Grounding Multilingual Multimodal LLMs With Cultural Knowledge

Jean de Dieu Nyandwi **Yueqi Song** **Simran Khanuja** **Graham Neubig**
{jeandedi, yueqis, skhanuja, gneubig}@andrew.cmu.edu
Carnegie Mellon University

https://neulab.github.io/CulturalGround/

Abstract

Multimodal Large Language Models excel in high-resource settings, but often misinterpret long-tail cultural entities and underperform in low-resource languages. To address this gap, we propose a data-centric approach that directly grounds MLLMs in cultural knowledge. Leveraging a large scale knowledge graph from Wikidata, we collect images that represent culturally significant entities, and generate synthetic multilingual visual question answering data. The resulting dataset, CulturalGround, comprises 22 million high-quality, culturally-rich VQA pairs spanning 42 countries and 39 languages. We train an open-source MLLM CulturalPangea on CulturalGround, interleaving standard multilingual instruction-tuning data to preserve general abilities. CulturalPangea achieves state-of-the-art performance among open models on various culture-focused multilingual multimodal benchmarks, outperforming prior models by an average of **+5.0%** without degrading results on mainstream vision–language tasks. Our findings show that our targeted, culturally grounded approach could substantially narrow the cultural gap in MLLMs and offer a practical path towards globally inclusive multimodal systems.

1 Introduction

Despite being trained on billions of image–text pairs, today's multimodal large language models (MLLMs) remain biased towards English and Western-centric training data (Ramaswamy et al., 2023; Vayani et al., 2024a; Yue et al., 2024; Liu et al., 2025). As a result, MLLMs often excel on high-resource languages, but even state-of-the-art models overlook or misinterpret non-Western cultural cues, especially long-tail entities (Liu et al., 2021b; Blasi et al., 2022; Ahia et al., 2023; AlKhamissi et al., 2024; Romero et al., 2024; Ananthram et al., 2024). Simply translating English data or increasing the size of the training corpora does not solve this problem. Translated data remain "Anglo-centric", and naively scaling up training corpora will not change the underlying biased distribution of the data (Yu et al., 2022; Tao et al., 2024; Gallegos et al., 2024).

Recent efforts emphasize the need for targeted, multicultural data curation to bridge this gap (Yue et al., 2024; Liu et al., 2025). MLLMs can only learn the knowledge they perceive: imbuing models with cultural understanding thus requires training data that explicitly incorporates diverse cultural entities, images, and linguistic contexts (Hershcovich et al., 2022; Li et al., 2024; Cahyawijaya et al., 2025). However, for long-tail entities, these data remain few and far between.

In this paper, we propose a novel approach to ground multilingual MLLMs with cultural knowledge from large-scale knowledge bases. We introduce a scalable pipeline for constructing culturally grounded multilingual multimodal data, curating rich visio-linguistic training data centered on regional cultural entities.

Our pipeline, as shown in Figure 1, consists of the following steps: **1) Cultural Concept Selection:** we first select culturally significant concepts from the large-scale knowledge resource Wikidata[1], leveraging its structured, multilingual knowledge graph (Vrandečić and Krötzsch, 2014); **2) Image Collection:** for each selected entity, we retrieve images from both Wikidata and Wikimedia Commons[2], increasing coverage and diversity beyond any single source; **3) Multilingual Factual VQA Generation:** utilizing a set of structured, language-specific templates, we generate multilingual factual Visual Question Answering (VQA) data based on each entity's Wikidata properties (e.g., *occupation*, *religion*) in 39 languages; **4) VQA Refine-**

[1] https://www.wikidata.org/
[2] https://commons.wikimedia.org/

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24199–24243
November 4-9, 2025

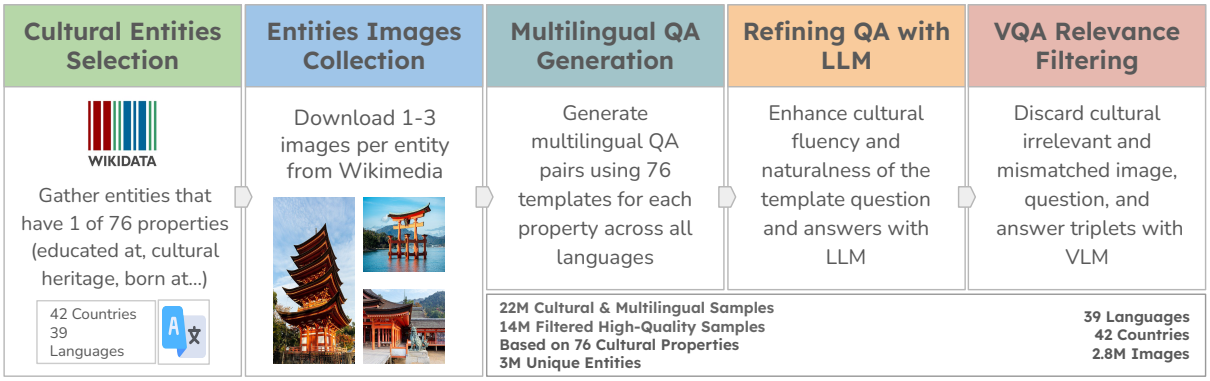

Figure 1: Our data curation pipeline involves gathering culturally relevant entities from the Wikidata knowledge base, creating several questions and answers about each entity, rephrasing them using an LLM, and filtering low-quality samples using a VLM.

ment: We refine these template-based VQAs with an LLM, prompting it to improve fluency, contextual richness, and cultural naturalness, without leaking the entity identity in the question; **5) Filtering:** To guarantee accurate cultural knowledge grounding, we use MLLMs to discard mismatched or culturally irrelevant VQA instances. Our pipeline results in CulturalGround, a high-quality multilingual multimodal dataset comprising 22 million VQA examples, explicitly curated to reflect diversity in cultural entities.

To evaluate the effectiveness of CulturalGround in imbuing models with cultural knowledge, we conducted experiments to train a multilingual MLLM CulturalPangea on CulturalGround. In experiments, CulturalPangea outperforms previous open-source MLLMs on multiple culture-focused vision–language benchmarks on average of **+5.0** across all benchmarks. On evaluation sets such as CVQA (Romero et al., 2024) and ALM-Bench (Vayani et al., 2024a), which specifically probe cross-cultural and multilingual understanding, CulturalPangea achieves state-of-the-art results among open models. Crucially, these gains in cultural competency do not come at the expense of general capability: our model retains competitive performance on standard vision–language tasks. These findings demonstrate that a data-centric strategy, carefully curating multilingual, culturally rich examples can substantially bridge the cultural gap in MLLMs.

2 Dataset

In this section, we describe our proposed data generation pipeline from Figure 1 in detail.

2.1 Problem Definition

Our goal is to construct a culturally grounded multimodal dataset suitable for training multilingual vision-language models. The core challenge is ensuring both cultural relevance and factual accuracy across diverse regions and languages.

Inputs. Our construction process takes four primary inputs:

- A structured knowledge base $\mathcal{K}$ (specifically Wikidata) containing entities E and their factual relationships
- A set of regions for which we would like to create data $R = \{r_1, \ldots, r_m\}$, each identified by unique IDs within $\mathcal{K}$
- A collection of target languages $L = \{l_1, \ldots, l_k\}$ representing the linguistic communities we would like to cover.
- A set of culturally meaningful properties $P = \{p_1, \ldots, p_n\}$ (e.g., *place of birth*, *occupation*) that indicate cultural relevance.

Output. We produce a dataset D consisting of triplets $(i_e, q_{e,p}^{(l)}, a_{e,p}^{(l)})$ where each triplet contains:

- i_e: An image depicting entity e
- $q_{e,p}^{(l)}$: A natural language question about property p of entity e in language l
- $a_{e,p}^{(l)}$: The corresponding factual answer in the same language l

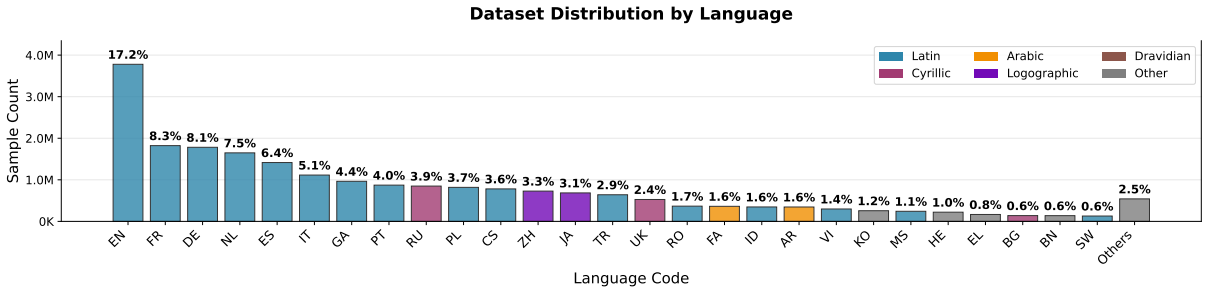

Figure 2: The distribution of samples across languages highlighting percentages of each language in CulturalGround.

2.2 Cultural Entity Selection

The first step toward constructing D is to identify a culturally representative subset of entities $E' \subseteq E$ from the knowledge base $\mathcal{K}$. For every target region $r \in R$ we map r to its unique Wikidata identifier (QID).[3]

$$E' = \{ e \in E \mid \exists r \in R,\, p \in P \text{ with } (e, p, r) \in \mathcal{K} \}$$

To guarantee multilingual coverage, we further require each $e \in E'$ to have a label or description in at least one language $l \in L$. This intersectional filtering step results in a culturally rich and linguistically diverse set of entities spanning people, places, foods, artifacts, and institutions (full statistics can be seen in Figures 20, 21, and 22 in the appendix).

2.3 Entity Images Collection

For every $e \in E'$ we assemble a set of images $\mathcal{I}_e = \{i_{e,1}, \ldots, i_{e,t_e}\}$ that will later appear in triplets $(i_e, q_{e,p}^{(l)}, a_{e,p}^{(l)})$. If $\mathcal{K}$ links an image via property P18, we include that as $i_{e,1}$.

To broaden visual coverage, we also obtain additional images from the Wikimedia Commons category page when available (Srinivasan et al., 2021). Commons categories often contain dozens of photographs or illustrations related to a given entity (for example, a notable person's category might include portraits from different events, while a landmark's category might contain images from various angles or time periods). This step enriches CulturalPangea 's training data with visual variability: the same cultural entity might appear in different settings or styles across its image set.

[3]A QID is a unique identifier beginning with "Q" followed by digits, e.g. https://www.wikidata.org/wiki/Q43649390.

2.4 Multilingual Question-Answer Generation

We generate factual QA pairs across languages by instantiating language-specific templates with knowledge-base facts. For every culturally selected entity $e \in E'$, property $p \in P$[4] with value v such that $(e, p, v) \in \mathcal{K}$, and target language $l \in L$ for which e is labelled in l, we create a question–answer pair $(q_{e,p}^{(l)}, a_{e,p}^{(l)})$.

Property-level QA. Let $T_p^{(l)}$ be the question template for property p in language l, and let $\mathrm{name}_l(\cdot)$ and $\mathrm{desc}_l(\cdot)$ denote the language-l label and (optional) short description returned by $\mathcal{K}$. We instantiate:

$$q_{e,p}^{(l)} = T_p^{(l)} \text{ (relevant reference to } e)$$

$$a_{e,p}^{(l)} = A_p^{(l)}(\mathrm{name}_l(e), v)$$

where $A_p^{(l)}$ is the answer template for property p in language l and v is the property's value.

Example (English): p = *place of birth* $\Rightarrow$ q: "Where was this person born?"; a: "{entity_name} was born in {property_value}." $\rightarrow$ "Albert Einstein was born in Ulm, Germany."

Entity-level QA. To capture identity and brief context, we include one entity-level pair per e using template $T_{\mathrm{id}}^{(l)}$:

$$q_{e,\mathrm{id}}^{(l)} = T_{\mathrm{id}}^{(l)} \text{ (relevant reference)}$$

$$a_{e,\mathrm{id}}^{(l)} = A_{\mathrm{id}}^{(l)}\,(\mathrm{name}_l(e), \mathrm{desc}_l(e))$$

.

Example (English): q: "What is the entity shown in the image?"; a: "{entity_name}, {entity_description}." $\rightarrow$ "The Taj Mahal, a 17th-century mausoleum in India."

[4]To avoid over-representation of popular entities with many Wikidata properties, we cap the number of properties used per entity at the country-specific median.

Each QA is initially generated in "fill-in-the-blank" style (Jiang et al., 2020) by inserting the entity's Wikidata facts into the QA templates. Together, the property-level questions and the general entity-level questions form a diverse initial factual QA set for each entity, covering both specific facts and broader contextual information. We then match the corresponding images of the entity to the questions generated for this entity, creating multiple candidate (image, question, answer) triplets $(i_{e,j}, q^{(l)}_{e,p}, a^{(l)}_{e,p})$ per entity. These templated pairs ensure wide coverage of factual attributes, but often lack linguistic variation and contextual nuance, motivating a subsequent refinement step.

2.5 Refining Multilingual VQA Data for Cultural Fluency

While the template-based QA generation ensures factual correctness, the resulting text is often formulaic or grammatically awkward, lacking the rich context of a human-written question or answer. We therefore employ a large language model (LLM) to rewrite and polish each question–answer pair $\big(q^{(l)}_{e,p}, a^{(l)}_{e,p}\big)$ for greater fluency and cultural naturalness.

$$(q'^{(l)}_{e,p},\ a'^{(l)}_{e,p}) \ = \ \mathsf{Refine}^{(l)}\big(q^{(l)}_{e,p},\ a^{(l)}_{e,p}\big).$$

In particular, we prompt strong open-source models (such as Qwen2.5-72B (Qwen et al., 2025) and Gemma3-27B (Gemma Team et al., 2025)) with the templated pair and specific instructions on how to improve it. The LLM is instructed to avoid any direct mention of the entity's name or identity in the question (ensuring that the refined question $q'^{(l)}_{e,p}$ remains answerable only via the image, with no textual leakage of the entity name) and to incorporate subtle contextual cues. For example, the model might say "in your country" or "this individual in India" rather than explicitly naming a person, or describe a shrine by its religion and notable features instead of giving its proper name.

We further direct the LLM to enrich each refined answer $a'^{(l)}_{e,p}$ with culturally relevant details while preserving the core factual content, often by adding a brief background or significance to the fact. For instance, a short answer like "He is the Prime Minister of India." can be expanded to "He is the 14th Prime Minister of India, serving as the country's highest political leader." Likewise, a generic question such as "Which public office does this person hold?" might be rewritten as "What is the highest political office that this individual currently holds in India?", adding clarity and specificity.

This step significantly improved the coherence and readability of the QA pairs, confirming that the refinement yields questions and answers that are more natural and informative yet still anchored in the original cultural context (see Figure 20) for examples).

2.6 Image-Text Relevance Filtering

After assembling the images with the rewritten QA pairs, we apply a final filtering step to ensure that each image is *meaningfully* relevant to its paired question and answer. Not every image retrieved from Wikimedia Commons precisely depicts the intended entity, and some (re)written QAs may not align with specific images (especially when multiple images exist per entity).

Concretely, for a candidate triplet $(i, q'^{(l)}, a'^{(l)})$ in language l, the VLM issues a binary alignment judgment

$$y\big(i, q'^{(l)}, a'^{(l)}\big) \in \{\texttt{true}, \texttt{false}\},$$

indicating whether the image both (i) *visually grounds* the entity/scene implicated by the QA and (ii) *matches the cultural context* described in $q'^{(l)}$ and $a'^{(l)}$ (e.g., region, heritage, religious or architectural attributes). We retain the triplet if $y(\cdot) = \texttt{true}$; otherwise it is discarded. We use Qwen-2.5VL(32B, 72B)-Instruct and Gemma-3(12B, 27B)-IT models for filtering, selecting models based on their strength for specific regions and languages.

This selective filtering ensures that **CulturalPangea**'s training data maintains a high level of image–text relevance, improving supervision quality. As a result, the final dataset primarily contains triplets where the image truly depicts the entity in question or closely relates to the QA content, which is crucial for grounding the model's understanding in visual evidence.

2.7 Dataset Statistics and Languages

Following the above curation steps, we compiled a large multilingual multimodal dataset of QA-image pairs. The resulting dataset comprises about 22M million high-quality image-question-answer triplets spanning 39 languages and 42 regions. As shown in Figure 2, Figure 19, Figure 22, and Figure 23, the dataset statistics demonstrate our focus on cultural diversity and long-tail entities.

Training Data		Model		Training	
Total CulturalGround (OE)	22M	Vision Encoder	CLIP-ViT-14	Batch Size	128
Sampled	13M	LLM	Qwen2-Instruct	Learning Rate	5×10^{-6}
Total CulturalGround (MCQs)	8M	Base Model	Pangea-7B	Epoch	1
Sampled	5M	Trainable Parts	Connector, LLM	GPU Hours (H100)	720

Table 1: CulturalPangea's training configurations.

3 Experimental Setup

3.1 Training

As a base model, we start from Pangea-7B (Yue et al., 2024), which couples a CLIP-based vision encoder (Radford et al., 2021) with a Qwen2-7B autoregressive language model (Bai et al., 2025). We keep the vision encoder frozen and fine-tune only the connector and the language model on our culturally grounded data. Key training details include:

- **Dataset**: We train on 13M open-ended VQA pairs sampled from CulturalGround 21M total samples, covering 39 culturally diverse languages and cultures. To improve model robustness, we also create 8M multiple-choices VQA samples grounded on collected cultural entities, and sample 5M to use in training. Multiple-choices question pipeline is expanded in Appendix B.3 and sampling details can be found in Appendix D.

