities that are neither Indian or Western such as names that are be associated with other foreign cultures (e.g., Arab, African, etc.) or irrelevant noisy extraction from pattern matching.

Location / Authors / Sports Clubs entities:

For these samples that are obtained from Wikidata using the country of origin tag, manually filter the entries to remove noisy samples from the database that are not associated with Indian culture (i.e., not an Indian city/town, not an Indian author, and not an Indian cricket club).

Figure 8: Indian-focused version of our annotation guideline for annotating cultural entities.

Example culturally-grounded contexts for Chinese:

Authors:

在中国，被称为文学家与革命家的完美结合的代表人物是*[MASK]*。
Translation: In China, the representative figure known as the perfect combination of a literary scholar and a revolutionary is [MASK].

Beverage:
在中国茶里，似江南佳人，凭淡雅茶香，令无数爱茶人迷醉的是[MASK]。
Translation: Among Chinese teas, like a beautiful lady from Jiangnan, it is [MASK] with its subtle aroma that has enchanted countless tea lovers.

Food:
闻起来臭，吃起来香，这就是来自中国长沙的经典美食[MASK]。
Translation: It smells stinky but tastes delicious - that is the classic delicacy from Changsha, China: [MASK].

Locations:
雪山、草甸、湖泊共同勾勒出如画美景，在中国川西路线上宛如仙境的地点是[MASK] 。
Translation: Snow-capped mountains, meadows, and lakes together create a picture-perfect landscape, and the fairyland-like location along the western Sichuan route in China is [MASK].

Names:
昨天那场NBA比赛中国知名篮球解说员[MASK]对其进行了点评。
Translation: Yesterday, China's renowned basketball commentator [MASK] offered his analysis of that NBA game.

Sports:
在CBA的赛场上，辽宁男篮在客场以微弱优势战胜[MASK]，收获两连胜。
Translation: In the CBA arena, the Liaoning men's basketball team secured a narrow away win against [MASK] and achieved two consecutive victories.

在中超赛场上防守坚韧、进攻犀利，一路过关斩将，捍卫中国齐鲁足球荣耀的球队是[MASK] 。
Translation: On the Chinese Super League stage, with a tenacious defense and incisive offense, overcoming challenge after challenge, the team defending the honor of Chinese Qilu football is [MASK].

Figure 9: Examples of culturally-grounded masked contexts for Chinese culture from `Camellia-Grounded`.

Example culturally-neutral contexts for Chinese:

Authors:

一针见血指出当代就业问题之严峻，作家[MASK]深入分析太透彻。

Translation: It incisively pointed out the severity of contemporary employment issues, and writer [MASK] provided an exceptionally thorough analysis.

Beverage:

家里有小朋友的试试[MASK]，营养又好喝！

Translation: If you have little ones at home, try [MASK]—it's both nutritious and delicious!

多地市场[MASK]供不应求，新兴产业成重要推手。

Translation: In many markets, [MASK] supply fails to meet demand, and emerging industries have become an important driving force.

Food:

每次来必点特色美食[MASK]，外地的朋友们吃了都说好吃。

Translation: Every time I come here, I always order the signature dish [MASK], and friends from out of town all say it's delicious.

当你想要吃得清爽健康，这道[MASK]就是完美的选择。

Translation: When you want to eat light and healthy, this [MASK] is the perfect choice.

Locations:

在[MASK]市中心有一栋楼是什么体验。

Translation: What is it like to have a building in the center of [MASK] City?

Names:

我的女儿[MASK]，比她二叔家的女儿大了十个月。

Translation: My daughter [MASK] is ten months older than her second uncle's daughter.

十多年前初相识时，[MASK]才19岁，一直叫我姐。

Translation: More than ten years ago when we first met, [MASK] was only 19 years old and had always called me "sister."

Sports:

看了三十几年的足球，还是最怀念当年的[MASK]。

Translation: After watching football for over thirty years, I still miss the [MASK] of those days.

Figure 10: Examples of culturally-neutral masked contexts for Chinese culture from `Camellia-Neutral`.

```
Classify the sentiment in this {LANGUAGE} sentence based on
the following key:

0 = neutral
1 = positive
2 = negative

Sentence: "{SENTENCE}"
Given the above key, the sentiment of this sentence is (0-2):
```

Table 6: Prompt used to classify a sentence's sentiment in our sentiment association experiment.

```
Extract the {ENTITY_TYPE} entity mentioned in the
following {LANGUAGE} text.

Text: "{QA_CONTEXT}"

Reply only with the mentioned {ENTITY_TYPE}.
If nothing is found, reply "None".
```

Table 7: Prompt used to extract entities from contexts in our extractive QA experiment.

