

Do You Know About My Nation? Investigating Multilingual Language Models' Cultural Literacy Through Factual Knowledge

Eshaan Tanwar¹ Anwoy Chatterjee¹ Michael Saxon^{2*} Alon Albalak^{3*}
William Yang Wang⁴ Tanmoy Chakraborty¹

¹Indian Institute of Technology Delhi ²University of Washington

³Lila Sciences ⁴University of California, Santa Barbara

{eshaantanwar2000, anwoychatterjee}@gmail.com

Abstract

Most multilingual question-answering benchmarks, while covering a diverse pool of languages, do not factor in regional diversity in the information they capture and tend to be Western-centric. This introduces a significant gap in fairly evaluating multilingual models' comprehension of factual information from diverse geographical locations. To address this, we introduce XNationQA for investigating the *cultural literacy* of multilingual LLMs. XNationQA encompasses a total of 49,280 questions on the geography, culture, and history of nine countries, presented in seven languages. We benchmark eight standard multilingual LLMs on XNationQA and evaluate them using two novel transference metrics. Our analyses uncover a considerable discrepancy in the models' accessibility to culturally specific facts across languages. Notably, we often find that a model demonstrates greater knowledge of cultural information in English than in the dominant language of the respective culture. The models exhibit better performance in Western languages, although this does not necessarily translate to being more literate for Western countries, which is counterintuitive. Furthermore, we observe that models have a very limited ability to transfer knowledge across languages, particularly evident in open-source models¹.

1 Introduction

Multilingual Large Language Models (LLMs) (Üstün et al., 2024; Achiam et al., 2023) show impressive performance on many languages across varied tasks. However, the best practices to evaluate them remain contested (Ahuja et al., 2023; Hada et al., 2023; Saha et al., 2023), with many criticising Western-centric (often Anglosphere) evalua-

tions (Held et al., 2023). For example, Faisal et al. (2021) demonstrated how multiple purportedly multilingual question-answering (QA) datasets disproportionately cover USA-related concepts and entities, with a similar question distribution to monolingual English benchmarks. This is because multilingual QA benchmarks are often derived from Western-centric English datasets (Longpre et al., 2021; Kassner et al., 2021; Dumitrescu et al., 2021). They, thus, fail to consider the cultural contexts where these languages are spoken (Naous et al., 2023). Blevins et al. (2022) proposed guidance techniques to create more inclusive benchmarks. To the best of our knowledge, there is no easy-to-evaluate large-scale parallel multilingual QA dataset that explicitly balances its distribution across a set of diverse cultures, capturing factual knowledge. LLMs are clearly able to access factual knowledge in multiple languages (Jiang et al., 2020; Kassner et al., 2021), but the relationship between language and information locality is poorly understood. To this end, and motivated by the education research literature, we produce a parallel knowledge test, *consisting of regionally differing factual information*, balanced across a set of related factual queries and languages.

In education theory, *cultural literacy* refers to the shared corpus of translinguistic cultural knowledge within a community, polity, or society on which efficient communication is built (Hirsch, 1983). By *translinguistic*, Hirsch (1983) meant that the information in the corpus is language-agnostic in its meaning. An interesting study conducted by Steffensen et al. (1979) in this regard found that American English speakers performed worse than Indian English speakers on a reading comprehension test of a story describing an Indian wedding, and vice versa on a story of an American one at equivalent reading levels because the context was more familiar to the speakers of the corresponding culture. Multilingual LLMs strive to create

*Work completed while affiliated with UC Santa Barbara.

¹The source code and XNationQA are available at <https://github.com/EshaanT/XNationQA>.

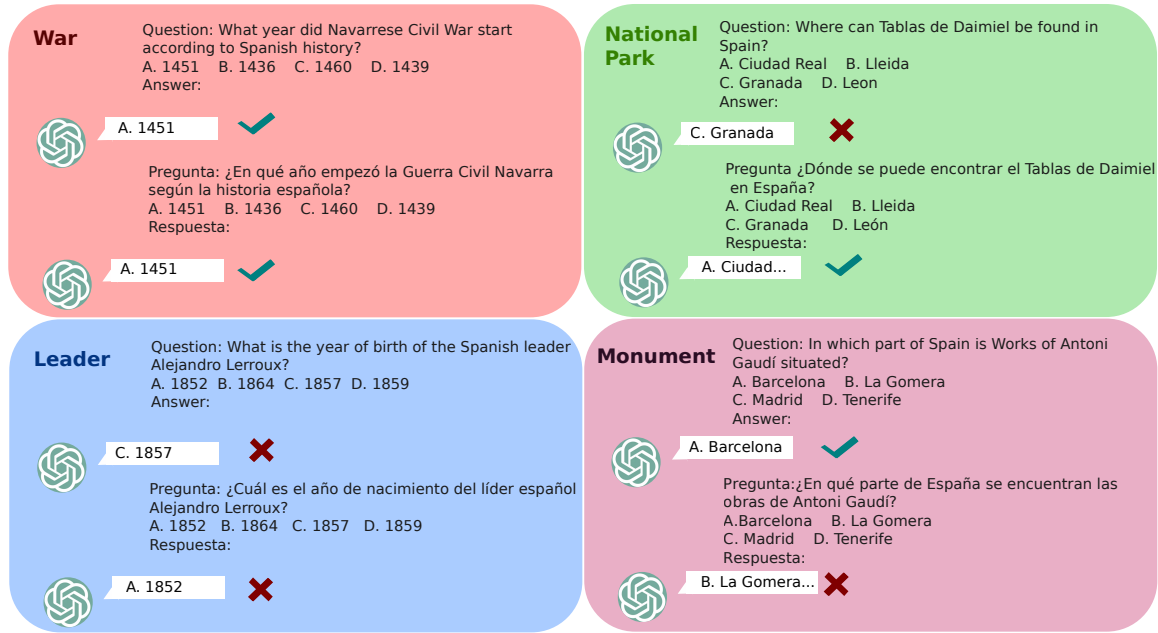


Figure 1: An example evaluating an LLM’s cultural literacy for Spain, in both English and Spanish. The model answers the *war* question correctly in both English and Spanish but fails on the *leader* question in both. Its performance on the *national park* and *monument* questions is language-dependent, highlighting inconsistencies in its cultural literacy across languages and topics.

technology that is inclusive of a diverse set of linguistic groups (Blasi et al., 2021). Therefore, they need to be culturally literate across the languages they cover to consistently answer translinguistic regional questions on specific topics concerning different nations.

Using cultural literacy as a framing device, we aim to investigate multilingual and multicultural knowledge in multilingual LLMs by answering the following research questions:

RQ1: *What disparities exist in the cultural literacy of various LLMs about different countries?*

RQ2: *How transferable is the cultural literacy of LLMs across different languages?*

RQ3: *How consistent are LLMs in recalling facts about nations across different languages?*

RQ4: *Does a systematic relationship between country and language exist in LLMs’ cultural literacy?*

To facilitate this investigation, we create XNationQA, a test set designed to assess cultural literacy through QA pairs that evaluate curricular knowledge for nine countries (Japan, India, China, Germany, Spain, Russia, Mexico, the USA, and the UK), fully translated into seven languages (Japanese, Hindi, Chinese, German, Spanish, Russian, and English). The dataset covers a set of four axes, namely: questions about wars, leaders, monuments, and national parks. Given culture’s nuanced and context-dependent nature, which can

vary across communities, languages, and sociopolitical environments, it is inherently challenging to define and quantify. To operationalize cultural literacy in our study, we draw on prior work in the monolingual (English-centric) setting that measures geographic and cultural erosion in language models via factual recall (Schwöbel et al., 2023; Zhou et al., 2022). Following this approach, we assess a model’s cultural literacy about a nation through its performance on questions concerning historically and geographically grounded facts specific to that nation. We categorise these facts into four domains: (1) *wars*, or the year of armed conflicts undertaken by the country, (2) *national parks*, or the site of nationally designated protected wilderness areas in each country, (3) *leaders*, or the date of birth of heads of state of the country, and (4) *monuments*, or the locations of UNESCO world heritage sites in the country. This framing allows us to systematically evaluate a model’s knowledge of culturally embedded facts, serving as a proxy for its cultural literacy. Figure 1 shows a working example of our task. We prompt the model to answer a question each on monuments, wars, leaders, and national parks. The example highlights the model’s inconsistent cultural literacy across languages. For instance, the model answers the war question correctly in both English and Spanish but fails the leader question in both. Meanwhile, its

performance on the national park and monument questions is language-dependent, answering correctly in one language but not the other.

We extensively evaluate eight commonly-used multilingual LLMs on XNationQA and find that:

- *LLMs show a disparity in cultural literacy across languages*, with models exhibiting varying knowledge. Models tend to perform the best in English, followed by other Western languages (German, Spanish, and Russian).
- *LLMs have poor transference of facts across languages*, with our novel coverage metrics showing that open-source models struggle to answer faithfully in all languages. When considering only Western languages, models tend to be more faithful, indicating a knowledge coverage disparity between Western and non-Western languages.
- *LLMs exhibit varying performance across nations*, with different models exhibiting different levels of literacy about different nations. Surprisingly, a nation’s native language is not the best-performing language even for questions on that nation, with Western languages performing the best.

We also observe that although the models generally perform better in Western languages, this does not translate to higher accuracy for Western countries; in fact, models sometimes exhibit greater cultural literacy for countries like India, China, and Japan than for Western nations such as Germany or Spain.

2 Related Work and Motivation

Multilingual Language Models. Since the advent of pre-trained transformer-based language models (Devlin et al., 2018), there has been a constant effort to develop multilingual LLMs that can understand and reason in multiple languages. These variants are trained on unsupervised training objectives using large multilingual corpora such as Oscar (Abadji et al., 2022), mC4 (Xue et al., 2021), and CulturaX (Nguyen et al., 2023), which are not parallel across languages. Therefore, the multilingual generalization ability of these models is a by-product of their ability to project different languages into a common representation space (Artetxe et al., 2019; Blevins et al., 2022). This representation ability directly depends on the datasets they are trained on (Deshpande et al., 2022). However, studies have shown them to be biased towards Western concepts due to their training data mixture (Naous et al., 2023; Cao et al., 2023).

Multilingual Benchmarks. Over the years, multiple multilingual benchmarks spanning various NLP tasks (Dac Lai et al., 2023) have been created to evaluate the multilingual abilities of language models. However, these benchmarks are commonly derived from monolingual English benchmarks (Clark et al., 2020; Dumitrescu et al., 2021), and, hence, tend to be biased in their coverage. Our initial analysis using multilingual QA datasets reveals these problems. For example, MKQA (Longpre et al., 2021), a popular closed-book QA dataset, is generated by translating 10,000 samples from Google’s Natural Questions dataset (Kwiatkowski et al., 2019) without any cultural or regional considerations. We find the resultant dataset to contain questions about popular Hollywood shows like ‘Modern Family’ and ‘Rick and Morty’, while it does not cover Bollywood or other regional shows. Hence, these datasets fail to highlight if the model is culturally literate across languages and regions. In contrast, XNationQA is large (49,280 questions) and covers multiple specific topics in parallel sets for nine nations.

Factual and Cultural Knowledge in Multilingual Models. Several recent studies have begun to probe cultural knowledge in multilingual settings (Fung et al., 2024; Shi et al., 2024). Keleg and Magdy (2023) explores multilingual factual knowledge by mining Wikidata triples in four languages; however, its reliance on language-linked labels restricts coverage to only entities available in all languages – a limitation that overlooks culturally important items, like the Mahatma Gandhi Marine National Park, and thus lacks uniform representation. XNationQA addresses such limitations by employing translation toolkits to construct a parallel corpus across seven languages, ensuring equitable coverage of historically and culturally significant content. While other works like BLEND (Myung et al., 2024) and CaLMQA (Arora et al., 2024) also emphasise cultural specificity, they differ from our work in scope and evaluation paradigm. BLEND focuses on subjective, everyday knowledge sourced from native speakers, while CaLMQA uses a generative format for long-form questions that necessitates high-budget human evaluation. In contrast, XNationQA targets objective, historically anchored factual knowledge that is widely documented (e.g., wars, national parks, monuments, leaders) through a multiple-choice setup that enables scalable and accessible benchmarking for future studies.