- **Multilingual mixture**: To preserve the base model's broad multilingual grounding and avoid catastrophic forgetting, we interleave 5.8M samples of PangeaInstruct English and multilingual data with the CulturalPangea samples during fine-tuning. We also include a filtered 90K samples of M3LS (Verma et al., 2023) to improve entity recognition. The details for M3LS entity linking data is further discussed in Appendix E.

- **Optimization**: We train with a small learning rate of 5e-6. This choice is motivated by recent observations that large vision–language models benefit from lower learning rates for cross-lingual transfer (Steiner et al., 2024). We train with a cosine decay schedule and a brief warm-up, tuning all connector and LLM parameters. Further training settings are highlighted in Table 1. We also experiment with checkpoints merging and we discuss this in Section 4.3.

Together, these settings allow for CulturalPangea to better retain its pretrained abilities while adapting to the culturally enriched data.

3.2 Evaluation Protocol

We evaluate CulturalPangea's multilingual and cultural competence on benchmarks from PangeaBench and related tasks. In particular, we use the following multimodal benchmarks: CVQA (Romero et al., 2024), MaRVL (Liu et al., 2021a), XM100 (Thapliyal et al., 2022), ALM-Bench (Vayani et al., 2024b), MERLIN (Anonymous, 2024), MaXM (Changpinyo et al., 2024), M3Exam (Zhang et al., 2023). Table 4 shows an overview of the benchmarks we evaluate on.

For baselines, we compare our model against other recent state of the art multimodal LLMs such as Llava-Next-7B (Liu et al., 2024), Molmo-7B-D (Deitke et al., 2024), Llama3.2-11B (Grattafiori et al., 2024), and mBLIP (Geigle et al., 2023), PaliGemma-3B(Beyer et al., 2024), AyaVision-8B (Dash et al., 2025), and Pangea (Yue et al., 2024).

4 Results and Analysis

4.1 Cultural Understanding Benchmarks

In Table 2 we show performance on culturally focused multimodal benchmarks including CVQA, XM100, ALMBench, MaRVL, and MERLIN. We can see that CulturalPangea achieves state-of-the-art accuracy on nearly all of these datasets and various question types(Figure 12, substantially improving over the base Pangea-7B and other competitive open models. These results verify that culturally grounded training data significantly enhances culture-specific visual reasoning. Overall, CULTURALPANGEA substantially advances the state of the art in culturally informed multimodal understanding.

4.2 General Multilingual and English Performance

Despite our attempts to specifically improve accuracy on culturally relevant phenomena, Cultural-

	Cultural Understanding			Entity Recognition	Multilingual VQA		Captioning	Average
Models	CVQA	MARVL	ALM	MERLIN	MAXM	M3EXAM	XM100	ALL
Llava-Next-7B	40.9	50.9	42.4	34.1	21.4	28.4	15.5	33.4
Molmo-7B-D	58.7	54.9	49.1	42.9	37.5	39.1	6.0	41.2
Llama3.2-11B	**69.6**	58.1	56.6	49.1	43.9	36.6	5.8	45.7
PaliGemma-3B	42.5	52.2	35.7	13.1	19.9	25.6	0.6	27.1
mBLIP-mT0-XL	37.5	66.7	36.9	15.8	36.8	25.0	6.8	32.2
AyaVision-8B	50.8	64.5	55.1	55.3	52.1	41.7	10.0	47.1
Pangea-7B	56.9	78.7	59.9	66.0	53.3	42.0	29.7	55.3
CulturalPangea	59.1	**80.3**	**63.5**	**81.1**	**53.9**	**46.7**	**36.9**	**60.3**
Δ over Pangea	+2.2	+1.6	+3.6	+15.1	+0.6	+4.7	+7.2	+5.0

Table 2: Multilingual performance comparison across models on cultural understanding (CVQA, MARVL, ALM), entity recognition (Merlin), multilingual VQA (MAXM, M3Exam), and captioning (XM100). For CVQA, we evaluate on 31 country-language pairs, 38 languages in ALMBench and for other benchmarks, we use all languages available in the respective datasets. The best-performing model on each dataset is in **bold** and the second best is underlined.

Pangea maintains strong general multimodal and multilingual capabilities. On M3Exam for example, CulturalPangea ouperforms Pangea-7B by +4.7 points and matches performance on MaXM, while surpassing most open models. Furthermore, CulturalPangea outperforms Pangea-7B on XM100(multilingual captioning benchmark) by a large margin.

CulturalPangea maintains excellent English performance not only in cultural settings, but also in general tasks as shown in Figure 3. Balancing different cultures, languages, while maintaining English and general skills is challenging (Chuang et al., 2025; Pouget et al., 2024) and we attribute this to our data curation pipeline that gathers entities in many regions and different languages in each region.

4.3 Analysis and Discussion

Cross-Cultural and Cross-Lingual Transfer. CulturalPangea demonstrates cross-cultural and lingual transfer on languages that are not in CulturalGround. To analyze transfer behaviors across cultures and languages, we compare performance with baseline on 17 languages from ALMBench. As shown in Table 3, our model consistently shows improvements over Pangea-7B. This trend suggests that the model effectively transfers knowledge to languages with limited training data, alleviating the typical drop-off seen in low-resource settings.

We hypothesize that these cross-lingual improvements may be a result of CulturalPangea 's multilingual, culturally grounded supervision strategy. During training, most entities are accompanied by QA examples in multiple languages, which encourages the model to align semantic representations across languages. This design can enable knowledge learned from one language to be readily applied to other languages. As a result, CulturalPangea leverages training signals from diverse languages to significantly boost accuracy on underrepresented languages, explaining its across-the-board gains on ALMBench.

Cultural Data Scaling and General Skill Preservation. As additional culturally grounded data is introduced during training, CULTURALPANGEA's performance steadily improves on all culture-sensitive benchmarks (Figure 4), while its general multilingual vision–language proficiency is concurrently preserved and even enhanced, as shown in Figure 5. This outcome indicates that our interleaved training strategy successfully avoided catastrophic forgetting by continuously mixing standard multilingual examples into the cultural fine-tuning process. In essence, this approach parallels replay-based continual learning, wherein revisiting earlier tasks helps maintain broad competence (Kirkpatrick et al., 2017; Rolnick et al., 2019).

Consistent with this, we observed a characteristic "dip-and-recovery" trajectory in the model's general performance during training: an initial drop when new cultural data was first introduced, followed by a rebound that ultimately exceeded the original baseline. Specifically, on general multilingual benchmarks like **M3Exam** and **MaXM**, performance temporarily decreased by 2-3% in early training steps before recovering to surpass baseline scores by +4.7 and +0.6 points respectively. These dynamics mirror the "stability gap" phenomenon reported in incremental learning (De Lange et al., 2022; Caccia et al., 2021).

Model	SC	AS	EG	YO	GU	BH	LA	SI	SA	DA	GL	AF	IC	AZ	SH	SK	FI	Avg
Pangea-7B	28.3	40.5	63.6	21.5	35.1	49.7	19.2	37.0	64.4	59.9	65.3	58.8	45.3	51.1	26.2	38.4	41.7	45.0
CulturalPangea-7B	39.4	50.9	68.3	25.8	39.1	53.2	22.0	39.6	66.8	62.2	66.9	60.3	46.5	52.2	26.8	39.0	42.1	48.3
Gain	**+11.1**	**+10.4**	**+4.7**	**+4.3**	**+4.0**	**+3.5**	**+2.9**	**+2.6**	**+2.4**	**+2.3**	**+1.5**	**+1.5**	**+1.3**	**+1.1**	**+0.7**	**+0.6**	**+0.4**	**+3.3**

Table 3: Cross-Cultural/Lingual Performance on ALM-Bench. Language codes: SC=Scots Gaelic, AS=Assamese, EG=Egyptian Arabic, YO=Yoruba, GU=Gujarati, BH=Bhojpuri, LA=Lao, SI=Sindhi, SA=Saudi Arabic, DA=Danish, GL=Galician, AF=Afrikaans, IC=Icelandic, AZ=Azerbaijani, SH=Shona, SK=Sanskrit, FI=Filipino.

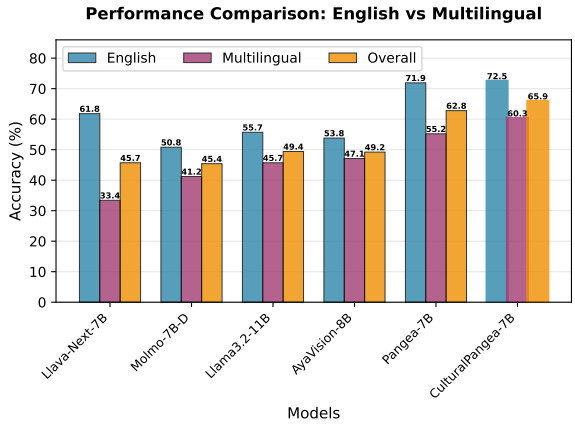

Figure 3: Overall performance comparison in english and multilingual

By the end of training, CulturalPangea not only acquires remarkably stronger culturally grounded capabilities (averaging +5.0% improvement across cultural benchmarks) but also achieves slightly higher overall VQA accuracy than the baseline, exemplifying a difficult-to-achieve equilibrium between new specialization and retained generalization, and highlighting the promise of CulturalGround and training approach.

Does CulturalGround Help Low-Resource Languages? CulturalPangea demonstrates substantial gains on ALMBench for languages with limited training data, with improvements most pronounced for resource-poor languages within CulturalGround. As shown in Figure 9, we observe absolute accuracy gains of 15.0% points on Sinhala, 10.9% on Hebrew, and 9.1% on Irish. Additional low-resource languages exhibit consistent improvements: Tamil (+6.3), Amharic (+5.3), Bengali (+4.8), and Telugu (+4.2). These gains occur without sacrificing performance elsewhere—nearly all languages improve, with only Norwegian (-0.5) showing negligible regressions. The largest improvements in traditionally underrepresented languages indicate that culturally-aware grounding effectively scales to the long tail of languages, enhancing multilingual inclusivity.

Which Cultural Domains Benefit Most? The model's gains vary substantially across cultural domains, with culturally rich categories showing the largest improvements. As shown in Figure 7 and Figure 10, **Heritage** achieves the highest relative improvement at 11.5% (from 64.4% to 71.8%), followed by **Media** at 10.6% and **Food** at 9.2%. Other culturally salient domains like **Architecture** (7.1%), **Economy** (7.0%), and **Music** (6.2%) also demonstrate substantial gains. These results indicate that domains requiring broad cultural knowledge and context benefit most from our approach. In contrast, generic visual domains show minimal improvement or slight regression. **Sketch** decreases by 0.6%, while **Meme** improves only marginally at 2.31%. Similarly, **Festivals** (0.73%) and **Religion** (2.08%) show limited gains. This pattern confirms that improvements concentrate in genuine cultural understanding areas, while abstract visual content or highly localized traditions remain challenging. The largest accuracy gains occur in well-represented cultural domains, reinforcing the value of targeting cultural knowledge in model training.

Performance Gains via Checkpoint Merging Following (Dash et al., 2025; Team et al., 2025; Li et al., 2025), we merge 5 strong CULTUREPANGEA checkpoints from different training stages using the **TIES** method, which recovers complementary model strengths often lost during continual training. Although linear and DARE-TIES variants show comparable results, TIES yields the highest average accuracy. As shown in Figure 8, using our strongest early checkpoint as the base outperforms using the original PANGEA-7B, evidence that the mixed-data regime had already mitigated catastrophic forgetting. The merged model improves mean accuracy by roughly +0.8 points over the best model, illustrating the value of checkpoint combination.

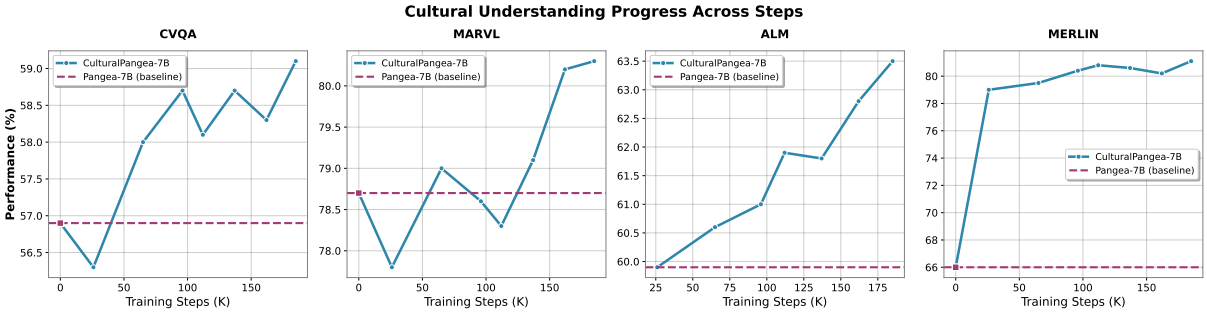

Figure 4: Training performance curves on four culture-centric benchmarks (CVQA, MaRVL, ALM-Bench, and MERLIN) show accuracy steadily rising as CulturalPangea is trained on more culturally grounded data versus the baseline model. Higher training step counts—*i.e.*, greater exposure to the CulturalGround consistently translate into improved accuracy.

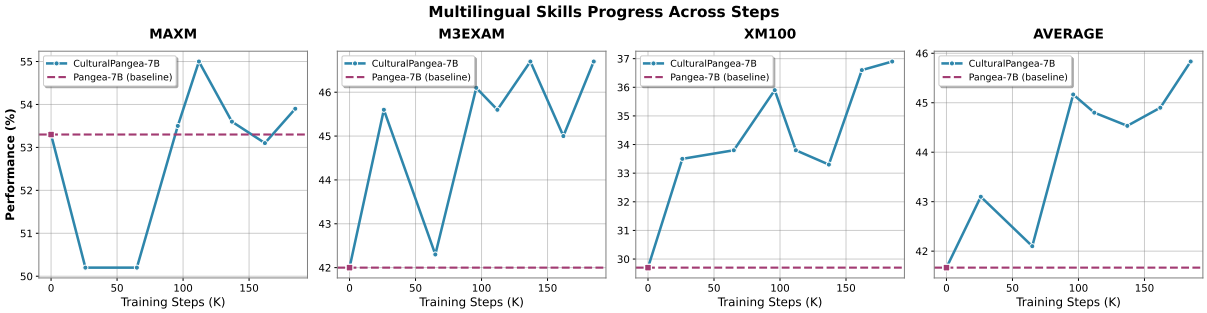

Figure 5: Training performance curves on three general multilingual benchmarks and their overall average show accuracy improvements as CulturalPangea is exposed on more data, compared to the Pangea-7B.

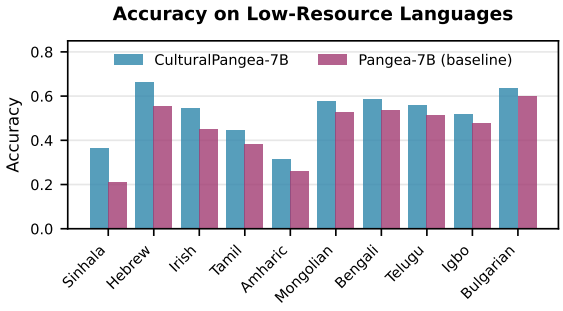

Figure 6: CulturalPangea achieves the largest gains on underrepresented languages, demonstrating effective scaling to the long tail of languages.

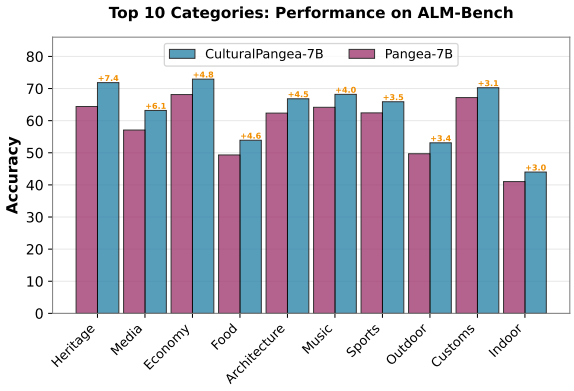

Figure 7: Cultural domains in ALMBench such as heritage and food show largest gains over general domains.

5 Related Work

5.1 Knowledge Bases for Cultural Grounding

Integrating structured knowledge bases into language models has proven effective for enriching factual and cultural understanding in text-only NLP. For example, works such as (Liu et al., 2020; Wang et al., 2021; Kim et al., 2022; Yang et al., 2024) have attempted to inject world facts in language model representations to provide domain-specific information via knowledge bases. These knowledge-enhanced models achieve better factual consistency and recall, but they operate purely in the textual domain.

