

GIMMICK

Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking

Florian Schneider¹, Carolin Holtermann², Chris Biemann¹, Anne Lauscher²

¹Language Technology Group, University of Hamburg

²Data Science Group, University of Hamburg

florian.schneider-1@uni-hamburg.de

Abstract

Large Vision-Language Models (LVLMs) have recently gained attention due to their distinctive performance and broad applicability. While it has been previously shown that their efficacy in usage scenarios involving non-Western contexts falls short, existing studies are limited in scope, covering just a narrow range of cultures, focusing exclusively on a small number of cultural aspects, or evaluating a limited selection of models on a single task only. Towards globally inclusive LVL research, we introduce GIMMICK, an extensive multimodal benchmark designed to assess a broad spectrum of cultural knowledge across 144 countries representing six global macro-regions. GIMMICK comprises six tasks built upon three new datasets that span 728 unique cultural events or facets on which we evaluated 20 LVLMs and 11 LLMs, including five proprietary and 26 open-weight models of all sizes. We systematically examine (1) regional cultural biases, (2) the influence of model size, (3) input modalities, and (4) external cues. Our analyses reveal strong biases toward Western cultures across models and tasks and highlight strong correlations between model size and performance, as well as the effectiveness of multimodal input and external geographic cues. We further find that models have more knowledge of tangible than intangible aspects (e.g., *food* vs. *rituals*) and that they excel in recognizing broad cultural origins but struggle with a more nuanced understanding.¹

1 Introduction

Recently, proprietary as well as open-weight Large Vision-Language Models (LVLMs) (OpenAI, 2023; Liu et al., 2023; Wang et al., 2024; Chen et al., 2023, *inter alia*) have attracted marked attention due to their broad applicability across various domains. Several large-scale holistic benchmarks (Duan et al., 2024; Yue et al., 2024; Fu et al.,

2023) demonstrate LVLMs’ remarkable performances in a wide range of multimodal tasks. However, most benchmarks concentrate on Western-centric English tasks, and multilingual benchmarks (Ahuja et al., 2024; Schneider and Sitaram, 2024) reveal a significant deterioration in performance on non-English tasks. While multilingualism is essential for globally equitable AI, *multiculturalism* (Gabriel, 2020; Adilazuarda et al., 2024) is equally crucial for models to reflect and respect the diverse cultural backgrounds of users worldwide. In this context, it has been shown that current LLMs (Myung et al., 2024; Chiu et al., 2024) and LVLMs suffer in tasks involving knowledge from non-Western cultures. However, the scope of existing multimodal cultural studies is still severely limited: Existing research often focuses only on specific concepts like food or dance (Winata et al., 2025; Burda-Lassen et al., 2025), covers a limited number of cultures (Uraierprasert et al., 2024; Baek et al., 2024), evaluates only a small selection of LVLM models (Cao et al., 2024; Nayak et al., 2024), or tests only a single combination of input modalities.

To address these gaps, we introduce GIMMICK, a comprehensive evaluation framework assessing 31 state-of-the-art models, ranging from proprietary LVLMs to open-weight LLMs and LVLMs of all sizes—from 500M to 78B parameters—across multiple model families. It comprises six tasks built on three novel datasets that contain 728 unique cultural events or facets (CEFs) from 144 countries in six global macro-regions and target both high-level and nuanced cultural knowledge through multimodal and unimodal tasks. Our VQA tasks span a total of 57 cultural aspects (see §B.2) Ultimately, GIMMICK enables us to answer four research questions:

(RQ1) Are there regional biases in LLMs’ and LVLMs’ cultural knowledge, and if so, which?

