Why are all LLMs Obsessed with Japanese Culture? On the Hidden Cultural and Regional Biases of LLMs

Joseba Fernandez de Landa[1] **Carla Perez-Almendros**[2] **Jose Camacho-Collados**[2]
[1] HiTZ Center - Ixa, University of the Basque Country EHU
[2] Cardiff University
joseba.fernandezdelanda@ehu.eus, CamachoColladosJ@cardiff.ac.uk

Abstract

LLMs have limitations when it comes to cultural coverage and competence, and in some cases, show specific cultural biases. Although prior studies have examined the cultural capabilities of LLMs, none have specifically investigated their regional preferences in generic culture-related questions. In this work, we propose a new dataset based on a comprehensive taxonomy of Culture-Related Open Questions (CROQ), with questions available in 24 languages. We evaluate LLMs by prompting them to answer questions from CROQ and provide a sample location. The results show that, contrary to previous cultural bias work, LLMs show a clear tendency towards countries such as Japan in their answers. Moreover, our results show that when prompting in languages such as English or other high-resource ones, LLMs tend to provide more diverse outputs. Low-resource languages, on the other hand, show more inclinations towards answering questions highlighting countries for which the input language is an official language. Finally, we also investigate at which point of LLM training this cultural bias emerges, with our results suggesting that the first clear signs appear after supervised fine-tuning, and not during pre-training. Dataset available at https://huggingface.co/datasets/HiTZ/CROQ.

1 Introduction

Specific cultural knowledge, values and behaviours are often rooted in certain regions or countries. In this sense, we would expect that people from different geographical locations will have different answers to the same cultural questions. However, the behaviour of Large Language Models (LLMs) to these types of culture-related input remains an open question. Extensive work has been done towards unveiling cultural biases in LLMs (Shi et al., 2024; Li et al., 2024; Bulté and Rigouts Terryn, 2025; Myung et al., 2024; Chiu et al., 2025; Arora et al., 2025; Ousidhoum et al., 2026), mainly showing biases in favour of Western or Anglocentric viewpoints, especially United States and Europe (Lee et al., 2024; Li et al., 2024; Hasan et al., 2024; Chiu et al., 2025; Arora et al., 2025; Bulté and Rigouts Terryn, 2025).

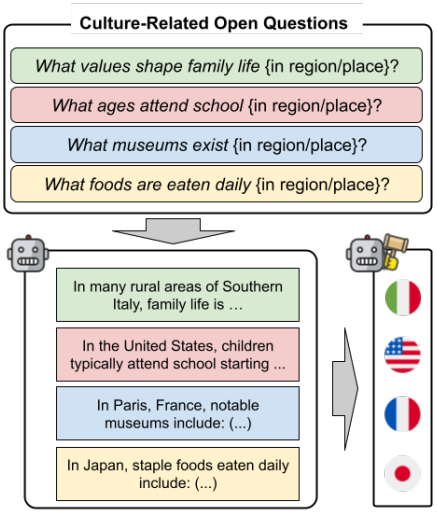


Figure 1: Framework for measuring cultural biases using CROQ dataset: (1) prompting models with cultural questions containing undefined locations, (2) collecting open responses from the specific model, and (3) identifying the referenced region(s) in the response.

Many of these works evaluate LLMs' responses against a set of predefined answers considered the gold standard (Romanou et al., 2024; Hasan et al., 2024; Arora et al., 2025; Chiu et al., 2025). Others evaluate the models' performance against human answers, covering a more or less wide spectrum of topics, cultures and countries/regions (Li et al., 2024; Shi et al., 2024; Myung et al., 2024; Bulté and Rigouts Terryn, 2025). In this work, we propose a novel framework to evaluate cultural regional biases in LLMs (Figure 1), by prompting different models with open ambiguous cultural questions while masking the geographic information. This strategy requires the model not only to answer the cultural question, but also to geographi-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 27628–27650
October 24-29, 2026 ©2026 Association for Computational Linguistics

cally anchor that answer by selecting a location. By encouraging the models to link their responses to a specific region, our framework surfaces internal cultural priors and exposes inherent biases.
We address the following Research Questions (RQ):

1 What regional cultural biases do LLMs have?
2 What is the impact of language in the bias?
3 At what training stage does the bias emerge?

To support this analysis, we construct CROQ, a dataset of 31,680 open cultural questions spanning 24 languages, 11 major topics, and 66 subtopics. We then prompt LLMs to answer questions from CROQ and requiring them to choose a specific location, therefore unveiling hidden cultural and regional biases. Finally, we use an LLM as-a-judge on the generated answers of each model to extract the geographic reference chosen by each model, which is later manually validated.

Our results indicate that regional preferences and model creativity are primarily determined by the input language. Notably, experiments across all models, languages and topics reveal a strong inclination for countries like Japan. Finally, further experiments suggest that these cultural and regional biases are induced predominantly during the post-training or instruction phase.

2 Related Work

Research on cultural knowledge and bias in LLMs is expanding rapidly (Adilazuarda et al., 2024; Liu et al., 2025a; Pawar et al., 2025), producing a variety of benchmarks, knowledge bases, and evaluation frameworks. While these resources provide important insights into multilingual and cultural capabilities, most are designed around closed-form answers and do not directly capture the leanings of models when responding to open-ended, ambiguous cultural questions.

Cultural knowledge is being benchmarked through question answering (QA) tasks. CulturalBench (Chiu et al., 2025) introduces human-authored multiple-choice questions covering 45 world regions and 17 topics, finding that LLMs fall far short of human performance, particularly in under-represented regions. Include (Romanou et al., 2024) compiles nearly 200k exam-style questions in 44 languages, targeting multilingual and regional knowledge. BLEnD (Myung et al., 2024) emphasizes everyday cultural knowledge (e.g., food, family, leisure) across 16 countries and 13 languages, revealing strong disparities between high- and low-resource languages. Following this lead, SemEval-2026 Task 7 (Ousidhoum et al., 2026) expands this framework to include 17 additional language–culture pairs. Additionally, NativQA (Hasan et al., 2024) and CaLMQA (Arora et al., 2025) use native-speaker based and culturally specific questions, respectively, across a wide range of topics, demonstrating heightened model failure rates for low-resource languages. Finally, Global MMLU (Singh et al., 2025) adapts MMLU benchmark by annotating questions for cultural sensitivity and expanding coverage to 42 languages. While these benchmarks reveal important weaknesses in factual cultural competence, their closed-answer format limits their ability to assess free-form generation.

Beyond QA, several works construct resources of cultural commonsense as knowledge bases. CANDLE (Nguyen et al., 2023) automatically extracts over one million assertions about food, rituals, and behaviours across 386 cultural groups, while CultureBank (Shi et al., 2024) curates community-grounded descriptors from TikTok and Reddit to capture lived experiences. These knowledge bases are useful for training and grounding models, but their assertional structures are primarily designed for factual recall rather than for evaluating models' tendencies in situations where multiple perspectives are valid.

A smaller but closely related line of work evaluates cultural adaptability and alignment. NormAd (Rao et al., 2025) measures whether models can adapt to diverse cultural norms across 75 countries using situational vignettes, showing that even state-of-the-art systems perform far below human levels, especially in abstract or implicit norm settings. CulFiT (Feng et al., 2025) introduces a training paradigm and evaluation dataset (GlobalCultureQA) for open-ended cultural QA, using fine-grained reward modeling to enhance cultural alignment. Similarly, Makieval (Zhao et al., 2025) proposes a framework for evaluating cultural awareness across 13 languages, 19 countries and regions, and six culturally salient topics, assessing open-ended generations by linking cultural entities in model outputs to Wikidata. While these approaches move beyond factual accuracy toward normative and generative evaluation, they often reduce outcomes to categorical labels (e.g., acceptable vs. unacceptable, aligned vs. misaligned) or optimize

models toward a predefined target norm, rather than characterizing the full range of plausible cultural positions a model may express.

In contrast to prior work, our framework explicitly targets open cultural questions for which no single correct answer exists, prompting the model to choose among all regions in the world. By eliciting multiple model generations, we evaluate the leaning of LLMs along diverse cultural topics and languages. This complements existing factual and normative benchmarks by shifting the focus from accuracy to distributional analysis, enabling a more nuanced understanding of cultural bias and alignment in LLMs. Moreover, we attempt to cover a wide range of cultural topics, encompassing areas covered in previous work into a single framework, as we detail in the following section.

3 Dataset Generation: CROQ

We introduce Culture-Related Open Questions (CROQ), a multilingual dataset designed to uncover the cultural tendencies of LLMs through open-ended yet culturally grounded questions. CROQ consists of culturally relevant prompts spanning 11 broad topics and 66 finer-grained subtopics, constructed consistently across topics to preserve structural uniformity and regional ambiguity. The dataset has been translated into 24 languages, covering a wide range of language families, typologies, and resource levels. It contains 1,320 questions (20 per subtopic) which, after translation, results in a total of 31,680 questions (details about the translation and validation in Appendix A).

Cultural Taxonomy. To make the open questions diverse and representative, we construct a taxonomy of 66 cultural subtopics grouped into 11 higher-level domains: ***Beliefs**, Values, and Identity* (🌱); ***Social** Structure and Daily Life* (🏠); *Knowledge, Communication, and **Education*** (🧠); *Cultural Expression and the **Arts*** (🎭); ***Food**, Drink, and Leisure* (🍽️); ***Geographic** Aspects* (🌍); ***Political** Aspects* (🏛️); ***Health** and Wellness* (🩺); ***Media** and Entertainment* (📺); ***History*** (📜); and ***Economy** and Industry* (💰). This taxonomy is intended to capture a broad and representative spectrum of cultural topics, encompassing as many culturally relevant dimensions as possible. It was developed by aggregating the topics from the datasets discussed in the related work and expanding them into a more comprehensive collection (more detail in Appendix B).

Open Question Generation. A central design choice in CROQ is the deliberate open-ended nature of each question. For each of the 66 cultural subtopics, we generate 20 questions, resulting in 1,320 queries in total. All questions follow the same basic pattern, avoiding direct references to specific countries or cultures, while remaining culturally meaningful. Questions are generated semi-automatically with *GPT-5.1* and manually reviewed to remove repetitions or accidental location references. Example questions per topic are shown below.

🌱 **Beliefs:**	What legends explain the land?
🏠 **Social:**	What is the role of neighbors?
🧠 **Education:**	What subjects are most valued?
🎭 **Arts:**	What traditional dances exist?
🍽️ **Food:**	What foods are for daily meal?
🌍 **Geography:**	What rivers influence settlements?
🏛️ **Politics:**	What political protests are common?
🩺 **Health:**	What exercise routines are common?
📺 **Media:**	What film industries exist?
📜 **History:**	What historical tours exist?
💰 **Economy:**	What commodity markets exist?

Multilingual Coverage. The dataset was initially constructed in English and then automatically translated and post-edited into 23 additional languages to ensure broad coverage across speaker populations, resource levels, and language families (full list in Appendix C). As a starting point, we included the 13 languages from the BLEND dataset (Myung et al., 2024), which range from low-resource (Amharic -am-, Assamese -as-, Azerbaijani -az-, Hausa -ha- and Sundanese -su-) to mid- and high-resource (Greek -el-, Indonesian -id-, Korean -ko-, Persian -fa-, Arabic -ar-, Chinese -zh-, Spanish -es-, English -en-) languages and provide wide geographical coverage.