Extending knowledge-base grounding to multimodal settings remains relatively rare. Prior work on knowledge-driven VQA has mostly used KBs in answering knowledge-intensive visual questions and focusing on general encyclopedic facts rather than culturally specific knowledge (Deng et al., 2025; Ding et al., 2022). In contrast, our work

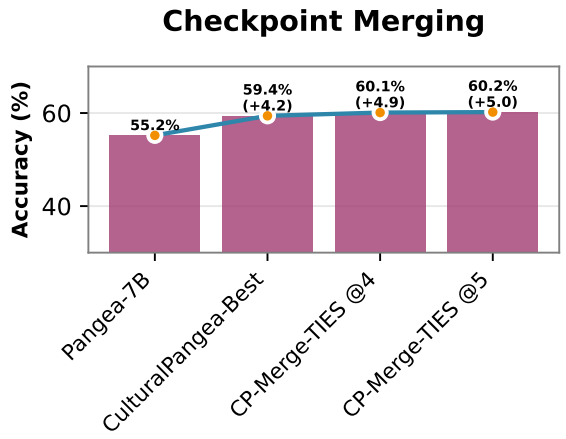

Figure 8: Accuracy improvements from merging checkpoints. CP stands for CulturalPangea.

uniquely leverages a structured knowledge graph to generate culturally diverse VQA pairs. To our knowledge, no previous study has used KBs to curate a multilingual multimodal QA dataset centered on cultural concepts, a gap our proposed pipeline addresses.

5.2 General-Purpose Multimodal LLMs

The last few years have seen rapid progress in general-purpose multimodal large language models. Open-source systems like LLaVA (Liu et al., 2023b) and its successors (Liu et al., 2024, 2023a) connect vision encoders with LLMs and are trained on visual instruction dataset, achieving impressive results on a wide range of visual tasks. More recent open models such as MoLMo (Deitke et al., 2024), Qwen2.5-VL (Bai et al., 2025), Phi-3 (Abdin et al., 2024), further close the gap to proprietary systems by streamlining multimodal training at scale and investing in high-quality datasets. However, it is important to note that none of these general models are explicitly optimized for cultural understanding. Their training data tend to be dominated by English and Western imagery which can lead to blind spots on region-specific content. In summary, general-purpose MLLMs provide a powerful foundation but still exhibit a cultural bias due to their data, leaving room for specialization in multicultural knowledge.

5.3 Multilingual and Culture-Aware Multimodal LLMs

To address these gaps, recent efforts have targeted multilingual and cultural understanding in multimodal models by using large-scale multilingual visual instruction datasets, translating existing datasets into a wide array of languages, and using synthetic multilingual instruction pipeline (Yue et al., 2024; Dash et al., 2025). Those works achieve strong performance and indicate a growing focus on true multilingual competence in VLMs.

However, multilingual capability alone does not guarantee cultural understanding (Pouget et al., 2024). A model trained on many languages may still suffer on culture-specific knowledge, especially if its data are translated or Western-centric. There is increasing recognition that targeted cultural data is needed beyond naive multilinguality. True cultural competence requires exposure to culture-specific content in both text and imagery. Our work follows this principle: rather than relying on pure translated captions, we curate QA pairs grounded in each culture's unique knowledge and visual context. This approach goes beyond prior multilingual setups by explicitly injecting structured cultural information into the multimodal training data, aiming to build models that are not only linguistically multilingual but also culturally knowledgeable.

6 Conclusion

We present a data-centric approach for mining cultural grounded multimodal data from public knowledge bases. CulturalPangea, a model trained on the resulting dataset demonstrates the effectiveness of the approach and outperforms prior open-source MLLMs on numerous cultural benchmarks such as CVQA, ALMBench, XM100, and MERLIN while preserving general and multilingual vision-language skills. Our findings show that deliberately curating culturally rich data is essential for creating more inclusive multimodal LLMs.

7 Limitations

Language and Cultural Coverage While CulturalGround covers 39 languages, its scope remains limited with respect to the full spectrum of world languages and cultures. Future work could potentially incorporate a larger set of languages and cultures to increase diversity and coverage. In addition, transfer learning techniques or multilingual adapters could be explored to improve cross-cultural generalization to languages and cultures not explicitly represented in existing training data.

Potential Bias in Data Distribution Despite our effort to promote linguistic and cultural diversity, the dataset might still reflect the underlying biases

in global knowledge bases. As shown in Table 16, there is still imbalance in the distribution of images, number of unique entities, and number of samples across languages and regions. Countries with higher Gross Domestic Product[5] (GDP), such as Japan, tend to have higher number of images, unique entities, and language samples. Higher-resourced languages, such as English, show similar patterns. These imbalance might still skew CulturalGround, leading CulturalPangea to perform better on well-represented languages and cultures. Addressing such imbalance remains a challenging yet important direction for future work.

Coverage of Cultural Knowledge Our work primarily focuses on grounding MLLMs with factual and entity-centric cultural knowledge, such as occupation and religion, via Wikidata. While this design enables structured scalable data generation, it does not represent the full spectrum of "cultural knowledge". Other forms of cultural knowledge, such as social norms, dialects, and implicit values, are not well represented in our dataset. Future work could potentially explore methods to incorporate other dimensions of cultural knowledge training data in a scalable way.

8 Acknowledgments

This work was supported in part by DSTA Singapore. We thank the CMU NeuLab and DSTA teams, and the ARR May cycle reviewers, for thoughtful feedback and discussions that substantially improved the paper.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Pulkit Agarwal, Settaluri Sravanthi, and Pushpak Bhattacharyya. 2024. IndiFoodVQA: Advancing visual question answering and reasoning with a knowledge-infused synthetic data generation pipeline. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1158–1176, St. Julian's, Malta. Association for Computational Linguistics.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Fakhraddin Alwajih, Samar Mohamed Magdy, Abdellah El Mekki, Omer Nacar, Youssef Nafea, Safaa Taher Abdelfadil, Abdulfattah Mohammed Yahya, Hamzah Luqman, Nada Almarwani, Samah Aloufi, and 1 others. 2025. Pearl: A multimodal culturally-aware arabic instruction dataset. *arXiv preprint arXiv:2505.21979*.

Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen McKeown. 2024. See it from my perspective: Diagnosing the western cultural bias of large vision-language models in image understanding. *arXiv preprint arXiv:2406.11665*.

Anonymous. 2024. Merlin: Multilingual evaluation of reasoning and logic in images. Unpublished manuscript.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, and 1 others. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, and 1 others. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. 2021. New insights on reducing abrupt representation change in online continual learning. *arXiv preprint arXiv:2104.05025*.

[5] https://en.wikipedia.org/wiki/Gross_domestic_product

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, and 1 others. 2025. Crowdsource, crawl, or generate? creating sea-vl, a multicultural vision-language dataset for southeast asia. *arXiv preprint arXiv:2503.07920*.

Soravit Changpinyo, Roberto Ponté, Ashish V Thapliyal, Yiyi Shen, Zhong Ding, and Radu Soricut. 2024. Pangeabench: A comprehensive benchmark for multilingual multimodal llms. *arXiv preprint arXiv:2410.16153*.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen tau Yih, Shang-Wen Li, and Hu Xu. 2025. Metaclip 2: A worldwide scaling recipe. *Preprint*, arXiv:2507.22062.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, and 1 others. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*.

Matthias De Lange, Gido van de Ven, and Tinne Tuytelaars. 2022. Continual evaluation for lifelong learning: Identifying the stability gap. *arXiv preprint arXiv:2205.13452*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Lianghao Deng, Yuchong Sun, Shizhe Chen, Ning Yang, Yunfeng Wang, and Ruihua Song. 2025. Muka: Multimodal knowledge augmented visual information-seeking. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9675–9686.

Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. 2022. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5089–5098.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *arXiv preprint arXiv:2307.06930*.

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

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

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-factr: Multilingual factual knowledge retrieval from pretrained language models. *arXiv preprint arXiv:2010.06189*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, and 1 others. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Yunshui Li, Yiyuan Ma, Shen Yan, Chaoyi Zhang, Jing Liu, Jianqiao Lu, Ziwen Xu, Mengzhao Chen, Minrui Wang, Shiyi Zhan, and 1 others. 2025. Model merging in pre-training of large language models. *arXiv preprint arXiv:2505.12082*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021a. Marvl: Multicultural reasoning over vision and language. *arXiv preprint arXiv:2109.13238*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021b. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2901–2908.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Steiner, Xiaohua Zhai, and Ibrahim M Alabdulmohsin. 2024. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. *Advances in Neural Information Processing Systems*, 37:106474–106496.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36:66127–66137.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. *Advances in neural information processing systems*, 32.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, and 1 others. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, and 1 others. 2025. Kwai keye-vl technical report. *arXiv preprint arXiv:2507.01949*.

Ashish V Thapliyal, Roberto Ponté, Shachi H Mullappilly, Ananya Srinivasan, Yiyi Shen, Zhong Ding, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2024a. All languages matter: Evaluating lmms on culturally diverse 100 languages. *Preprint*, arXiv:2411.16508.

Zaid Vayani, Zhong Ding, Yiyi Shen, and Radu Soricut. 2024b. All languages matter: Evaluating lmms on culturally diverse 100 languages. *arXiv preprint arXiv:2411.16508*.

Yash Verma, Anubhav Jangra, Raghvendra Kumar, and Sriparna Saha. 2023. Large scale multi-lingual multi-modal summarization dataset. *arXiv preprint arXiv:2302.06560*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, and 1 others. 2024. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *arXiv preprint arXiv:2410.12705*.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and 1 others. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel J. Candes, and Tatsunori Hashimoto. 2024. Synthetic continued pretraining. *ArXiv*, abs/2409.07431.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*.

Wenxuan Zhang, Da Yin, Zhong Ding, and Radu Soricut. 2023. M3exam: A multilingual, multimodal exam-style qa benchmark. *arXiv preprint arXiv:2306.05179*.

A Additional Analysis

A.1 Multimodal Entity Recognition

We evaluate CULTURALPANGEA on MERLIN, a benchmark for multilingual multimodal entity recognition and linking. MERLIN is a challenging testbed constructed from news articles paired with images, featuring over 7,000 named entity mentions (linked to 2,500 Wikidata entities) across 5 languages (Hindi, Japanese, Indonesian, Vietnamese, Tamil). This benchmark specifically targets scenarios where textual context alone can be ambiguous, and visual context provides crucial disambiguation an important evaluation for culturally diverse, multilingual settings. On this benchmark shown in Table 10, our model achieves strong results, significantly outperforming the base Pangea-7B model on open-ended entity recognition (i.e., freely identifying the correct entity from multimodal context).

Evaluation Methodology We employ four evaluation metrics with increasing levels of tolerance to comprehensively assess model performance. The strictest metric, **Exact Match**, requires predictions to exactly match the target entity name, demanding precise recall of Wikipedia titles. More tolerant metrics better capture semantic understanding and practical utility. The **+Alias** metric accepts predictions that match any known alias of the target entity, where aliases are sourced from the Wikidata entity's "Also known as" field. For instance, when the target is "Narendra Modi," a model receives credit for predicting any of the common variants: "Modi," "Narendra Bhai," "Narendra Damodardas Modi," "Narendrabhai Damodardas Modi," "Narendrabhai," "Modiji," "Modi Ji," or "NaMo." This reflects real-world usage where entities are referenced through multiple names and honorifics. The **+Target** metric accepts cases where the target entity name appears anywhere within the prediction text, accommodating longer descriptive outputs. For example, if the target is "Taj Mahal" and the model predicts "The famous Taj Mahal monument in Agra," this would be considered correct. We deliberately avoid checking if the prediction appears in the target (prediction-in-target) to prevent false positives that could arise from generic terms. Finally, **All Methods** combines exact matching, alias matching, and target-in-prediction checking to provide the most comprehensive evaluation of model understanding.

A.2 Performance Gains from Merging Different Checkpoints

We experiment with various checkpoint merging methods such as linear (Wortsman et al., 2022), TIES (Yadav et al., 2023), DARE-TIES (Yu et al., 2024). All methods improve performance over our best model and baseline and we don't rule out that neither is the best as also observed in (Li et al.,

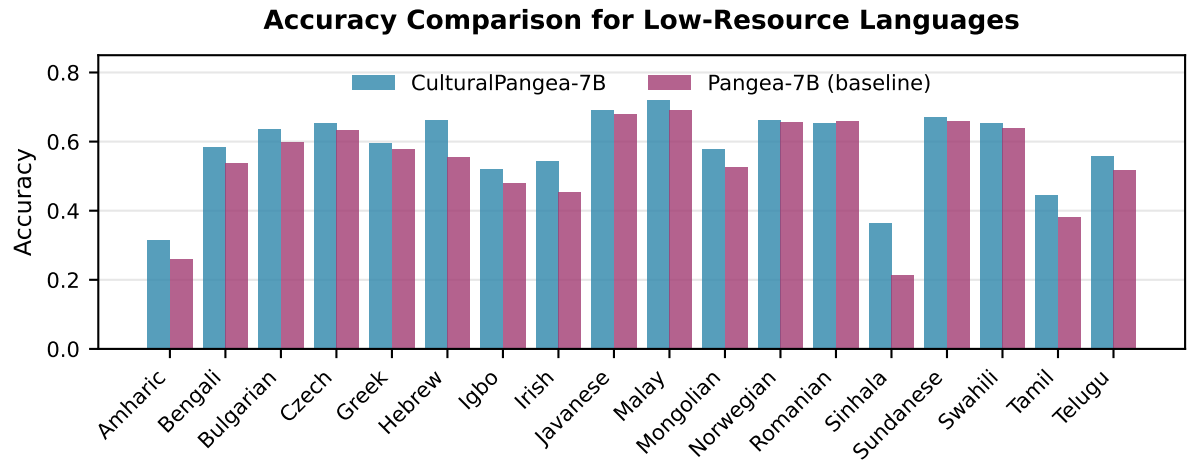

Figure 9: Accuracy improvements on ALM-Bench for low-resource languages with CulturalGround. CulturalPangea achieves the largest gains on underrepresented languages (e.g., Sinhala +13.54), demonstrating that CulturalGround effectively scales to the long tail of languages. Performance on other languages remains stable or improves slightly, with only Norwegian and Romanian showing negligible negative changes.

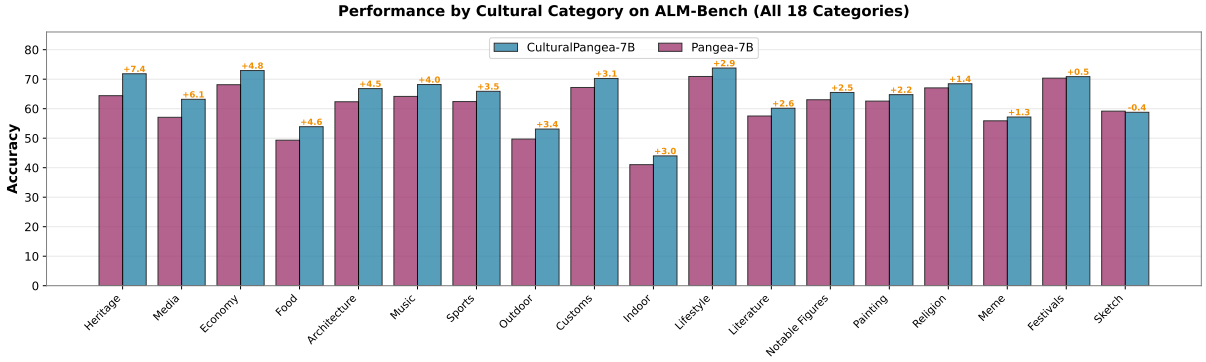

Figure 10: Absolute accuracy gains of CulturalPangea over the baseline across 18 ALM-Bench cultural domains. Improvements cluster in culture-rich categories (Media, Heritage, Music), while Sketch and Meme offer minimal or negative change.

2025). Performance across the board improves with number of checkpoints up to a certain ceiling, where the choice of merging methods, number of checkpoints, and weight assigned to each checkpoint only makes marginal difference. Figure 11 shows the comparison and gains between Pangea-7B(baseline), CulturalPangea (best checkpoint), and merged model with 4, and 5th latest checkpoints.

B Additional Details on Data Curation

B.1 Cultural Relevant Wikidata Properties and Templates QA

Table 15 presents a comprehensive catalog of culturally relevant Wikidata properties employed in our entity extraction process. This table includes the specific properties used to identify cultural entities within Wikidata's knowledge graph, along with the corresponding template questions and template answers that form the foundation of our multilingual VQA generation pipeline.

B.2 Refining Multilingual QA Data for Cultural Fluency

The prompt used in refining and rephrasing multilingual QA data is shown in Figure 13.