For the most complex tasks, we observe consis-

¹<http://github.com/floschne/gimmick>

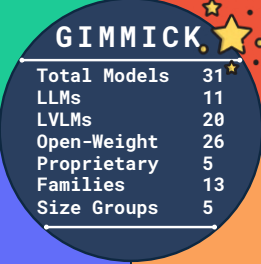
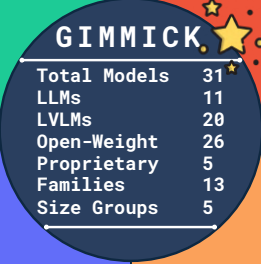
Cultural Image VQA					Cultural Origin QA			
Type	Open-Ended	Samples	2233		Type	Multi. Choice	Samples	982/759
Input	I+T	Images	1928		Input	I+T, T, I	Images	6857
Label	Answer Word	Cult. Events	635		Label	Choice Letter	Cult. Events	728
Score	Accuracy	Countries	144		Score	Accuracy	Countries	144
Cultural Video VQA					Cultural Knowledge QA			
Type	Open-Ended	Samples	1809		Type	Open/Long Form	Samples	728
Input	V+T	Videos	1809		Input	I+T, T, I	Images	6857
Label	Answer Word	Cult. Events	553		Label	Title/Desc.	Cult. Events	635
Score	Accuracy	Countries	139		Score	Judge Score	Countries	144

Figure 1: An overview of the GIMMICK benchmark and its tasks.

tent cultural regional biases (up to 14.72pp difference between instances targeting Western Europe & North America vs. Sub-Saharan Africa; §5.1) – even for the largest models. For less complex tasks, these differences flatten out.

(RQ2) To what degree does model size influence performance? We show that increasing the number of parameters significantly boosts performance on complex tasks, with larger models exhibiting less regional biases (§5.2). Still, even the largest models still struggle with nuanced cultural understanding.

(RQ3) How do input modalities affect cultural understanding? We observe that providing input in multiple modalities typically leads to the best results, as models leverage the cultural cues present in the visual inputs we provide (§5.3). Interestingly, on text-only tasks, LVLMs perform consistently worse than their LLM backbones, indicating a loss of cultural knowledge during integration training.

(RQ4) What is the influence of external cultural cues? We demonstrate that providing country information consistently guides the models towards better answers, especially for regions for which the models perform poorly (§5.4). Overall, with GIMMICK, we hope to encourage more research on culturally-aware and more globally-inclusive AI.

2 Related Work

Multicultural LLM Benchmarks. Naous et al. (2024) introduce CAMEL, a dataset that contrasts Arab and Western cultures to measure cultural biases in LLMs through extrinsic and intrinsic evaluations on core NLP tasks. With CultureAtlas, Fung et al. (2024) introduced an approach for massively multicultural knowledge acquisition and benchmarking of 5 LLMs from Wikipedia articles on cultural topics. BLEnD (Myung et al., 2024) is a large benchmark to evaluate LLMs’ everyday knowledge across diverse cultures and from 16

BENCHMARK	#M	#DS	#T	#S	#C	#R	MODS
SEA-VQA	2	1	1	1,999	8	1	T+I
Urailertprasert et al. (2024)							
WorldCuisines	18	1	2	1.15M	189	6	T+I
Winata et al. (2025)							
CROPE	17	1	1	1,060	6	3	T+I
Nikandrou et al. (2025)							
CulturalVQA	8	1	1	2,378	11	5	T+I
Nayak et al. (2024)							
Ananthram et al. (2025)	10	–	3	–	2	2	T+I
GlobalRG	12	2	2	3,591	51	6	T+I
Bhatia et al. (2024)							
MOSAIC-1.5K	4	1	1	1,500	–	6	T+I
Burda-Lassen et al. (2025)							
FoodieQA	8	1	3	1,839	1	1	T+I
Li et al. (2024)							
Cao et al. (2024)	1	–	3	–	5	3	T+I
K-VISCUIT	13	1	1	657	1	1	T+I
Back et al. (2024)							
CVQA	8	1	1	10,374	30	6	T+I
Romero et al. (2024)							
CulturalBench	30	2	1	6,135	45	6	T+I
Chiu et al. (2024)							
GIMMICK (ours)	31	3	6	7,239	144	6	T+I V+T T, I

Table 1: A comparative overview of recent benchmarks assessing cultural knowledge of LVLMs. The abbreviations in the columns stand for the (combined) number of: (unique) **M**odels, **D**atasets, **T**asks, **S**amples, **C**ountries, or **R**egions contained. The **M**odalities column lists the input modalities—Text, Image, Video—contained.

countries in 13 different languages. (Mukherjee et al., 2024) test four popular LLMs with culturally sensitive and non-sensitive prompts on both sensitive and neutral datasets. Instead of assessing models’ intrinsic cultural knowledge, (Bhatt and Diaz, 2024) focuses on the extrinsic evaluation of cultural competence, e.g., in user-interaction, in two text generation tasks, open-ended question answering, and story generation of 6 LLMs.