To further enhance diversity, we added several of the world's most widely spoken languages with varying degrees of representation in CommonCrawl (Hindi -hi-, French -fr-, Bengali -bn-, Portuguese -pt-, Russian -ru-, German -de-, Japanese -ja-, Swahili -sw-). Additionally, we included all co-official but low-resource languages of Spain (Basque -eu-, Galician -gl-, Catalan -ca-)[1], which have relatively small speaker populations.

4 Evaluation Framework

We propose a novel method to uncover cultural biases in LLMs. We achieve this by prompting

[1]We use ISO 639-1:2002 codes for each language.

models with CROQ and requiring them to explicitly select a country or region. Subsequently, a secondary model is employed to systematically extract the geographic information inferred from the responses, allowing us to analyze potential cultural biases (see Figure 1).

4.1 Regional Question Grounding

To probe cultural priors in LLMs, we leverage the set of culture-related open questions (Section 3) and require the model to openly generate responses to these questions. Each question contains a locational placeholder (*in region/place*) that the model must implicitly resolve when producing an answer. We further prompt the model to be brief and to select a specific location (example below), thereby requiring it to tacitly choose a region or cultural background to ground its response:

> *What values shape family life* {in region/place}?
>
> Be brief. Choose yourself the place.

Each LLM evaluated receives the same set of 1,320 open questions in different languages (see Section 3 for more details). The prompt used to guide the model varies by language. To reduce prompt variability, all questions follow a standardized template, and no explicit regional cues are provided. This ensures that any regional or cultural assumptions arise from the model's internal priors rather than prompt design.

4.2 Processing Model Responses

To process the open-ended answers generated by the LLMs, we employ a secondary model as-a-judge to interpret the responses and extract the relevant information required to associate each question with a specific set of countries. This process is designed to uncover potential cultural biases by tasking the judge model with identifying up to five regions or places mentioned or implied in each response; if no such information can be reliably determined, the model is prompted to explicitly indicate that no inference is possible (non-answered responses). Following an evaluation of various prompting strategies, we selected the optimal approach, which achieved an accuracy of 98% over 264 evaluation items, the details of which, including the prompts and additional technical information, can be found in Appendix D.

4.3 Comparison of Frontier LLMs

For our evaluation, we compare a wide range of multilingual LLMs, including both frontier closed- and open-weight models. All models are accessed through OpenRouter (temp.=0.7 and top-p=0.9) and evaluated across the 24 languages in our dataset. The models include: gpt-4o-mini, gemini-2.5-flash, claude-3.5-haiku, llama-4-maverick, command-r-08-2024, magistral-small-2506, deepseek-v3.2-exp, qwen3-next-80b-a3b-instruct.

4.4 Analysis Metrics

For our subsequent analysis, we devise three metrics that can help analyse the output of each LLM.

Diversity. We define diversity as the number of distinct countries or regions referenced across all model outputs. Higher values indicate broader geographic coverage of cultural references.

Entropy. We compute normalized entropy over the distribution of referenced places to capture how evenly cultural references are distributed, with higher scores indicating greater balance.

Raw counts. We additionally report country or region frequency counts to provide complementary and interpretable evidence.

5 RQ1: What Regional Cultural Biases Do LLMs Have?

Models favor input language countries: Table 1 summarizes the regional distribution of country references across models for the 24 languages. Across all models, references overwhelmingly concentrate on *own* countries (ranging from 43% for Command-r to 78% for GPT), confirming a strong alignment between the output language and its associated regions. This pattern is consistent and robust, suggesting that language choice alone strongly constrains the cultural priors adopted by the models.

Exogenous references are dominated by a small set of regions: The data reveal that for non-own country references, all models display a similar pattern of global salience. On average, Japan (preferred country for six out of the eight evaluated models) and the United States are the most frequently referenced nations, followed by India, China, and France. References to other countries are considerably less frequent and vary by model.

Model	Own (%)	NA	🇯🇵	🇺🇸	🇮🇳	🇨🇳	🇫🇷	🇮🇹	🇬🇧	🇩🇪	🇲🇽	🇧🇷	🇷🇺	🇪🇬	Div	Ent
GPT	24,763 (.78)	1,136	811	**944**	451	237	289	265	128	150	134	234	53	87	82	0.40
Gemini	20,172 (.64)	1,853	**1,493**	**1,493**	673	473	439	344	310	271	164	125	75	240	113	0.50
Claude	20,585 (.65)	2,063	**1,601**	1,200	510	560	210	218	138	172	340	349	122	141	95	0.48
Llama	20,775 (.66)	887	**2,701**	1,074	453	524	587	265	190	173	75	222	61	132	102	0.49
Command-r	13,707 (.43)	4,815	936	**2,064**	567	512	504	389	515	309	209	259	113	142	**120**	**0.59**
Magistral	23,040 (.73)	2,485	**1,754**	1,254	496	560	632	392	264	390	315	339	163	211	88	0.47
Qwen	22,572 (.71)	1,502	**1,861**	483	249	434	191	170	60	158	222	67	58	60	79	0.40
DeepSeek	21,907 (.69)	877	**2,104**	1,006	541	608	351	415	390	499	308	205	110	267	116	0.52

Table 1: Frontier Model outputs for 24 languages. **Own** counts references to countries in which the language is an official language. **NA** denotes missing responses. Top referenced countries are Japan, USA, India, China, France, Italy, United Kingdom, Germany, Mexico, Brazil, Russia, Egypt. **Div** and **Ent** represent the average for diversity and entropy by model. Bold values indicate the highest count for each model, excluding self-references, as well as the highest diversity and entropy values (two right-most columns).

Differences among models: Although the distribution across mentioned exogenous countries (Japan, United States, India, and China) is similar, some differences emerge. Five models favor Japan, while two emphasize the US, and one is tied. Command-R stands out as the most US-biased model and the one most likely to refuse an answer, producing in general the most diverse outputs with the highest entropy. Command-R, DeepSeek, and Gemini have higher output diversity (over 113 different countries mentioned), whereas Qwen, GPT, and Magistral show lower diversity and entropy (fewer than 88 countries mentioned in their outputs – more details on individual model responses in Appendix E.2).

Topic Analysis: Table 2 presents the top countries mentioned by the models across the 11 general topics in our dataset. Japan and the US remain salient across all categories, followed by other frequently mentioned countries. However, some variation occurs depending on the topic which may reflect global trends. For example, in *Geography* and *Economy*, there is a noticeable shift in prominence between the US and Japan, where US takes the lead in number of references, and *Politics* and *History* are the only two topics where Japan is relegated to third and fourth positions. Countries such as Greece, China and France stand out in very specific topics, namely *Beliefs*, *Politics* and *History*, respectively. Or other countries such as South Korea, which is barely mentioned in most topics, emerge as the third most mentioned country in *Media and Entertainment*. Mentions of other countries also fluctuate across topics, but their overall counts remain substantially lower (data in Appendix E.3).

Topic	Top Countries
1. Beliefs	🇯🇵 🇬🇷 🇮🇳 🇺🇸 🇨🇳 🇸🇦 🇪🇬 🇮🇱 🇮🇹 🇫🇷
2. Social	🇯🇵 🇺🇸 🇮🇱 🇮🇹 🇵🇭 🇨🇳 🇩🇪 🇫🇷 🇮🇳 🇬🇧
3. Education	🇯🇵 🇺🇸 🇬🇷 🇨🇳 🇫🇷 🇮🇳 🇩🇪 🇪🇬 🇮🇹 🇬🇧
4. Arts	🇯🇵 🇺🇸 🇮🇳 🇫🇷 🇮🇹 🇧🇷 🇨🇳 🇲🇽 🇪🇸
5. Food	🇯🇵 🇺🇸 🇮🇳 🇲🇽 🇨🇳 🇫🇷 🇧🇷 🇪🇸 🇫🇷
6. Geography	🇺🇸 🇯🇵 🇮🇳 🇨🇳 🇫🇷 🇮🇹 🇲🇽 🇧🇷 🇩🇪 🇮🇱
7. Politics	🇺🇸 🇨🇳 🇯🇵 🇫🇷 🇬🇧 🇩🇪 🇮🇳 🇸🇦 🇷🇺
8. Health	🇯🇵 🇺🇸 🇨🇳 🇮🇳 🇸🇪 🇷🇺 🇧🇷 🇫🇷
9. Media	🇯🇵 🇺🇸 🇰🇷 🇬🇧 🇨🇳 🇫🇷 🇩🇪 🇧🇷 🇮🇳
10. History	🇺🇸 🇫🇷 🇮🇳 🇯🇵 🇮🇹 🇬🇧 🇨🇳 🇬🇷 🇩🇪
11. Economy	🇺🇸 🇯🇵 🇨🇳 🇪🇺 🇸🇬 🇩🇪 🇮🇳 🇰🇷 🇬🇧

Table 2: Top 10 countries most frequently referenced by all Frontier API models for each topic in the taxonomy. 24 languages, excluding own country mentions.

Summary of findings and discussion. All eight frontier models exhibit a clear bias toward languages referring to their own regions of origin. When associations to own language-region pairs are isolated, this bias becomes more pronounced. Consequently, all models consistently favour Japan or the United States, while regions such as India and China receive comparatively less emphasis, and references to other regions are negligible. This pattern suggests an uneven regional representation in frontier model outputs, with a strong concentration on a small set of dominant regions.

6 RQ2: What is the Impact of Language in Regional Cultural Bias?

Salient regions are overrepresented across all languages: As established in the previous section, models strongly associate languages with their

primary regions or countries of origin. To quantify this effect and explore its variation, we conducted a per-language analysis. Table 3 shows that beyond self-referential patterns, the models exhibit a stable hierarchy of non-own country references. Japan and the United States are the most frequently referenced countries across languages. Notably, these references appear even in languages with limited cultural or geographic ties to these countries. This finding aligns with prior observations that LLMs tend to overrepresent culturally salient entities.

	Own	🇯🇵	🇺🇸	🇮🇳	🇨🇳	Div	Ent
en	31%	**2,893**	*	*	367	157	0.69
zh	48%	**2,064**	1460	642	*	142	0.57
hi	70%	94	**104**	*	92	56	0.14
es	66%	**843**	588	130	173	136	0.65
ar	69%	**504**	396	182	85	108	0.71
fr	56%	**832**	557	215	236	149	0.54
bn	79%	73	**151**	*	59	68	0.20
pt	73%	**479**	442	167	161	114	0.42
ru	36%	**1,423**	657	265	431	146	0.60
id	76%	**411**	246	129	133	78	0.26
de	70%	**485**	237	169	203	122	0.38
ja	59%	*	**736**	434	219	114	0.36
ko	59%	1,064	**1,308**	574	663	112	0.47
fa	57%	473	**544**	263	194	117	0.48
sw	78%	230	**389**	85	79	89	0.50
ha	51%	30	**99**	30	44	49	0.10
az	75%	123	**212**	81	64	78	0.43
am	77%	43	**106**	17	19	38	0.13
su	79%	**167**	146	49	76	62	0.11
el	70%	**330**	307	161	135	106	0.36
as	83%	41	**53**	*	42	41	0.30
ca	72%	213	**339**	170	224	121	0.49
gl	75%	**314**	297	129	143	113	0.37
eu	78%	132	**144**	48	66	68	0.37

Table 3: Model outputs for the 24 languages. **Own**: percentage of references to countries in which the language is an official language; **Flags**: references to each country (top 4 displayed); * mark references to own countries; **Div** and **Ent** represent the average for diversity and entropy for all models. Colours indicate the top three and bottom three results.