B.3 Multilingual MCQ and True/False Question Generation Pipeline

To create culturally grounded VQA examples, we developed an entity-driven question generation pipeline that produces both multiple-choice and true/false questions across languages. Each culturally significant entity (identified via Wikidata) served as a seed for generating question–answer

Dataset	Lang/Regions	#Samples	#Images	Q Types	Eval Metric	Description
Benchmark / Evaluation Datasets						
ALM-Bench (2024)	100L/73 countries	22.7K	2.7K	MCQ&OE	LLM-Judge	Global cultural knowledge across 19 domains
CVQA (2024)	31L/30 countries	10K	4.5K	MCQ&OE	Accuracy	Under-represented countries, local experts
MaXM (2023)	7L/General	2.1K	335	OE	Relaxed-Acc	Translation-based VQA (no cultural focus)
M3EXAM (2023)	9L/General	12.3K	2.8K	MCQ	Accuracy	Multilingual exams (no cultural focus)
XM100 (2022)	36L/Global	3.6K	100	Caption	CIDEr	Multilingual captioning
MERLIN (2024)	5L/4 countries	7.1K	4.2K	OE	Accuracy	Multilingual and cultural entity recognition
Training Datasets						
WorldCuisines-train(2024)	30L/189 countries	1.08M	6K	MCQ&OE	–	Large-scale global cuisine training data
IndiFoodVQA (2024)	1L/India	16.7K	16K	MCQ	–	Knowledge-infused Indian food VQA
SEA-VL(2025)	6L/11 SE Asia	1.28M	1.28M	Caption	–	Authentic SEA cultural visual scenes & short captions
PEARL-train(2025)	1L/19 Arab countries	309K	12.6K	MCQ&OE	–	Arabic cultural instruction dataset
CulturalGround (2025)	**39L/42 regions**	**30M**	**2.8M**	**MCQ&OE**	**–**	**Massive multilingual diverse cultural VQA**

Table 4: Comparison of culturally diverse vision-language datasets. Our CulturalGround dataset (highlighted) provides the largest-scale multilingual cultural training data. We compare with WorldCuisines-train (Winata et al., 2024), IndieFoodVQA (Agarwal et al., 2024), SEA-VL (Cahyawijaya et al., 2025), and PEARL-train (Alwajih et al., 2025)

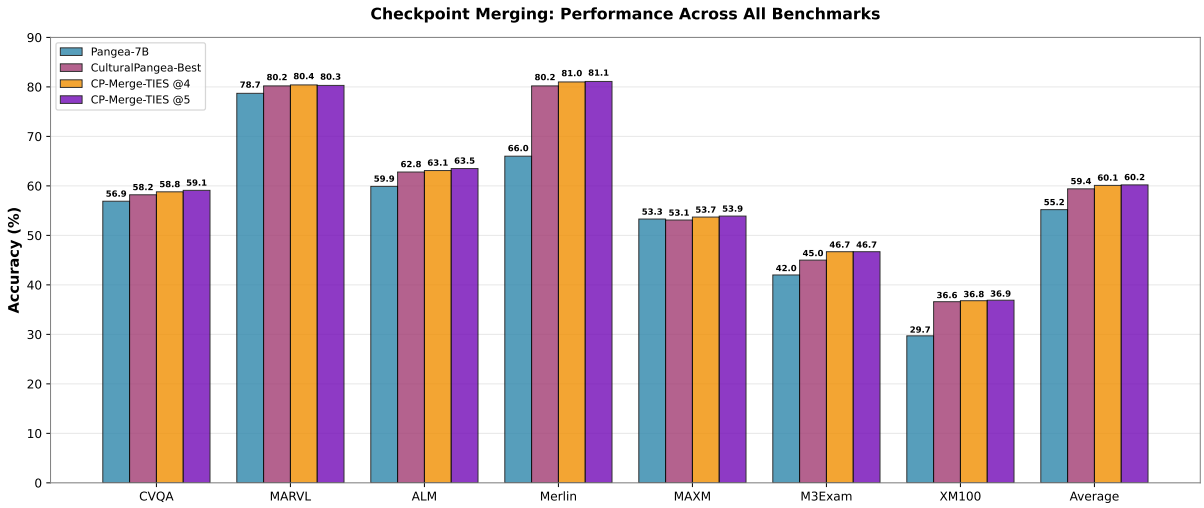

Figure 11: Accuracy improvements from merging checkpoints across all evaluation benchmarks. CK stands for CulturalPangea.

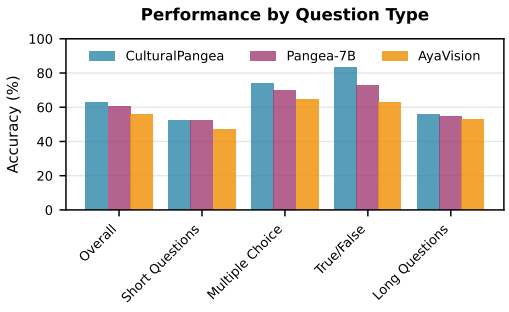

Figure 12: Performance on different types of questions in ALMBench: multiple choices, open-ended, and true/false.

pairs in multiple languages. Using a multilingual large language model (Gemma-3-27B and Qwen-2.5-Instruct), we automatically generated questions based on each entity's metadata (i.e., a known factual attribute of that entity). In particular, we prompted the model to generate a four-option multiple-choice question (MCQ) in the target language for each entity, and for format diversity we also created a yes/no true/false question for a subset of entities. This approach yielded roughly 8 million initial question–answer samples, with approximately 60% in MCQ format and 40% in true/false format, providing a balanced mix of question types. Moreover, each entity was covered in several languages (e.g., Hindi, Chinese, Arabic, Swahili), ensuring that the dataset captured culturally authentic knowledge across diverse linguistic contexts.

Prompt Design and Multilingual Strategy. We carefully crafted the generation prompts to preserve each question's cultural context and authenticity. The system prompt instructed the model to act as

a "cultural expert," following detailed guidelines with in-context examples in the target language to demonstrate the desired question format. Specifically, the prompt emphasized using natural, engaging phrasing in the local language; respecting cultural sensitivities and naming conventions; avoiding revealing the answer in the question; ensuring correct grammar and clarity; varying the question style beyond simple identification; and aiming for a moderate difficulty level that tests deeper cultural knowledge. These measures kept the questions grounded in each entity's cultural background and prevented trivial or overly direct queries. We also tailored the prompts for each target language by providing example Q&A pairs in that language's script and style. This multilingual prompting strategy enabled the model to produce fluent, culturally nuanced questions in every target language without losing authenticity.

Multiple-Choice Question Generation. For MCQs, the model was prompted to produce a question followed by four answer options (A, B, C, D), where option A was the correct answer[6] (a true fact about the entity) and options B, C, and D were plausible but incorrect distractors drawn from the same cultural or topical domain. All answer choices were written to be similar in style and length and culturally sensible, so as to avoid any obvious giveaway of the correct option. In addition, the model was asked to provide a brief explanation justifying why option A is correct, incorporating relevant cultural or historical context. This process yielded rich multiple-choice items, for instance, a Hindi question about a historical dance might present one correct cultural tradition and three other plausible traditions as answer choices, all phrased naturally in Hindi script.

True/False Question Generation. We adopted a similar strategy for true/false items. The model generated a concise factual statement about the entity and labeled it as either true or false, effectively forming a yes/no question. To ensure these binary questions were non-trivial, we directed the model to leverage the entity's attributes when crafting statements, for example, introducing a plausible but incorrect detail when a false statement was required so that answering the question would require specific knowledge of the entity's background. Each true/false question was written in the target language with culturally appropriate wording (for instance, a Swahili prompt might ask whether a certain landmark is located in a particular region, with the answer marked "False" if that region is incorrect).

Relevance Filtering. All generated questions underwent rigorous filtering to ensure accuracy and cultural validity. We filtered out any outputs that were irrelevant and factually incorrect. From the roughly 8 million candidates initially generated, we retained only the best, non-trivial questions, resulting in a final CULTURALGROUND dataset of about 6 million culturally MCQs samples. This final collection preserved a balanced mix of question formats and, coupled with our culturally-informed prompt design, ensured that the questions remained authentic to each culture while covering a broad range of entities and properties. Overall, this multilingual MCQ/true–false generation pipeline allowed us to inject diverse cultural knowledge into the VQA data in a high-quality manner, which was crucial for training the CULTURALPANGEA model.

Finally, we provide the full prompt templates used for MCQ generation in Figure 14 and True/False in Figure 15.

C Cultural Domains and Long-Tail Entity Coverage in CulturalGround

Long-tail entities emphasis. CulturalGround treats long-tail entities as the first-class citizen. As shown in Figure 23, most entities in the dataset have only a small number of connections, such as very few incoming or outgoing links in the knowledge graph or minimal Wikipedia backlinks. Consequently, median link counts are extremely low, indicating that the typical entity in *CulturalGround* is sparsely connected. This distribution is highly right-skewed, with a long tail: a handful of entities are linked to many others, but the vast majority are referenced only a few times. Such patterns underscore the dataset's focus on culturally specific, niche entities that lie beyond the well-connected head of popular or globally known concepts. Notably, a substantial portion of the entities included have no dedicated Wikipedia page at all(see Figure 24. The prevalence of entries lacking Wikipedia articles further highlights the dataset's extension into culturally important yet under-documented regions of knowledge, reinforcing the long-tail presence.

Regional domain emphasis. Beyond its long-

[6]Option A is always the correct answer for output parsing consistency, during training we use shuffled options to ensure model does not overfit or prefer one option over the others.

tail connectivity profile, *CulturalGround* spans a broad array of domains and regions, surfacing localized cultural themes that vary systematically by country. For example, India's entries are dominated by heritage and settlement categories (e.g., numerous Hindu temples and villages), reflecting the country's rich corpus of historical sites and settlements; the Netherlands highlights vernacular architecture and art categories, aligning with its distinctive building traditions and artistic heritage; and Portugal's distribution is led by ecclesiastical heritage categories, indicative of its deep-rooted religious and historical legacy. Similarly, China's data feature prominent infrastructure and administrative categories (such as railway stations and provincial divisions), echoing an emphasis on civic infrastructure and governance; Brazil is characterized by an abundance of cultural media and monument categories, mirroring its vibrant media culture and iconic landmarks; and Japan's prominent categories include religious architecture (temples and shrines) and education (schools and universities), underscoring the country's spiritual traditions and academic legacy. This diversity of region-specific dominant categories underscores *CulturalGround*'s broader goal of surfacing underrepresented cultural knowledge by grounding the dataset in regionally significant, long-tail cultural entities.

D Balancing Regions and Languages With Hybrid Sampling

Figure 19, Table 16, and Table 17 provide comprehensive statistics on the training data across regions and languages. While preparing the training data, to mitigate the imbalance between high-resourceful and low-resourceful regions and languages, we employ hybrid temperature-based sampling (Arivazhagan et al., 2019), where we first sample data by regions to cap high-resourceful regions without affecting smaller regions, and then finally apply smooth language sampling to balance languages. We use temperature of 4.0 for region sampling and 1.5 for language sampling. We use small temperature in language sampling to keep cross-lingual associations between entities.

E M3LS Dataset Creation

M3LS is a large-scale multi-lingual, multi-modal summarization dataset introduced by (Verma et al., 2023). It contains news articles with images in 20 languages (sourced from BBC News over a decade), paired with professionally annotated summaries. We leverage this dataset to automatically create an **entity-centric** data source for improving cross-lingual entity recognition.

Automated Entity Extraction Pipeline: For each news article in five high-diversity languages (Hindi, Tamil, Indonesian, Japanese, Vietnamese) plus English, we use a specialized LLM prompt to identify the article's most central entity. The prompt provides the article's title, summary, first paragraph, and any image captions, and asks the model to return: (1) the main entity mention (in the original language), (2) its corresponding English Wikipedia page title, and (3) a brief justification. This process is fully automated and yields a structured JSON output per article containing the extracted entity and its English Wikipedia reference. By querying an LLM in this way, we effectively perform cross-lingual entity linking: mapping non-English entity mentions to a canonical English Wikipedia title.

Data Curation and Filtering: The raw generation produced approximately 203K candidate Q&A pairs (image + question-answer) across the six languages. Each instance uses the article's image and asks a question like "What is the Wikipedia page title that corresponds to the entity 'X' mentioned in the news article 'Y'?", with the answer being the English Wikipedia title for entity X. We then removed 4,575 instances that overlapped with the MERLIN dataset (to avoid test contamination) and reduced the exceedingly large English portion to 50K examples. This resulted in a final training set of ∼90K high-quality multi-modal instances, which is about 45% of the initially generated data. Each example thereby provides a multilingual visual context and a linked entity label, serving as valuable weak supervision for entity recognition and linking. This M3LS-derived resource augments our training data and helps the model learn to recognize entities in multiple languages and modalities.

The full prompt used for generating entity data is shown in Figure 18.

F Contamination Analysis

Most cultural multimodal benchmarks such as CVQA and ALMBench use images from Wikimedia. We thus analyze potential test contamination due to training on CulturalGround, which includes images sourced from Wikimedia. CVQA

and ALMBench contain a very small fraction of test images that also appear in the training corpus (Table 5). Removing these overlapping test cases produces negligible changes in accuracy (under 0.53% absolute on either benchmark), indicating minimal impact of leakage.

For multimodal entity recognition benchmark MERLIN, overlap exists only at the *entity-name* level; images (from BBC News) and questions differ from training. Despite 46% entity-name overlap, most of CulturalPangea-7B's overall improvement over Pangea-7B (15.7 points) remains on the clean subset; the overlap advantage accounts for only about 3.7 points (Table 6). In addition, strong cross-lingual/cross-cultural transfer (e.g., gains on ALMBench languages unseen in training) shown in Section 4.3 further supports that improvements stem from generalization rather than memorization.[7]

G Breakdown Results

G.1 MaRVL

We show the performance of different models on the MaRVL benchmark in Table 7.

G.2 M3Exam

We show the performance of different models on the M3Exam benchmark in Table 8.

G.3 MAXM

We show the performance of different models on the MAXM benchmark in Table 9.

G.4 MERLIN

We show the performance of different models on the MERLIN benchmark in Table 10.

G.5 XM100

We show the performance of different models on the XM100 benchmark in Table 11.

G.6 CVQA

We show the performance of different models on the CVQA benchmark in Table 12.

G.7 ALMBench

We show the performance of different models on the ALMBench benchmark in Table 14.

[7]CVQA/ALMBench contamination was computed on an early model trained on early CulturalGround containing 2.3M samples in 6 languages: English, Hindi, Vietnamese, Japanese, Indonesian, Tamil; MERLIN results use the final model and CultureGround

H CultureGround Samples

See Figure 20 for examples drawn from CulturalGround.

Bench	Size	Images	Wiki-Images	Overlap Images	Train-Test Overlap	Train Affected	Acc (w/o overlap)
CVQA	10.3K	5.2K	1,771	10	0.6%	78	0.53% (57.1→56.6)
ALMBench	22.7K	2.3K	361	6	1.7%	52	0.09% (62.5→62.4)

Table 5: Overlap analysis for Wikimedia-sourced benchmarks. Overlap is computed between benchmark test images and CulturalGround training images. Removing overlapping test items yields negligible accuracy changes. The dataset used for CVQA and ALMBench contamination analysis is early version containing 2.3M samples in 6 languages: English, Hindi, Vietnamese, Japanese, Indonesian, Tamil

Model	Overall Acc	Cont. Acc	Clean Acc	Cont. % (entity)	Gap
Pangea-7B	65.6%	69.3%	62.4%	46%	+6.9%
CulturalPangea-7B	81.3%	87.0%	76.4%	46%	+10.6%
Δ (CulturalPangea − Pangea)	+15.7	+17.7	+14.0	–	+3.7

Table 6: MERLIN contamination at the entity-name level. Images/questions do not overlap; only some entity names appear in CulturalGround. Most of the overall gain persists on the clean subset, indicating genuine generalization.

Model	**English**	**Multi**	**Indonesian**	**Swahili**	**Tamil**	**Turkish**	**Chinese**
Llava-Next-7B	62.8	50.9	52.2	50.6	50.5	50.4	50.6
Molmo-7B-D	65.3	54.9	61.1	49.6	49.6	52.2	62.2
Llama3.2-11B	64.5	58.1	62.7	52.4	54.0	61.6	59.5
PaliGemma-3B	56.5	52.2	53.4	49.6	50.5	56.3	51.3
mBLIP mT0-XL	67.3	66.7	64.9	64.8	69.7	68.1	65.9
AyaVision-8B	73.8	64.5	66.6	53.1	59.7	73.7	69.3
Pangea-7B	87.0	78.7	80.8	74.2	69.8	84.6	84.1
CulturalPangea-7B	**89.0**	**80.3**	**83.6**	**74.6**	**73.2**	**85.4**	**84.5**

Table 7: Performance of selected models on the MaRVL benchmark across different languages.

Model	**English**	**Multi**	**Afrikaans**	**Chinese**	**Italian**	**Portuguese**	**Thai**	**Vietnamese**
Llava-Next-7B	36.5	28.4	28.2	25.4	37.8	27.0	23.7	28.4
Molmo-7B-D	57.1	39.1	35.6	56.4	49.4	40.2	27.4	25.9
Llama3.2-11B	51.8	36.6	42.3	46.4	45.8	28.4	26.4	30.2
PaliGemma-3B	36.0	25.6	26.4	24.7	32.2	24.3	27.2	19.0
mBLIP mT0-XL	22.8	25.0	16.0	25.6	33.7	21.2	22.4	31.0
AyaVision-8B	56.2	41.7	47.2	49.2	55.4	37.8	30.4	30.2
Pangea-7B	61.4	42.0	52.1	49.2	54.9	43.3	32.9	19.8
CulturalPangea-7B	**58.0**	**46.7**	**52.8**	**55.2**	**57.9**	**46.0**	**34.9**	**33.6**

Table 8: Performance of selected models on the M3Exam dataset across different languages.

Model	**English**	**Multi**	**French**	**Hindi**	**Hebrew**	**Romanian**	**Thai**	**Chinese**
Llava-Next-7B	54.9	21.4	33.7	16.2	10.7	15.5	18.3	33.9
Molmo-7B-D	52.9	37.5	45.5	33.5	30.7	28.9	46.3	40.4
Llama3.2-11B	55.3	43.9	48.1	50.4	41.8	36.6	56.7	30.0
PaliGemma-3B	47.9	19.9	8.0	36.5	19.3	13.4	31.3	10.8
mBLIP mT0-XL	44.7	36.8	36.0	42.7	28.9	30.3	56.3	26.4
AyaVision-8B	49.4	52.1	56.1	63.8	57.5	50.7	34.7	49.8
Pangea-7B	53.5	53.3	43.9	53.5	59.3	45.8	67.2	50.2
CulturalPangea-7B	**55.3**	**53.9**	**45.5**	**50.4**	**62.9**	**48.6**	**68.3**	**48.0**

Table 9: Performance of selected models on the MAXM dataset across different languages.