3 XNationQA Dataset

XNationQA is a parallel multilingual dataset encompassing factual knowledge related to nine countries and seven languages. XNationQA is designed to evaluate the cultural literacy of LLMs across languages by testing their knowledge of nation-specific information. Each instance within XNationQA includes an objective-type question focusing on a specific domain, with one correct and three incorrect options. In total, the dataset contains 49,280 questions, spanning 1,760 factual entities. Figure 3 shows the distribution of the entities across nine countries we covered.

Mining Entities. We began by first mining factual entities specific to a country. We used English Wikipedia pages to obtain lists for four specific domains: (i) leaders of a country, (ii) national parks in a country, (iii) UNESCO sites in the nation, and (iv) wars the nation has participated in. This method was applied to extract entities for nine countries: Japan, India, China, Germany, Spain, Russia, Mexico, the USA, and the UK (see Appendix A.3 for further details).

After extracting the entity list, we used Wikidata to mine additional information. This included information about the start year of a war, the birth year of a leader, and the administrative location of a national park and the UNESCO site. Information about the entities without an appropriate Wikidata entry was filled out manually. This process produced the final entity-answer pairs used to construct the questions.

Constructing Question Templates and Options. After extracting a nation’s entity-answer pair, we manually created four prompt templates for each domain and country in English. Each domain’s template was designed to ask a specific question, either about the year or about the location (refer to Table 6 of Appendix A.1 for details on the prompt templates for each domain).

To generate the desired objective-type questions, we generated three incorrect answers by sampling three random administrative areas of a country where the UNESCO site and national park are not located. Additionally, for year-type questions, we generated incorrect answers by randomly adding or subtracting a number between 5 and 10 from the correct year. These incorrect answers, along with the correct answer, are paired with the manually created question templates associated with an entity

Score	DE	ES	HI	RU	JA	ZH	AVG
Cos-sim	0.94	0.94	0.87	0.90	0.87	0.87	0.89
BLEU	50.80	67.45	68.48	43.80	52.30	45.64	54.74
H-Eval	4.85	4.75	4.57	4.25	4.50	4.41	4.55

Table 1: Validation of dataset quality using human evaluation (H-Eval), BLEU score, and cosine-similarity (Cos-sim). The queries seem to be semantically aligned (see Table 5 of Appendix A.2 for the languages corresponding to the ISO codes).

to form multiple-choice questions. Hence, for each entity, we generate four prompts. This approach is designed to account for any prompt sensitivity or bias in the LLMs.

Expansion to Other Languages. We use translation toolkits to create the parallel multilingual corpus. This is done because not all entities in Wikidata have labels in multiple languages. Hence, to create XNationQA, all the generated templates and entity-option pairs were translated using Google Translate and GPT-4, respectively, into Hindi, Spanish, Chinese, Japanese, Russian, and German. We filled out the entity and the options in the templates in their respective languages to generate the relevant objective question. This resulted in a geographically diverse, parallel multilingual factual knowledge corpus, XNationQA.

Validation of Translation. To validate the quality of XNationQA, we employ back-translation (Miyabe and Yoshino, 2015) and semantic similarity (c.f. Table 1). We also conduct human evaluation on a subset of queries to further ensure that the translation quality is preserved across languages.

- (i) **Semantic Similarity:** To ensure that queries in different languages have the same semantic meaning as their English counterparts, we compute the cosine similarity between the English queries and their translations. We use multilingual sentence transformers to extract embeddings, which are then used to compute cosine similarity. Our dataset has an average similarity score of 0.89, suggesting cross-lingual consistency and meaning preservation.
- (ii) **Back-Translation:** In this study, we randomly selected 1,000 queries from XNationQA for each of the translated languages, i.e., 6,000 queries in total, across the various topics in our dataset. These queries were back-translated into English using Google Translate and then compared to

their original English versions. We observed an average BLEU score of 54.74 which indicates that the translations preserve the concepts of the original queries.

- (iii) **Human Evaluation:** To assert the grammatical correctness and overall clarity of our dataset, 1,000 queries (same as in back-translation) from each language in XNationQA, i.e., 6,000 queries in total, were evaluated by language experts. The experts assigned a score of 1 to 5 based on grammar, fluency, and coherence to assess the quality of the dataset. We found that all samples at least got a score of 4 with an average score of 4.55, indicating the high quality of our dataset (see Appendix B for further details on the human evaluation procedure).

4 Problem Definition and Experimental Setup

Our dataset \mathcal{D} spans L languages and covers a set of nation-specific factual entities \mathcal{E} . For each entity $e \in \mathcal{E}$ in our dataset, we have a set of four manually created questions q^l and options o^l where language $l \in L$. The task is to generate the correct answer a^l from the options. To do so, we formalize our prompting setup as generating the output \hat{y}^l conditioned upon the question and option, i.e., $\hat{y}^l = \arg\max P(y|q^l \oplus o^l)$. We then match the generated output \hat{y}^l with the correct answer a^l across all languages l and entities e , to check for transferability of cultural literacy.

We experiment with eight commonly-used instruction-tuned multilingual LLMs, specifically the 7 and 13-billion versions of LLaMA-2-Chat (Touvron et al., 2023), 8-billion Meta-LLaMA-3-Instruct (Dubey et al., 2024), 7-billion Bloomz (Yong et al., 2023), 7-billion Mistral-Instruct (Jiang et al., 2023), 7-billion Mixtral (Jiang et al., 2024), 13-billion Aya (Üstün et al., 2024) and GPT-4 (Achiam et al., 2023). For the GPT-4 model, due to budgetary constraints, we sample one question for each entity in every language. In total we evaluated GPT-4 on 12,320 questions. These models have different mixtures of languages in their training corpus (refer to Appendix A.4 for more information). Further, we also extend our analyses to Meta-LLaMA-3.1-8B-Instruct (Meta AI, 2024) and Qwen3 (8B and 14B) (Yang et al., 2025) models — the findings of which are detailed in Appendix D.

5 Results and Analyses

5.1 Cultural Literacy Across Languages

Table 2 presents the accuracy of eight multilingual LLMs on XNationQA across languages, averaged over countries. The evaluation reveals notable disparities in the cultural literacy of the LLMs. The models, ranked by average accuracy, are GPT-4 (72%), Mixtral-8x7B (60%), LLaMA-3-8B-Instruct (59%), Mistral-7B-Instruct (42.6%), LLaMA-2-7B-Chat (41%), LLaMA-2-13B-Chat (40%), Aya (32%), and Bloomz (30%).

A key finding is that the performance of the models varies significantly with both the topic and the language of the query. Models are generally more proficient at recalling locations for *monuments* and *national parks* than they are at recalling specific years for *wars* or *leaders* (see Appendix A.7 for a detailed breakdown). Across all topics, English is typically the best-performing language.

Surprisingly, models specifically trained for multilingual alignment, such as Bloomz (46 languages) and Aya (101 languages), underperform compared to the models from LLaMA and Mistral families of comparable size, which have a much smaller span of languages in their pre-training data. In fact, Bloomz and Aya show near-random performance on date-recall tasks related to wars and leaders. The performance variation across languages for the same country is further analyzed using standard deviation in Appendix C.

5.2 Western vs. Non-Western Languages

Our analysis reveals a significant performance gap between Western (English, German, Spanish, Russian) and non-Western (Hindi, Japanese, Chinese) languages, as shown in Table 2. Models are consistently more culturally literate when queried in Western languages, a finding that is statistically significant (see Table 7 in Appendix for the statistical test results). While the composition of pre-training data partially explains this trend, it does not account for all anomalies. For instance, the underperformance of the LLaMA-2 series in Hindi and that of the Mixtral series in Hindi, Japanese, and Chinese is expected, as these languages are underrepresented in their respective training data. However, some results are counterintuitive. Mixtral, for example, despite being primarily trained on English, German, and Spanish, performs substantially better in Russian than in other non-primary languages like Hindi, Japanese or Chinese. Sim-

Language Model	EN	DE	ES	HI	RU	JA	ZH	AVG	AVG _W	AVG _{NW}
Monuments										
Bloomz-7B1	57.30	45.64	51.51	21.00	13.08	17.88	44.04	35.78	41.88	27.64
LLaMA-2-7B-Chat	82.03	75.98	76.25	3.83	36.30	37.72	34.43	49.50	67.64	25.31
Mistral-7B-Instruct	88.26	80.07	81.41	29.27	65.30	26.25	33.54	57.73	78.76	29.68
Meta-LLaMA-3-8B-Instruct	59.16	88.52	87.90	53.91	57.38	55.43	63.08	66.48	73.24	57.46
LLaMA-2-13B-Chat	85.41	77.76	80.16	15.39	54.80	37.19	31.49	54.60	74.53	28.02
Aya	64.15	62.81	59.52	35.68	30.34	34.34	32.83	45.67	54.20	34.29
GPT-4	96.09	95.73	96.44	87.90	92.02	92.17	80.78	89.88	92.08	86.94
Mixtral-8x7B	93.33	85.50	88.52	20.46	82.12	37.19	37.19	63.47	87.37	31.60
Leaders										
Bloomz-7B1	22.25	26.83	20.08	21.67	24.00	26.58	26.00	23.92	23.29	24.76
LLaMA-2-7B-Chat	56.42	46.58	27.42	8.50	23.00	28.33	28.00	31.18	38.35	21.62
Mistral-7B-Instruct	32.83	34.92	33.67	22.92	32.83	19.08	28.42	29.24	33.56	23.47
Meta-LLaMA-3-8B-Instruct	80.25	82.92	80.42	57.58	63.08	35.50	28.67	61.20	76.67	40.57
LLaMA-2-13B-Chat	52.83	30.42	30.25	18.25	26.42	25.42	27.50	30.15	34.98	23.70
Aya	23.33	22.42	21.08	22.08	23.42	18.75	19.92	21.57	22.56	20.25
GPT-4	62.33	62.33	63.33	47.33	55.67	48.00	48.33	55.33	60.92	47.87
Mixtral-8x7B	89.17	85.25	85.92	9.33	79.33	32.75	28.08	58.55	84.92	23.39
Wars										
Bloomz-7B1	32.93	25.11	29.04	33.96	28.12	28.79	39.24	31.02	28.80	33.98
LLaMA-2-7B-Chat	56.27	46.07	42.25	24.93	34.49	39.27	35.94	39.89	44.77	33.38
Mistral-7B-Instruct	48.19	45.11	46.46	25.78	44.51	30.74	35.87	39.52	46.07	30.78
Meta-LLaMA-3-8B-Instruct	71.74	64.09	58.64	51.17	53.54	50.74	49.43	57.05	62.00	50.78
LLaMA-2-13B-Chat	47.59	40.37	34.63	20.75	37.96	37.25	35.38	36.28	40.14	31.13
Aya	30.59	27.05	27.58	26.17	26.24	26.81	24.68	27.02	27.87	25.88
GPT-4	74.08	69.41	68.98	57.79	67.85	60.34	59.07	65.36	70.08	59.06
Mixtral-8x7B	76.77	73.65	70.64	26.27	66.29	49.04	47.52	58.60	71.84	40.94
National Parks										
Bloomz-7B1	48.04	40.06	46.30	25.95	15.49	20.08	25.11	31.58	37.47	23.72
LLaMA-2-7B-Chat	74.63	64.80	66.12	11.42	33.14	30.66	27.85	44.09	59.67	23.31
Mistral-7B-Instruct	77.33	72.94	67.39	19.50	51.11	13.90	25.69	46.84	67.19	19.70
Meta-LLaMA-3-8B-Instruct	64.22	79.60	80.02	51.32	56.77	39.11	40.96	58.86	70.15	43.80
LLaMA-2-13B-Chat	75.53	68.97	66.70	10.15	42.49	31.92	26.27	46.01	63.42	22.79
Aya	50.32	45.30	44.71	36.21	33.30	32.66	24.42	38.13	43.41	31.09
GPT-4	97.25	94.93	94.29	83.09	89.64	74.63	60.04	84.84	94.03	72.58
Mixtral-8x7B	91.01	82.98	84.25	19.93	73.26	31.08	24.84	58.19	82.88	25.27

Table 2: Model accuracy on XNationQA across seven languages, averaged over all nine countries. The AVG_W and AVG_{NW} columns show mean accuracy for Western (EN, DE, ES, RU) and non-Western (HI, JA, ZH) languages respectively. While GPT-4 demonstrates the strongest overall cultural literacy, other competitive models like Mixtral-8x7B show a significant drop in performance for non-Western languages (see Table 5 for the ISO codes).

ilarly, Russian outperforms Hindi for LLaMA-3, even though the model is purportedly optimized more for Hindi. These discrepancies, alongside GPT-4’s varied performance across Western and non-Western languages for certain domains, and the sub-par performance of open-source models in non-Western languages, raise concerns about their inclusivity for a diverse global user base (Blasi et al., 2021).