Bias is consistent on exogenous references across languages: When isolating non-own-country references, a consistent global hierarchy emerges across all languages. Japan and the United States remain the most frequently cited countries, followed by India, China, and France. This suggests that cross-country mentions are driven more by a model's inherent priors of cultural salience than by contextual relevance to the specific language, underscoring a systematic overrepresentation of a narrow set of culturally dominant entities.

Low-resource languages produce more self-referential outputs: A key factor moderating this bias is the resource level of the language. Comparing languages by the volume of available training data reveals a clear pattern. Higher-resource languages (e.g., *en, zh, es, fr, ru*) show a lower proportion of own-country references and higher output diversity, with a significant skew toward globally dominant regions like Japan and the United States. Lower-resource languages (e.g., *su, as, am, ha, eu*), demonstrate a markedly higher proportion of own-country references (or an increase in non-responses) and lower diversity scores. This pattern suggests a link between training data volume and cultural framing: for languages with less comprehensive corpora, models default to safer, more self-referential, or less diverse outputs (more detail in Appendix E.1).

Pretraining data and diversity metrics correlate: To quantify the relationship between output diversity and language resource availability, we perform a statistical analysis correlating output diversity with the amount of CommonCrawl data available for each language, used as a proxy for pretraining resources (see Figure 2). Across all frontier models, output diversity correlates positively with CommonCrawl coverage. Spearman's correlation score, $r_s = 0.843$, reveals a strong rank-based association; being statistically significant (p-value < 0.001). Analyses at the individual model level further confirm this trend for the majority of systems (see Appendix E.4). Together, these results indicate that higher-resource languages benefit from more varied cultural representations, whereas lower-resource languages tend to elicit more self-referential or constrained outputs, reflecting reduced diversity in generated content.

Summary of findings and discussion. Overall, these findings demonstrate that language models exhibit strong language-aware referencing behaviour. However, once self-references are excluded, models consistently rely on the same narrow set of countries across languages. A particularly notable result is the disproportionate prominence of Japan, which emerges as one of the most frequently referenced exogenous countries across nearly all languages. United States is also highly salient, although this is less surprising given results from previous work and its global influence and the dominance of English as a high-resource language. Taken together, these patterns suggest

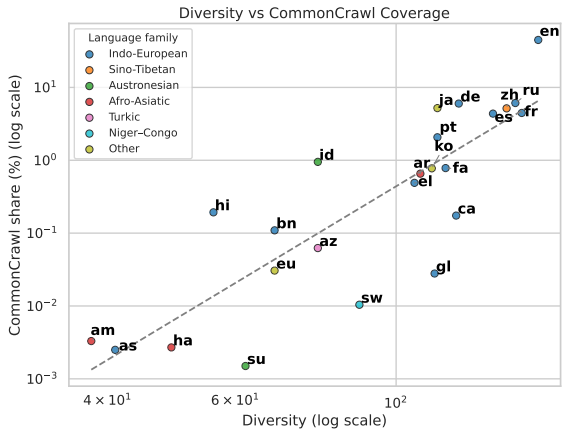


Figure 2: Correlation between CommonCrawl data share and output diversity (mean of models) across 24 languages (log-scale).

	Diversity			Entropy		
Model	**Base**	SFT	**Inst.**	**Base**	SFT	**Inst.**
Llama-3.1-8B	109	-	**174**	**0.78**	-	0.66
Llama-3.1-70B	117	-	**174**	**0.79**	-	0.65
gemma-2-9b	95	-	**136**	**0.70**	-	0.71
gemma-2-27b	80	-	**154**	**0.77**	-	0.73
Mistral-7B	91	-	**160**	**0.81**	-	0.63
Qwen2.5-7B	147	-	**150**	**0.71**	-	0.62
Qwen2.5-32B	**142**	-	118	**0.65**	-	0.61
Qwen2.5-72B	**144**	-	127	**0.66**	-	0.60
Olmo-3-7B	114	**145**	124	**0.74**	0.66	0.64
Olmo-3-7B-Think	114	121	**127**	**0.74**	0.66	0.68
OLMo-2-7B	102	151	**158**	**0.84**	0.63	0.64
OLMo-2-32B	88	151	**161**	**0.77**	0.65	0.71

Table 4: Diversity and normalized entropy scores for base, SFT (for OLMo) and instruct model variants. For each model, the highest diversity and entropy are bolded.

that the composition of training data, either during pretraining or post-training, plays a central role in shaping models' culturally biased referencing behaviour.

7 RQ3: At What Stage of LLM Training Do Regional Cultural Biases Emerge?

To identify the training stage at which regional cultural biases emerge, we extend our analysis to open-weight *base* and *instruction-tuned* LLMs. By comparing models before and after post-training, we aim to disentangle biases learned during large-scale pretraining from those introduced or amplified during supervised fine-tuning and alignment.

7.1 Experimental Setting

Comparison Models. To assess whether instruction tuning itself introduces or amplifies bias, we analyze open models providing both base and instruct variants across multiple enterprises and parameter scales, using default settings. This evaluation is performed only on English and includes Llama-3.1 (8B, 70B) (Dubey et al., 2024), Qwen2.5 (7B, 32B, 72B) (Yang et al., 2024), Gemma-2 (9B, 27B) (Team et al., 2024) and Mistral-7B-v0.3 (7B) (Jiang et al., 2023). In addition, we analyze OLMo-2 (7B, 32B) (Team, 2025a) and OLMo-3 (7B, 7B-Think) (Team, 2025b), which provide base, supervised fine-tuned (SFT), and fully aligned instruction models. This allows us to disentangle the effects of instruction tuning from those of subsequent supervised fine-tuning.

Prompting. Due to model availability, we restrict all experiments to English prompts. Given the nature of base models that are explicitly trained to predict the next tokens, we rephrase the question prompt utilised in our experiments so the sentence can be completed rather than answered. Everything else remains unchanged and both base and instruct models are given the same questions as input.[2]

7.2 Findings

Figure 3 contrasts the distributions of country references produced by base and instruction-tuned models, while Table 4 shows the diversity and entropy scores for each model variant.

Cultural distributions are more balanced on Base Models: Base models (Figure 3, left) exhibit a more balanced geographic coverage. While the United States remains prominent, substantial references are also made to Japan, India, China, and several European countries. This pattern is consistent across architectures and model sizes, suggesting that pretraining alone yields a comparatively diffuse set of cultural associations. This is further reflected on the entropy scores, which are consistently lower after instruction-tuning for all models.

Cultural leaning is stronger after Instruction-Tuning: Despite providing a higher number of countries overall, instruction-tuned models (Figure 3, right) display a pronounced concentration of references to a narrow set of regions. Across all examined model families, instruction tuning sharply increases alignment with the United States and Japan while reducing references to most other

[2]In Appendix H, we provide the prompt and an analysis of the impact of the prompt in base and instruct models.

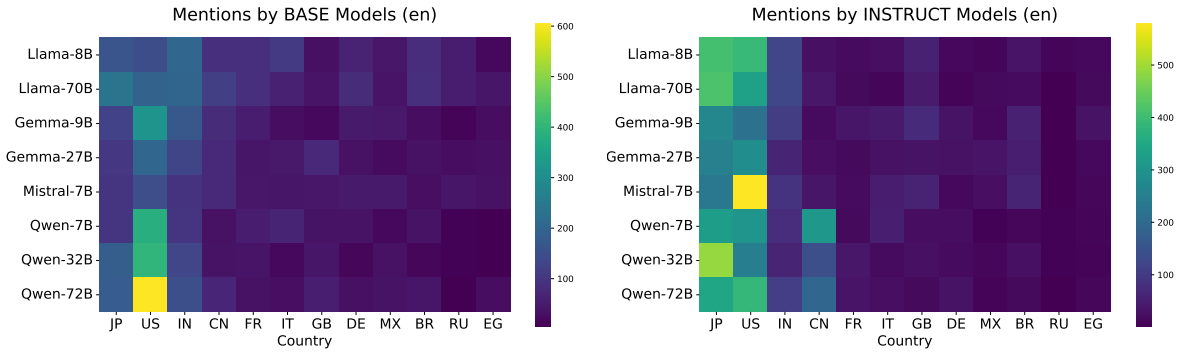


Figure 3: Number of outputs from Base and Instruct models of Llama-3.1, Gemma-2, Mistral and Qwen2.5 variants over the English questions. Countries included: Japan (JP), USA (US), India (IN), China (CN), France (FR), Italy (IT), UK (GB), Germany (DE), Mexico (MX), Brazil (BR), Russia (RU), Egypt (EG).

countries. This convergence toward culturally dominant regions occurs even in models developed outside Western contexts, indicating that post-training induces a homogenization of cultural perspectives rather than merely reflecting model origin.

Supervised fine-tuning as a primary driver of cultural bias: To further assess whether instruction tuning itself introduces or amplifies cultural bias, we analyze OLMo models, which provide base, supervised fine-tuned (SFT), and instruction-aligned variants. In Figure 4 we observe that the most substantial shift in country reference distributions occurs during supervised fine-tuning: SFT sharply increases concentration on a small number of dominant regions (most notably the United States and Japan) while reducing entropy across all model outputs (see Table 4). Subsequent instruction alignment only marginally mitigates these effects and does not recover the more balanced distributions observed in base models. These results indicate that alignment-induced cultural bias primarily originates from supervised fine-tuning data rather than pretraining, and that later instruction tuning largely preserves, rather than corrects, these learned cultural priors (see Appendix G for further details).

Summary of findings and discussion. Overall, base models distribute references more evenly across regions, exhibiting lower concentration of regions despite limited diversity. Instruction-tuned models, by contrast, show substantially more concentrated cultural biases. These findings suggest that post-training, rather than pretraining, plays a decisive role in shaping the dominant cultural perspectives expressed by LLMs (more details and results in Appendix F). The results from OLMO suggest that this behaviour starts appearing specifically after supervised fine-tuning. Finally, we checked the instructions provided to LLMs that could potentially amplify these biases without clear evidence (see Appendix I).

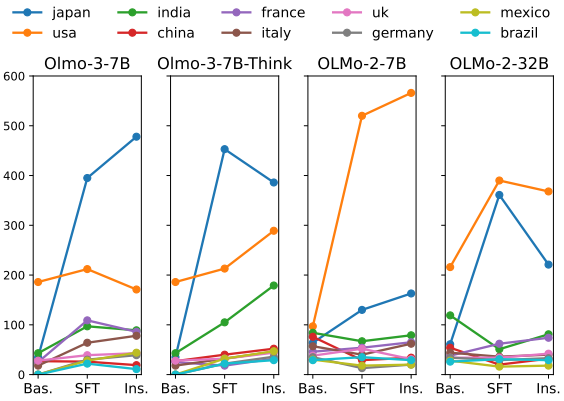


Figure 4: Number of outputs mentioning top-referenced countries from the Base, SFT, and Instruct variants of the named OLMo models.

8 Conclusions

Ensuring cultural awareness and diversity in LLMs is an open research question in NLP. However, measuring the impact of certain decisions is hard given all the cultural aspects that are involved in model responses. In this paper, we created a dataset of open questions based on a taxonomy of eleven cultural domains and sixty-six subtopics. We used this dataset to test eight frontier LLMs by prompting them in twenty-four languages. Instead of testing the models about their cultural knowledge, we let them answer the questions openly by only requiring them to provide regional examples of their choice.