Prompt for Improving Fluency in Cultural Templates QAs

System Prompt:
You are a cultural expert specializing in creating high-quality, culturally sensitive questions and answers about diverse entities from around the world. Your goal is to create natural-sounding questions and factually accurate answers that respect cultural nuances while maintaining value about important properties of entities such as location, category, administrative territory, and other key attributes.

User Prompt:
Given this entity and context in {language_name}:
Entity: {label}
Description: {description}
Entity Region(Country): {region}
Original Question: {question}
Original Answer: {answer}

Task: Create both a natural question and an answer for a visual question answering dataset focused on cultural recognition of entities in multilingual contexts.

For the question:
1. Maintain the precise property being asked about in the original question (like location, category, administrative territory, awards, etc.)
2. Use natural, conversational phrasing in authentic {language_name} that a native speaker would use
3. Do NOT reveal specific details about the entity in the question unless absolutely necessary
4. Ensure cultural sensitivity and respect for local naming conventions and terminology
5. Make the question grammatically correct, clear, and unambiguous
6. Phrase it as if someone is looking at an image of this entity and asking about it
7. Avoid awkward phrasing or other template artifacts

For the answer:
1. Ensure complete factual accuracy based on the provided information
2. Use natural language appropriate for {language_name} with proper cultural context
3. Include key factual details from the original answer and leave out any unnecessary information
4. When appropriate, provide brief additional cultural context or significance of the entity
5. Make sure to include the full entity name and keep the answer around the property being asked about
6. Avoid vague phrases - be specific and informative. Ensure the answer is clear, concise, relevant, don't include unnecessary details.
7. It is best to avoid adding new information unless it is a well-known fact about the entity that enhances understanding.
8. We are grounding the model to cultural knowledge, so it is really important to be accurate and keep answers factually correct.
9. The region/country of the entity {region} is provided to you for context, so please don't confuse the entity with other regions or countries.

Format your response exactly as:
Q: [your reformulated question]
A: [your reformulated answer]

Figure 13: Refining QA Prompt

Cultural Entity Multiple-Choice Question Generation

System Prompt:

You are a cultural expert who creates high-quality multiple-choice questions about entities while preserving cultural context, and authenticity.

User Prompt:

Given this entity and context in {language_name}:
Entity: {label}
Description: {description}
Original Question: {question}
Original Answer: {answer}
Region/Country: {region}
Task: Create a multiple-choice question with four options (A, B, C, D) based on the cultural entity.

For the multiple-choice question:

1. Maintain the original topic but use natural, engaging phrasing in {language_name}
2. NEVER reveal specific details about the entity in the question unless necessary
3. Respect cultural context and sensitivity
4. Make it grammatically correct, clear, and culturally relevant
5. Vary question formats beyond basic identification
6. Create questions with appropriate difficulty level
7. Aim for questions that test deeper cultural knowledge
8. Keep the entity as the grounding point

For the options:

1. Option A should ALWAYS be the correct answer
2. Create three plausible but incorrect options (B, C, D)
3. All options should be culturally accurate, sensible, and realistic
4. Ensure all options are similar in length and format
5. All incorrect options should be from the same general category
6. Options should represent meaningful distinctions but yet plausible and challenging within the cultural/regional context

For the explanation:

1. Briefly explain why option A is correct
2. Include relevant cultural or historical context
3. Keep explanation concise but informative (1–3 sentences)
4. Keep the entity as the grounding point

Example format will be provided based on language context.

Create a multiple-choice question for this entity in exactly this format:

Q: [your multiple-choice question]
A) [correct answer]
B) [plausible incorrect option]
C) [plausible incorrect option]
D) [plausible incorrect option]
Correct: A
Explanation: [brief explanation with cultural context]

Figure 14: Refining QA Prompt

Cultural Entity True/False Question Generation

System Prompt:
You are a cultural expert who creates clear and culturally-sensitive true/false statements/questions about entities that test understanding while preserving authenticity.

User Prompt:
Given this entity and context in {language_name}:
Entity: {label}
Description: {description}
Original Question: {question}
Original Answer: {answer}
Region/Country: {region}

Task: Create a true/false statement based on the cultural entity.

For the statement/question:
1. Create either a clear statement OR a yes/no question about the entity in {language_name}
2. Mix between statements and questions for variety
3. Make it unambiguous - clearly either true or false
4. Test meaningful cultural knowledge, not trivial details
5. Respect cultural sensitivity
6. Keep the entity as the central focus
7. Vary between true and false answers for diversity

For the explanation:
1. Briefly explain why the statement is true or false
2. Include relevant cultural or historical context
3. Keep it concise (1-2 sentences)

Example format will be provided based on language context.

Create a true/false item for this entity in exactly this format:
Statement: [your true/false statement]
Answer: [True/False]
Explanation: [brief explanation]

OR

Question: [your true/false question]
Answer: [True/False]
Explanation: [brief explanation]

Figure 15: True/False Question Generation Prompt Template

Prompt for Evaluating VQA Sample Quality and Image-Entity Alignment

System Prompt:
You are an expert at evaluating whether images match with cultural entities and their descriptions. You assess alignment between visual content and textual information.

User Prompt:
Evaluate this VQA sample for quality and alignment.

Entity Information:
Label: {label}
Description: {description}
Region/Country: {region}
Language: {language}
Question: {question}
Answer: {answer}

Your task is to determine:
1. Does the image show or reasonably represent the entity described?
2. Are there any quality issues with this sample?

Common issues to check for:
1. Image completely unrelated to the entity (e.g., entity is about a person, but image is of animal. Or entity is about park but image show city, or entity is about a person but image is of a building)
2. Mixed languages in question or answer
3. Obvious factual errors in the answer that you can confirm and very sure about
4. Question and answer mismatch
5. Corrupted or incomplete answer

If you are not sure about the answer:
1. Treat sample as match and no issue
2. We are mostly concerned with the image being completely irrelevant to the entity and we understand some models may not know some long-tail entities
3. So unless there is a clear mismatch or quality issue in rephrased question/answer, treat it as match

Other considerations:
1. If the question asks about education or birth place or other entity properties, treat it as a match if the image is related to the entity, even if it does not show the specific property
2. If the image is not provided, treat it as match unless the answer is clearly unrelated to the entity or has problematic issues mentioned above
3. I repeat, if you are not sure about your answer and can not confirm it which might happen alot with long-tail entities, treat it as match and no issue

Format your response exactly as (Notice and keep the line breaks):
MATCH: [True/False]
ISSUE: [None/ImageMismatch/MixedLanguage/FactualError/QAMismatch/Unclear]
EXPLANATION: [Brief explanation of your assessment]

Figure 16: VQA Quality Evaluation Prompt

Prompt for Evaluating MCQ Quality and Cultural Alignment

System Prompt:
You are an expert at evaluating MCQ quality and cultural alignment. You assess whether questions match with cultural entities and check for quality issues.

User Prompt:
Evaluate this MCQ sample for quality and alignment.

Entity Information:
Label: {label}
Description: {description}
Region/Country: {region}
Language: {language}
Question Type: {question_type}
Question: {question}
Options: {options_text}
Correct Answer: {correct_answer}
Explanation: {explanation}

Your task is to determine:
1. Does the image (if present) reasonably represent the entity described?
2. Is the question culturally relevant to the specified region?
3. Are there any quality issues with this MCQ?

Common issues to check for:
1. Image completely unrelated to the entity or question
2. Question not relevant to the cultural context or region
3. Incorrect answer or poor explanation
4. Mixed languages in question, options, or explanation
5. Poorly formed question or confusing options
6. Factual errors you can confirm

Guidelines:
1. If you are not sure about cultural relevance or correctness, treat it as acceptable
2. Focus on obvious mismatches and clear quality issues
3. For questions without images, focus on cultural relevance and question quality
4. Consider regional context when evaluating cultural appropriateness

Format your response exactly as:
MATCH: [True/False]
CULTURALLY_RELEVANT: [True/False]
ISSUE: [None/ImageMismatch/CulturalMismatch/IncorrectAnswer/MixedLanguage/ PoorQuestion/FactualError/Other]
EXPLANATION: [Brief explanation of your assessment]

Figure 17: MCQ Quality Evaluation Prompt

Prompt for Entity Extraction from Multilingual News Articles

System Message:
You are a specialized entity extraction AI assistant that identifies the most important entities in news articles across multiple languages. Your task is to analyze news content and extract the central entity, providing both its name in the original language and its standard English Wikipedia title.

User Prompt:
Below is a news article in {LANGUAGE}. Analyze it and identify the most relevant entity mentioned.

Article Information:
Article Title: {title}
Article Summary: {summary}
Article First Paragraph: {first_paragraph}
Image Captions: {image_captions} (if available)
Keywords: {keywords} (if available)

Please extract the following information:
1. The most relevant entity mentioned in the article (in {LANGUAGE})
2. The Wikipedia page title for this entity (in English)
3. A brief justification for why this is the most relevant entity

Format your response as JSON with the following keys:
- entity_mention: The entity name in {LANGUAGE}
- wikipedia_title: The English Wikipedia title for this entity
- justification: Your explanation

Examples of entity extraction from different languages:
Example 1 (Vietnamese):
Article Title: Amabie, 'bua yem' chong Covid-19 cua nguoi Nhat
Entity Name: Covid-19
English Wikipedia Title: COVID-19

[MORE EXAMPLES...]
Additional Guidelines:
If there are multiple important entities, choose the most central one to the article.

Figure 18: Entity Extraction Prompt

Models	Multi	Hindi	Tamil	Indonesian	Japanese	Vietnamese
Exact Match						
Llava-Next-7B	34.1	30.1	13.0	42.4	41.9	43.1
Molmo-7B-D	42.9	30.8	15.8	54.7	53.8	59.5
PaliGemma-3B	13.1	12.6	2.5	22.0	13.7	15.0
mBLIP-mT0-XL	15.8	10.8	11.0	19.2	14.9	23.3
AyaVision-8B	55.3	55.4	40.5	55.3	62.6	62.9
Pangea-7B	66.5	67.6	55.2	68.8	73.1	67.6
CulturalPangea-7B	**81.1**	**77.0**	**76.2**	**80.6**	**85.8**	**85.7**

Models	Multi	Hindi	Tamil	Indonesian	Japanese	Vietnamese
Exact Match + Alias						
Llava-Next-7B	38.8	33.0	14.7	50.0	46.3	50.1
Molmo-7B-D	47.8	33.2	17.2	63.3	59.3	66.0
PaliGemma-3B	19.1	15.0	2.6	31.9	20.4	25.8
mBLIP-mT0-XL	20.1	14.2	12.9	26.5	20.9	26.0
AyaVision-8B	57.5	56.9	41.3	59.0	65.1	65.0
Pangea-7B	71.6	71.2	58.9	77.4	79.5	71.2
CulturalPangea-7B	**84.8**	**79.3**	**78.9**	**86.6**	**90.1**	**89.3**

Models	Multi	Hindi	Tamil	Indonesian	Japanese	Vietnamese
Exact Match + Target in Prediction						
Llava-Next-7B	50.7	48.5	22.3	67.8	57.2	57.7
Molmo-7B-D	56.5	40.6	23.4	73.5	69.2	75.9
PaliGemma-3B	16.1	17.8	4.3	25.3	15.4	17.8
mBLIP-mT0-XL	17.5	12.5	12.7	20.9	16.8	24.5
AyaVision-8B	69.9	66.3	58.9	73.3	73.9	77.3
Pangea-7B	73.1	73.3	61.1	79.1	78.7	73.3
CulturalPangea-7B	**83.5**	**81.2**	**78.9**	**83.7**	**87.0**	**86.9**

Models	Multi	Hindi	Tamil	Indonesian	Japanese	Vietnamese
Exact Match + Alias + Target in Prediction						
Llava-Next-7B	60.4	57.0	31.8	77.6	66.1	69.3
Molmo-7B-D	66.5	52.3	32.9	85.2	77.6	84.7
PaliGemma-3B	25.1	23.4	5.7	40.3	25.1	31.1
mBLIP-mT0-XL	23.4	17.1	16.4	30.0	24.5	28.9
AyaVision-8B	78.3	73.4	68.2	83.4	82.1	84.3
Pangea-7B	80.8	80.1	68.6	88.7	86.3	80.1
CulturalPangea-7B	**88.4**	**84.0**	**84.4**	**91.0**	**91.7**	**91.0**

Table 10: Performance results across all models and evaluation metrics on cultural vision tasks across five languages. Four evaluation methods are used with increasing compassion: Exact Match requires predictions to exactly match the target title; +Alias also accepts predictions matching any alias of the entity(alias names from Wikidata); +Target also accepts cases where the target appears anywhere in the prediction; and All Methods combines exact match, alias matching, and target-in-prediction. Results demonstrate that all models benefit significantly from more lenient evaluation criteria, as exact matching requires precise title recall while alias and target-in-prediction methods better capture semantic understanding. CulturalPangea-7B consistently achieves the highest performance across all metrics and languages.

Models	English	Multi	Arabic	Bengali	Czech	Danish	German	Greek	Spanish	Persian	Finnish	Filipino
Llava-Next-7B	92.4	15.5	5.0	0.0	27.3	22.6	23.8	2.4	59.6	4.8	9.3	6.2
Molmo-7B-D	4.7	6.0	5.5	2.7	2.8	7.6	9.5	2.6	8.3	7.2	2.0	2.4
Llama3.2-11B	23.1	5.8	0.0	0.0	0.2	15.3	4.4	0.5	14.9	0.0	2.3	11.0
PaliGemma-3B	24.7	0.6	0.0	0.0	0.6	1.0	3.1	0.0	0.5	0.0	0.0	0.1
mBLIP mT0-XL	101.0	6.8	7.7	2.7	11.2	6.9	5.8	7.7	19.0	11.5	4.7	5.9
AyaVision-8B	10.3	10.0	7.8	7.7	9.3	7.9	11.7	6.0	14.5	15.5	2.0	2.5
Pangea-7B	93.2	29.7	39.6	25.1	27.5	32.4	42.7	11.9	96.5	34.7	8.6	13.8
CulturalPangea	90.8	36.9	41.2	29.6	39.4	30.9	51.7	21.3	82.9	38.5	3.5	10.0

Models	French	Hebrew	Hindi	Croatian	Hungarian	Indonesian	Italian	Japanese	Korean	Maori	Dutch	Norwegian	Polish
Llava-Next-7B	61.9	3.0	3.9	7.4	16.7	20.5	36.3	1.1	10.1	2.4	50.7	27.2	21.8
Molmo-7B-D	18.5	10.6	2.5	2.2	1.5	25.8	12.5	1.2	4.5	1.0	8.5	5.6	3.7
Llama3.2-11B	20.0	0.0	0.0	0.6	18.3	1.4	22.1	0.0	0.0	2.3	35.6	0.6	1.1
PaliGemma-3B	1.3	0.0	0.0	0.3	0.4	0.3	0.4	0.0	0.0	2.3	3.1	0.5	0.6
mBLIP mT0-XL	10.9	9.4	2.0	2.2	7.6	8.9	4.8	1.1	4.5	4.2	11.1	7.2	11.2
AyaVision-8B	15.1	20.4	3.5	4.3	2.5	27.0	19.0	3.6	11.4	0.2	12.0	5.9	13.2
Pangea-7B	71.0	33.3	30.4	13.9	9.6	64.6	54.4	0.5	20.4	0.6	56.3	53.3	33.8
CulturalPangea	77.7	42.5	28.4	14.2	6.1	57.7	58.3	0.7	23.1	0.6	63.6	58.3	48.0

Models	Portuguese	Quechua	Romanian	Russian	Swedish	Swahili	Telugu	Thai	Turkish	Ukrainian	Vietnamese	Chinese
Llava-Next-7B	48.2	0.0	14.9	20.2	31.9	0.6	0.0	0.0	0.0	0.1	0.0	1.1
Molmo-7B-D	8.5	0.2	5.4	12.5	4.9	0.0	0.4	0.0	3.8	3.8	20.5	0.0
Llama3.2-11B	24.5	0.0	10.5	0.5	9.1	6.5	0.0	0.0	0.0	0.0	0.0	0.0
PaliGemma-3B	2.8	0.1	1.1	0.0	2.0	0.0	0.0	0.9	0.0	0.0	0.4	0.0
mBLIP mT0-XL	10.8	0.6	4.3	7.9	11.0	7.5	7.4	0.0	9.9	4.0	6.6	0.0
AyaVision-8B	11.5	0.2	23.1	21.6	6.0	1.4	3.6	0.0	14.7	14.9	25.2	4.0
Pangea-7B	82.6	0.0	41.7	62.0	30.8	45.8	0.0	0.0	0.0	0.0	0.0	0.9
CulturalPangea	87.5	0.8	43.4	66.1	23.0	46.8	27.6	0.0	44.3	39.9	84.1	0.5

Table 11: XM100 benchmark performance (%) of all multimodal models across 37 languages.