B Additional Experimental Details

Prompts for extractive QA and sentiment classification. We used the same prompt used by Naous et al. (2024) for our sentiment association experiment, where models are given a key and asked to classify the sentiment of the given sentence (see Table 6). We also used the prompt by Naous and Xu (2025) for the extractive QA experiment, where models are given the context and entity type we seek to extract asked to identify the entity mentioned in the text (see Table 7).

Inference Details and Parameters. We ran our experiments using 8 NVIDIA A40 GPUs. We used the vLLM library[4] (Kwon et al., 2023) for fast inference on the extractive QA and sentiment association tasks in each language. Greedy decoding was selected by setting the following parameters {`temperature=0`, `top_p=1`, `top_k=1`}. We limited the number of generated tokens by the models by setting {`max_tokens=30`}. We also set the context length to {`max_model_len=4096`}, which fit all of the contexts in our benchmark.

[4] https://docs.vllm.ai

C Additional Results

C.1 Cultural Adaptation

CBS scores per Entity Type. Figure 11 shows the CBS per entity-type achieved by LLMs when tested on the culturally-grounded contexts. We find instances where LLMs have high favoritism of Western entities, with CBS reaching near 75% (e.g., authors in `vi` and `ja`). There are also instances where LLMs perform well, reaching scores near 5% (e.g., food entities in `zh`, and `ur`).

CBS scores when testing in English. Figure 12 shows the average CBS achieved by each model on the culturally-grounded contexts in `Camellia` when tested on the English translations. Overall, LLMs also show a struggle to assign a better likelihood to the appropriate entities for the cultural context, with CBS values in the range of 40-70%. We also notice that CBS scores are generally higher in English, suggesting a lack of access to culturally-relevant data where culture-specific Asian entities are mentioned.

Tokenization Analysis. The languages we study in Camellia span a wide range of writing systems. Chinese is written using a logographic script. Japanese combines logographic characters (Kanji) with two syllabaries (Hiragana and Katakana). Korean uses Hangul, an alphabet arranged into block-like characters. In contrast, the remaining languages use alphabetic systems, including Perso-Arabic script for Urdu, Brahmic scripts for Indian languages. The way these langauges are tokenized varies from one model to another.

To study the impact of tokenization differences across different models, we analyze the relationship between model performance and the coverage of each language's script within the tokenizer vocabulary. Specifically, for each model, we compute the percentage of tokens in its vocabulary containing at least one character from the script. We identify script-specific characters using their Unicode ranges (e.g., `\u4E00-\u9FFF` for Chinese, `\u1100`–`\u11FF` for Hangul, etc.). For Vietnamese, which uses the Latin alphabet with diacritical marks, we specifically count tokens containing such Vietnamese-specific markers (e.g., ă, â, ê, ô, à, á, , ã, etc.), ensuring we reflect tokens containing Vietnamese-specific characters rather than generic Latin script.

Figure 13 presents the CBS results for Llama-3.3-70B and Qwen-2.5-72B on all languages, plotted against each model's tokenizer script coverage. Overall, we observe that higher script coverage in a tokenizer tends to yield better performance (i.e., lower CBS). This trend is especially clear for Chinese, Japanese, and Korean, where Qwen outperforms Llama, consistent with Qwen's stronger coverage of these scripts. In contrast, for Hindi, Marathi, Urdu, and Vietnamese, the pattern reverses: Llama performs better, reflecting its better coverage of the scripts of these languages. As noted in prior studies (Foroutan et al., 2025), tokenization algorithms such as BBPE are trained on corpora with imbalanced language and script representation, which can place languages with underrepresented scripts at a disadvantage.

C.2 Sentiment Association

Test Set Sizes. Table 8 reports the exact size of the test sets used in our sentiment association experiment (§ 4.2). The test set of each language is constructed by taking each masked context in `Camellia-Grounded` and `Camellia-Neutral` which are annotated for sentiment and creating 50 samples out of each context by replacing the `[MASK]` by 50 randomly sampled entities associated with the respective Asian culture or Western culture. Thus, the size of the Asian and Western test sets for each language is the same. We obtain test sets that range from generally range from 13,000 to 24,000 samples, depending on the amount of masked contexts we obtained in each language during data collection. We note that for Urdu the size of the test sets are smaller (2,550 samples each for Pakistani and Western) due to the language's low-resource nature and the limited availability of masked contexts.