The results show an inclination of models to provide cultural examples related to regions for

which the prompted language is an official language. Not only that, but there is a similarity on the type of example that models choose, preferring countries such as Japan, United States or India overwhelmingly with respect to other choices across the world. When analysing the potential causes, the results indicate that this concentration of answers emerges during post-training. While this is a stage traditionally thought to guide models toward providing more diverse and unbiased responses, it appears to also increase this cultural bias into a different direction. For future work, it would be interesting to think of new methods to take this behaviour into account, in order to better understand the post-training effects in model output diversity as a whole, in particular when it comes to different languages and regions.

Limitations

Our study has several limitations that should be considered when interpreting the results.

1) Due to the lack of open data in frontier multilingual LLMs, the comparison between base and instruction-tuned models is conducted only in English, which limits the extent to which these findings generalize to other languages. In addition, base and instruction-tuned models differ inherently in training objectives and interaction style, which may influence generation behaviour independently of the factors analyzed here. Nevertheless, this comparison provides a first step toward analyzing the intersection between base and instruction-tuned models, which we plan to explore further in future work.

2) All experiments are conducted using default generation settings, and we do not explore the robustness of our results to alternative decoding parameters. We deliberately chose this approach in order to focus on the settings employed by most users, enabling us to compare across different languages and models – an additional exhaustive exploration of decoding parameter configurations would incur high generation and analysis costs.

3) We rely primarily on an automatic judge rather than human evaluation for all languages. While we validate the judge on a subset of languages, automatic evaluation may fail to capture subtle linguistic or cultural distinctions. Nevertheless, given that our expertise is limited to a small number of languages, conducting large-scale human evaluation would not be feasible, making automatic evaluation a necessary choice which we believe does not affect the reliability of the results as a whole.

4) Our evaluation covers only 24 languages, representing a limited subset of global linguistic diversity. We further restrict some analyses to official languages of countries, which may not accurately reflect real-world language use, as many widely spoken languages extend beyond official or national boundaries. Moreover, referring to countries as proxies for cultural contexts may lead to oversimplifications.

5) Finally, we do not assess the impact of the observed behaviours on downstream tasks, leaving open the question of how these findings translate to applied or real-world scenarios.

Ethical considerations

In this work, we use LLMs to semi-automatically generate candidate questions for CROQ (GPT-5.1, see Sec. 3 for more details) and as-a-judge models (gpt-4o-mini, Sec. 4.2) to extract information from model responses. Because this process may introduce model biases, all generated questions underwent manual review to remove repetitions, accidental location references, and other artifacts. This human oversight was particularly important given that our work explicitly examines cultural biases in LLM outputs.

The cultural taxonomy underlying CROQ was developed by aggregating topics from existing benchmarks and expanding them into a comprehensive framework of 11 domains and 66 subtopics. Although we designed this taxonomy to capture a broad spectrum of culturally relevant dimensions, we acknowledge that cultural phenomena exist on a continuum and cannot be exhaustively categorized. This structural limitation means our analysis, while broad, remains grounded in particular cultural frameworks rather than offering a universal view of human culture.

In addition, we made deliberate efforts to ensure broad linguistic coverage by including 24 languages spanning different resource levels, language families, and geographical regions. This enables the study of cultural biases across varied linguistic contexts, including communities often underrepresented in NLP. However, 24 languages represent only a small fraction of the world's linguistic diversity, and our selection inevitably privileges certain communities and perspectives.

Consequently, this work remains affected by the

broader inequalities that shape language technology research. We encourage future work to extend cultural evaluation frameworks beyond country-level language boundaries and better represent underrepresented communities.

Our results, particularly the concentration of model outputs towards culturally dominant regions like Japan and the United States, have implications for global users of these systems. Models trained to favour certain cultural perspectives may provide inadequate or inappropriate responses for users from other backgrounds. We hope this work contributes to awareness among both model developers and users about the need for more balanced and inclusive cultural training.

Acknowledgements

Joseba Fernandez de Landa is supported by Grant DeepThought (PID2024-159202OB-C21), funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU, as well as the ALIA Models Development Project, Resolution of the SEDIA 19/08/2024, within the framework of the National Language Technologies Plan – ENIA 2024, funded by MTDFP, PRTR, and the EU – NextGenerationEU.

Jose Camacho-Collados is supported by a UKRI Future Leaders Fellowship.

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.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bram Bulté and Ayla Rigouts Terryn. 2025. Llms and cultural values: The impact of prompt language and explicit cultural framing. *Computational Linguistics*, pages 1–85.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. CulFiT: A fine-grained cultural-aware LLM training paradigm via multilingual critique data synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22413–22430, Vienna, Austria. Association for Computational Linguistics.

Md Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2024. Nativqa: Multilingual culturally-aligned natural query for llms. *arXiv preprint arXiv:2407.09823*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. 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 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing*

Systems, volume 37, pages 84799–84838. Curran Associates, Inc.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025a. 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.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, Yen-Sung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, and 1 others. 2025b. Olmotrace: Tracing language model outputs back to trillions of training tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 178–188.

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.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1907–1917, New York, NY, USA. Association for Computing Machinery.

Nedjma Ousidhoum, Junho Myung, Carla Perez-Almendros, Jiho Jin, Amr Keleg, Meriem Beloucif, Yi Zhou, Rodrigo Agerri, Vladimir Araujo, Naomi Baes, James Barry, Joanne Boisson, Nancy F. Chen, Christine de Kock, Aleksandra Edwards, Joseba Fernandez de Landa, Mohamed Fazli Imam, Huda Hakami, Shu-Kai Hsieh, and 11 others. 2026. SemEval-2026 task 7: Everyday knowledge across diverse languages and cultures. In *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, pages 3823–3837, San Diego, California, USA. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–98.

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, Albuquerque, New Mexico. Association for Computational Linguistics.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, and 1 others. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

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.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

OLMo Team. 2025a. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

OLMo Team. 2025b. Olmo 3. *arXiv preprint arXiv:2501.00656*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A Hedderich. 2025. Makieval: A multilingual automatic wikidata-based framework for cultural awareness evaluation for llms. *arXiv preprint arXiv:2505.21693*.

A Validation of Dataset Translations

Given the scale of the CROQ dataset (31,680 total questions across 24 languages), we took deliberate steps to ensure the high quality and semantic integrity of the translations. Because the core 1,320 questions are intentionally designed to be short, direct, and structurally simple (as shown in Section 3 and Figure 1), the translation process is highly

straightforward. Contemporary machine translation systems are highly effective at handling such short-form text without loss of context. Accordingly, we utilized *GPT-5.1* for the translation process.

To verify this, we analyzed a representative sample of 6 (zh, ko, es, ca, gl, eu) out of the 23 translated languages, explicitly including low-resource languages, and conducted a validation check with native speakers. This involved a manual evaluation of 400 questions in total. The main conclusion from this analysis is that the automated translations are accurate and highly consistent across the dataset. While native speakers noted that a small fraction of the prompts exhibited minor artificiality (occasional phrasing that sounds slightly unnatural or different from how a native speaker would typically talk) the translations remain entirely correct and semantically faithful to the original prompts. This validation confirms that the minor stylistic variations do not impact the clarity of the questions, making the translation pipeline fully sufficient and reliable for the scope of this benchmark.

B Topics and subtopics in the Taxonomy

We construct a comprehensive taxonomy comprising 66 fine-grained cultural subtopics, which are organized into 11 higher-level cultural domains. The taxonomy was designed to capture a broad and diverse range of cultural phenomena while maintaining sufficient granularity to support detailed analysis. Each higher-level domain aggregates thematically related subtopics, enabling both coarse-grained and fine-grained examination of cultural dimensions. The complete taxonomy, including definitions and hierarchical relationships between domains and subtopics, is presented in Table 5.

We conducted a semi-automatic evaluation of our proposed categories and subcategories, and we analyzed the datasets introduced in related work. Figure 5 illustrates the extent to which each general topic in our taxonomy is covered by existing datasets. The analysis shows that many datasets address only a limited subset of the general topics defined in our taxonomy. Although NativQA (Hasan et al., 2024) and CaLMQA (Arora et al., 2025) cover all general topics, their coverage across subtopics is uneven, with notable imbalances within individual topics. Our goal is not to propose a definitive or exhaustive taxonomy of cultural topics, but rather to cover as broad a range of cultural topics as possible, in order to prompt LLMs with maximal cultural diversity.

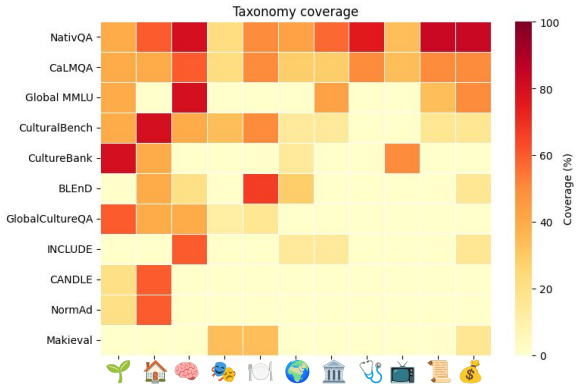


Figure 5: Taxonomy coverage for Related Work datasets. Percentage of subtopics covered in each higher-level domain proposed in our taxonomy for the named datasets.

C Selected languages analysis

Table 6 shows the coverage of languages in our work, detailing the distribution of the 24 languages included in our dataset. For each language, the table reports its relative proportion within the Common Crawl snapshot CC-MAIN-2025-43, together with an estimate of the total number of speakers (in millions), derived from publicly available Wikipedia statistics. In addition, we provide the corresponding language family for each language, as well as the ISO 639-1:2002 two-letter language code used consistently throughout the paper and accompanying resources. This information offers a concise yet comprehensive view of the linguistic diversity and representativeness of the dataset analyzed in this work.

Figure 6 illustrates the relationship between the selected languages (excluding English), showing the correlation between number of speakers and their share in CommonCrawl. The broad dispersion across the plot demonstrates that CROQ covers a wide range of linguistic scenarios, spanning more than six distinct language families.