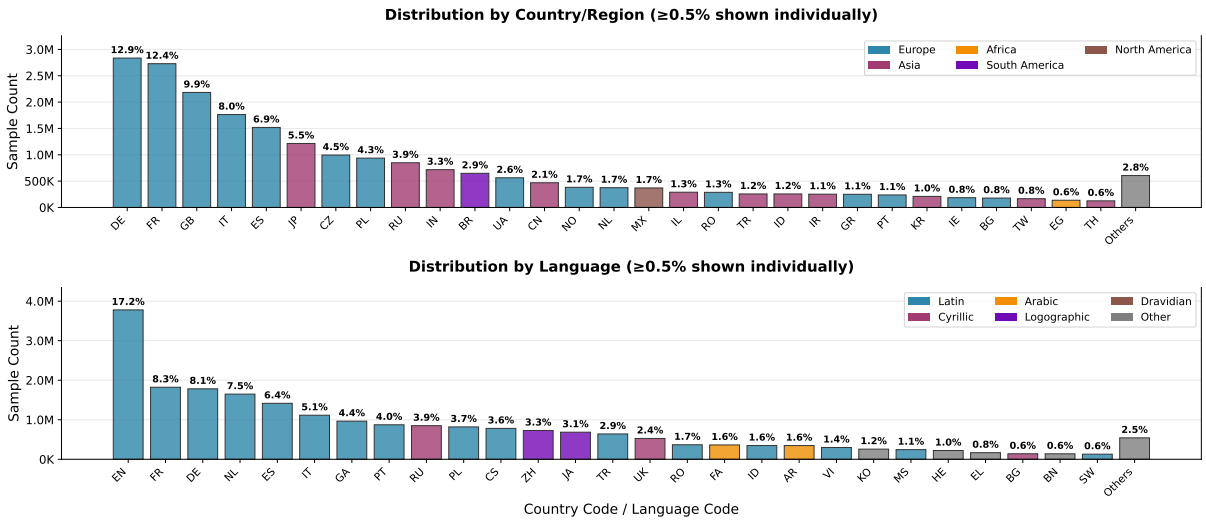

Figure 19: The distribution of languages and countries in CulturalGround

Models	Average	Brazil-Portuguese	Bulgaria-Bulgarian	China-Chinese	Ethiopia-Amharic	India-Bengali	India-Hindi	India-Tamil	India-Telugu	India-Urdu	Indonesia-Indonesian
Llava-Next-7B	41.3	62.3	41.5	51.1	29.5	31.1	N/A	28.8	28.0	N/A	42.2
Molmo-7B-D	58.8	69.0	54.9	66.2	58.1	61.9	51.7	61.2	58.5	50.5	52.9
Llama3.2-11B	69.6	74.6	64.2	73.6	68.4	76.9	68.2	80.4	80.5	54.6	65.8
PaliGemma-3B	43.3	53.9	39.1	53.7	24.8	46.2	N/A	46.0	43.5	N/A	45.4
mBLIP mT0-XL	37.7	44.4	38.0	39.9	35.9	36.4	N/A	44.2	39.0	N/A	37.4
Aya-Vision-8B	51.8	66.8	44.7	65.2	29.4	49.8	62.7	44.4	47.5	47.3	56.7
Pangea-7B	57.5	72.9	53.6	74.0	35.9	59.4	74.6	51.9	54.5	59.5	62.1
CulturalPangea	59.1	71.5	56.6	70.7	39.7	59.1	74.6	54.2	58.5	66.4	61.2

Models	Indonesia-Javanese	Indonesia-Sundanese	Ireland-Irish	Japan-Japanese	Kenya-Swahili	Malaysia-Malay	Mexico-Spanish	Mongolia-Mongolian	Nigeria-Igbo	Norway-Norwegian	Pakistan-Urdu
Llava-Next-7B	38.7	35.5	42.6	32.5	46.2	45.7	51.4	33.3	35.0	56.9	36.6
Molmo-7B-D	53.9	55.0	64.4	42.9	73.3	54.6	53.6	51.9	53.0	54.8	67.1
Llama3.2-11B	60.6	64.0	76.4	54.2	79.1	72.1	66.6	54.5	61.5	66.9	78.7
PaliGemma-3B	41.4	33.0	34.4	43.3	44.0	44.1	47.4	29.2	32.0	52.2	44.9
mBLIP mT0-XL	37.4	31.0	35.3	30.0	45.1	40.6	44.9	29.2	30.5	42.8	40.3
Aya-Vision-8B	48.2	46.5	47.2	48.3	55.0	57.0	57.6	28.5	34.7	53.2	50.9
Pangea-7B	49.2	53.0	56.4	48.3	64.1	59.7	61.9	42.0	46.5	64.2	66.2
CulturalPangea	52.9	58.0	54.9	52.2	67.0	58.4	56.3	42.6	43.5	62.9	71.3

Models	Romania-Romanian	Russia-Russian	Singapore-Chinese	South Korea-Korean	Spain-Spanish	Sri_Lanka-Sinhala	Indonesia-Minangkabau	Egypt-Egyptian Arabic	France-Breton	India-Marathi
Llava-Next-7B	52.3	53.5	44.8	43.4	63.5	29.8	40.2	33.5	27.4	N/A
Molmo-7B-D	63.6	61.5	69.3	65.2	70.1	68.0	54.6	56.7	44.2	N/A
Llama3.2-11B	76.8	74.5	80.7	73.8	81.4	72.4	68.9	68.5	49.4	N/A
PaliGemma-3B	50.3	53.5	48.6	61.0	60.1	31.6	39.8	40.4	29.9	N/A
mBLIP mT0-XL	43.7	42.0	36.8	38.3	53.5	31.1	34.7	31.0	23.5	N/A
Aya-Vision-8B	62.8	66.3	66.8	74.4	74.5	28.9	52.4	51.5	34.4	50.8
Pangea-7B	64.2	74.0	66.0	70.3	72.6	39.1	47.4	49.3	34.3	56.4
CulturalPangea	69.2	70.0	74.1	69.0	73.6	39.1	53.8	52.7	37.5	59.4

Table 12: CVQA benchmark performance (%) of multimodal models across 31 language-country pairs in local-languages.

Models	Average	Brazil-Portuguese	Bulgaria-Bulgarian	China-Chinese	Ethiopia-Amharic	India-Bengali	India-Hindi	India-Tamil	India-Telugu	India-Urdu	Indonesia-Indonesian
Llava-Next-7B	56.1	61.3	50.9	58.8	52.9	60.8	N/A	61.4	60.5	N/A	48.5
Molmo-7B-D	47.7	65.8	45.6	68.5	31.6	47.9	N/A	36.4	41.5	N/A	50.5
Llama3.2-11B	61.0	72.9	54.4	72.0	41.9	62.9	N/A	66.4	66.5	N/A	63.6
PaliGemma-3B	54.2	59.5	49.3	54.9	52.6	59.1	N/A	66.0	62.5	N/A	49.3
mBLIP mT0-XL	40.8	45.4	39.1	43.7	34.2	43.0	N/A	46.0	41.0	N/A	38.1
Aya-Vision-8B	61.8	68.7	54.7	63.0	55.6	62.6	73.1	69.2	69.0	69.1	56.3
Pangea-7B	64.9	72.9	60.1	67.2	60.7	67.1	76.1	71.0	68.0	72.3	60.4
CulturalPangea	66.2	71.1	60.4	62.7	63.2	69.9	76.1	72.0	70.5	73.2	65.0

Models	Indonesia-Javanese	Indonesia-Sundanese	Ireland-Irish	Japan-Japanese	Kenya-Swahili	Malaysia-Malay	Mexico-Spanish	Mongolia-Mongolian	Nigeria-Igbo	Norway-Norwegian	Pakistan-Urdu
Llava-Next-7B	48.1	49.0	66.6	40.9	71.1	54.9	51.1	44.2	53.0	57.2	67.1
Molmo-7B-D	45.1	39.5	43.6	44.8	47.6	51.7	55.1	35.9	36.0	49.2	46.8
Llama3.2-11B	48.8	54.0	57.4	58.1	61.5	69.2	64.7	41.0	39.5	65.9	65.7
PaliGemma-3B	48.1	46.0	58.3	44.8	59.7	54.9	51.7	43.4	46.0	55.2	67.6
mBLIP mT0-XL	39.1	32.5	37.4	34.0	50.2	41.6	34.7	33.9	39.5	43.1	45.4
Aya-Vision-8B	54.2	53.5	62.9	50.2	72.5	61.3	57.6	47.8	57.0	57.9	70.8
Pangea-7B	57.2	56.0	72.7	45.8	77.2	62.5	61.6	52.9	59.5	64.9	72.2
CulturalPangea	57.2	61.0	72.7	50.2	80.2	65.4	60.7	52.6	57.0	64.2	75.0

Models	Romania-Romanian	Russia-Russian	Singapore-Chinese	South Korea-Korean	Spain-Spanish	Sri_Lanka-Sinhala	Indonesia-Minangkabau	Egypt-Egyptian Arabic	France-Breton	India-Marathi
Llava-Next-7B	62.6	58.5	62.3	60.0	67.6	59.1	51.4	54.7	37.5	N/A
Molmo-7B-D	52.0	63.5	66.0	56.9	66.7	31.6	43.4	43.8	29.6	N/A
Llama3.2-11B	75.5	74.5	73.6	73.1	83.3	51.1	58.2	56.7	36.3	N/A
PaliGemma-3B	60.9	56.0	59.4	58.3	61.0	62.2	43.4	51.2	37.3	N/A
mBLIP mT0-XL	43.7	41.0	43.9	41.4	51.9	48.0	38.6	42.9	30.4	N/A
Aya-Vision-8B	63.9	66.0	70.3	67.2	69.5	64.4	58.2	57.6	43.7	68.8
Pangea-7B	71.9	68.5	71.7	66.6	75.2	70.6	56.9	59.1	45.2	69.3
CulturalPangea	69.9	66.5	75.0	73.4	73.9	69.8	60.2	63.1	50.1	70.8

Table 13: CVQA benchmark performance (%) of multimodal models across 31 language-country pairs in English.

Models	English	Multi	Amharic	Bengali	Bulgarian	Chinese (Simpl.)	Chinese (Trad.)	Czech	Dutch	French
Llava-Next-7B	68.0	42.4	14.0	18.0	50.0	56.0	52.0	51.0	56.0	59.0
Molmo-7B-D	77.0	49.1	32.0	41.0	50.0	61.0	62.0	54.0	56.0	64.0
Llama3.2-11B	78.5	56.6	17.4	57.1	58.0	64.5	53.8	63.5	65.2	72.5
PaliGemma-3B	50.0	35.3	16.3	28.4	34.5	41.4	36.1	39.8	38.3	41.3
mBLIP mT0-XL	72.0	36.9	21.0	22.0	39.0	49.0	40.0	35.0	43.0	65.0
AyaVision-8B	71.3	55.1	8.8	34.3	53.2	64.8	56.9	66.4	70.7	74.3
Pangea-7B	71.5	59.9	25.7	53.2	59.8	67.3	60.9	63.2	64.0	75.9
CulturalPangea	75.9	63.5	31.3	58.4	63.5	70.0	63.2	65.4	62.9	74.8

Models	German	Greek	Hebrew	Hindi	Igbo	Indonesian	Irish	Italian	Japanese	Javanese
Llava-Next-7B	60.0	28.0	32.0	30.0	37.0	57.0	21.0	64.0	47.0	47.0
Molmo-7B-D	58.0	39.0	37.0	44.0	34.0	57.0	45.0	58.0	61.0	53.0
Llama3.2-11B	67.4	62.9	57.0	66.6	28.8	71.8	35.6	78.5	59.0	44.9
PaliGemma-3B	45.6	40.2	30.1	34.7	14.1	42.6	20.6	45.4	39.7	34.8
mBLIP mT0-XL	58.0	29.0	28.0	29.0	32.0	40.0	24.0	59.0	39.0	34.0
AyaVision-8B	69.3	69.6	64.1	64.6	22.6	70.9	18.9	74.9	68.7	61.6
Pangea-7B	71.4	57.9	55.3	58.1	47.9	67.6	45.2	66.2	69.3	67.9
CulturalPangea	73.1	59.5	66.2	62.4	51.9	70.2	54.3	76.6	71.1	69.1

Models	Korean	Malay	Mongolian	Norwegian	Persian	Polish	Portuguese	Romanian	Russian	Sinhala
Llava-Next-7B	41.0	57.0	28.0	57.0	33.0	57.0	62.0	52.0	51.0	18.0
Molmo-7B-D	52.0	57.0	36.0	50.0	42.0	58.0	54.0	50.0	61.0	39.0
Llama3.2-11B	54.4	70.8	24.8	66.2	59.0	66.3	66.3	66.7	62.2	37.0
PaliGemma-3B	41.1	39.3	16.9	41.0	37.6	38.0	41.4	40.4	32.6	13.8
mBLIP mT0-XL	30.0	44.0	25.0	34.0	28.0	33.0	60.0	34.0	44.0	21.0
AyaVision-8B	65.0	70.8	22.3	53.9	64.3	65.7	66.8	66.0	65.1	17.9
Pangea-7B	60.8	69.1	52.4	65.5	56.1	66.0	61.7	65.7	63.3	21.0
CulturalPangea	63.4	72.0	57.7	66.2	62.4	67.6	64.8	65.4	64.8	36.2

Models	Spanish	Sundanese	Swahili	Tamil	Telugu	Thai	Turkish	Ukrainian	Urdu	Vietnamese
Llava-Next-7B	59.0	37.0	23.0	18.0	18.0	23.0	43.0	55.0	25.0	48.0
Molmo-7B-D	58.0	50.0	41.0	23.0	24.0	40.0	51.0	50.0	44.0	50.0
Llama3.2-11B	66.6	39.4	41.0	44.1	51.8	59.2	64.0	63.8	54.5	64.5
PaliGemma-3B	38.3	32.2	37.9	25.8	32.7	42.5	46.6	47.1	31.2	36.5
mBLIP mT0-XL	60.0	38.0	33.0	26.0	30.0	26.0	31.0	28.0	28.0	27.0
AyaVision-8B	65.0	56.1	26.4	36.6	34.4	40.6	71.9	74.0	39.9	69.2
Pangea-7B	65.4	65.8	63.1	38.2	50.9	62.3	67.2	68.3	62.0	67.3
CulturalPangea	69.4	67.1	65.2	44.5	55.8	66.2	73.1	70.3	59.9	64.8

Table 14: ALMBench benchmark performance (%) of all multimodal models across 39 languages.

CulturalGround(French)

Marie Déa

Entity Info
Language: French
Entity name: Marie Déa
Country: France
Wikidata ID: Q3292505
Question type: Property-level question
Property: P1412 (languages spoken / written)
Template question: Quelle(s) langue(s) cette entité parle-t-elle ou écrit-elle ?(Which language(s) does this entity speak or write)
Template answer: Marie Déa parle ou écrit français, langue romane.(Marie Déa speaks or writes French, a Romance language)

Question
Quelle langue parle ou écrit l'actrice française que vous voyez sur cette photo ? (Which language does the French actress you see in this photo speak or write?)

Answer
Marie Déa parle et écrit en français, la langue romane de France. (Marie Déa speaks and writes French, the Romance language of France.)

Figure 20: Sample from in CulturalGround. Template question and answer are created from our curation pipeline and further

CulturalGround(Indonesian)

Nina Nesbitt

Entity Info

Language: Indonesian
Entity name: Nina Nesbitt
Country: United Kingdom
Wikidata ID: Q510361
Question type: Property-level question
Property: P27 (country of citizenship)
Template question: Negara manakah yang mengakui entitas ini sebagai warganya?
(Which country recognizes this entity as its citizen?)
Template answer: Nina Nesbitt diakui sebagai warga negara Britania Raya, negara di Eropa Barat Laut.
(Nina Nesbitt is recognized as a citizen of United Kingdom, country in north-west Europe.)

Question

Kalau melihat gambar ini, Nina Nesbitt berasal dari negara mana ya?
(Where is Nina Nesbitt from?)

Answer

Nina Nesbitt adalah warga negara Britania Raya, sebuah negara di Eropa Barat Laut.
Britania Raya dikenal dengan sejarah dan budayanya yang kaya, serta merupakan rumah bagi banyak musisi dan seniman berbakat seperti Nina Nesbitt.
(Nina Nesbitt is a singer-songwriter from the United Kingdom, specifically Scotland. She is known for her indie pop music and has gained recognition as a Scottish musical artist.)

CulturalGround(Spanish)

Vázquez Sounds

Entity Info

Language: Spanish
Entity name: Vázquez Sounds
Country: Mexico
Wikidata ID: Q2485066
Question type: Property-level question
Property: P495 (country of origin)
Template question: ¿Cuál es el país de origen de esta entidad?(What is the country of origin of this entity?)
Template answer: Vázquez Sounds se originó en México, país de América del Norte.(Vázquez Sounds originated in Mexico, a country in North America.)

Question

¿En qué país se originó este grupo musical?
(In which country did this music group originate?)

Answer

Vázquez Sounds se originó en México, un país de América del Norte.
(Vázquez Sounds originated in Mexico, a country in North America.)

CulturalGround(Spanish)

castro

Entity Info
Language: Spanish
Entity name: castro
Country: Spain
Wikidata ID: Q5050823
Question type: Property-level question
Property: P2596 (culture)
Template question: ¿Con qué cultura está asociada esta entidad?
(Which culture is this entity associated with?)
Template answer: El *castro* está asociado a la cultura castreña, una cultura arqueológica ibérica.
(The *castro* is associated with the Castro culture, an Iberian archaeological culture.)

Question
¿A qué cultura o tradición arqueológica se asocia normalmente este tipo de asentamiento?
(What culture built these kinds of settlements?)