5.3 Transferability of Cultural Knowledge Across Languages

While the previous analysis focused on accuracy in individual languages, a crucial aspect of true multilingual proficiency is knowledge transferability.

For a model to be considered culturally literate, it should be able to answer factual questions consistently, regardless of the language used. To quantify this, we introduce two novel metrics: *Total Coverage* and *Smooth Coverage*.

Total Coverage (TC). This metric evaluates the consistency of complete factual recall. To measure how consistently a model knows a specific fact across multiple languages, we first define *coverage* (C_l^d) for a given language l and domain d as the set of all entities in d for which the model correctly answers at least three of the four associated questions in the given language. Then, similar to Qi et al. (2023), we define *Total Coverage* (TC^d) as

the ratio of entities covered across *all* tested languages to the total number of entities covered in *any* language:

$$TC^d = \frac{|\bigcap_{l \in L} C_l^d|}{|\bigcup_{l \in L} C_l^d|} \quad (1)$$

TC^d directly measures a model’s ability to transfer knowledge. A high TC^d score indicates strong knowledge transference, even if overall accuracy is modest. We evaluate TC^d under four distinct scenarios to probe different aspects of this transferability:

- (i) **Total Coverage (All), $TC^d(All)$.** This measures cultural transferability across all seven languages in XNationQA, providing a holistic view of a model’s multilingual alignment on our dataset.
- (ii) **Total Coverage (Pre-training data), $TC^d(Pre - Train)$.** To provide a fairer assessment of models like LLaMA and Mistral, where some languages are known to be underrepresented in their training data, this metric calculates TC only on the languages well-represented in each model’s pre-training corpus.
- (iii) **Total Coverage (West), $TC^d(W)$.** Given the observation from Table 2 that models perform better in Western languages, this metric quantifies their knowledge transferability specifically within this language group (English, German, Spanish, Russian).
- (iv) **Total Coverage (English Non-Western), $TC^d(Eng - NW)$.** Since English is the dominant language for all models, this scenario measures how well knowledge transfers from English to the non-Western languages in our dataset.

The results, presented in Table 3, reveal a significant lack of knowledge transfer in most models. The $TC(All)$ score for open-source models is notably low, often near 5%, indicating that a vast majority of their cultural knowledge is not consistently accessible across languages. While the $TC(Pre - Train)$ scores show an expected increase (averaging 10.4 \times), the gap remains substantial. We also observe a strong bias towards Western languages (Naous et al., 2023); for instance, Mixtral and Mistral show a nearly 30 \times and 18 \times jump in transference for $TC(W)$, respectively, compared to $TC(All)$. This jump can be attributed to the language distribution in their pre-

Model	TC^d (All)	TC^d (Pre-Train)	TC^d (W)	TC^d (Eng-NW)
Monuments				
Bloomz-7B1	0.00	0.00	6.08	0.89
LLaMA-2-7B-Chat	0.38	6.51	26.46	0.82
Mistral-7B-Instruct	6.34	78.16	57.20	7.34
Meta-Llama-3-8B-Instruct	14.18	30.26	34.59	18.95
LLaMA-2-13B-Chat	2.27	8.71	44.87	2.40
Aya	6.01	6.01	21.90	10.70
GPT-4	62.59	62.59	81.88	72.56
Mixtral-8x7B	2.16	83.33	72.43	2.59
Leaders				
Bloomz-7B1	3.75	3.75	9.84	9.70
LLaMA-2-7B-Chat	0.48	3.40	6.95	1.05
Mistral-7B-Instruct	2.29	54.62	41.54	4.03
Meta-Llama-3-8B-Instruct	10.25	44.80	48.75	12.17
LLaMA-2-13B-Chat	6.67	14.12	24.71	8.28
Aya	16.95	16.95	30.69	22.45
GPT-4	28.46	28.46	50.83	33.48
Mixtral-8x7B	2.46	86.59	73.05	2.57
Wars				
Bloomz-7B1	5.21	5.21	11.39	13.50
LLaMA-2-7B-Chat	3.47	9.31	16.67	8.53
Mistral-7B-Instruct	3.68	42.25	35.40	5.29
Meta-Llama-3-8B-Instruct	21.22	37.09	38.81	27.81
LLaMA-2-13B-Chat	2.75	12.05	18.16	5.07
Aya	5.28	5.28	10.79	10.77
GPT-4	47.23	47.23	70.76	52.36
Mixtral-8x7B	9.45	71.26	59.33	11.30
National Parks				
Bloomz-7B1	0.29	0.29	7.72	0.65
LLaMA-2-7B-Chat	0.48	5.60	21.54	0.52
Mistral-7B-Instruct	1.41	67.08	43.45	1.78
Meta-Llama-3-8B-Instruct	7.94	27.74	31.54	11.45
LLaMA-2-13B-Chat	0.49	6.55	34.25	0.53
Aya	5.17	5.17	25.17	8.39
GPT-4	44.35	44.35	86.42	47.76
Mixtral-8x7B	1.08	77.88	60.35	1.11

Table 3: Total Coverage (TC) scores across the four domains, evaluated under different language scenarios. $TC(All)$ shows overall transferability, which is low for most open-source models. Scores improve for scenarios limited to pre-training languages ($TC(Pre - Train)$) and Western languages ($TC(W)$), indicating a strong data bias. GPT-4 consistently outperform other models, especially in the all-language scenario.

training data, as we saw a similar jump when comparing $TC(All)$ and $TC(Pre - Train)$ for these models. Interestingly, other models show stronger transference for $TC(W)$ than $TC(Pre - Train)$ indicating a bias towards Western languages, as also discussed in earlier section. The consistently low $TC(Eng - NW)$ scores further highlight the poor alignment between English and non-Western languages. Our pairwise analysis in Appendix A.5 corroborates this, showing that Western language pairs have significantly higher TC scores.

Smooth Coverage (SC). The binary nature of TC (an entity is either covered or not) can be overly strict, as it fails to credit models for partial knowledge (e.g., answering 2 of 4 questions correctly). To address this, we introduce *Smooth Coverage (SC)*, a more nuanced metric that uses a fuzzy set

extension. First, for each entity e in the domain d , we define its membership score, $m_l^{(e)}$, as the proportion of correctly answered questions for that entity in language l :

$$m_l^{(e)} = \frac{1}{P} \sum_{p=1}^P \mathbb{I}(y_{pred} = y_{truth})_{l,p}^{(e)} \quad (2)$$

where, P is the total number of prompts for a question on the entity e , y_{pred} and y_{truth} are the model generated and ground-truth answers respectively, and $\mathbb{I}(\cdot)$ is the indicator function. This score reflects the degree of knowledge and converts C_l^d into a fuzzy set. We then define SC as the average ratio of the minimum membership score to the maximum score for each entity across all languages:

$$SC_l^d = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \frac{\min_{l \in L} m_l^{(e)}}{\max_{l \in L} m_l^{(e)} + \epsilon} \quad (3)$$

where, \mathcal{E} is the set of entities in domain d , and ϵ is a small constant to prevent division by zero. This metric rewards models for consistent, even if partial, knowledge.

Model	M	L	W	N
Bloomz-7B1	0.71	3.66	5.03	0.84
LLaMA-2-7B-Chat	1.24	1.33	7.50	1.64
Mistral-7B-Instruct	6.94	2.58	5.56	1.95
Meta-LLaMA-3-8B-Instruct	20.13	14.16	24.11	12.06
LLaMA-2-13B-Chat	4.27	7.33	4.68	2.59
Aya	7.02	7.83	4.47	5.34
GPT-4	61.92	23.99	39.94	43.97
Mixtral-8x7B	5.07	3.75	13.27	2.48

Table 4: Smooth Coverage (SC) scores across all languages and domains. GPT-4 demonstrates the most robust cross-lingual transference. In contrast, Bloomz, despite its linguistically diverse training, shows poor transference across languages for all domains (M: Monuments, L: Leader, W: War, and N: National Park).

As reported in Table 4, the SC scores confirm the trends observed with TC. While the scores are numerically higher because they account for partial knowledge, the overall conclusion remains unchanged. GPT-4 demonstrates the strongest cross-lingual robustness, while models like Bloomz, despite their linguistically diverse training data, score the lowest. This finding reinforces that current open-source models are still largely inadequate at providing faithful and consistent answers in multi-lingual scenarios.

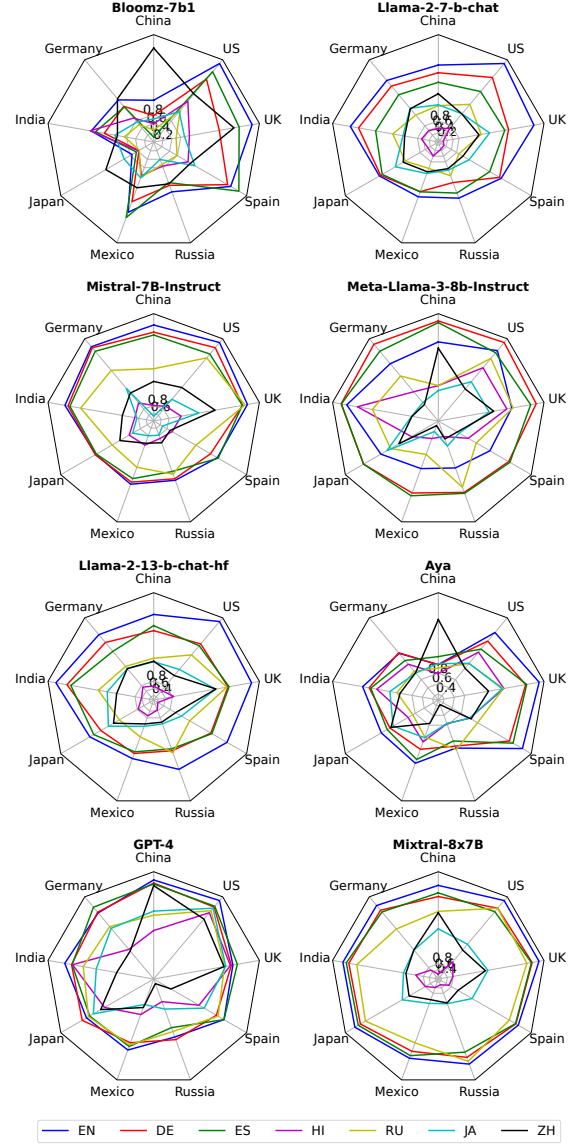


Figure 2: A detailed breakdown of model accuracy across different languages for each of the nine countries in XNationQA (see Table 5 for ISO codes). Each radar plot represents a model, with axes for the nine countries and colored lines for the seven languages. The plots demonstrate the significant performance disparities across both nations and languages. For instance, most models show stronger performance in English (blue line) across all countries, while struggling with non-Western languages like Hindi (magenta line). The irregular shapes of the plots for models like Bloomz or LLaMA-2 highlight inconsistent cultural knowledge across different nations.

5.4 Disparities in Cultural Literacy Across Nations

While we previously analyzed the performance of the models aggregated by language, let us now look into how cultural literacy varies across different nations, revealing significant inconsistencies.

As visualized in the radar plots in Figure 2, no single model demonstrates uniform cultural knowledge across all nine countries. The asymmetrical shapes of the plots clearly indicate that a model’s proficiency in one nation’s cultural facts does not guarantee similar proficiency in another’s, even when queried in the same language.

Performance on Leaders. A topic-by-topic analysis reveals further nuance. For questions about the birth year of *leaders* (see Appendix Figure 5), Mixtral-8x7B is the most consistent performer across different nations. However, this consistency is largely confined to Western languages; its performance drops significantly in non-Western languages, a trend observed across all models. LLaMA-3-8B-Instruct also shows reasonable consistency, particularly in English, German, and Spanish, though it notably struggles with facts about Mexico.