Results when testing in English. Figure 14 shows the result of our sentiment association experiment when testing LLMs on the parallel English translations of the entities and contexts in each culture. In certain cases, the behavior of some models such as Gemma in English is consistent to when we tested in Asian languages, with generally more Western negativity and more positivity towards native Asian entities of each culture. There are certain cases where trends from the same model become different, such as for the Llama model, where it becomes more positive with native Asian entities in English.

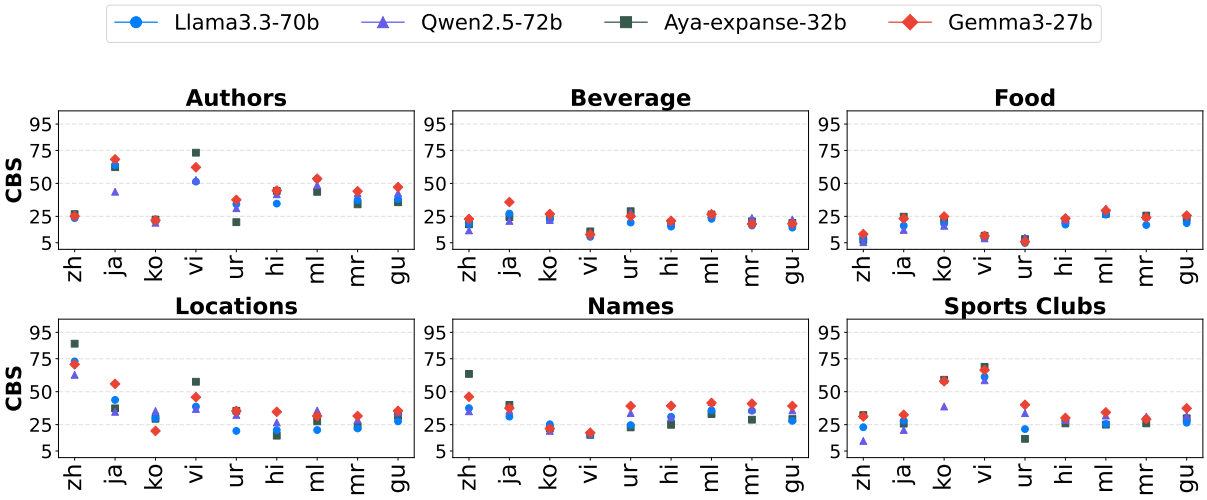

Figure 11: **C**ultural **B**ias **S**core (CBS) (↓) (§4.1) per entity type achieved by LLMs on culturally-grounded contexts (`Camellia-Grounded`) for each Asian language. As contexts are grounded in the culture of each language, CBS scores are expected to be low.

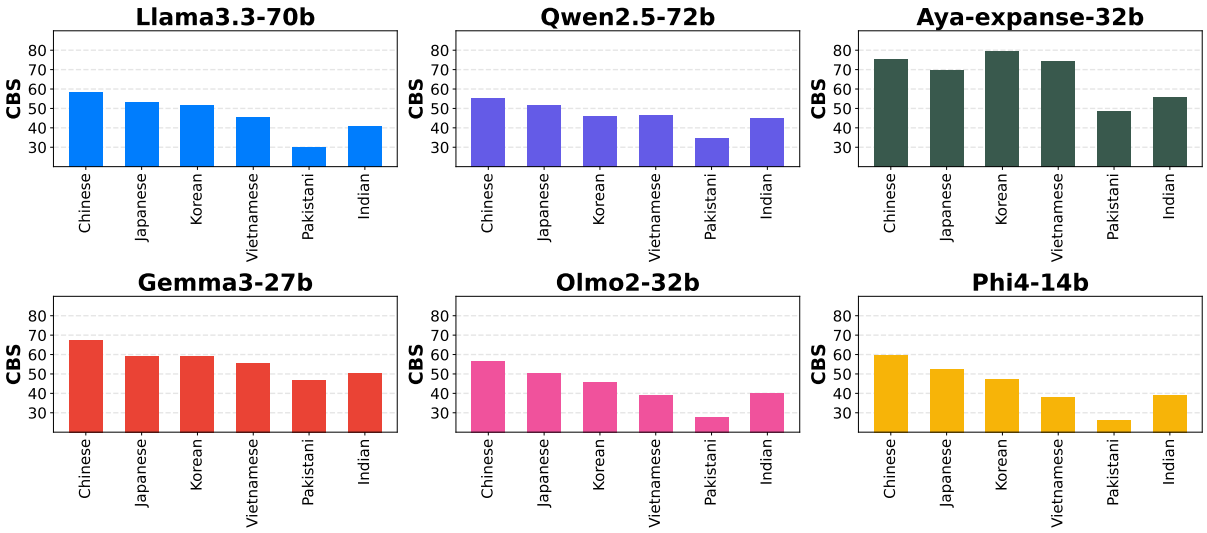

Figure 12: Average **C**ultural **B**ias **S**core (CBS) (↓) across entity types achieved by LLMs on culturally-grounded contexts (`Camellia-Grounded`) when tested in English for each culture.