Answer
El *castro* está asociado a la cultura castreña, una cultura arqueológica propia del noroeste de España. Estos asentamientos fortificados son característicos de la época prerromana y reflejan las formas de vida de las comunidades locales antes de la llegada del Imperio Romano.
(This *castro* is associated with the Castro culture, an archaeological culture primarily found in northern Spain. These fortified settlements were often constructed before the Roman period and represent a significant part of the region's pre-Roman history.)

CulturalGround(English)

Christ Church

Entity Info

Language: English
Entity name: Christ Church
Country: United Kingdom
Wikidata ID: Q19362878
Question type: Property-level question
Property: P140 (religion or worldview)
Template question: Which religion is associated with this entity?
Template answer: Christ Church is associated with Anglicanism, a Christian tradition developing out of the practices, liturgy and identity of the Church of England.

Question

What Christian tradition does this church belong to?

Answer

Christ Church belongs to the Anglican tradition, which developed from the practices and identity of the Church of England. As part of the Anglican faith, Christ Church observes a blend of Catholic and Protestant traditions and is common throughout the United Kingdom.

CulturalGround(German)

Muhammad-Bey-Abu-adh-Dhahab-Moschee

Entity Info

Language: German
Entity name: Muhammad-Bey-Abu-adh-Dhahab-Moschee
Country: Egypt
Wikidata ID: Q6916436
Question type: Property-level question
Property: P1435 (heritage designation)
Template question: Welchen Denkmalstatus hat diese Entität?
(What heritage designation does this entity have?)
Template answer: Muhammad-Bey-Abu-adh-Dhahab-Moschee hat den Denkmalstatus Teil einer UNESCO-Welterbestätte, Element einer UNESCO-Welterbestätte.
(The Muhammad Bek Abu El Dahab Complex has the heritage designation of being part of a UNESCO World Heritage Site, a component forming part of a UNESCO World Heritage Site.)

Question

Welchen Denkmalstatus hat diese Moschee?
(What special cultural recognition has this mosque received?)

Answer

Die Muhammad-Bey-Abu-adh-Dhahab-Moschee hat den Denkmalstatus als Teil einer UNESCO-Welterbestätte.
(The Muhammad Bek Abu El Dahab Complex has been recognized as part of a UNESCO World Heritage Site, highlighting its significant cultural and historical importance in Egypt.)

CulturalGround(Turkish)

Filibe Saat Kulesi(Clock Tower of Plovdiv)

Entity Info

Language: Turkish
Entity name: Filibe Saat Kulesi(Clock Tower of Plovdiv)
Country: Bulgaria
Wikidata ID: Q37939027
Question type: Entity-level question
Template question: Bu görüntüde gösterilen varlık nedir?(What is the entity shown in this image?)
Template answer: Bu görüntüde gösterilen varlık Filibe Saat Kulesi'dir, .(The entity shown in the image is Clock Tower of Plovdiv, .)

Question

Bu fotoğraftaki yapı ne olarak bilinir?(What is the structure in this photo known as?)

Answer

Bu fotoğraftaki yapı Filibe Saat Kulesi'dir. Bulgaristan'ın Filibe şehrinde yer alan bu kule, şehrin sembol yapılarından biridir ve Osmanlı döneminden kalmıştır. (The structure in this photo is the Plovdiv Clock Tower. It is located in Plovdiv, Bulgaria, and is one of the city's landmark buildings, dating back to the Ottoman period.)

CulturalGround(Japanese)

Shinnakagawa-machi Station

Entity Info

Language: Japanese
Entity name: 新中川町停留場(Shinnakagawa-machi Station)
Country: Japan
Wikidata ID: Q11501118
Question type: Entity-level question
Template question: この画像に表示されているものは何ですか？?(What is the entity shown in this image?)
Template answer: この画像に表示されているのは新中川町停留場です。長崎県長崎市にある長崎電気軌道の路面電車停留場.(The entity shown in the image is Shin Nakagawa-Machi Station, tram station in Nagasaki, Nagasaki prefecture, Japan.)

Question

画像に映っている停留場はどこにありますか (What type of transportation facility is shown in this image?)

Answer

画像に映っているのは新中川町停留場で、長崎県長崎市にある長崎電気軌道の路面電車停留場です。 (Shin Nakagawa-Machi Station is a tram station located in Nagasaki, Nagasaki prefecture, Japan.)

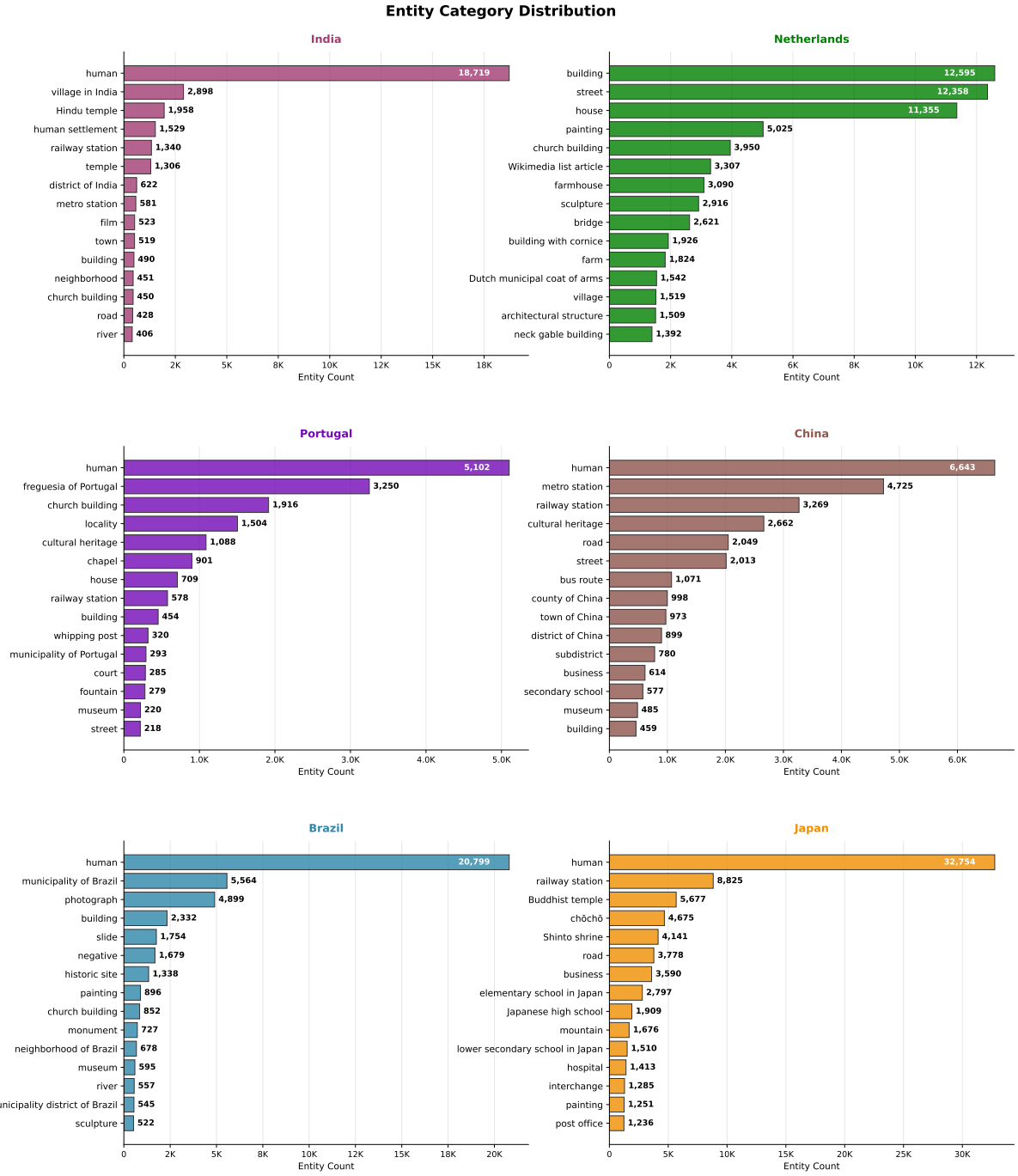

Figure 21: Entity–category distributions for six representative countries reveal distinct regional cultural emphases. India focuses on heritage and settlements (e.g., Hindu temples, villages); the Netherlands highlights vernacular architecture and art (e.g., historic buildings, streets, paintings); Portugal surfaces ecclesiastical heritage and protected cultural assets (e.g., church buildings, chapels, registered heritage sites); China emphasizes infrastructure and administrative categories (e.g., metro and railway stations, county- and town-level entities); Brazil shows strong cultural media and monument presence (e.g., photographs, negatives, historic sites); and Japan combines religious architecture with education-related categories (e.g., Buddhist temples, Shinto shrines, schools). Bar length shows entity counts, with longer bars indicating more frequent categories within each country.

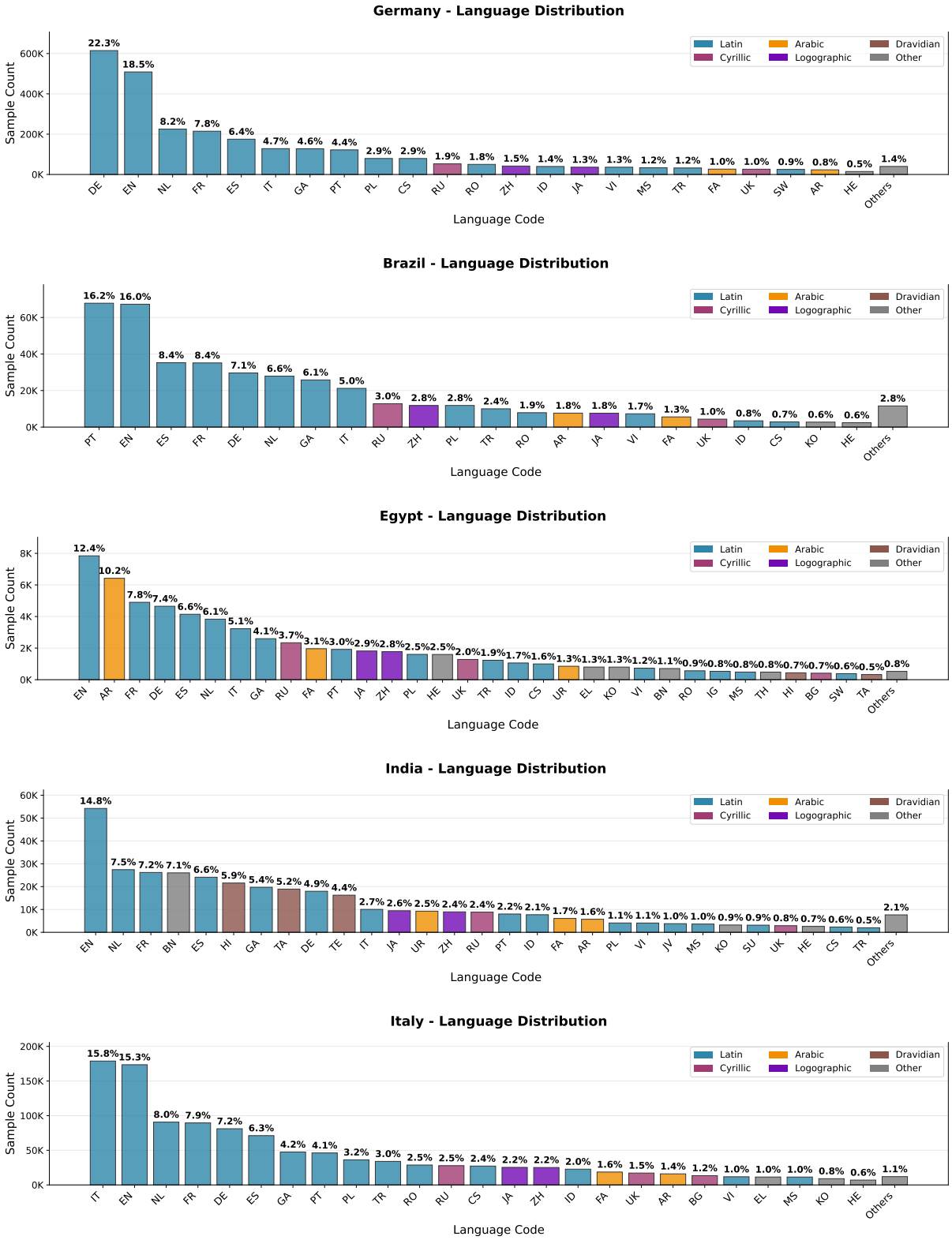

Figure 22: Language distribution across five countries showing the diversity of multilingual content in CulturalGround. Each plot displays languages with less than 0.5% representation, with colors indicating script families. The percentages above each bar indicate the proportion of samples in each language within the respective country's subset.

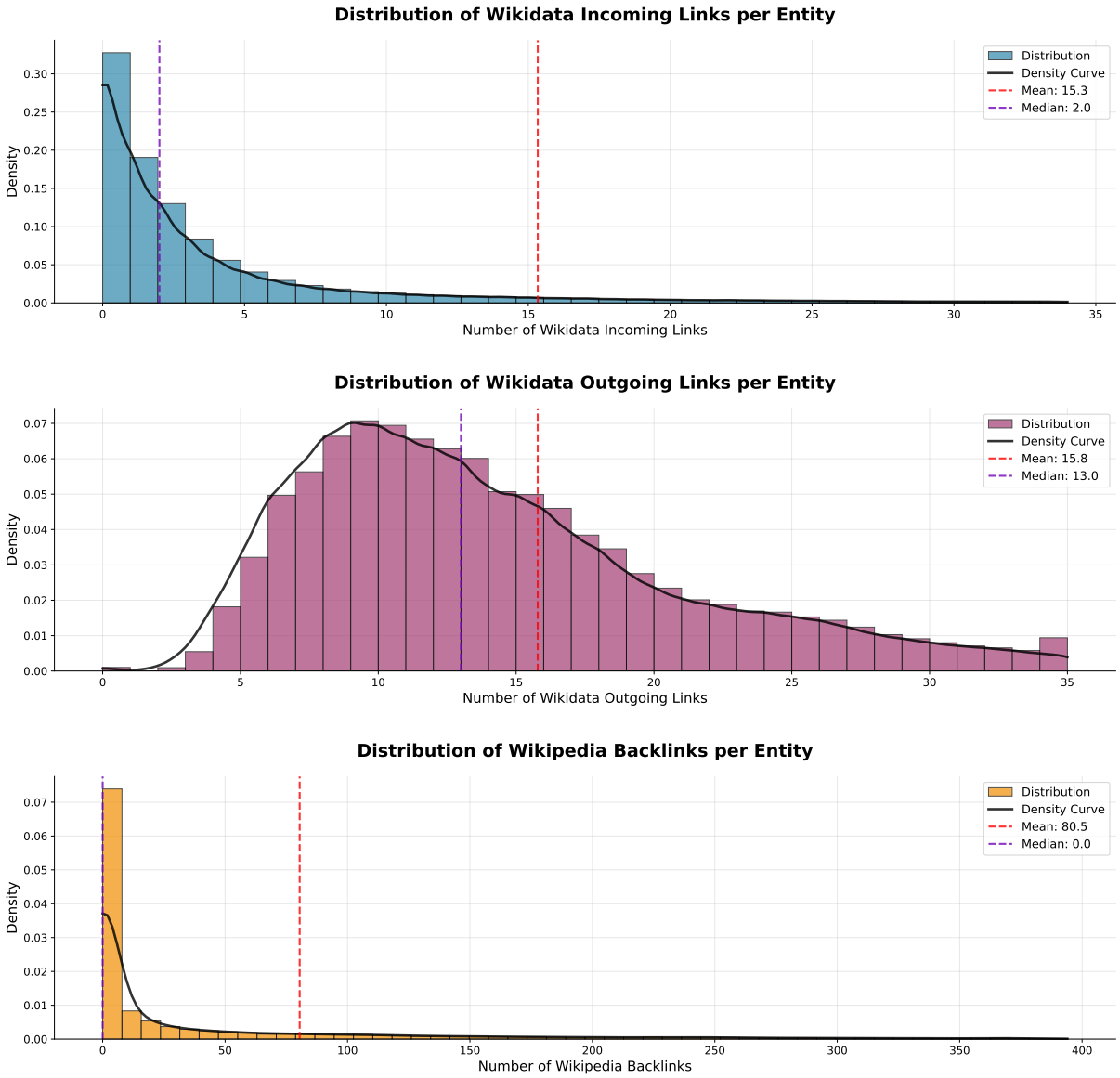

Figure 23: Distribution of entity connectivity in CulturalGround, demonstrating our commitment to treating long-tail entities as first-class citizens. The plots show (top) incoming Wikidata links, (middle) outgoing Wikidata links, and (bottom) Wikipedia backlinks across all countries. Unlike previous datasets that focus primarily on highly popular entities, CulturalGround includes substantial representation of entities with few or no links, as evidenced by the significant density at lower links values.