Performance on Monuments. When recalling the location of *monuments* (c.f. Appendix Figure 6), most models again perform well in Western languages. The exceptions are Aya and Bloomz, which struggle more broadly. The LLaMA-2 family exhibits a specific weakness, showing low performance on facts related to Japan, Mexico, and China, even when queried in those nations’ dominant languages. Among all models, only GPT-4 maintains a relatively uniform level of literacy across both Western and non-Western languages like Japanese and Hindi for this topic.

Performance on National Parks and Wars. The highest degree of variability is observed for questions about *national parks* (Appendix Figure 8) and *wars* (Appendix Figure 7). For these topics, performance is highly inconsistent, fluctuating significantly between both nations and languages for nearly all models. The only notable exception is GPT-4’s relatively stable and high performance on national park locations when queried in English, German, and Spanish.

5.5 Cultural Knowledge vs. Linguistic Competence

Finally, we analyze model performance averaged across all languages to see how culturally literate models are about different nations overall (see Appendix Figure 13). This perspective reveals a surprising trend that contrasts with our earlier language-based findings.

Performance by Nation. When viewed by country, the performance hierarchy is largely consistent with previous results: GPT-4 exhibits the strongest cultural literacy (nearly 80% accuracy for the US, China, and India), followed by LLaMA-3-8B-Instruct and Mixtral-8x7B (above 60% for the same countries). The LLaMA-2 series remains in the moderate 40 – 60% range, while Aya and Bloomz show the lowest literacy, often below 40% for most countries.

Decoupling of Knowledge and Language. Most interestingly, the strong performance divide between Western and non-Western languages does *not* translate into a similar divide between Western and non-Western *nations*. As shown in Table 2, models often demonstrate deeper cultural knowledge of countries like China, India, and Japan than of Western nations like Germany or Spain. This occurs even while the models perform poorly in the native languages of those non-Western nations (e.g., Chinese, Hindi). This decoupling of linguistic competence from cultural knowledge is a key finding of our study. It suggests that a model can be highly knowledgeable about a particular nation’s facts, even if it cannot express that knowledge effectively in that nation’s primary language.

6 Conclusion

In this work, we introduced XNationQA, a large-scale multilingual benchmark designed to evaluate the cultural literacy of LLMs beyond the typical Western-centric scope. Our analysis of eight models revealed significant inconsistencies in their factual knowledge across languages and nations. We found a strong performance bias towards Western languages, though this did not always translate to better knowledge of Western countries. Models specifically designed for broad language support, like Aya and Bloomz, struggled with factual recall, particularly with dates. Furthermore, our novel transference metrics showed that open-source models have a severe limitation in transferring cultural knowledge across languages, highlighting a critical gap between them and the proprietary LLMs. These findings underscore the need for more culturally inclusive training and evaluation methods to create truly global and equitable language models.

Limitations

This work attempts to present a more inclusive approach towards benchmarking the cultural lit-

eracy of models across nations and demonstrates the disparity in LLMs’ performance with variation in language. However, it has three main limitations. First, due to computational constraints, we could not experiment with larger multilingual LLMs. Consequently, while our work benchmarks a range of widely-used models, the cultural literacy of colossal models remains an important direction for future investigation.

Second, the design of XNationQA has limitations in scope and structure. Our selection of nations based on widely spoken languages resulted in no coverage of countries from Africa and South America. Furthermore, while the dataset is parallel across languages, the number of factual entities for each domain (e.g., wars, monuments, etc.) naturally varies between nations, an inherent constraint of nation-specific factual benchmarking approach.

Finally, this study’s scope is focused on performance benchmarking rather than an in-depth analysis of the models’ training data. Such an analysis could provide insights into our findings, such as why models like Bloomz and Aya exhibit lower cultural literacy despite their linguistically inclusive training. Investigating the topic distribution and biases in their publicly available training corpora, like that of BLOOM (Laurençon et al., 2022), is a valuable avenue for future research.

Ethical Statement

This work aims to evaluate the cultural literacy of LLMs across multiple languages and nations by probing their knowledge of historically and culturally significant facts. We take care to ensure that our dataset is balanced, factual, and respectful of diverse cultures, avoiding stereotypes or biased representations. All cultural content is sourced from publicly available, reputable references such as encyclopedias, official historical records, and recognized heritage listings.

We acknowledge that cultural knowledge is complex and dynamic, and our work does not attempt to capture the full richness of any culture or community. Instead, it focuses on well-documented factual knowledge as a proxy for measuring models’ cross-cultural understanding. We encourage future research to incorporate a wider range of cultural perspectives and to evaluate the social impact of deploying LLMs in multicultural contexts.

Acknowledgements

Anwoy Chatterjee gratefully acknowledges the support of the Google PhD Fellowship. Tanmoy Chakraborty acknowledges the support of the Anusandhan National Research Foundation (DST/INT/USA/NSF-DST/Tanmoy/P-2/2024) and Rajiv Khemani Young Faculty Chair Professorship in Artificial Intelligence. The authors acknowledge the support of Google GCP Grant.

References

- Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. [Towards a Cleaner Document-Oriented Multilingual Crawled Corpus](#). *arXiv e-prints*, arXiv:2201.06642.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Krithika Ramesh, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. [Mega: Multilingual evaluation of generative ai](#). *ArXiv*, abs/2303.12528.
- Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2024. Calmqa: Exploring culturally specific long-form question answering across 23 languages. *arXiv preprint arXiv:2406.17761*.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. [On the cross-lingual transferability of monolingual representations](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Damián E. Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. [Systematic inequalities in language technology performance across the world’s languages](#). *ArXiv*, abs/2110.06733.
- Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. [Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models](#). *ArXiv*, abs/2205.11758.
- Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. [Assessing cross-cultural alignment between chatgpt and human societies: An empirical study](#). *ArXiv*, abs/2303.17466.
- Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in tyologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

- Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.
- Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. [When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, and Viorica Patraucean. 2021. [Liro: Benchmark and leaderboard for romanian language tasks](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.
- Fahim Faisal, Yinkai Wang, and Antonios Anastasopoulos. 2021. Dataset geography: Mapping language data to language users. *arXiv preprint arXiv:2112.03497*.
- Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.
- Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. [Are large language model-based evaluators the solution to scaling up multilingual evaluation?](#) In *Findings*.
- William B. Held, Camille Harris, Michael Best, and Diyi Yang. 2023. [A material lens on coloniality in nlp](#). *ArXiv*, abs/2311.08391.
- Erik D Hirsch. 1983. Cultural literacy. *The American Scholar*, pages 159–169.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. [X-FACTR: Multilingual factual knowledge retrieval from pre-trained language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.
- Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. [Multilingual LAMA: Investigating knowledge in multilingual pretrained language models](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.
- Amr Keleg and Walid Magdy. 2023. Dlama: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. *arXiv preprint arXiv:2306.05076*.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.
- Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. Mlqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.
- Meta AI. 2024. Introducing llama 3.1: Our most capable models to date. <https://ai.meta.com/blog/meta-llama-3-1/>. Preprint / Technical Report.
- Mai Miyabe and Takashi Yoshino. 2015. [Evaluation of the validity of back-translation as a method of assessing the accuracy of machine translation](#). *2015 International Conference on Culture and Computing (Culture Computing)*, pages 145–150.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

- Tarek Naous, Michael Joseph Ryan, and Wei Xu. 2023. [Having beer after prayer? measuring cultural bias in large language models](#). *ArXiv*, abs/2305.14456.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Démoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.
- Jirui Qi, Raquel Fern'andez, and Arianna Bisazza. 2023. [Cross-lingual consistency of factual knowledge in multilingual language models](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2023. [Branch-solve-merge improves large language model evaluation and generation](#). *ArXiv*, abs/2310.15123.
- Pola Schwöbel, Jacek Golebiowski, Michele Donini, Cedric Archambeau, and Danish Pruthi. 2023. [Geographical erasure in language generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12310–12324, Singapore. Association for Computational Linguistics.
- Weiyang Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*.
- Margaret S Steffensen, Chitra Joag-Dev, and Richard C Anderson. 1979. A cross-cultural perspective on reading comprehension. *Reading research quarterly*, pages 10–29.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Zheng Xin Yong, Hailey Schoelkopf, Niklas Muenighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vasilina Nikoulina. 2023. [BLOOM+1: Adding language support to BLOOM for zero-shot prompting](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.
- Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2022. [Richer countries and richer representations](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2074–2085, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Prompt Templates

Table 6 shows a few question templates used in our dataset creation. We translate the question template along with the entity-answer tuples and options to create XNationQA.

A.2 ISO Codes

Table 5 contains the ISO codes of the seven languages used in our dataset.

A.3 Selecting Languages and Entity Distribution

We started our study by selecting the three most widely spoken languages in the world (Mandarin, Spanish and English). One nation from each of these languages (China, Spain, US) was selected in the beginning. We then expanded our selection to include a combination of diverse (Hindi, Russian) and similar languages (German and Japanese), with one nation where they are widely spoken (India, Russia, Germany). This resulted in a set of

Language	ISO 639-1 code	Family
English	EN	IE: Germanic
German	DE	IE: Germanic
Hindi	HI	IE: Indo-Iranian
Chinese	ZH	Sino-Tibetan
Russian	RU	IE: Balto-Slavic
Spanish	ES	IE: Italic
Japanese	ZH	Japonic

Table 5: List of languages and their ISO codes used in our experiments.

seven nations and languages. We further included Mexico and UK to expand on the cultures we cover. Figure 3 shows the distribution of entities over the nine countries.

A.4 Pre-Training Datasets of the LLMs Used in Our Study

- (i) **Aya:** Aya is an instruction-tuned mT5 model(Xue et al., 2020) that supports 101 languages, including all seven languages used in our study.
- (ii) **Bloomz:** Bloomz supports 46 languages, including programming languages. It covers all seven languages used in our study.
- (iii) **Mistral:** The Mistral family of models primarily focuses on European languages, covering English, French, Italian, German, and Spanish. The exact distribution of its training dataset is unknown, so its coverage of other languages is unknown.
- (iv) **LLaMA-2:** LLaMA-2’s pre-training dataset predominantly consists of English, with 89.7% of the data in English. The remaining 8.38% comprises unknown languages (mainly programming languages), and about 2% consists of non-English languages. Except for Hindi, all other languages in our dataset seem to be represented.
- (v) **LLaMA-3:** LLaMA-3 builds upon the multilingual capabilities of LLaMA-2, with 8% of its training dataset dedicated to multiple languages. Although LLaMA-3 has been trained on a wide array of languages, it is specifically optimized and safety-tuned for eight languages: English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai.

- (vi) **GPT-4:** The exact details of its training dataset are not publicly available.

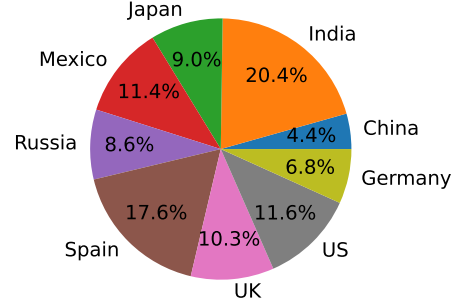


Figure 3: Distribution of entities across the nine nations considered in our study.

A.5 Pairwise Total Coverage Scores of Models

Figure 4 represents the pairwise *TC* of Bloomz-7B1, LLaMA-2-7B-Chat, Mistral-7B-Instruct, LLaMA-2-13B-Chat, and Aya. We see that the model generally sees a better *TC* score in Western languages compared to non-Western ones, highlighting that the models struggle in non-Western languages. Mistral-7B-Instruct and LLaMA-2-13B-Chat show relatively better performance among the models evaluated, but none as strong as GPT-4, Mixtral or Meta-LLaMA-3 that have been discussed in the main text.

LLaMA-2 series shows very poor *TC* for all language pairs consisting of Hindi, indicating the lack of consistency and understanding in Hindi. Aya while showing low absolute performance (c.f. Table 2), still shows overall reasonable *TC* across languages.

A.6 Evaluating Disparity Across National Park and War Questions

Figure 7 shows the performance of models on war-type questions. We note that models have different literacy on these facts with no model performing similarly across all countries or languages. Dates about wars Russia has taken part in seem to be difficult for the model to recall, except for Mixtral and Meta-LLaMA-3.

Figure 11 shows the performance of models on the location of national park-type questions. Here also we note that models tend to have different literacy on these facts with no model performing similarly. Except for GPT-4 for Western languages, there is no homogeneous performance across countries.