D Further Evaluation Details

To process the open-ended answers produced by the LLMs, we employ a secondary model as a judge to interpret the responses and extract the relevant information. The objective is to associate each question with a set of countries in order to uncover potential cultural biases. Therefore, the judge

Domains	Subtopics
1. Beliefs, Values, and Identity	**Religion & Spirituality**: Beliefs and practices about the divine, morality, and afterlife. **Values & Ethics**: Core beliefs about right/wrong, community, and individual roles. **Symbolism**: Visual or material symbols that carry cultural meaning. **Mythology & Folklore**: Narratives that explain origins, values, or natural phenomena. **Gender Roles**: Cultural expectations related to gender identity and expression.
2. Social Structure and Daily Life	**Family & Social Roles**: Structure of family and communal relationships. **Community Life**: How people relate to neighbors and groups. **Manners & Etiquette**: Everyday norms for respectful behavior and interaction. **Work Culture**: Attitudes and habits related to labor, time, and productivity. **Conflict Resolution**: How societies handle disputes or disagreements.
3. Knowledge, Communication, and Education	**Language & Communication**: Spoken, written, and non-verbal systems of communication. **Education Systems**: Ways knowledge is transmitted and valued. **Technology & Tools**: Use of tools and innovation in daily and cultural life. **Knowledge Transmission**: How wisdom and skills are preserved across generations. **Science & Philosophy**: Intellectual traditions shaping discovery and thought.
4. Cultural Expression and the Arts	**Music**: Rhythmic and melodic expression of culture **Dance**: Movement as performance, ritual, or celebration **Art**: Visual cultural expression (painting, sculpture, etc.) **Literature & Storytelling**: Written and oral traditions conveying culture and values **Theater & Performance**: Live dramatic cultural expression **Film & Cinema**: Storytelling through modern moving-image media **Crafts & Handicrafts**: Traditional material culture and artisanal skills **Jewelry & Body Decoration**: Ornamentation for identity, beauty, or tradition **Architecture & Housing**: Building styles and structures shaped by culture and climate
5. Food, Drink, and Leisure	**Food & Cuisine**: Traditional dishes, cooking styles, and dietary customs **Traditional Beverages**: Culturally significant drinks **Recreation & Leisure**: Activities for enjoyment and relaxation **Sports & Games**: Physical activities and competitions enjoyed culturally **Life-Cycle Rituals**: Ceremonies marking stages of life **Festivals & Celebrations**: Seasonal, national, or spiritual gatherings expressing shared identity
6. Geographic Aspects	**Climate & Environment**: How weather and landscape shape cultural practices, housing, and food **Topography & Land Use**: How people interact with physical geography like mountains, rivers, or plains **Rural vs Urban Culture**: Differences in lifestyle, work, and values between city and countryside **Regional Identity**: Sub-national cultural traits tied to a specific region **Borders & Territory**: How borders influence cultural blending or division **Migration & Diaspora**: Movements shaping identity **Relationship with Nature**: Beliefs and practices related to land, animals, and ecology
7. Political Aspects	**Government Systems**: Political structures influencing law, rights, and daily life **Law & Legal Traditions**: Cultural expectations of justice, rights, and punishment **Civic Values & Participation**: How people engage with politics, voting, and national identity **Freedom of Expression**: Cultural and political limits on speech or art **Colonial History & Influence**: Impact of colonization on culture, identity, and systems **Power & Hierarchy**: How authority is structured culturally and politically **Human Rights & Social Movements**: How cultures address equity, justice, and activism
8. Health & Wellness	**Health Practices**: Traditional, spiritual, or modern approaches to healing and care **Public Health**: Collective strategies for health and safety **Wellbeing & Lifestyle**: Balancing mental, physical, and social health **Birth & Reproductive Health**: Fertility, childbirth, and parenting norms and practices
9. Media & Entertainment	**Local Media**: Print, radio, and TV shaping cultural narratives **Digital Media**: Online platforms influencing communication, identity, and activism **Entertainment Industries**: Mass production of cultural content **News & Information**: Flow of journalism and narratives in society **Popular Culture**: Icons, trends, and shared cultural references **Gaming & Interactive Media**: Games, esports, and digital interactivity
10. History	**Historical Events**: Major turning points shaping collective identity **Historical Figures**: Leaders, thinkers, and cultural icons **Cultural Memory**: How societies remember, teach, and reinterpret their past **Colonialism & Resistance**: Struggles of domination and liberation **National Narratives**: Shared stories of origin, destiny, and identity **Heritage & Preservation**: Protecting, curating, and interpreting cultural pasts
11. Economy & Industry	**Key Industries**: Dominant sectors shaping work and identity **Trade & Exchange**: Movement of goods, ideas, and culture **Labor & Employment**: Work systems, professions, and class structures **Wealth & Inequality**: How resources and opportunities are distributed **Economic Growth & Development**: Paths to modernization and sustainability **Globalization & Cultural Economy**: Markets influencing culture and identity

Table 5: Overview of the taxonomy's domains (11) and subtopics (66), with brief definitions for each category.

model is tasked with identifying up to five regions or places mentioned or implied in each answer. If no such information could be reliably determined, the model is prompted to explicitly indicate that no inference is possible (non-answered responses).

We evaluated three prompting strategies for the judge model: (1) *Explicit*, in which only regions explicitly mentioned in the text were extracted; (2)

Code	Language	Family	Speakers (M)	CC %
en	English	Indo-European	1,528	44.8292
zh	Chinese	Sino-Tibetan	1,184	5.1598
hi	Hindi	Indo-European	609	0.1930
es	Spanish	Indo-European	558	4.3672
ar	Arabic	Afro-Asiatic	335	0.6549
fr	French	Indo-European	312	4.4605
bn	Bengali	Indo-European	284	0.1093
pt	Portuguese	Indo-European	267	2.0696
ru	Russian	Indo-European	253	6.1083
id	Indonesian	Austronesian	252	0.9505
de	German	Indo-European	134	6.0060
ja	Japanese	Other	126	5.2018
ko	Korean	Other	82	0.7754
fa	Persian	Indo-European	127	0.7814
sw	Swahili	Niger–Congo	87	0.0104
ha	Hausa	Afro-Asiatic	94	0.0027
az	Azerbaijani	Turkic	24	0.0624
am	Amharic	Afro-Asiatic	60	0.0033
su	Sundanese	Austronesian	32	0.0015
el	Greek	Indo-European	13	0.4898
as	Assamese	Indo-European	24	0.0025
ca	Catalan	Indo-European	9	0.1741
gl	Galician	Indo-European	2.4	0.0279
eu	Basque	Other	0.8	0.0306

Table 6: Distribution of the selected 24 languages (ISO 639-1:2002) included on our dataset by share of data in CommonCrawl (CC-MAIN-2025-43) and estimated number of total speakers (in millions) based on Wikipedia sources.

	Basic eval.			Thorough eval.		
Lang.	*Exp.(1)*	*Inf.(2)*	*Mid.(3)*	*Exp.(1)*	*Inf.(2)*	*Mid.(3)*
en	93.94	92.42	**100**	92.42	89.39	**100**
es	89.39	90.91	**96.97**	77.27	78.79	**96.97**
ca	93.94	96.97	**100**	56.06	46.97	**96.97**
eu	84.85	**96.97**	**96.97**	59.09	62.12	**86.36**
Avg.	90.53	94.32	**98.48**	71.21	69.32	**95.08**

Table 7: Evaluation Accuracy results for LLM as-a-judge in Basic and Thorough evaluations. System prompt types: (1) *Explicit*, (2) *Infer* and (3) *Middle*.

Infer, which allowed the model to extract regions that were either explicitly stated or reasonably inferred; and (3) *Middle*, a balanced approach that prioritized explicit mentions while permitting limited, non-speculative inference when necessary (see Figures 7 and 8).

For the evaluation, we conducted a manual assessment of 264 items (one per subtopic across four languages), covering both high-resource (*en*, *es*) and low-resource (*ca*, *eu*) settings. Ground truth labels for these items are established via manual human annotation. Accuracy is measured as the percentage of instances where the judge model's extractions perfectly matches this human-annotated ground truth.

We considered two evaluation granularities within this setup:

Basic evaluation: The objective is to extract the country or countries referenced in the text. If a country is explicitly mentioned, that country should be extracted. If a specific sub-national region is mentioned instead, the judge must extract either the general country name or the specific region, and both are considered correct.

Thorough evaluation: The objective shifts toward finer-grained geographic specificity, and the model is required to adhere strictly to the geographic entity explicitly mentioned in the output; if a country is mentioned, the model must extract the country, and if a specific sub-national region is mentioned, the model must extract that specific sub-national region.

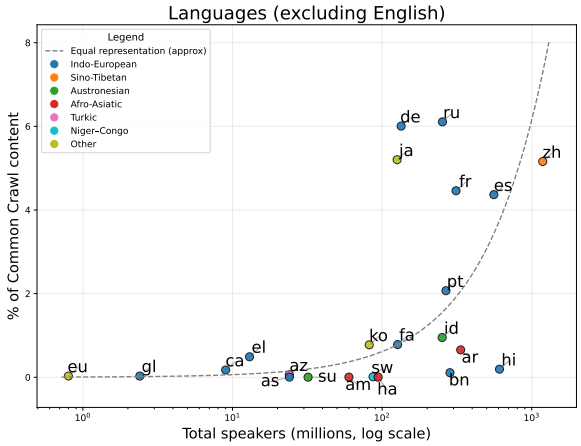


Figure 6: Distribution of the selected 23 languages (English excluded for a more clear visualisation) by share of data in CommonCrawl (CC-MAIN-2025-43) and estimated number of speakers based on Wikipedia sources.

Text: **"In Japan, staple foods eaten daily..."**
Basic Eval: Output must be -> Japan
Thorough Eval: Output must be -> Japan

Text: **"In Paris, notable museums include..."**
Basic Eval: Paris *or* France are accepted.
Thorough Eval: Only Paris is accepted.

Text: **"In the Basque Country*, the food is..."**
**transboundary region between Spain and France.*
Basic Eval: Spain, France *or* Basque Country.
Thorough Eval: Only Basque Country.

To account for this finer-grained geographic specificity, we report the comparative results of both setups in Table 7. The manual evaluation showed that the *(3) Middle* prompting strategy yielded the most accurate and consistent results. Based on this assessment, we adopted this prompt

in all subsequent analyses, using gpt-4o-mini from OpenRouter with default settings as the underlying judge model.

E Frontier models outputs for 24 languages

As an initial analysis, we examined the outputs of frontier models across seven widely spoken languages and identified a pronounced concentration in their regional references. Specifically, we analyzed the top three most frequently mentioned countries or regions for each language-model pair (see Table 8). Models overwhelmingly anchor their references to the primary country associated with each language, such as the U.S. for English, China for Chinese, France or Canada for French, and Japan for Japanese. This shows that models correctly internalize the conventional linguistic–national associations. Secondary references vary somewhat across models, but still follow intuitive cultural or regional links: for example, Arabic prompts frequently bring up Egypt, Saudi Arabia, and the UAE, while Spanish prompts often reference Mexico, Spain, and Peru.

Figure 9 presents the outputs (excluding models' own references) generated by the gpt-4o-mini, gemini-2.5-flash, claude-3.5-haiku, llama-4-maverick, command-r-08-2024, magistral-small-2506, qwen3-next-80b-a3b-instruct, and deepseek-v3.2-exp model variants across the complete dataset covering 24 languages. The evaluation spans content associated with twelve countries: Japan (JP), United States (US), India (IN), China (CN), France (FR), Italy (IT), United Kingdom (GB), Germany (DE), Mexico (MX), Brazil (BR), Russia (RU), and Egypt (EG). The left figure reports country references aggregated by model, highlighting how different model variants distribute references across countries. The right figure reports country references aggregated by language, illustrating how country mentions vary across the linguistic dimension.

E.1 Analysis by language

The results in Table 9 illustrate consistent patterns in how frontier models associate languages with countries. Across all 24 languages, references to countries where the language is an official language (*Own*) dominate the model's outputs. This behaviour suggests a heavy reliance on language–country co-occurrence signals, likely reflecting strong correlations within the training data. While such associations are linguistically plausible, their prevalence indicates limited decoupling between a language and its primary national context.

Beyond these self-referential patterns, the model exhibits a stable hierarchy regarding non-official country references. Japan and United States are the most frequently referenced nations across all languages, followed by India, France, and China. Importantly, these references appear even in languages with limited cultural or geographic ties to these nations, suggesting that cross-country mentions are driven more by general global prominence than by language-specific relevance. This finding aligns with prior observations that multilingual models tend to overrepresent culturally salient entities regardless of the linguistic context.

Examining the *Own* and *NA* columns reveals distinct behaviours based on language resource availability. The *Own* count is highest for low-resource languages such as Assamese, Sundanese, and Bengali, indicating a strong tendency for the model to ground these languages almost exclusively in their home countries. In contrast, high-resource languages like English, Chinese, and Russian show comparatively lower *Own* counts, reflecting the model's ability to distribute references more broadly. Additionally, while *NA* (missing response) counts are generally low, spikes in languages like Hausa and Hindi indicate occasional gaps in model generation for specific low-resource contexts.