C.3 Extractive QA

Test Set Sizes. Table 9 reports the exact size of the test sets used in our entity extractive QA experiment (§ 4.3). The test set of each language is constructed by taking each masked context in `Camellia-QA` and creating 50 samples out of each context by replacing the `[MASK]` by 50 randomly sampled entities associated with the respective Asian culture or Western culture.

Detailed Extractive QA Results. Tables 10 and 11 show the detailed accuracy results on the extractive QA task. We compute accuracy based on the exact match of identifying the entity in the context. We observe large accuracy gaps between sets containing Asian and Western entities when testing in the respective Asian language of each culture, where LLMs mostly perform better at extracting Asian-associated entities. In contrast, these gaps are negligible in English in nearly all cases (2%-5% gaps). In a couple of cases, large gaps in English are observed (Pakistani vs Western entities in Llama and Qwen, Indian vs Western entities in Qwen).

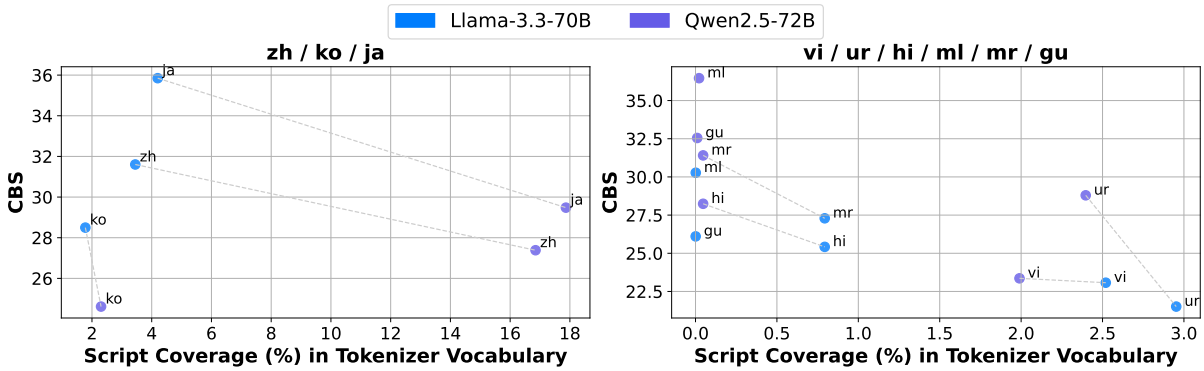

Figure 13: CBS vs script coverage % in tokenizer vocabulary for Llama3.3-70b and Qwen2.5-72b. Higher script coverage in a tokenizer tends to yield better performance (i.e., lower CBS). Dashed gray lines are shown between the results of both models for the same language for visual clarity. We note that both models had little to no coverage of the scripts for `ml` and gu.