Domain	Question Template	Entity-Answer example
Year of War	Question: As per American history, when did {War} happen? {Option} Answer: Question: What year did {War} start according to American history? {Option} Answer: Question: Historical accounts from American indicate {War} occur in? {Option} Answer: Question: Can you tell me according to American history the date when {War} took place? {Option} Answer:	(American Revolutionary War, 1775)
Location of national park or monument	Question: Where can {Location} be found within China? {Option} Answer: Question: In which part of China is {Location} situated? {Option} Answer: Question: Where in China is {Location} located? {Option} Answer: Question: In which area of China can {Location} be found? {Option} Answer:	(Wuyi Mountains National Park, Fujian)
Birth year of Leader	Question: When was German leader {NAME} born? {Option} Answer: Question: What is the birthyear of the German leader {NAME}? {Option} Answer: Question: On what year was the German leader {NAME} born? {Option} Answer: Question: What is the year of birth of the German leader {NAME}? {Option} Answer:	(Olaf Scholz, 1958)

Table 6: Question Template in English along with an example entity-answer. We create the option using three negative samples and the true answer to fill the {Option}.

Model	M	L	W	N
Bloom-7B1	1.9e-47	0.38	0.002	1.23e-59
LLaMA-2-7B-chat	2.82e-160	4.16e-39	1.20e-38	2.83e-195
Mistral-7B-Instruct	4.96e-113	1.324e-5	3.06e-243	3.77e-189
Meta-LLaMA-3-8B	1.145e-31	3.90e-86	1.82e-32	1.02e-81
LLaMA-2-13B-chat	1.83e-32	7.35e-15	9.82e-12	4.67e-193
Aya	3.73e-45	0.45	0.02	1.06e-21
Mixtral-8x7B	1.09e-132	2.67e-156	1.47e-105	3.99e-247
GPT-4	1.25e-13	4.23e-9	2.45e-12	3.57e-51

Table 7: p-values for significance testing of the hypothesis that models perform better in Western languages. M (Monuments), L (Leader), W (War) and N (National Park). Significance is decided by taking $\alpha = 0.05$ and statistically significant results has been marked bold.

A.7 Model-wise Analysis of Results

Here we dive deeper into analysing each model for a specific use case and along the country-language axes as shown in Figures 9, 10, 11 and 12. The analysis is as follows:

- (i) **Bloomz-7B1:** We note that for war and leader domain questions, the model shows modest to poor performance in most cases, even worse than random in some instances. Despite being trained on a diverse multilingual corpus, Bloomz struggles with XNationQA compared to LLaMA or Mistral models of the same size. For questions related to monuments and

national parks, the model performs better in English, German, Spanish, and Chinese but shows low performance in other languages.

- (ii) **LLaMA-2-7B-Chat:** We consistently see that in all domains the model performs badly in the Hindi language (with accuracy lower than 20%). It seems to be most literate in Western-European languages which consistently outperform other languages for all countries and domains. Its coverage in Russian is still moderate compared to Western European languages. While it does show a decent level of literacy about a nation in the nation’s native language, they are still not the best language to prompt the model in.
- (iii) **Mistral-7B-Instruct:** It shows modest performance on leader domain questions, but for other domains, we note it to be culturally literate. However, it is biased towards Western languages and shows only modest performance in other languages, especially Hindi.
- (iv) **Meta-LLaMA-3-8B-Instruct:** It is the only model below the size of 10 billion to be liter-

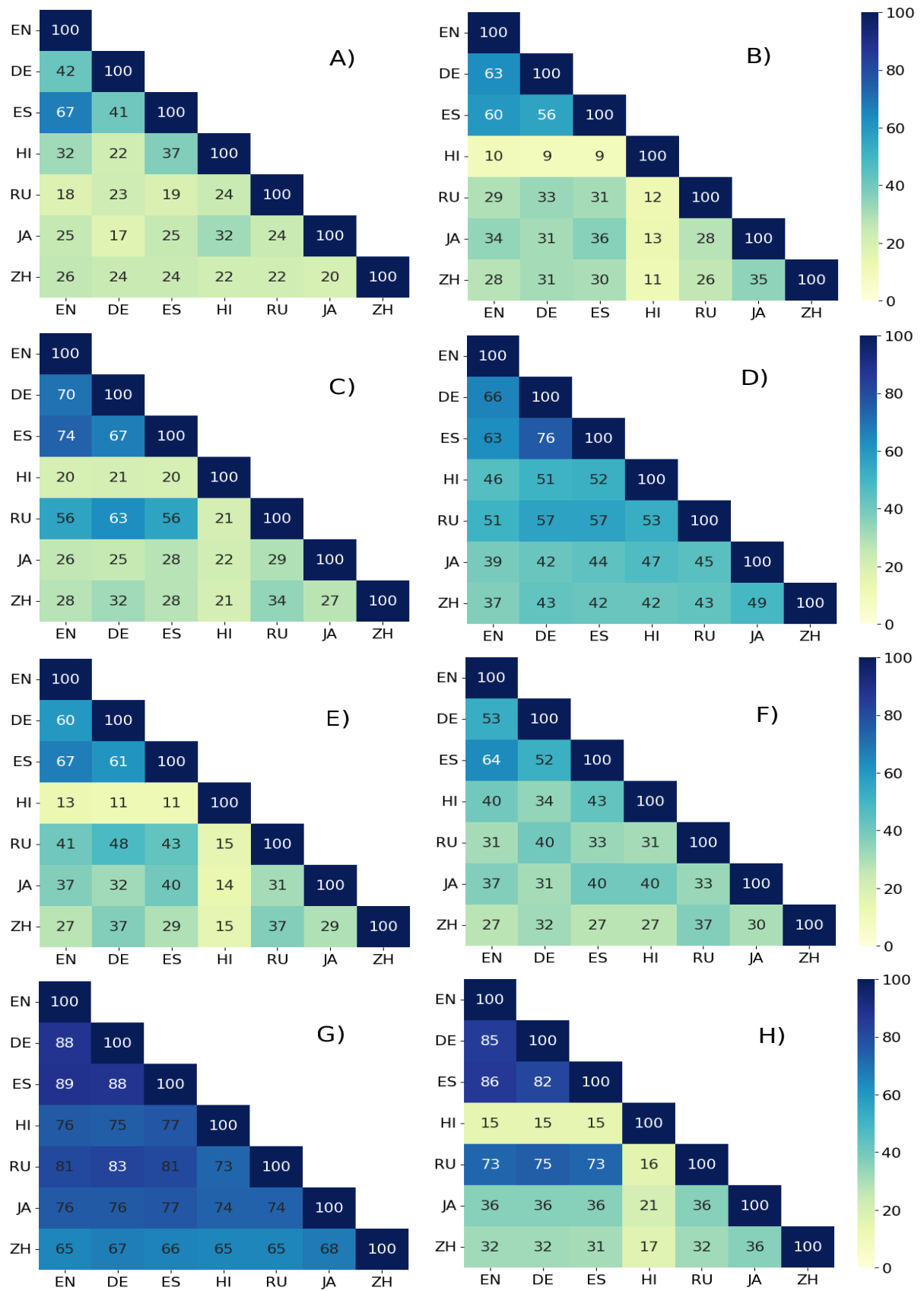


Figure 4: Heatmaps for pairwise TC of all language pairs, for – (A) Bloomz-7B1, (B) LLaMA-2-7B-Chat, (C) Mistral-7B-Instruct, (D) Meta-LLaMA-3-8B-Instruct, (E) LLaMA-2-13B-Chat, (F) 13-billion Aya, (G) GPT-4 and (H) Mixtral-8x7B.

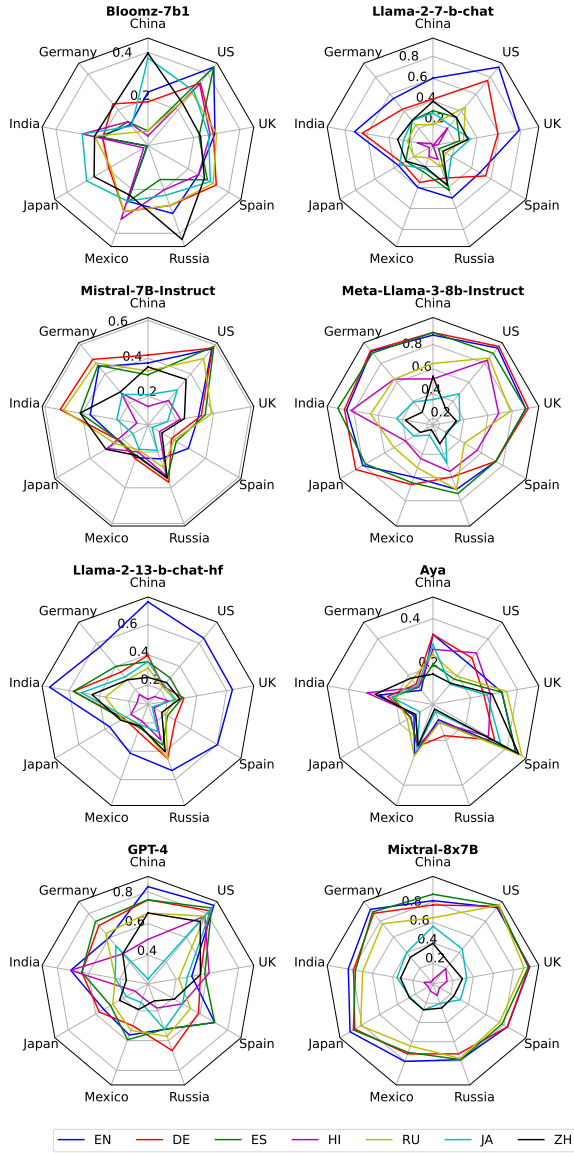


Figure 5: Accuracy of models in different languages for questions on the birth year of leaders of different nations (see Table 5 for ISO code).

ate to some degree in all the languages covered by us, even Hindi. It still seems to be more proficient in Western languages than others.

- (v) **LLaMA-2-13B-Chat:** Similar to the 7-billion variant of LLaMA-2, this model also struggles to perform well in Hindi and has considerable disparity between Western-European languages and other languages.
- (vi) **Aya:** Like Bloomz, Aya has been trained on a diverse multilingual corpus, and similar to Bloomz it also struggles to perform well in war and leader domain questions. Its performance in other domains is better, but still not

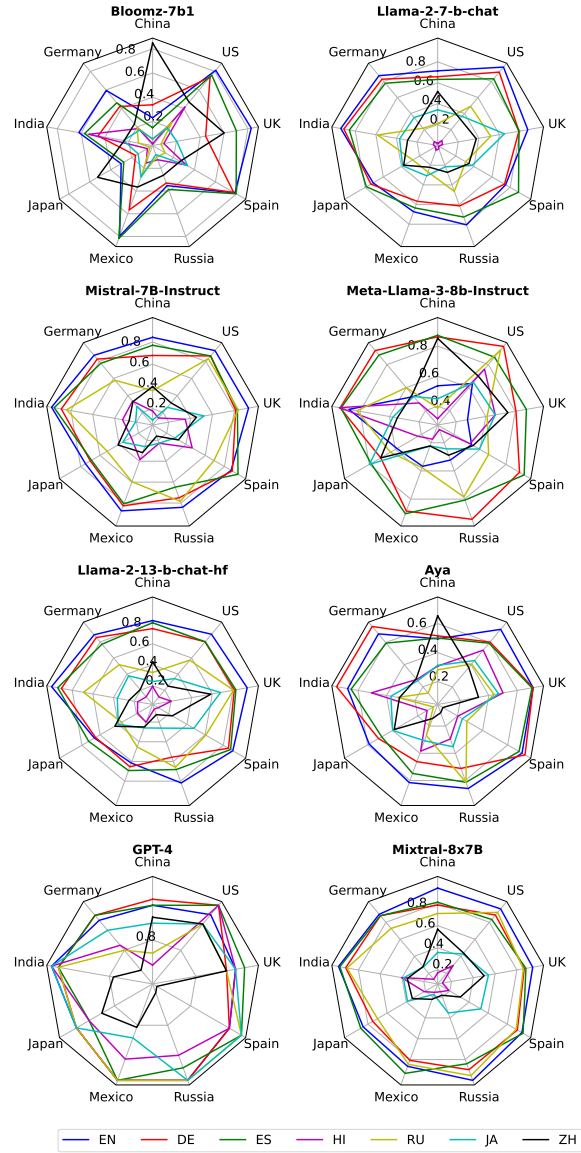


Figure 6: Accuracy of models in different languages for questions on the location of UNESCO sites of different nations (see Table 5 for ISO code).

comparable to LLaMA-2-13B-Chat which is of a similar size.