The *O+N* column, which aggregates *Own* and *NA* percentages, further highlights this divide. Languages such as Bengali, Sundanese, Amharic, Basque, and Hindi exhibit the highest *O+N* values, reflecting a concentration of outputs that are either biased toward the home country or result in non-responses—traits characteristic of low-resource processing. Conversely, high-resource languages like English, French, and Russian display lower *O+N* values. This indicates a more diversified output distribution, consistent with languages that possess richer cross-country associations in the training corpus. Overall, a clear trend emerges: low-resource languages tend to yield highly concentrated or missing responses (high *O+N*), while high-resource languages yield more balanced references reflecting broader cultural associations.

Finally, the analysis of the relationship between normalized diversity and entropy scores may indicate a correlation with language resource availability. In this context, diversity measures how

```
You are a precise geographical context classifier. Your task is to determine which nation,
country or state(s) a given text is explicitly referring to. Follow these rules strictly:

   1. Do not infer nations, countries or states based on general knowledge, stereotypes, or
   assumptions beyond what is stated in the text.

   2. If multiple nations, countries or states are mentioned, list all of them. Not more than 5.

Keep your answers concise and limited to the nation, country or state names. Translate the final
answers into English.
```

Figure 7: Chosen System prompt, *Middle* (3), given to the Judge model

```
You are a geographical context classifier. Identify the nation, country, or state(s) explicitly mentioned in the text.

Rules:
1. Only list countries/states that appear in the text. Do not infer.
2. If multiple countries/states are mentioned, list up to 5.
3. Only list nations, countries, or states; skip regions or cities unless needed for clarity.

Answer with country names only, in English, separated by commas.
```

```
You are a geographical context classifier. Identify the nation, country, or state(s) the text refers to.

Rules:
1. If no country is mentioned, infer the most likely one from context.
2. If multiple countries/states are mentioned, list up to 5.
3. Only list nations, countries, or states; skip regions or cities unless needed for clarity.

Answer with country names only, in English, separated by commas.
```

Figure 8: System prompt given to the Judge model (top: *Explicit (1)* ; bottom: *Infer (2)*)

	Models							
Language								
English (en)								
Chinese (zh)								
Spanish (es)								
Arabic (ar)								
French (fr)								
Russian (ru)								
Japanese (ja)								

Table 8: Top three countries most frequently referenced by model for each high-resourced language.

many unique countries are referenced at all, while normalized entropy captures the evenness of the distribution of mentions across those referenced entities. High-resource languages (e.g., English, Chinese, French, Russian) exhibit both high diversity and moderate-to-high entropy, indicating that they not only reference a wider array of nations but also distribute their focus more uniformly. In contrast, low-resource languages (e.g., Hindi, Bengali, Amharic, Sundanese) demonstrate low diversity and entropy; they reference fewer unique countries, and those references are heavily concentrated on a single entity (typically *own* countries) rather than being spread out. This suggests that the model's capacity to generate broad, evenly distributed cultural references is heavily conditioned by the availability of the underlying training data.

E.2 Measuring Cultural Diversity in Model Outputs

Table 10 reports diversity scores, the number of distinct geographic regions referenced across 24 languages and eight frontier models. A key aspect of this metric is that it measures model diversity across different regions per language, meaning that

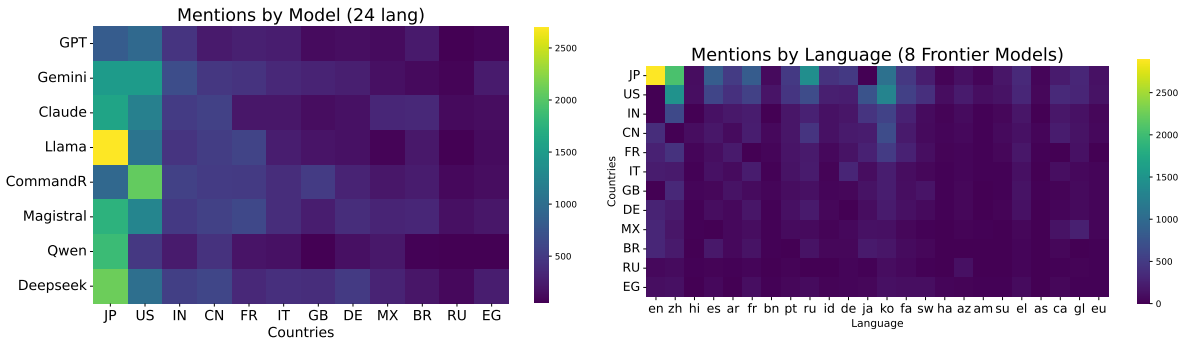


Figure 9: Outputs (without *own* references) from gpt-4o-mini, gemini-2.5-flash, claude-3.5-haiku, llama-4-maverick, command-r-08-2024, magistral-small-2506, qwen3-next-80b-a3b-instruct and deepseek-v3.2-exp variants over the full dataset of 24 languages. Countries: Japan (JP), USA (US), India (IN), China (CN), France (FR), Italy (IT), UK (GB), Germany (DE), Mexico (MX), Brazil (BR), Russia (RU), Egypt (EG). Data from Tables 1 and 9 respectively.

Lang.	**Own** (%)	**NA**	O+N	JP	US	IN	CN	FR	IT	GB	DE	MX	BR	RU	EG	**Div**	**Ent**
en	3,223 (.31)	192	.32	**2,893**	*2,028**	695*	367	265	223	*323**	277	309	314	57	104	157	0.69
zh	5,110 (.48)	258	.51	**2,064**	1,460	642	*4,781**	407	193	319	210	165	202	80	119	142	0.57
hi	7,430 (.70)	1,706	.87	94	**104**	*7,423**	92	37	13	38	21	14	11	14	11	56	0.14
es	6,935 (.66)	152	.67	**843**	588	130	173	110	128	48	100	*2,852**	186	23	58	136	0.65
ar	7,325 (.69)	499	.74	**504**	396	182	85	197	82	136	63	59	64	21	*1,645**	108	0.71
fr	5,956 (.56)	136	.58	**832**	557	215	236	*4,656**	225	85	162	95	137	44	104	149	0.54
bn	8,219 (.79)	1,189	.89	73	**151**	*2,415**	59	35	17	41	20	13	16	10	13	68	0.20
pt	7,719 (.73)	358	.76	**479**	442	167	161	147	146	94	124	85	*6,199**	30	50	114	0.42
ru	3,831 (.36)	173	.38	**1,423**	657	265	431	199	214	118	196	71	143	*3,827**	90	146	0.60
id	8,055 (.76)	870	.85	**411**	246	129	133	59	24	35	46	29	48	10	30	78	0.26
de	7,359 (.70)	71	.70	**485**	237	169	203	145	304	48	*6,927**	62	53	45	39	122	0.38
ja	6,213 (.59)	375	.62	*5,584**	**736**	434	219	267	98	144	98	133	193	21	47	114	0.36
ko	6,191 (.59)	1,580	.74	1,064	**1,308**	574	663	508	227	229	214	123	160	97	124	112	0.47
fa	6,015 (.57)	722	.64	473	**544**	263	194	266	101	107	124	55	89	59	105	117	0.48
sw	8,196 (.78)	144	.79	230	**389**	85	79	106	55	156	72	19	48	17	106	89	0.50
ha	5,385 (.51)	3,270	.82	30	**99**	30	44	25	13	21	16	3	9	1	37	49	0.10
az	7,960 (.75)	253	.78	123	**212**	81	64	43	34	47	48	23	10	119	29	78	0.43
am	8,088 (.77)	1,133	.87	43	**106**	17	19	13	19	9	6	12	5	8	22	38	0.13
su	8,331 (.79)	995	.88	**167**	146	49	76	48	14	15	32	10	7	6	18	62	0.11
el	7,393 (.70)	253	.72	**330**	307	161	135	173	124	141	132	46	36	34	83	106	0.36
as	8,732 (.83)	140	.84	41	**53**	*2,900**	42	18	6	17	8	2	4	4	6	41	0.30
ca	7,622 (.72)	158	.74	213	**339**	170	224	*306**	103	31	45	142	50	17	51	121	0.49
gl	7,967 (.75)	88	.76	**314**	297	129	143	135	58	77	68	259	292	23	21	113	0.37
eu	8,266 (.78)	903	.87	132	**144**	48	66	*674**	37	39	40	38	15	15	13	68	0.37
Avg.	6,980 (.66)	651	.72	**577**	414	197	170	153	102	87	92	77	82	33	56	99	0.40

Table 9: Frontier Model outputs for 24 languages. **Own**: counts references to countries in which the language is an official language. **NA**: denotes missing responses. **O+N**: percentage of the sum of Own and NA references relative to the total for the language. Referenced countries: Japan (JP), USA (US), India (IN), China (CN), France (FR), Italy (IT), United Kingdom (GB), Germany (DE), Mexico (MX), Brazil (BR), Russia (RU), Egypt (EG). **Div**: diversity scores. **Ent**: normalized entropy values. Bold values indicate the highest count for each language, excluding self-references. An asterisk (*) marks references to own countries.

for each language in our set, we calculate how many unique countries or regions are represented in the model's cultural references. A higher score indicates that a model's outputs for a given language draw from a broader, more globally varied set of cultural contexts, rather than being concentrated in one or two regions.

Averaged over languages, we observe large model-level differences: CommandR achieves the highest overall diversity, followed by DeepSeek and Gemini, while Mistral, GPT, and Qwen exhibit substantially lower scores. Language-level variation is also pronounced. High-resource languages such as English, French, Russian, and Spanish consistently receive broadest cultural coverage across models. In contrast, several low-resource or regionally concentrated languages (including Amharic, Assamese, Hausa, Hindi, and Bengali) remain well below the global mean, indicating narrower cultural framing in model outputs.

Beyond aggregate trends, we find substantial cross-model variance within the same language,

suggesting that cultural diversity is not determined by language alone. For example, Chinese ranges from 98 (Qwen) to 188 (CommandR), Bengali from 20 (Qwen) to 116 (LLaMA), and Hindi from 18 (Claude) to 107 (DeepSeek). Such differences often exceed average inter-language gaps, underscoring the impact of model-specific training and alignment choices. Overall, these results show that cultural diversity in multilingual generation is highly uneven, favoring high-resource languages, but also that these disparities are model-dependent rather than unavoidable.

Language									AVG
en	160	**171**	159	135	170	145	152	165	157
zh	127	165	125	157	**188**	137	98	143	142
hi	33	69	18	58	52	60	47	**107**	56
es	112	131	127	141	149	121	134	**169**	136
ar	89	126	88	101	**137**	95	110	115	108
fr	143	154	135	130	**174**	143	137	173	149
bn	36	75	28	**116**	115	64	20	90	68
pt	91	134	105	78	**143**	117	109	139	114
ru	141	156	147	149	**179**	120	118	158	146
id	72	**118**	89	56	53	56	62	115	78
de	94	113	**150**	118	137	119	132	112	122
ja	90	127	129	127	**130**	94	92	123	114
ko	73	120	145	99	99	**149**	75	137	112
fa	98	142	91	101	**167**	115	88	136	117
sw	68	109	87	110	**134**	85	46	75	89
ha	34	74	31	**86**	34	11	35	**86**	49
az	77	**109**	56	79	77	61	55	**109**	78
am	19	53	43	**74**	43	11	25	33	38
su	41	77	64	30	**100**	82	26	77	62
el	87	100	109	101	**162**	91	75	127	106
as	31	30	22	**97**	30	30	23	64	41
ca	102	115	140	126	**171**	97	105	114	121
gl	96	120	109	88	**163**	89	100	140	113
eu	55	**115**	94	83	64	28	36	73	68
ALL	82	113	95	102	120	88	79	116	99

Table 10: Diversity scores by language for different Frontier API models. ALL: per-model average across languages. AVG: per-language average. Highlighted results: above average results.