Multicultural LVLm Benchmarks. Bhatia et al. (2024) introduced the GlobalRG benchmark, which comprises two tasks: retrieving culturally diverse images for universal concepts from 50 countries and grounding culture-specific concepts within images from 15 countries. Karamolegkou et al. (2024) proposed a culture-centric evaluation benchmark investigating the reliability of LVLms as visual as-

sistants for blind people in a culturally diverse setting. Using the CulturalVQA (Nayak et al., 2024), the authors assessed geo-diverse cultural understanding of nine “1st-Gen” LVLMs on a curated dataset of 2,378 VQA pairs representing cultures from 11 countries and five cultural aspects. CulturalBench (Chiu et al., 2024) is a dataset of 1,227 human-written and human-verified questions for evaluating LLMs’ cultural knowledge, covering 45 global “regions”. Nikandrou et al. (2025) propose CROPE, a VQA benchmark designed to probe the knowledge of culture-specific concepts and evaluate the capacity for cultural adaptation through contextual information featuring over 1M data points across 30 languages and dialects. See Table 1 for an overview and a comparison or related work with GIMMICK.

Multilingual Multicultural LVLMM Benchmarks. Several studies evaluate the cultural awareness and capabilities of LVLMMs in a multilingual setting. Geigle et al. (2025) extensively benchmarked state-of-the-art LVLMMs across multiple multilingual and multicultural datasets, including MaRVL (Liu et al., 2021), XM3600 (Thapliyal et al., 2022) and MaXM (Changpinyo et al., 2023), M5B-VGR and M5B-VLOD (Schneider and Sitaram, 2024), CVQA (Romero et al., 2024) Winata et al. (2025) created WorldCuisines, a large-scale benchmark for multilingual and multicultural VQA on global cuisines. However, in GIMMICK, we focus on the English language, considering English performance as an upper bound.

3 The GIMMICK Benchmark

Cultural Benchmark Positioning Adilazuarda et al. (2024) surveyed 90+ recent papers on cultural awareness in LLMs and found that *none* explicitly define “culture”. Instead, these studies evaluate models on datasets capturing only specific cultural aspects, which the authors organize into two dimensions: *demographic* and *semantic* proxies (with seven and five subsets, respectively). In GIMMICK, we adopt the proposed taxonomy by using countries and regions as *demographic* cultural proxies. Our tasks span all five *semantic* proxies: “emotions and values”, “food and drink”, “social and political relations”, “basic actions and technology”, and “names”. We implement primarily “black-box” generative and discriminative probing approaches.

UNESCO Intangible Cultural Heritage. All tasks in GIMMICK are based on high-quality open-







REGION	ABBRV.	#C	#CEF
Arab	 A	18	76
Asia & Pacific	 AP	35	226
Eastern Europe	 E	25	150
Latin-America & Caribbean	 LAC	28	98
Subsaharian Africa	 SA	40	73
Western Europe & North America	 W	23	149
<i>Unique</i>		144	728

Table 2: Regions within GIMMICK. #C and #CEF stand for the number of Countries and CEFs related to the respective region. Some CEFs may span multiple regions.

access data from the UNESCO Intangible Cultural Heritage (ICH) project², which aims to safeguard cultural traditions and practices vital to the identity and heritage of communities worldwide while honoring cultural diversity. Intangible cultural heritage encompasses oral traditions, performing arts, rituals, festive events, traditional craftsmanship, and cultural knowledge. The open-access dataset is structured as a knowledge graph, where most nodes represent cultural events or facets (CEFs; e.g., *Yukitsumugi*, a silk fabric production technique from Japan³), with additional nodes including countries, regions, case studies in which the CEFs occur. For GIMMICK, we extract the CEFs, each together with their title, description, associated macro-regions and countries, and several images depicting different aspects of the CEF. Moreover, each CEF is detailed in one or more YouTube videos. In total, GIMMICK contains 728 CEFs from 144 countries represented by 6,887 images and 993 videos⁴. While most CEFs (88.60%) are associated with one country, some are associated with two or more countries. The UNESCO ICH project groups the countries into six global macro-regions⁵, which we adopt in this work. Throughout the paper—including all figures and tables—we use the region abbreviations listed in Table 2.