- (vii) **GPT-4:** It is the most culturally literate model in multilingual setup, with considerably higher performance than all other models. There is an interesting observation though – it struggles with questions about Germany if asked in Russian, especially for the monuments and national parks domains.

B Human Evaluation of Translation Quality in XNationQA

To evaluate the translation quality of queries in our dataset, we recruited university students aged between 20 to 25 who were either native speakers

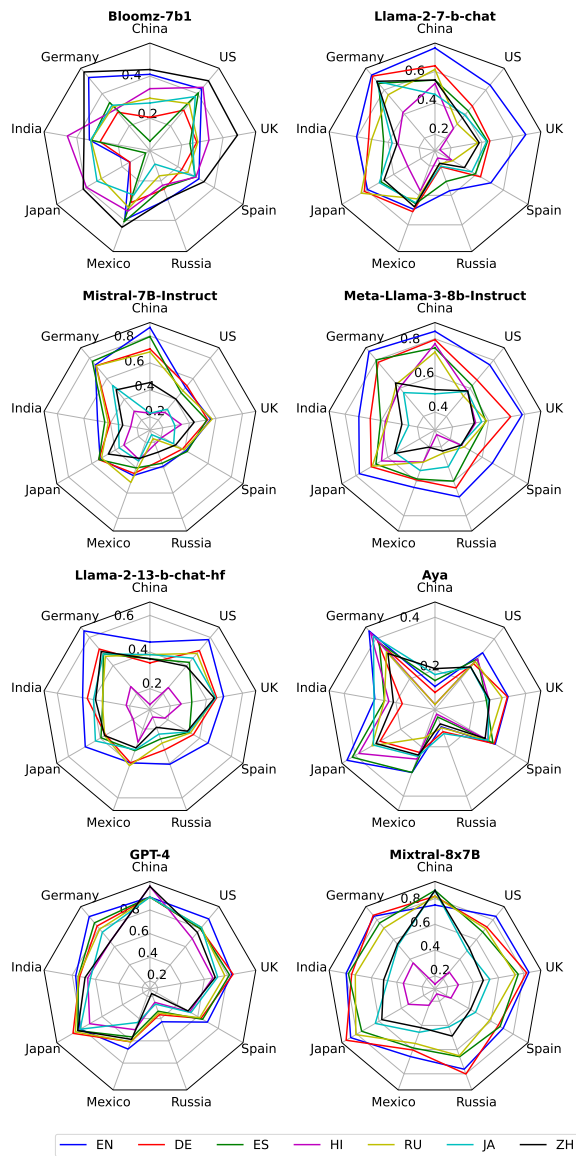


Figure 7: Accuracy of models in different languages (indicated in different colours) for questions on the date of wars in different nations.

of the language or had cleared language proficiency tests. We recruited a total of six language experts, one for each language. They were provided with the following Annotation Guidelines and Scoring Criteria.

Annotation Guidelines: Each translation is evaluated based on the following criteria:

- **Grammar** – Proper use of verbs, tense, agreement, and punctuation.
- **Fluency** – The sentence should sound natural and idiomatic.
- **Cohesion** – The sentence should be logically structured and clear.

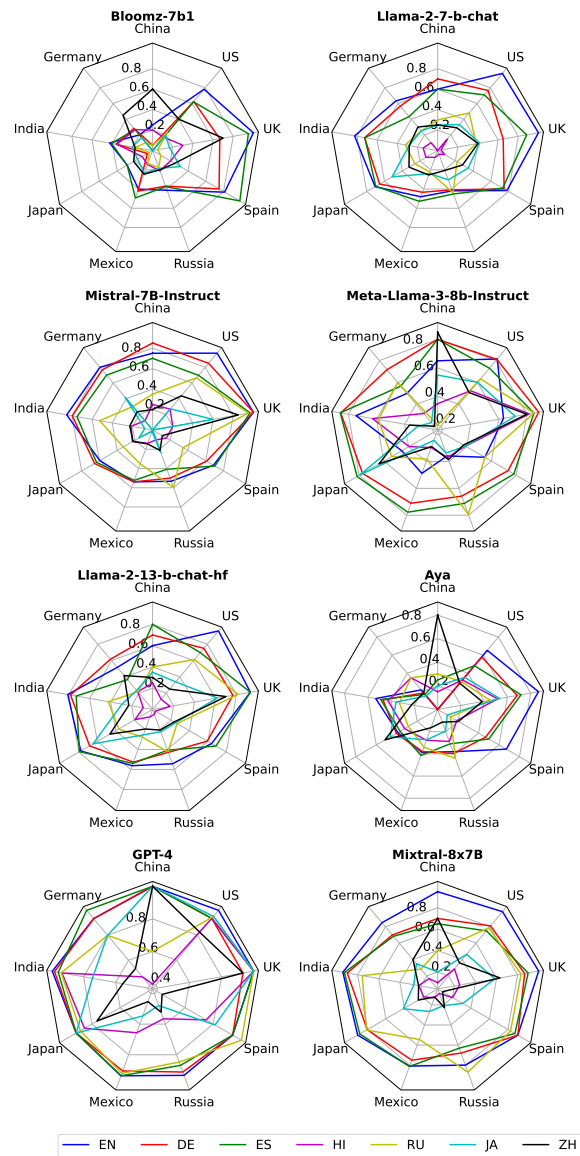


Figure 8: Accuracy of models in different languages (indicated in different colours) for questions on the location of national parks of different nations.

Scoring Criteria: Each translation is rated on a scale from 1 to 5:

- **5 (Excellent)** – No grammatical errors. The sentence follows correct conjugations and agreements. It is natural, idiomatic, and fluent, resembling native speech.
- **4 (Good)** – Minor grammatical mistakes (e.g., incorrect prepositions, small conjugation errors), but they do not affect understanding. The sentence is still readable and sounds natural.
- **3 (Acceptable)** – Noticeable grammar mistakes, such as verb tense inconsistencies or

incorrect word order. The sentence is understandable but slightly unnatural.

- **2 (Poor)** – Multiple grammatical mistakes significantly affecting readability. The sentence sounds unnatural or contains awkward phrasing.
- **1 (Unacceptable)** – The sentence is not grammatically correct and is difficult or impossible to understand. Severe errors break fluency, making the translation unusable.

C Evaluating the Variation of Model Performance Across Languages

To study the disparity in a model’s performance between languages on facts about the same country, we analyze the standard deviation of the accuracy of the models across different languages. A lower standard deviation indicates better homogeneous performance across languages. However, we observe high variance (c.f. Table 8), indicating that multilingual LLMs struggle to transfer their cultural literacy across languages. In many instances, the standard deviation exceeds 20, indicating a severe disparity in performance transference across languages.

D Evaluation of Qwen3 and LLaMA-3.1 Models

Table 9 presents the performance of Meta-Llama-3.1-8B-Instruct (Meta AI, 2024), Qwen3-8B (Yang et al., 2025), and Qwen3-14B (Yang et al., 2025). Similar to other models, we observe that they achieve stronger results on Western languages. Moreover, while Meta-Llama-3.1-8B-Instruct performs best among them, its performance still lags behind GPT-4.

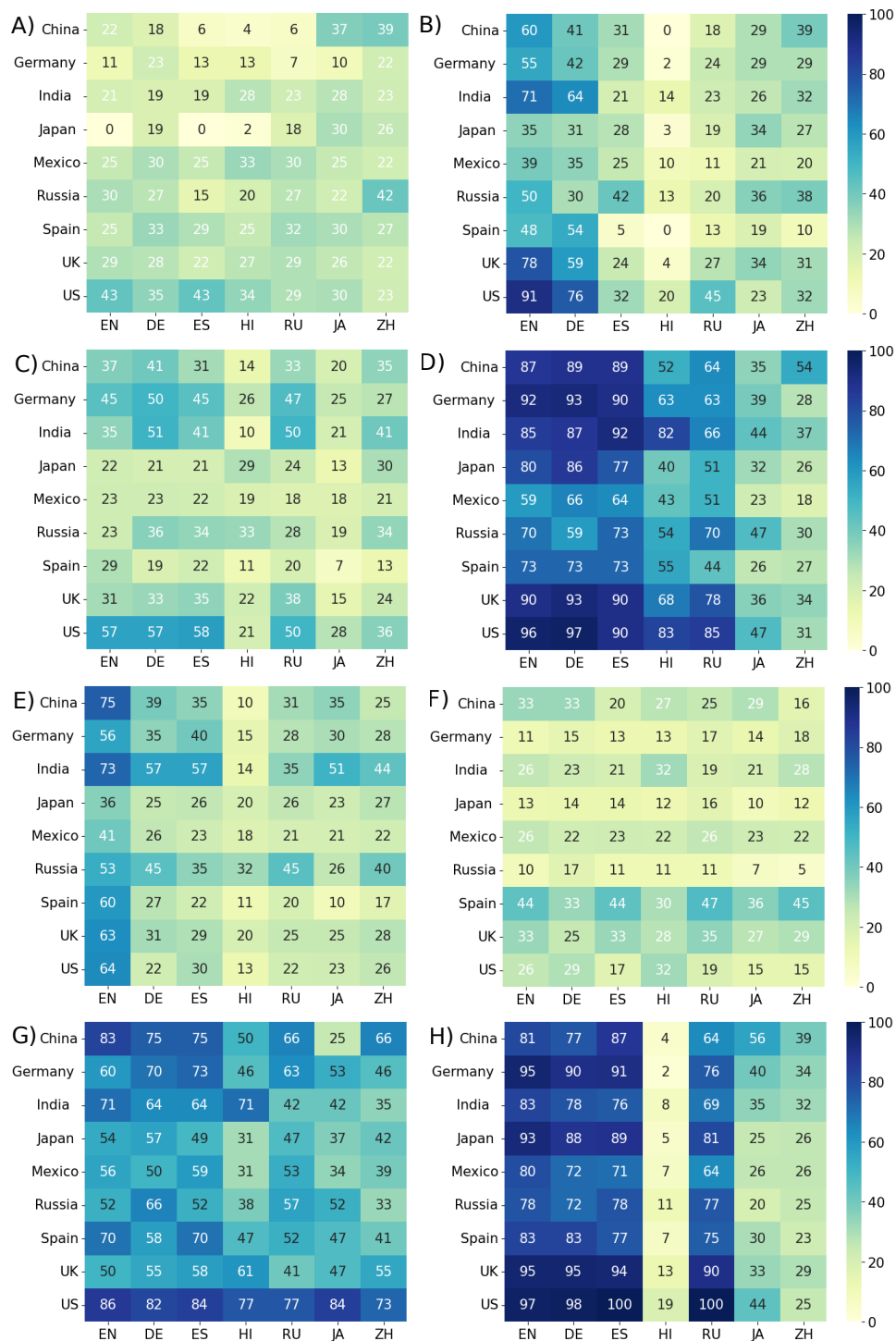


Figure 9: Heatmap of accuracy on leader domain, for – (A) Bloomz-7B1, (B) LLaMA-2-7B-Chat, (C) Mistral-7B-Instruct, (D) Meta-LLaMA-3-8B-Instruct, (E) LLaMA-2-13B-Chat, (F) 13-billion Aya, (G) GPT-4 and (H) Mixtral-8x7B.