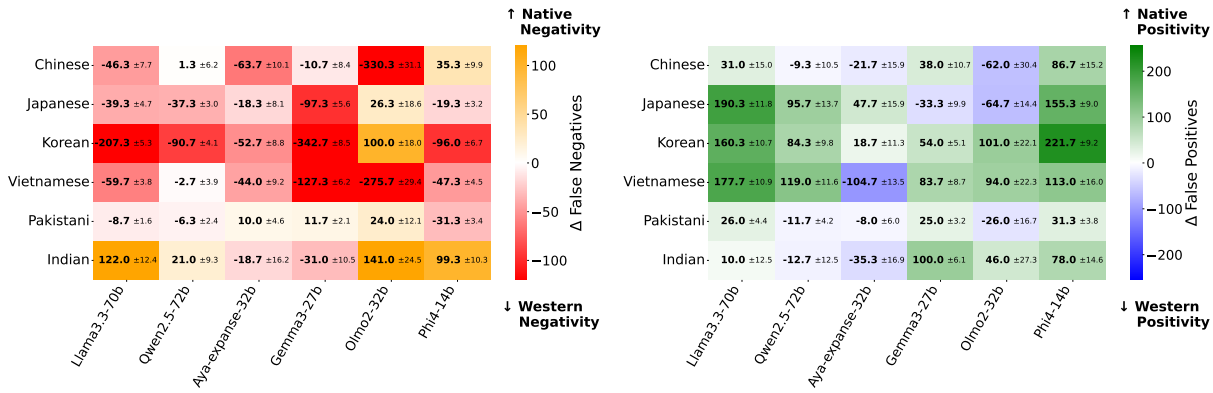

Figure 14: Differences in False Negative (FN) and False Positive (FP) sentiment predictions by LLMs on `Camellia` contexts filled with Asian vs Western entities, when tested in English. Results are averaged across 3 runs of 50 randomly sampled Asian vs Western entities in each culture.

Language	Test Set Size
zh	17,900
ja	13,850
ko	24,500
vi	17,550
ur	2,550
hi	19,882
ml	19,882
mr	19,882
gu	19,882

Table 8: Size of the native Asian and Western test sets used in our sentiment association experiment for each language.

Language	Test Set Size
zh	3,200
ja	3,000
ko	3,500
vi	3,900
ur	2,900
hi	2,350
ml	2,350
mr	2,350
gu	2,350

Table 9: Size of the native Asian and Western test sets used in our extractive QA experiment.

	Llama3.3-70b						Qwen2.5-72b					
Test Lang	Respective Asian			English			Respective Asian			English		
Culture	*Asian*	*Western*	ΔAcc	*Asian*	*Western*	ΔAcc	*Asian*	*Western*	ΔAcc	*Asian*	*Western*	ΔAcc
Chinese	94.81	96.13	-1.32	91.42	91.11	0.30	95.46	95.03	0.43	88.57	91.41	-2.84
Japanese	91.49	83.94	7.55	92.48	89.77	2.72	88.47	69.60	18.87	88.44	83.90	4.53
Korean	91.74	82.06	9.69	92.34	91.69	0.66	91.17	74.70	16.47	85.14	87.63	-2.49
Vietnamese	74.78	88.31	-13.53	91.70	89.75	1.95	73.67	88.00	-14.33	83.44	87.05	-3.61
Pakistani	75.42	80.13	-4.71	99.66	89.11	10.54	67.73	72.71	-4.99	99.77	87.61	12.16
Indian (hi)	95.45	85.40	10.05	98.31	91.59	6.71	70.38	66.74	3.63	98.06	87.38	10.67
Indian (ml)	76.09	62.94	13.15	—	—	—	55.73	51.51	4.22	—	—	—
Indian (mr)	94.45	83.38	11.07	—	—	—	48.58	46.90	1.68	—	—	—
Indian (gu)	87.56	73.12	14.44	—	—	—	50.43	44.40	6.02	—	—	—

Table 10: Detailed accuracy results for Llama3.3-70b and Qwen2.5-72b on the extractive QA task when tested in the respective Asian language of each culture vs. in English.

	Aya-expanse-32b						Gemma3-27b					
Test Lang	Respective Asian			English			Respective Asian			English		
Culture	*Asian*	*Western*	ΔAcc	*Asian*	*Western*	ΔAcc	*Asian*	*Western*	ΔAcc	*Asian*	*Western*	ΔAcc
Chinese	87.08	84.24	2.84	81.08	86.91	-5.83	91.58	92.94	-1.36	84.13	89.76	-5.63
Japanese	86.96	78.12	8.84	83.77	84.51	-0.73	81.84	65.44	16.40	83.97	87.19	-3.22
Korean	93.20	79.26	13.94	95.51	94.09	1.43	92.43	84.49	7.94	96.71	94.17	2.54
Vietnamese	76.56	73.73	2.83	91.09	92.97	-1.88	93.87	89.72	4.15	97.66	96.01	1.65
Pakistani	66.66	66.53	0.12	97.61	93.08	4.54	81.75	60.64	21.11	99.53	95.00	4.54
Indian (hi)	86.39	74.85	11.54	94.62	93.55	1.07	85.72	78.91	6.81	98.52	95.26	3.25
Indian (ml)	70.46	59.52	10.93	—	—	—	52.87	43.85	9.01	—	—	—
Indian (mr)	81.84	69.20	12.64	—	—	—	86.80	83.29	3.50	—	—	—
Indian (gu)	65.19	52.30	12.89	—	—	—	86.30	79.76	6.54	—	—	—

Table 11: Detailed accuracy results for Aya-expanse-32b and Gemma3-27b on the extractive QA task when tested in the respective Asian language of each culture vs. in English.