3.1 Datasets and Tasks

We created three novel multimodal datasets that serve as the foundation for six tasks designed to evaluate the cultural knowledge of models. See Figure 1 for an overview of the different tasks.⁶

²<https://ich.unesco.org>

³More examples including images are shown in §A.2.1

⁴We provide licensing details in §A.1

⁵We provide a comprehensive list in Table 4 in §A.3

⁶Sample counts per task & region are shown in §A.3.1

3.2 Cultural Image VQA

In the Cultural Image VQA (CIVQA) task, models are presented with an image depicting a CEF and a question that relates to a particular CEF aspect (see §B.1 for examples). Models are evaluated based on answer correctness. To create the data for CIVQA, we couple synthetic data generation with a two-stage annotation process.

Synthetic Data Generation. Building on the high-quality UNESCO ICH data, we applied synthetic data generation by prompting GPT-4o⁷ to construct the basis for our dataset. Each VQA pair is related to a CEF and consists of an image depicting one aspect of the CEF, a question related to the CEF and the image, and an answer. Maximizing the quality of the generated silver data, we applied extensive prompt engineering combining techniques such as Few-Shot, Chain-of-Thought, ReAct (Wei et al., 2022; Zhang et al., 2023; Zheng et al., 2024; Sahoo et al., 2024) to craft the prompt. Key aspects of the prompt are a role description, a general task description, detailed annotation guidelines, a step-by-step strategy, an expected output format, few-shot examples, and the information of the target CEF (see §B.4 for the full prompt). We then generated silver VQA pairs for each of the 6,827 images contained in the ICH data source, which resulted in 17,369 pairs. Afterward, we automatically removed pairs where 1) the question contained words that introduce subjectiveness or ambiguity (“could”, “should”, “maybe”, etc.); 2) the answer contained abstract words that are hard to depict visually; and 3) where the answer is not a substring of the description of the related CEF. This way, we obtained 9,900 silver VQA samples related to 5,517 images from all 728 CEFs.

Annotation Process. Opting for high-quality VQA pairs as well as cultural diversity, we devised a two-stage annotation process with 18 trained experts from various cultural backgrounds covering all six regions (see Table 8 in §B.5). Each silver pair was evaluated using two questionnaires—one with seven question-related requirements and another with four answer-related requirements. Questions had to target the CEF and image content directly, require cultural knowledge, and depend on visual evidence (Chen et al., 2024a). Answers needed to be clear, objective, concise, and depictable. For details on the annotation process, see §B.5.

In the first round, we annotated each sample once, resulting in 4,114 samples, of which 2,826 (68.69%) met all criteria. In the second round, five annotators re-evaluated these, retaining only samples with concordant approval. This process finally yielded 2,233 samples for 1,928 images from 728 CEFs across 144 countries in six global regions.

3.3 Cultural Video VQA

In this task, models are evaluated on questions relating to videos instead of single images, again employing accuracy as the metric. To this end, we extend CIVQA in two steps: synthetic data generation and quality annotation.

Synthetic Data Generation. First, we adjusted the CIVQA questions by replacing the term “image” with “video”. We then coupled the question with a short video clip, for which we started from the CEF’s associated YouTube video. We ensured that the shortened clip contains relevant information for answering the question as follows: From each video, we extracted one frame per second, and computed image embeddings for both the frames and the CIVQA image, using DINOv2⁸ (Oquab et al., 2024; Darcet et al., 2024). We then identified the frame that best matches the original image by calculating Cosine similarity. We selected this frame as the center (at $t = 0$) for a 10-second clip⁹ (from $t = -5$ to $t = 5$). We only include clips with a best-matching frame similarity > 0.5 , which we found to yield high-quality instances based on a manual inspection of random samples. Overall, this procedure resulted in 2,001 silver samples.

Annotation Process. For additional quality control, a trained expert annotated 20% of the silver data (400 samples). Each sample was evaluated using a three-item questionnaire¹⁰ assessing whether (1) the video contained frames resembling the CEF image, (2) it clearly answered the question, or (3) neither condition was met. Overall, 95% of the annotated samples were accepted. For closer inspection, we stratified the annotated samples into four similarity bins, revealing that roughly 10% of those in the lower bins ($[0.5, 0.75]$) were rejected, while nearly all, i.e., 99% and 100%, in the higher bins ($[0.75, 1.0]$) were retained. The residual 5% label noise was considered acceptable based on further manual analysis. Notably, we found that of

⁷gpt-4o-2024-08-06

⁸facebook/dinov2-with-registers-large

⁹We do not include the audio stream in our clips.