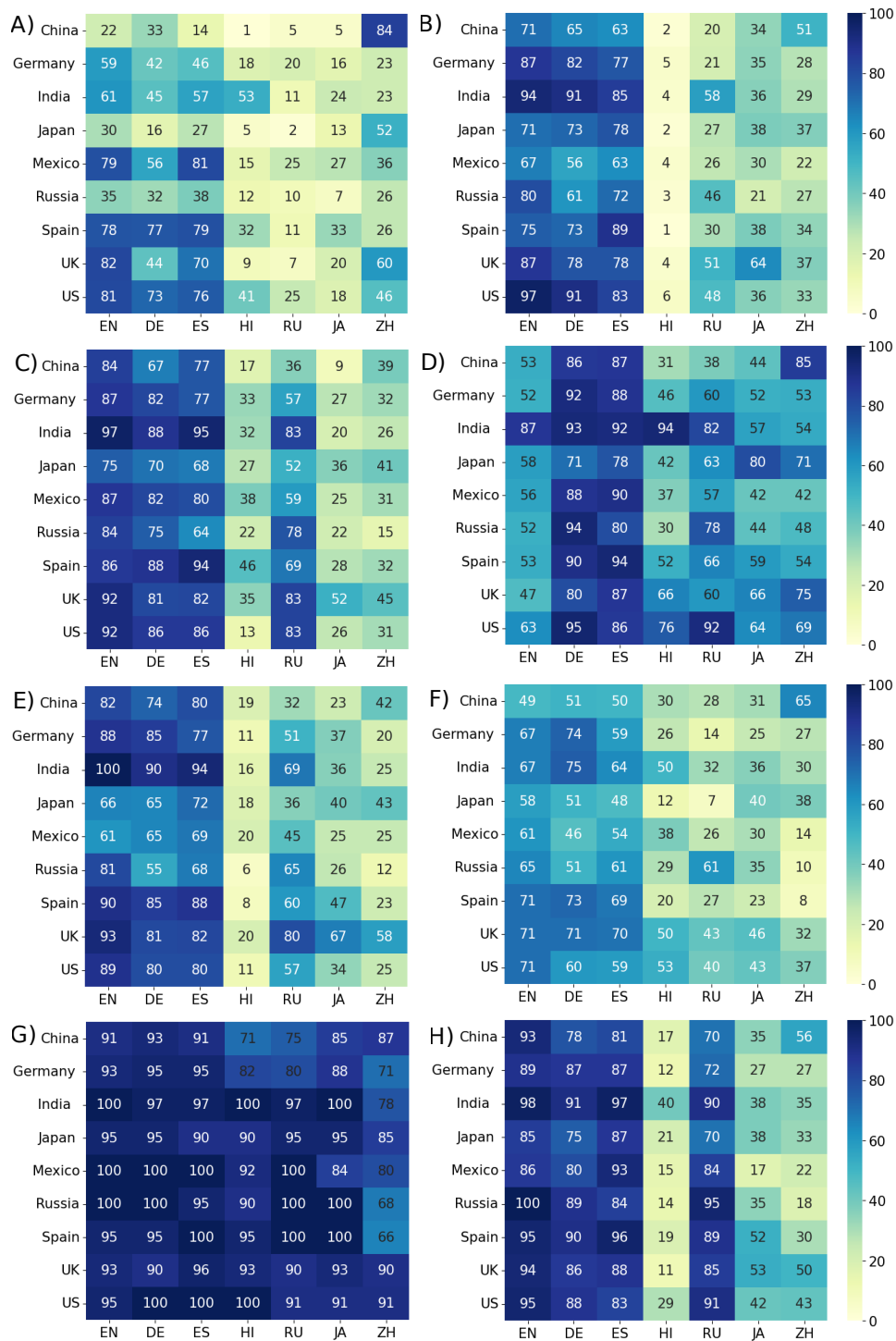


Figure 10: Heatmap of accuracy on monuments domain, for – (A) Bloomz-7B1, (B) LLaMA-2-7B-Chat, (C) Mistral-7B-Instruct, (D) Meta-LLaMA-3-8B-Instruct, (E) LLaMA-2-13B-Chat, (F) 13-billion Aya, (G) GPT-4 and (H) Mixtral-8x7B.

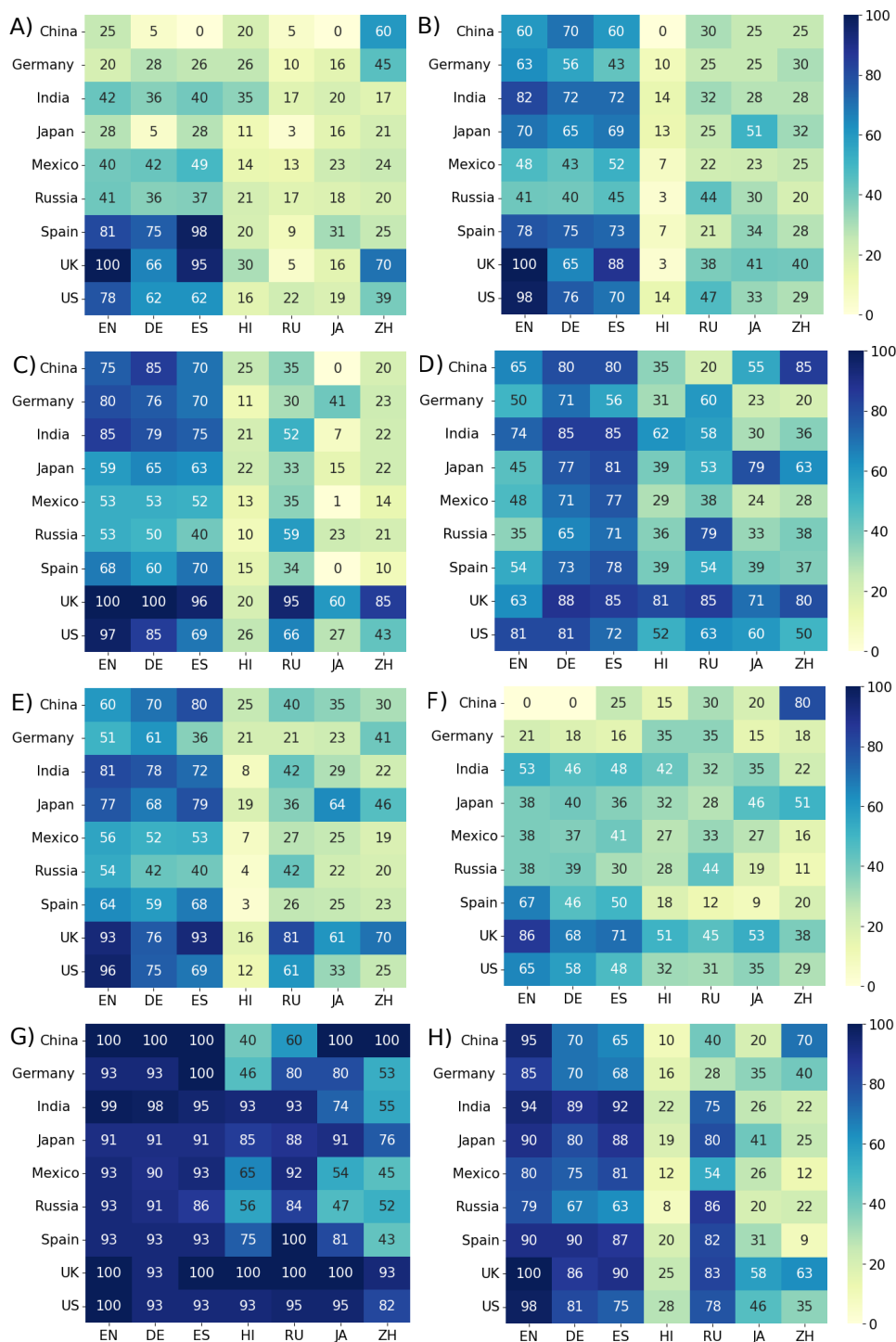


Figure 11: Heatmap of accuracy on national park domain, for – (A) Bloomz-7B1, (B) LLaMA-2-7B-Chat, (C) Mistral-7B-Instruct, (D) Meta-LLaMA-3-8B-Instruct, (E) LLaMA-2-13B-Chat, (F) 13-billion Aya, (G) GPT-4 and (H) Mixtral-8x7B.

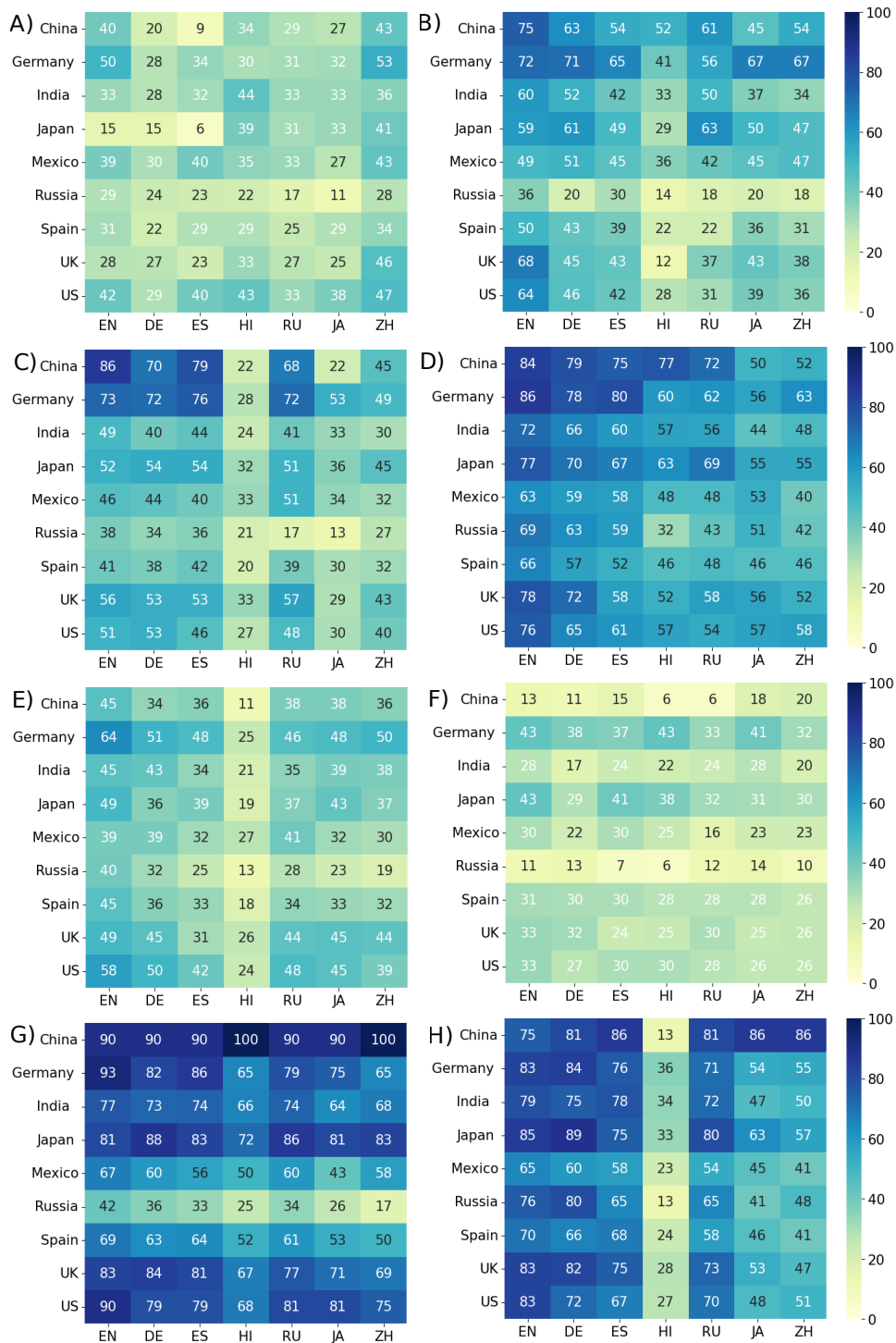


Figure 12: Heatmap of accuracy on war domain, for – (A) Bloomz-7B1, (B) LLaMA-2-7B-Chat, (C) Mistral-7B-Instruct, (D) Meta-LLaMA-3-8B-Instruct, (E) LLaMA-2-13B-Chat, (F) 13-billion Aya, (G) GPT-4 and (H) Mixtral-8x7B.