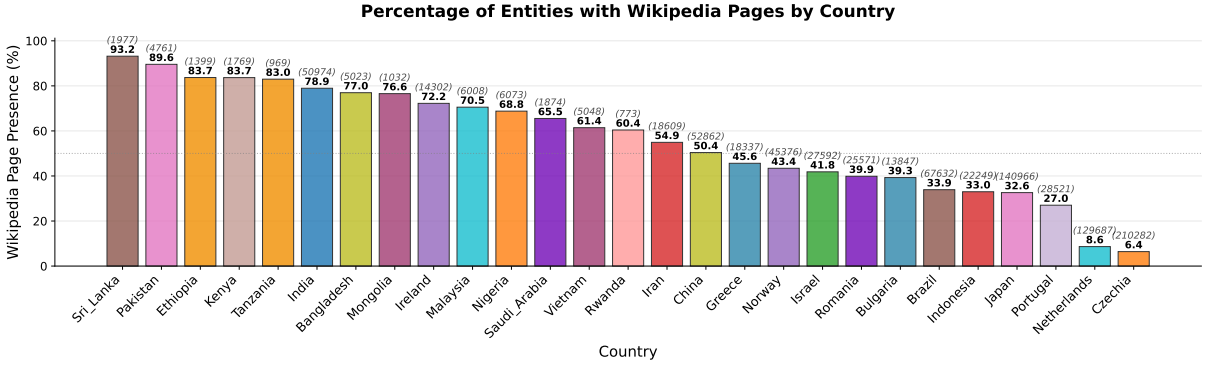

Figure 24: The proportion of entities that have Wikipedia page in some regions in CulturalGround. The majority of entities don't have Wikipedia presence, confirming that our data curation pipeline captures long-tail entities

Property	Label	Description	Template question	Answer template
P17	country	sovereign state of this item; don't use on humans	Which sovereign state does this entity belong to?	{entity_label} belongs to the sovereign state of {property_value}.
P2596	culture	culture associated with this entity	Which culture is this entity associated with?	{entity_label} is associated with {property_value} culture.
P172	ethnic group	ethnic group of which an entity is a member	Which ethnic group is this entity a member of?	{entity_label} is a member of the {property_value} ethnic group.
P140	religion	religion of a person, organization or religious building, or associated with this subject	Which religion is associated with this entity?	{entity_label} is associated with the {property_value} religion.
P1559	name in native language	name of the entity in its native or original language	What is the name of this entity in its native language?	The name of {entity_label} in its native language is {property_value}.
P37	official language	language designated as official by this item	Which language is officially designated by this entity?	{entity_label} officially designates the language {property_value}.
P103	native language	language or languages a person has learned from early childhood	Which language did this entity learn from early childhood?	{entity_label} learned {property_value} from early childhood.
P825	dedicated to	subject is dedicated to this entity	What is this entity dedicated to?	{entity_label} is dedicated to {property_value}.
P149	architectural style	architectural style of a building or structure	What is the architectural style of this entity?	{entity_label} is built in the {property_value} architectural style.
P1435	heritage designation	heritage designation of a cultural site or monument	What heritage designation does this entity have?	{entity_label} has the heritage designation of {property_value}.
P1268	represents	entity, concept or subject represented by this item	What does this entity represent?	{entity_label} represents {property_value}.
P837	day in year for periodic occurrence	date in the year when a periodic event occurs	On which day of the year does this event occur?	{entity_label} occurs on {property_value}.
P495	country of origin	country of origin of this item (creative work, food, phrase, product, etc.)	Which country is the origin of this entity?	{entity_label} originated in {property_value}.
P407	language of work or name	language associated with this creative work	Which language is associated with this entity?	{entity_label} is associated with the {property_value} language.
P1412	languages spoken, written or signed	language(s) that a person speaks or writes, including the native language(s)	Which language(s) does this entity speak or write?	{entity_label} speaks or writes {property_value}.
P2936	language used	language widely used in this place or by this organization	Which language is used by this entity?	{entity_label} uses the language {property_value}.
P131	located in the administrative territorial entity	the item is located on the territory of the following administrative entity	Within which administrative territorial entity is this entity located?	{entity_label} is located in {property_value}.
P276	location	location of the item, physical object or event	Where is this entity located?	{entity_label} is located in {property_value}.
P625	coordinate location	geocoordinates of the subject	What are the coordinates of this entity?	{entity_label} is located at coordinates {property_value}.
P706	located in/on physical feature	located on the given landform or body of water	On which physical feature is this entity located?	{entity_label} is located on {property_value}.
P206	located in or next to body of water	body of water on or next to which a place is located	Which body of water is this entity located in or next to?	{entity_label} is located in or next to {property_value}.
P30	continent	continent of which the subject is a part	On which continent is this entity located?	{entity_label} is located on the continent of {property_value}.
P170	creator	maker of this creative work or other object	Who is the creator of this entity?	{entity_label} was created by {property_value}.
P86	composer	person(s) who wrote the music for this song, musical work, or opera	Who composed the music for this entity?	The music for {entity_label} was composed by {property_value}.
P162	producer	person(s) who produced the film, musical work, theatrical production, etc.	Who produced this entity?	{entity_label} was produced by {property_value}.
P136	genre	creative work's genre or an artist's field of work	What genre is associated with this entity?	{entity_label} is associated with the {property_value} genre.
P571	inception	time when an entity begins to exist; for date of official opening use P1619	When was this entity established or founded?	{entity_label} was established in {property_value}.
P585	point in time	date or point in time when an event occurred	When did this event occur?	{entity_label} occurred on {property_value}.
P1269	facet of	topic of which this item is an aspect	This entity is a facet of which broader topic?	{entity_label} is a facet of {property_value}.
P19	place of birth	most specific known birth location of a person, animal or fictional character	Where was this entity born?	{entity_label} was born in {property_value}.
P20	place of death	most specific known death location of a person, animal or fictional character	Where did this entity die?	{entity_label} died in {property_value}.
P27	country of citizenship	the object is a country that recognizes the subject as its citizen	Which country recognizes this entity as its citizen?	{entity_label} is recognized as a citizen of {property_value}.
P569	date of birth	date on which the subject was born	When was this entity born?	{entity_label} was born on {property_value}.
P570	date of death	date on which the subject died	When did this entity die?	{entity_label} died on {property_value}.
P36	capital	primary city of a country, state or other type of administrative territorial entity	What is the capital of this entity?	The capital of {entity_label} is {property_value}.
P1376	capital of	administrative division of which the municipality is the governmental seat	Of which administrative division is this entity the capital?	{entity_label} is the capital of {property_value}.
P47	shares border with	countries or administrative subdivisions that this item borders	Which entity does this share a border with?	{entity_label} shares a border with {property_value}.
P106	occupation	occupation of a person	What is the occupation of this entity?	{entity_label}'s occupation is {property_value}.
P39	position held	position or public office currently or formerly held	Which position does this entity hold?	{entity_label} holds the position of {property_value}.
P102	member of political party	the political party of which this politician is or has been a member	Which political party is this entity a member of?	{entity_label} is a member of the {property_value} party.
P166	award received	award or recognition received by a person, organisation or creative work	Which award did this entity receive?	{entity_label} received the following award(s): {property_value}.
P800	notable work	notable scientific, artistic or literary work among subject's works	What is a notable work by this entity?	A notable work by {entity_label} is {property_value}.
P1303	instrument	musical instrument that a person plays	Which musical instrument does this entity play?	{entity_label} plays the {property_value}.
P641	sport	sport that the subject participates in or is associated with	In which sport does this entity participate?	{entity_label} participates in {property_value}.
P54	member of sports team	sports teams that the subject represents or represented	Which sports team does this entity represent?	{entity_label} represents the sports team {property_value}.

(continued)

Property	Label	Description	Template question	Answer template
P69	educated at	educational institution attended by subject	Which educational institution did this entity attend?	{entity_label} studied at {property_value}.
P512	academic degree	academic degree that the person holds	Which academic degree does this entity hold?	{entity_label} holds the academic degree of {property_value}.
P101	field of work	specialization of a person or organization	In which field does this entity work?	{entity_label} works in the field of {property_value}.
P108	employer	person or organization for which the subject works	Who employs this entity?	{entity_label} is employed by {property_value}.
P937	work location	location where persons were active	In which location was this entity active?	{entity_label} was active in {property_value}.
P31	instance of	class of which this subject is a particular example	This entity is an example of which class?	{entity_label} is an instance of {property_value}.
P279	subclass of	this item is a class (subset) of that item	Of which broader class is this entity a subclass?	{entity_label} is a subclass of {property_value}.
P361	part of	object of which the subject is a part	Which larger entity is this entity part of?	{entity_label} is part of {property_value}.
P527	has part	part of this subject	Which parts does this entity include?	{entity_label} consists of {property_value}.
P138	named after	entity or event that inspired the subject's name	What is this entity named after?	{entity_label} was named after {property_value}.
P577	publication date	date when a work was first published or released	When was this entity published?	{entity_label} was published on {property_value}.
P1619	date of official opening	date when a place or organization officially opened	When was this entity officially opened?	{entity_label} was officially opened on {property_value}.
P740	location of formation	location where a group or organization was formed	Where was this entity formed?	{entity_label} was formed in {property_value}.
P159	headquarters location	location where an organization's headquarters is situated	Where are the headquarters of this entity located?	The headquarters of {entity_label} is located in {property_value}.
P793	significant event	significant or notable events associated with the subject	What significant events are associated with this entity?	Notable events associated with {entity_label} include {property_value}.
P463	member of	organization or club to which the subject belongs	Which organization is this entity a member of?	{entity_label} is a member of {property_value}.
P190	twinned administrative body	twin towns or sister cities	Which city is twinned with this entity?	{entity_label} is twinned with {property_value}.
P530	diplomatic relation	diplomatic relations of the country	With which countries does this entity maintain diplomatic relations?	{entity_label} maintains diplomatic relations with {property_value}.
P176	manufacturer	manufacturer or producer of this product	Who manufactures this entity?	{entity_label} is produced by {property_value}.
P178	developer	organisation or person that developed the item	Who developed this entity?	{entity_label} was developed by {property_value}.
P127	owned by	owner of the subject	Who owns this entity?	{entity_label} is owned by {property_value}.
P137	operator	entity that operates the equipment, facility, or service	Who operates this entity?	{entity_label} is operated by {property_value}.
P449	original network	network the radio or television show was originally aired on	On which network was this show originally aired?	{entity_label} was originally aired on the {property_value} network.
P264	record label	brand associated with the marketing of music recordings	Under which record label is this entity's music released?	{entity_label}'s music is released under {property_value}.
P364	original language of film or TV show	language in which a film or performance work was originally created	In which language was this entity originally created?	{entity_label} was originally created in {property_value}.
P180	depicts	entity visually depicted in an image or work	What does this entity depict?	This entity depicts {property_value}.
P921	main subject	primary topic of a work	What is the main subject of this entity?	The main subject of {entity_label} is {property_value}.
P1433	published in	larger work that a given work was published in	In which larger work was this entity published?	{entity_label} was published in {property_value}.
P413	position played on team / speciality	position or specialism of a player on a team	What position does this entity play?	{entity_label} plays in the position {property_value}.
P1923	participating team	teams that participated in an event	Which teams participated in this event?	Teams participating in {entity_label} include {property_value}.
P1001	applies to jurisdiction	territorial jurisdiction that an institution or law applies to	Under which jurisdiction does this entity operate?	{entity_label} operates under the jurisdiction of {property_value}.

Table 15: Culturally relevant Wikidata properties with corresponding template questions and answers used for multilingual VQA generation.

Country/Region	Entities	Images	Template QA	Open-Ended	MCQ	Open-Ended$_F$	MCQ$_F$
Germany	332,650	350,828	2,752,048	2,835,679	965,541	1,506,438	426,272
France	268,298	276,983	2,676,838	2,729,262	941,466	1,435,627	528,449
United Kingdom	175,486	328,906	1,355,577	2,183,466	891,282	1,319,135	469,302
Italy	128,821	222,351	1,133,463	1,763,658	745,977	1,323,626	653,884
Spain	124,280	216,019	985,241	1,519,295	616,304	906,943	545,056
Japan	82,690	145,843	793,759	1,214,762	483,233	799,963	431,739
Czechia	110,384	198,223	636,978	994,864	401,437	679,115	380,160
Poland	98,577	131,155	753,750	936,799	361,028	529,669	328,143
Russia	119,158	180,253	613,822	848,540	343,834	628,558	311,416
India	29,574	72,683	365,804	717,067	218,854	542,516	270,301
Brazil	38,575	68,775	419,684	648,164	257,966	479,162	236,749
Ukraine	57,665	100,367	367,819	562,770	224,044	421,096	207,434
China	38,435	68,858	288,524	468,916	200,950	365,277	187,660
Norway	27,632	47,615	255,226	382,264	146,757	273,697	118,463
Netherlands	72,709	72,709	375,078	375,020	119,563	225,651	114,602
Mexico	12,224	29,724	184,998	370,152	113,682	271,408	122,758
Israel	19,689	33,731	183,099	289,430	124,912	233,556	105,840
Romania	15,408	26,451	196,705	287,122	109,326	194,952	104,126
Indonesia	9,026	22,060	145,832	256,309	66,731	148,594	79,859
Turkey	13,610	23,876	163,963	256,350	107,366	183,648	99,250
Iran	12,930	32,496	114,996	252,235	80,307	194,867	103,478
Greece	9,975	24,887	125,163	250,048	76,779	172,912	95,873
Portugal	19,733	35,229	155,542	237,166	94,069	162,184	93,708
South Korea	8,809	15,175	149,796	209,911	71,649	123,550	65,233
Ireland	9,115	22,856	86,838	185,033	58,225	146,654	72,337
Bulgaria	7,167	17,315	94,452	177,989	54,002	129,713	64,048
Taiwan	12,644	33,410	71,483	166,306	54,930	142,712	70,085
Egypt	3,920	9,596	63,237	136,891	43,655	104,816	48,698
Thailand	5,837	15,037	58,397	125,292	39,345	101,078	49,959
Pakistan	2,851	6,973	38,005	76,927	24,085	59,778	29,507
Malaysia	3,858	9,788	38,208	79,684	24,666	63,484	31,065
Nigeria	2,519	6,368	42,080	77,164	21,339	53,213	25,948
Bangladesh	3,659	9,236	29,253	62,700	20,382	51,071	25,715
Vietnam	3,230	5,744	37,035	58,513	24,297	43,626	21,855
Singapore	1,752	4,298	23,619	54,281	17,059	41,825	19,176
Saudi Arabia	948	2,292	17,759	35,046	10,772	26,547	13,087
Kenya	1,120	2,763	17,251	36,337	11,412	29,164	14,657
Ethiopia	880	2,163	14,244	29,976	9,551	23,713	10,955
Sri Lanka	1,066	2,651	14,643	29,484	8,861	22,177	10,913
Tanzania	592	1,454	11,966	26,332	8,451	17,689	11,589
Mongolia	542	1,306	12,482	23,604	6,900	16,429	8,765
Rwanda	572	1,393	7,332	15,693	5,157	11,850	5,821
Total	**1,888,610**	**2,879,840**	**15,871,989**	**21,986,501**	**8,206,146**	**14,207,683**	**6,613,935**

Table 16: Dataset statistics by country/region. The dataset contains culturally significant entities from Wikidata with 1-3 images per entity and questions generated from 76 cultural properties. Unique entities shown are those with images, text-only entities excluded. F in last two columns means filtered data.

Language	Open-Ended	MCQ	Open-Ended$_F$	MCQ$_F$
en (English)	3,778,963	1,369,758	2,501,144	1,152,830
fr (French)	1,822,466	668,153	1,181,935	530,004
de (German)	1,782,256	626,116	1,083,314	469,522
nl (Dutch)	1,648,445	602,869	1,053,835	487,091
es (Spanish)	1,415,511	508,136	878,913	412,530
it (Italian)	1,114,458	430,928	745,316	347,233
ga (Irish)	964,614	357,266	615,712	282,814
pl (Polish)	818,624	312,878	511,913	245,297
ru (Russian)	849,610	336,357	553,662	277,540
pt (Portuguese)	872,402	324,938	542,464	244,671
cs (Czech)	781,353	285,846	480,799	233,627
ja (Japanese)	685,032	267,259	441,822	215,680
zh (Chinese)	728,825	286,369	491,016	236,206
tr (Turkish)	640,652	246,485	415,126	194,963
uk (Ukrainian)	526,988	208,179	346,493	172,357
ro (Romanian)	366,781	141,767	242,138	105,055
fa (Persian)	362,570	145,847	241,236	115,227
id (Indonesian)	347,249	130,057	223,098	100,871
ar (Arabic)	346,263	134,798	229,576	110,000
vi (Vietnamese)	298,369	118,273	199,562	87,990
ko (Korean)	256,574	104,499	172,769	84,691
he (Hebrew)	221,549	91,434	150,173	71,602
ms (Malay)	243,026	93,024	161,397	69,462
el (Greek)	166,436	64,092	102,493	50,733
bg (Bulgarian)	139,184	55,340	92,766	45,780
bn (Bengali)	137,984	48,763	95,023	46,212
ur (Urdu)	97,025	37,085	65,567	33,166
hi (Hindi)	77,997	27,260	57,202	29,295
sw (Swahili)	128,935	46,746	77,641	32,988
ta (Tamil)	75,908	27,264	53,259	26,670
th (Thai)	85,927	33,369	58,433	30,558
te (Telugu)	55,477	20,068	38,837	20,015
jv (Javanese)	58,164	21,218	39,747	19,933
su (Sundanese)	30,857	10,840	21,238	10,583
ig (Igbo)	23,854	8,278	16,154	7,729
si (Sinhala)	16,828	6,687	12,407	6,306
mn (Mongolian)	13,495	5,605	9,650	4,682
am (Amharic)	3,975	1,627	2,704	1,483
no (Norwegian)	1,875	668	1,149	539
TOTAL	**21,986,501**	**8,206,146**	**14,207,683**	**6,613,935**

Table 17: Data distribution statistics by language, showing unfiltered and filtered counts for Open-Ended and Multiple-Choice Question (MCQ) items.