E.3 Topic Analysis in Model Outputs

Table 11 presents the countries most frequently mentioned by the models across the 11 general topics in our dataset. It reports the corresponding counts for these leading countries, as well as the models' own-country mentions.

On a model-by-model analysis of each topic (Table 12), the systemic bias emerges once again across all eight frontier AI models, with Japan and the United States dominating as the most frequently referenced countries across the eleven evaluated topics. Specifically, Japan consistently takes the top position in culturally and socially oriented domains such as Beliefs, Education, Arts, Food, and Health, while the United States leads in structural and global topics like Geography, Politics, History, and the Economy. Notably, this US-Japan centricity represents a strong cross-model consensus, appearing almost equally in Western models like OpenAI, Gemini, and Anthropic, as well as in Chinese models like Qwen and DeepSeek. Other countries, including India, China, and France, appear only sporadically as secondary or tertiary references, highlighting a profound homogenization in the ways current LLMs exhibit biases toward particular countries and regions.

E.4 Correlation between CC and output diversity

We examined, across all frontier models, the relationship between each language's proportion of Common Crawl (CC) data and the diversity of outputs generated in that language for each model. Table 13 reports Pearson's r and Spearman's r_s correlations, along with their associated p-values.

Across all models combined (Avg_all), we observed a Pearson correlation of 0.498 ($p = 0.013$), indicating a moderate positive linear relationship, and a Spearman correlation of 0.843 ($p < 10^{-6}$), showing a very strong rank correlation. These results suggest that languages with a higher proportion of CC data tend to exhibit greater diversity in model outputs.

When examining individual models, the correlations varied but largely supported this trend. For example, GPT-4o-mini showed the strongest Pearson correlation (0.577, $p = 0.003$) and a high Spearman correlation (0.774, $p < 10^{-5}$), indicating both linear and rank-based relationships between data availability and diversity. Qwen3-next-80b-a3b-instruct also displayed high correlations (Pearson 0.529, Spearman 0.808), while Gemini-2.5-flash and Claude-3.5-haiku exhibited moderate correlations. Deepseek-v3.2-exp and Magistral-small-2506 presented slightly lower but still statistically significant correlations, reinforcing the overall pattern.

Two models, however, showed weaker correlations. LLaMA-4-maverick had a Pearson correlation of 0.383 ($p = 0.064$) and Spearman 0.650 ($p = 5.84 \times 10^{-4}$), while Command-R-08-2024 showed Pearson 0.329 ($p = 0.116$) and Spearman 0.629 ($p = 9.87 \times 10^{-4}$). Although the Spearman correlations remain significant, the lower Pearson correlations suggest that the linear relationship is less pronounced for these models, possibly reflecting model-specific factors or limitations in handling low-resource languages.

Topic	Top-10 Countries	Own
1. Beliefs	1,506 586 549 400 332 318 282 278 189 147	10,801
2. Social	1,594 455 145 144 126 117 117 104 85 79	11,571
3. Education	1,086 453 268 267 169 165 157 153 149 136	12,279
4. Arts	2,028 1,191 768 621 516 381 337 304 263 262	23,032
5. Food	1,980 515 413 344 340 309 270 208 175 137	16,007
6. Geography	1,494 822 563 471 421 368 366 341 302 264	17,591
7. Politics	1,344 730 589 395 362 360 287 247 174 164	17,585
8. Health	1,068 361 220 210 159 111 83 70 64 59	10,474
9. Media	1,330 1,304 391 272 200 194 167 116 109 103	17,112
10. History	1,024 761 521 455 398 362 348 341 332 245	14,916
11. Economy	1,388 803 546 373 323 319 270 226 205 191	16,188

Table 11: Top 10 countries most frequently referenced by all Frontier API models for each topic in the taxonomy for 24 languages. Raw counts for each top country as well as own country mentions. **Own**: counts references to countries in which the language is an official language.

Topic								
1. Beliefs								
2. Social								
3. Education								
4. Arts								
5. Food								
6. Geography								
7. Politics								
8. Health								
9. Media								
10. History								
11. Economy								

Table 12: Top three countries most frequently referenced by Frontier API models for each topic in the taxonomy for 24 languages, excluding own country mentions.

Model	Pearson r	Pearson p	Spearman r_s	Spearman p
Qwen	0.529	0.008	0.808	1.80×10^{-6}
Gemini	0.483	0.017	0.798	3.00×10^{-6}
DeepSeek	0.425	0.038	0.775	8.90×10^{-6}
GPT	0.577	0.003	0.774	9.14×10^{-6}
Magistral	0.426	0.038	0.770	1.07×10^{-5}
Claude	0.452	0.027	0.760	1.63×10^{-5}
Llama	0.383	0.064	0.650	5.84×10^{-4}
Command-r	0.329	0.116	0.629	9.87×10^{-4}
Avg_all	0.498	0.013	0.843	2.31×10^{-7}

Table 13: Correlation between CommonCrawl percentage and output diversity for Frontier models. Pearson's r and Spearman's r_s are reported along with their p-values.

Overall, these results indicate that the amount of possibly available training data for each language may be an important factor influencing output diversity. Languages with less CC data tend to yield less diverse outputs, suggesting a potential reduction in creative or varied responses when data is scarce. This pattern is robust across most models, highlighting the influence of data coverage on multilingual model behaviour.

F Open-Weight Base and Instruct Models Analysis Over English Dataset

Table 14 presents a comparative analysis of cultural leanings exhibited by base and instruction-tuned Open Weighted LLMs when responding to open-ended cultural questions in English. Responses are categorized by their alignment with references to specific countries, allowing us to examine how pretraining and instruction tuning affect the geographic distribution of model outputs. We additionally report diversity (Div) and normalized entropy (Ent) scores to quantify the spread and balance of cultural references.

Base models exhibit broader cultural distributions. Across all model families, base models demonstrate relatively more balanced distributions

Model		NA	JP	US	IN	CN	FR	IT	GB	DE	MX	BR	RU	EG	Div	Ent
Llama-3.1-8B	Base	354	164	144	**200**	85	87	106	30	62	42	82	54	17	109	0.78
	Inst.	14	**407**	391	121	25	18	22	55	14	10	32	9	14	174	0.66
Llama-3.1-70B	Base	783	**237**	194	199	116	85	61	36	80	37	85	51	40	117	0.79
	Inst.	13	**417**	333	125	35	15	9	42	7	15	19	2	15	174	0.65
gemma-2-9b	Base	30	124	**317**	169	80	53	26	18	45	43	26	12	25	95	0.70
	Inst.	19	**269**	217	104	18	33	41	69	29	14	55	2	29	136	0.71
gemma-2-27b	Base	28	100	**202**	128	71	38	43	75	31	23	33	24	29	80	0.77
	Inst.	9	250	**285**	58	23	16	27	30	26	31	49	1	14	154	0.73
Mistral-7B	Base	52	98	**146**	93	73	41	38	41	46	45	28	40	31	91	0.81
	Inst.	14	233	**580**	86	34	18	45	57	14	22	58	1	11	160	0.63
Qwen2.5-7B	Base	265	98	**385**	98	32	51	64	34	35	18	34	9	7	147	0.71
	Inst.	16	**321**	304	74	310	15	50	20	21	5	11	5	10	150	0.62
Qwen2.5-32B	Base	248	183	**399**	131	35	36	19	39	15	31	15	8	5	142	0.65
	Inst.	12	**489**	247	58	140	37	18	24	17	11	25	4	7	118	0.61
Qwen2.5-72B	Base	182	180	**606**	150	67	31	28	55	29	34	40	6	25	144	0.66
	Inst.	10	343	**386**	106	192	31	22	15	29	6	16	1	11	127	0.60
Average	Base	243	148	299	146	70	53	48	41	43	34	43	26	22	116	0.73
	Instruct	13	341	343	92	97	23	29	39	20	14	33	3	14	149	0.65

Table 14: Comparison of named base and instruct models over English questions (data used for Figure 3). NA: denotes missing responses. Referenced countries: Japan (JP), USA (US), India (IN), China (CN), France (FR), Italy (IT), United Kingdom (GB), Germany (DE), Mexico (MX), Brazil (BR), Russia (RU), Egypt (EG). Bold values indicate the highest count for each country per model. Div: diversity scores. Ent: normalized entropy values.

of cultural references. While the United States is frequently the most represented region, other countries (particularly Japan, India, China, and several European nations) appear with substantial frequency. This pattern is reflected in higher normalized entropy values, indicating a comparatively more diverse set of cultural alignments than their instructed counterparts.

Instruction tuning induces strong cultural concentration. In contrast, instruction-tuned variants exhibit a pronounced shift toward a narrow set of cultural references. Across all examined model families, instruction tuning results in a sharp increase in references to Japan and the United States, often making these the two dominant regions by a large margin. On average, references to Japan and the US increase in instruction-tuned models, while references to India, China, Russia, Egypt, and Latin American countries decline substantially. This concentration is accompanied by a consistent reduction in normalized entropy, indicating a loss of cultural diversity. Notably, this trend holds across different architectures and parameter scales, including models developed in non-Western contexts (e.g., Qwen with China), suggesting that instruction tuning exerts a homogenizing effect that overrides differences introduced during pretraining.

Diversity metrics corroborate qualitative trends. The observed redistribution is quantitatively supported by the diversity and entropy metrics. While diversity scores (Div) increase for instruction-tuned models due to higher total counts concentrated in fewer regions, entropy consistently decreases, reflecting a more peaked and less balanced distribution.

Possible implications. Taken together, these results demonstrate that instruction tuning systematically reduces cultural diversity in model outputs, steering responses toward a limited set of culturally dominant perspectives, particularly those associated with the United States and Japan. This finding has important implications for the deployment of instruction-tuned models in cross-cultural or global applications, where preserving diverse cultural viewpoints may be critical.

G Models' Analysis Across Post-training Stages

Table 15 reports a controlled comparison between base, supervised fine-tuned (SFT), and instruction-aligned (Inst.) variants of open-weight models that provide multiple alignment stages. The analysis focuses exclusively on English prompts, enabling us to isolate the effect of alignment from multilingual prompting effects. By including OLMo-2 (7B,

Model		NA	JP	US	IN	CN	FR	IT	GB	DE	MX	BR	RU	EG	Div	Ent
Olmo-3-7B	Base	608	35	**186**	43	27	25	18	28	13	12	11	7	9	114	0.74
	SFT	85	**395**	212	97	26	109	64	39	30	28	22	26	10	145	0.66
	Inst.	41	**478**	171	89	19	86	78	43	39	44	11	27	5	124	0.64
Olmo-3-7B-Think	Base	608	35	**186**	43	27	25	18	28	13	12	11	7	9	114	0.74
	SFT	88	**453**	213	105	40	18	32	32	22	32	22	2	13	121	0.66
	Inst.	200	**386**	289	179	52	32	46	44	36	47	29	2	25	127	0.68
OLMo-2-7B	Base	109	64	**97**	84	76	46	57	38	35	30	29	26	35	102	0.84
	SFT	252	130	**520**	67	29	54	40	52	13	18	35	4	13	151	0.63
	Inst.	53	163	**566**	79	34	65	62	31	20	20	29	11	18	158	0.64
OLMo-2-32B	Base	155	61	**216**	119	54	38	44	34	34	28	26	17	9	88	0.77
	SFT	33	361	**390**	51	20	62	36	34	29	16	31	2	16	151	0.65
	Inst.	44	221	**368**	81	32	74	40	42	33	18	29	8	6	161	0.71

Table 15: Comparison of Base, SFT, and Instruct variants over English questions. NA: denotes missing responses. Referenced countries: Japan (JP), USA (US), India (IN), China (CN), France (FR), Italy (IT), United Kingdom (GB), Germany (DE), Mexico (MX), Brazil (BR), Russia (RU), Egypt (EG). Bold values indicate the highest count for each country per model. Div: diversity scores. Ent: normalized entropy values.