Country Model	US	China	India	UK	Japan	Germany	Russia	Mexico	Spain
Monuments									
Bloomz-7b1	51.93 ± 23.55	24.05 ± 26.89	39.80 ± 18.07	42.08 ± 28.19	21.07 ± 15.99	32.30 ± 15.57	23.38 ± 11.96	46.29 ± 24.76	48.47 ± 26.86
LLaMA-2-7B-chat	56.85 ± 32.05	44.17 ± 24.04	57.06 ± 32.36	57.70 ± 26.98	47.14 ± 26.27	48.25 ± 31.01	44.81 ± 26.47	38.87 ± 21.98	48.98 ± 28.92
Mistral-7B-Instruct	59.97 ± 31.97	47.52 ± 27.06	63.69 ± 32.38	67.63 ± 21.04	53.04 ± 17.28	56.98 ± 24.03	51.95 ± 27.87	58.10 ± 24.20	63.61 ± 25.67
Meta-LLaMA-3-8B-Instruct	78.42 ± 12.33	61.08 ± 22.62	80.36 ± 15.75	69.20 ± 12.21	66.61 ± 12.08	57.30 ± 24.21	61.36 ± 21.45	59.34 ± 20.28	67.35 ± 16.40
LLaMA-2-13B-chat	54.02 ± 28.50	50.80 ± 25.36	61.65 ± 32.55	69.31 ± 22.56	48.93 ± 18.09	53.33 ± 29.10	45.29 ± 27.55	44.64 ± 19.47	57.82 ± 30.42
Aya	52.38 ± 11.52	43.73 ± 12.96	51.11 ± 16.86	55.13 ± 14.45	36.79 ± 18.11	42.22 ± 22.37	44.97 ± 19.10	39.01 ± 15.25	42.01 ± 26.08
GPT-4	95.83 ± 3.86	85.42 ± 8.05	95.92 ± 7.17	92.86 ± 2.19	92.14 ± 3.64	79.68 ± 21.43	93.51 ± 10.83	93.96 ± 7.65	93.20 ± 11.05
Mixtral-8x7B	67.86 ± 25.99	62.03 ± 25.21	70.24 ± 27.99	67.30 ± 27.96	58.75 ± 24.89	57.94 ± 31.36	62.50 ± 35.23	57.14 ± 33.89	67.69 ± 30.52
Leader									
Bloomz-7b1	34.37 ± 6.99	19.35 ± 13.76	23.47 ± 3.50	26.59 ± 2.77	13.93 ± 11.87	14.64 ± 5.54	26.70 ± 8.15	27.57 ± 3.50	29.20 ± 3.19
LLaMA-2-7B-chat	46.11 ± 25.30	31.55 ± 17.59	36.22 ± 20.71	37.20 ± 22.42	25.88 ± 10.29	30.24 ± 14.99	33.16 ± 11.90	23.49 ± 10.17	21.64 ± 19.71
Mistral-7B-Instruct	44.44 ± 14.23	30.65 ± 8.89	35.97 ± 13.88	28.77 ± 7.77	23.42 ± 5.12	38.33 ± 10.22	30.10 ± 6.09	21.15 ± 1.99	17.65 ± 6.85
Meta-LLaMA-3-8B-Instruct	76.03 ± 24.05	67.56 ± 20.10	70.92 ± 20.47	70.44 ± 23.54	56.62 ± 22.94	67.26 ± 24.52	58.16 ± 14.15	46.82 ± 17.78	53.57 ± 19.57
LLaMA-2-13B-chat	29.05 ± 15.14	36.01 ± 18.25	47.70 ± 17.38	32.14 ± 13.35	26.64 ± 4.65	33.57 ± 11.67	39.80 ± 8.53	25.00 ± 7.01	24.37 ± 15.70
Aya	22.30 ± 6.50	26.49 ± 5.76	24.74 ± 4.21	30.36 ± 3.40	13.41 ± 1.89	14.76 ± 2.21	11.05 ± 3.58	23.94 ± 1.82	40.34 ± 5.93
GPT-4	80.95 ± 4.42	63.10 ± 18.29	56.12 ± 13.99	52.78 ± 6.30	45.67 ± 8.52	59.05 ± 9.87	50.34 ± 10.45	46.21 ± 10.36	55.46 ± 10.83
Mixtral-8x7B	69.37 ± 34.96	58.63 ± 26.87	55.10 ± 26.99	64.58 ± 34.36	58.67 ± 34.88	61.43 ± 33.31	52.04 ± 28.85	50.00 ± 26.65	54.62 ± 30.31
Wars									
Bloomz-7b1	39.58 ± 5.78	29.22 ± 10.96	34.69 ± 4.55	30.47 ± 7.25	26.33 ± 12.41	37.32 ± 9.33	22.45 ± 5.84	35.87 ± 5.36	28.91 ± 3.66
LLaMA-2-7B-chat	41.31 ± 10.99	58.12 ± 8.83	44.50 ± 9.50	41.22 ± 15.10	51.66 ± 10.83	63.05 ± 10.14	22.85 ± 7.31	45.50 ± 4.39	35.18 ± 9.75
Mistral-7B-Instruct	42.42 ± 9.40	56.49 ± 24.37	37.89 ± 7.93	46.68 ± 10.50	46.84 ± 8.23	60.84 ± 16.54	27.10 ± 9.14	40.53 ± 6.70	34.99 ± 7.03
Meta-LLaMA-3-8B-Instruct	61.78 ± 6.96	70.13 ± 12.47	58.13 ± 8.80	61.30 ± 9.38	65.78 ± 7.56	69.83 ± 10.63	51.93 ± 12.07	53.34 ± 7.37	52.13 ± 6.86
LLaMA-2-13B-chat	44.16 ± 9.90	34.42 ± 9.98	37.11 ± 7.29	41.03 ± 7.96	37.46 ± 8.56	47.78 ± 10.86	25.96 ± 8.24	34.86 ± 4.88	33.25 ± 7.40
Aya	28.81 ± 2.35	13.31 ± 4.92	23.81 ± 3.58	28.43 ± 3.45	35.38 ± 5.23	38.79 ± 4.07	11.17 ± 2.72	24.53 ± 4.54	29.26 ± 1.53
GPT-4	79.54 ± 6.02	93.51 ± 4.11	71.43 ± 4.41	76.53 ± 6.61	82.39 ± 4.79	78.33 ± 9.53	31.07 ± 7.77	56.83 ± 7.30	59.33 ± 6.72
Mixtral-8x7B	60.23 ± 17.63	73.05 ± 24.55	62.64 ± 16.80	63.45 ± 19.24	69.44 ± 18.05	66.01 ± 16.65	55.90 ± 21.42	50.08 ± 13.36	53.80 ± 15.92
National Parks									
Bloomz-7b1	43.15 ± 22.78	16.43 ± 19.95	29.96 ± 10.43	54.76 ± 35.01	16.70 ± 9.41	24.76 ± 10.25	27.56 ± 9.69	29.69 ± 13.11	48.66 ± 32.61
LLaMA-2-7B-chat	52.94 ± 27.82	38.57 ± 23.41	47.12 ± 25.45	53.81 ± 30.65	46.95 ± 21.53	36.19 ± 17.65	32.38 ± 14.22	31.81 ± 15.33	45.54 ± 27.04
Mistral-7B-Instruct	59.50 ± 25.86	44.29 ± 29.93	49.14 ± 29.51	79.52 ± 27.58	40.23 ± 20.14	47.62 ± 25.69	36.96 ± 17.38	32.14 ± 20.54	37.28 ± 27.26
Meta-LLaMA-3-8B-Instruct	66.01 ± 11.78	60.00 ± 22.99	62.09 ± 20.36	79.29 ± 8.16	63.03 ± 15.82	36.90 ± 21.67	51.63 ± 18.15	45.54 ± 19.77	53.79 ± 15.49
LLaMA-2-13B-chat	53.40 ± 28.04	48.57 ± 19.77	47.92 ± 27.22	70.48 ± 24.44	56.09 ± 20.98	36.90 ± 14.65	32.61 ± 15.85	34.71 ± 18.00	38.62 ± 23.32
Aya	42.86 ± 13.39	24.29 ± 25.13	40.44 ± 9.90	59.29 ± 15.68	39.29 ± 7.23	22.86 ± 7.90	30.36 ± 10.76	31.92 ± 8.06	32.14 ± 20.66
GPT-4	93.32 ± 4.98	85.71 ± 23.21	87.30 ± 15.09	98.10 ± 3.01	87.82 ± 5.08	66.67 ± 33.24	73.29 ± 18.62	76.56 ± 19.30	83.04 ± 17.90
Mixtral-8x7B	63.42 ± 24.37	52.86 ± 28.27	60.48 ± 32.37	72.38 ± 23.67	60.71 ± 28.82	49.05 ± 23.48	50.00 ± 29.42	49.00 ± 29.10	58.93 ± 34.04

Table 8: Accuracy and standard deviation of accuracy across the seven languages covered by XNationQA on the countries covered. All models show high variation in accuracy over languages. Results with standard deviation higher than ten are marked with red color.

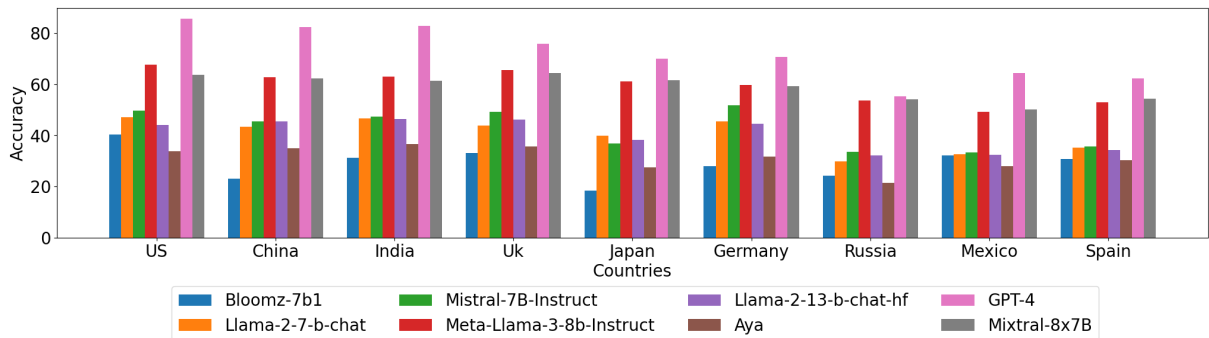


Figure 13: Accuracy of models across nations averaged over language. We note that models like GPT-4 demonstrate relatively low performance for Mexico, Spain, and Russia as compared to India and China even though the native languages of these nations (Hindi and Chinese) are poorly performing (c.f. Table 2).

Lang Model	EN	DE	ES	HI	RU	JA	ZH	AVG	AVG _W	AVG _{NW}
Monuments										
Meta-LLaMA-3.1-8B-Instruct	37.01	73.31	78.47	51.25	68.95	53.74	58.81	60.22	64.44	54.59
Qwen3-8B	21.44	59.96	55.87	17.70	35.05	41.99	26.96	37.00	43.08	28.89
Qwen3-14B	57.92	61.21	77.14	53.65	46.17	29.89	42.62	52.66	60.61	42.06
Leaders										
Meta-LLaMA-3.1-8B-Instruct	72.75	76.50	76.58	49.75	58.83	50.00	31.50	59.42	71.17	43.75
Qwen3-8B	12.00	40.17	48.33	30.25	43.25	15.33	14.08	29.06	35.94	19.88
Qwen3-14B	63.75	59.58	59.08	31.67	50.25	3.75	11.67	39.96	58.17	15.68
Wars										
Meta-LLaMA-3.1-8B-Instruct	66.47	64.84	62.68	55.63	60.80	58.07	54.04	60.36	63.70	55.90
Qwen3-8B	40.93	54.75	57.29	46.88	55.45	47.27	41.25	49.12	52.11	45.13
Qwen3-14B	67.85	63.42	63.74	48.97	59.74	48.09	41.57	56.20	63.69	46.21
National Parks										
Meta-LLaMA-3.1-8B-Instruct	39.48	57.45	71.19	46.62	58.77	41.70	40.38	50.80	56.73	42.89
Qwen3-8B	26.69	57.51	54.12	20.56	26.69	27.96	19.56	33.30	41.25	22.70
Qwen3-14B	72.04	57.03	77.54	44.82	50.26	21.72	30.50	50.56	64.22	32.34

Table 9: Accuracy of Meta-LLaMA-3.1 and Qwen3 models across languages and domains. Columns AVG_W and AVG_{NW} show average performance over Western (EN, DE, ES, RU) and non-Western (HI, JA, ZH) languages respectively. Meta-LLaMA-3.1-8B-Instruct consistently outperforms Qwen3 models across most domains and languages, except in some cases where Qwen3-14B shows higher scores in EN and ES.