32B) and OLMo-3 (7B, 7B-Think), which expose intermediate SFT checkpoints, the table allows us to disentangle the impact of supervised fine-tuning from that of final instruction alignment.

Base models: higher entropy and weaker geographic priors. Across all architectures and scales, base models exhibit the highest entropy values and comparatively balanced country distributions. While some geographic preferences are already present (notably toward the US), these preferences remain moderate and are spread across a broader set of countries. For example, OLMo-2-7B Base achieves the highest entropy in the table, indicating minimal concentration on any single country. However, base models also show higher rates of missing or underspecified responses (NA), particularly in OLMo-3, reflecting lower controllability rather than alignment-induced bias.

Supervised fine-tuning sharply amplifies cultural bias. The transition from Base to SFT introduces the largest shift in geographic skew across all models. SFT variants consistently display dramatic increases in counts for a small number of dominant countries (United States and Japan) accompanied by substantial drops in normalized entropy. For instance, OLMo-2-7B SFT increases US-associated outputs, while entropy drops. Similar patterns appear in both OLMo-3 variants, where Japan becomes the dominant country after SFT. These results indicate that supervised alignment data strongly reinforces cultural priors present in instruction datasets, amplifying bias beyond what is observed in pretraining alone.

Instruction tuning moderates but does not eliminate alignment bias. Instruction-tuned models partially attenuate the extreme peaks introduced during SFT, but do not recover the diversity of base models. In several cases, entropy remains close to or below SFT levels, suggesting that dominant-country preferences become stabilized during alignment. Overall, instruction tuning acts as a smoothing mechanism rather than a corrective one.

Diversity and entropy: misleading signals. While diversity scores (Div) often increase after SFT and instruction tuning, this increase does not correspond to more balanced outputs. Instead, higher diversity frequently co-occurs with lower entropy, reflecting broader but highly uneven country coverage dominated by a few regions.

H Prompt Analysis for RQ3

H.1 Prompting for Base Models

Since foundational Large Language Models (LLMs) are originally trained with a self-supervised next-token prediction objective, they are not inherently designed to perform specific QA tasks without guidance. Unlike instruction-tuned models, base models often lack the alignment necessary to interpret a raw query as a command to provide a factual answer. To address this limitation, we employ In-Context Learning (ICL), a method that leverages explicit reasoning over background knowledge to guide the model's inference process (Brown et al., 2020). By providing the model with

relevant context and demonstrations within the input window, we effectively transform the generative completion task into a structured QA format, allowing the base model to infer the desired output pattern without the need for parameter updates.

```
Question: Who wrote '1984'?
Answer: George Orwell.

Question: What is the chemical symbol for
water?
Answer: H2O.

Question: What is the dominant religion
(in region/place)?
Answer: I'll choose {Country1}. The dominant
religion in {Country1} is {Country1_Religion}.

Question: What sacred texts are studied
(in region/place)?
Answer: I'll choose {Country2}. In {Country2},
the sacred texts studied depend primarily on
the religious community, as the country has
a diverse religious landscape.

Question: {Input Question}
Answer:
```

Figure 10: Prompt given to the Base models in order to follow instructions.

We operationalize this approach using the prompt structure illustrated in Figure 10. To ensure the model adheres to the correct output format, the prompt includes both general QA demonstrations and task-specific examples that explicitly condition the model to respond with a country or geographic region. The template is constructed with strategic placeholders to facilitate this: two slots are reserved for preselected countries—including a specific attribute for religion in the first instance to demonstrate cultural reasoning, while a final placeholder is designated for the target culture-related question. This few-shot setup serves to constrain the model's generation, ensuring it focuses on the relevant cultural entity.

To maintain a rigorous evaluation, we implement a post-processing filter once the judge model extracts the mentioned regions or countries from the output. We specifically identify instances where the LLM simply replicates the example regions provided in the few-shot prompt rather than generating an original response to the cultural query. These cases are classified as "Non-Answered," as they represent a failure of the model to generalize beyond the immediate context. Consequently,

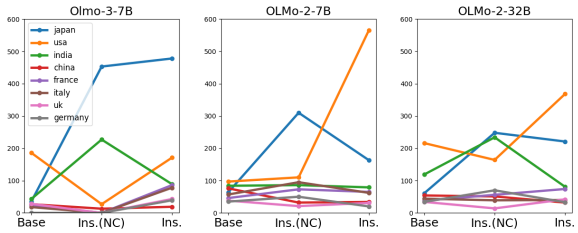


Figure 11: Outputs from the Base and Instruct variants of *Olmo3-7B*, *Olmo2-7B* and *Olmo2-32B* models. The Instruct models are shown in two modes: *Ins.(NC)*, which uses the same prompt as the Base model, and *Ins.*, which uses the standard chat template.

all reported counts and statistics in our results are derived exclusively from countries that were not introduced as examples in the prompt, ensuring our findings reflect the model's internal cultural associations.

H.2 Impact of the prompt

To assess the role of prompt formulation in eliciting instruction-following behaviour, we conduct controlled experiments in which prompts designed for base models are applied to instruction-tuned models. We examine whether such prompt mismatches lead to systematic differences in generation, focusing on the instruction-tuned variants of OLMo-2 (7B, 32B) and OLMo-3 (7B).

To evaluate how prompting affects Base versus Instruct models, we prompt the Instruct variants with the same input as the Base models (*Ins.(NC)*) across *Olmo3-7B*, *Olmo2-7B* and *Olmo2-32B* variants (see Figure 11). The 32B model consistently shows more stable, yet also more biased, behaviour than the 7B models, suggesting greater robustness of bigger models to prompt-formatting differences. Across settings, the USA and Japan frequently receive higher values across all settings, reflecting biases similar to those observed in prior models. Overall, while the prompt can influence responses, the main conclusion holds: cultural biases remain largely consistent, particularly toward Japan, the USA, and India, reinforcing the hypothesis that such biases stem from the models' training data.

I Surface Analysis of the instructions

To understand how post-training instructions affect cultural bias, we investigated whether a model's tendency to reference specific countries correlates with the geographic distribution of its training

Model Family	Post-Training Stage / Dataset	JP	US	IN	CN	FR	IT	GB	DE
OLMo-2	SFT (*tulu-3-sft-olmo-2-mixture*)	8,816	**10,188**	6,046	8,123	**9,668**	5,407	6,655	4,235
	DPO (*olmo-2-1124-7b-preference-mix*)	7,389	**7,839**	4,591	7,140	**8,110**	3,986	4,742	4,147
	DPO (*olmo-2-0325-32b-preference-mix*)	14,754	**15,660**	9,090	14,192	**16,128**	7,957	9,525	8,227
	RLVR (*RLVR-GSM*)	252	**466**	261	331	**622**	250	209	242
OLMo-3-7B-Think	SFT (*Dolci-Think-SFT-7B*)	21,009	10,261	22,002	**133,953**	**29,418**	12,408	19,987	9,525
	DPO (*Dolci-Think-DPO-7B*)	**7,837**	2,315	2,588	**4,399**	3,478	1,595	1,921	1,441
	RLVR (*Dolci-Think-RL-7B*)	743	298	703	**1,359**	**903**	443	477	373
OLMo-3-7B	SFT (*Dolci-Instruct-SFT*)	20,883	24,325	17,243	**52,509**	26,091	12,153	41,682	**42,349**
	DPO (*Dolci-Instruct-DPO*)	3,031	**7,011**	2,753	3,684	**4,006**	2,152	2,874	1,687
	RLVR (*Dolci-Instruct-RL*)	338	**365**	194	**386**	298	204	200	120

Table 16: Frequency of explicit country mentions across the SFT, DPO, and RLVR datasets for OLMo-2 and OLMo-3 models available instructions. Referenced countries: Japan (JP), USA (US), India (IN), China (CN), France (FR), Italy (IT), United Kingdom (GB), Germany (DE). **Bold**: Two highest counts per dataset

data. An initial qualitative analysis using OLMotrace (Liu et al., 2025b) yielded no significant insights. Therefore, we conducted a broader, quantitative experiment leveraging all publicly available post-training data for the OLMo-2 and OLMo-3 model families (Team, 2025a,b). Using a regular-expression-based approach, we quantified the raw appearances of our top eight countries across the entire post-training dataset (Table 16). This alignment allows us to examine whether country distributions in the instruction-tuning data are associated with regional patterns in the model responses (Table 15).

For the OLMo-2 models, the data reveals an inconsistent relationship between dataset frequency and model outputs. A direct correlation is observable regarding the United States, which maintains the highest (or near) representation across the SFT and DPO datasets, such as 15,660 mentions in the 32B preference mix, and consistently emerges as the most referenced region in both the SFT and Instruct model outputs. However, this correlation breaks down entirely for other highly represented regions. France exemplifies this disconnect, as it is frequently the most heavily represented country (of our selection) within the OLMo-2 instruction sets, even surpassing the US in the 7B and 32B DPO sets. Although France appears frequently in the training data, it is underrepresented in the model-generated outputs (Table 15), yielding only 74 mentions in the OLMo-2-32B Instruct variant compared to 368 for the US and 65 vs 566 in OLMo-2-7B. Conversely, Japan shows a distinct increase in generation behaviour, consistently ranking as the second most referenced country in the final outputs despite having only a modest presence in the post-training datasets, where it consistently appears less frequently than France and similar to China.

This divergence between dataset composition and generation behaviour is particularly evident in the OLMo-3 models. Within the SFT datasets, China accounts for the largest number of mentions, reaching 133,953 in *Dolci-Think-SFT-7B* and 52,509 in *Dolci-Instruct-SFT-7B*, exceeding the counts for any other country in the analysis. However, this prominence is not reflected in the model outputs, with China receiving only 26 mentions in the SFT variant and 19 in the Instruct variant, as in the generation outputs. At the same time, Japan and India become substantially more prominent in generated content. Although both countries appear relatively infrequently in the OLMo-3 post-training datasets and rank among the lower-frequency countries in several training stages, they receive some of the highest levels of representation in model outputs. For example, in OLMo-3-7B Instruct, Japan records the highest output count (478 mentions), considerably exceeding the US count (171 mentions), despite the US appearing more frequently in the DPO data and both China and several European countries, including Germany and the UK, appearing more frequently in the SFT data.

Overall, these empirical results demonstrate that raw frequency count within post-training instructions are not sufficient to explain the cultural and regional biases observed in model outputs. While certain heavy priors (like the US in OLMo-2) appear to persist, massive influxes of specific country data (such as China in OLMo-3) can be entirely attenuated during training, while countries (such as Japan) can become amplified. We must conclude by noting that this experiment did not clarify the main ***Why*** question of the paper, which continues to remain open for future investigation.