¹⁰cf. §C for details.

the 20 rejected samples, only 9 were unanswerable based on the video, while the remaining 11 exhibited only a suboptimal frame match w.r.t. the CIVQA image. The final GIMMICK CVVQA dataset contains 1,809 samples (see §C.1 for examples) linked to 553 CEFs from 139 countries.

3.4 Cultural Origin QA

With Cultural Origin QA (COQA), we test a model’s ability to capture coarse-grained cultural knowledge. Given a CEF’s images, title, or both, the models must select its cultural origin (multiple-choice). We refer to the task as COQA_R when the origin is a region and as COQA_C when it is a country.

Dataset Construction. The COQA dataset contains all 728 CEFs from UNESCO ICH. To ensure that each instance corresponds to a unique origin, we replicate each CEF N times—where N represents the number of associated regions (for COQA_R) or countries (for COQA_C). For COQA_R, three negatives are randomly sampled from the remaining pool. Negatives for COQA_C drawn from those within the same region as the target country.

Input Modalities and Prompts. The COQA tasks support multiple input configurations alongside the task prompt. In the text-only setting, only the title of the CEF is provided, whereas in the “image-only” setting, *all* images associated with the CEF are included. Both the title and the images are used in the text-image setting. Examples and complete prompts for all variations are shown in §D.2.

3.5 Cultural Knowledge QA

In GIMMICK Cultural Knowledge QA (CKQA), we evaluate whether current AI models capture fine-grained cultural knowledge. The dataset supports two open-answer tasks: naming (CKQA_N) and describing (CKQA_D). For CKQA_N, the ground truth corresponds to the title of the CEF, while for CKQA_D, it is the detailed description. For both tasks, we leverage all 728 CEFs from UNESCO ICH. As with COQA, CKQA supports multiple input configurations: text-only, “image-only”, and text+image. We provide examples and prompts for all variations in §E.1.

4 Experimental Setup

Models and Inference. We evaluate a total of 31 models, including five proprietary LVLMS, 15 open-weight LVLMS, and 11 open-weight LLMs—the backbones of the respective LVLMS—covering

GROUP	PARAMETERS (B)	LLMS	LVLMS
S	0.5 – 4	5	5
M	7 – 11	3	6
L	26 – 38	2	2
XL	72 – 78	1	2
Closed	unknown	0	5
<i>Total</i>		11	20

Table 3: The size groups we define for result aggregation according to models’ number of parameters.

9 LVLMS and 4 LLM model families. The sizes of the open-weight models vary, categorized as small, medium, large, and extra-large (see Table 3). A comprehensive list of models is provided in Table 6 in §A.4. For our experiments, we download open weights from the respective Huggingface (Wolf et al., 2020) repositories (see Table 6) and generate responses employing greedy decoding. For proprietary models, we use the official Python SDKs. More details are reported in §F.

Metrics. For the CIVQA, CVVQA, and COQA tasks, we report relaxed answer accuracy, for which we consider a generated answer correct if it starts with the ground truth answer. For CKQA_D and CKQA_N, due to their generative nature, we use GPT-4o¹¹ in an “LVLMS-as-a-Judge” (Zheng et al., 2023; Xiong et al., 2024) setup to judge responses with a score $s \in [0, 100]$. Where $s = 0$, $s = 50$, and $s = 100$ indicate *completely incorrect or irrelevant*, *partially correct or relevant*, and *perfectly correct and complete* answers, respectively.

Video Processing. The 10-second video clips from CVVQA do not contain an audio stream, and we only use the visual information. Following established praxis (e.g., Wang et al., 2024), we extract one frame per second from the videos and provide them to the models as input alongside the textual prompt. Specifics about the image and video processing of the individual models are documented in the code.

5 Results and Analyses

In this section, we present a series of in-depth analyses based on the outcomes of our benchmark. We show aggregated results: open-weight models are grouped and averaged by parameter size, and proprietary models are averaged together (see Table 3). We provide the complete numerical results for all tasks and models in §G. In the following, we use abbreviations for regions, as defined in Table 2.

¹¹gpt-4o-2024-11-20

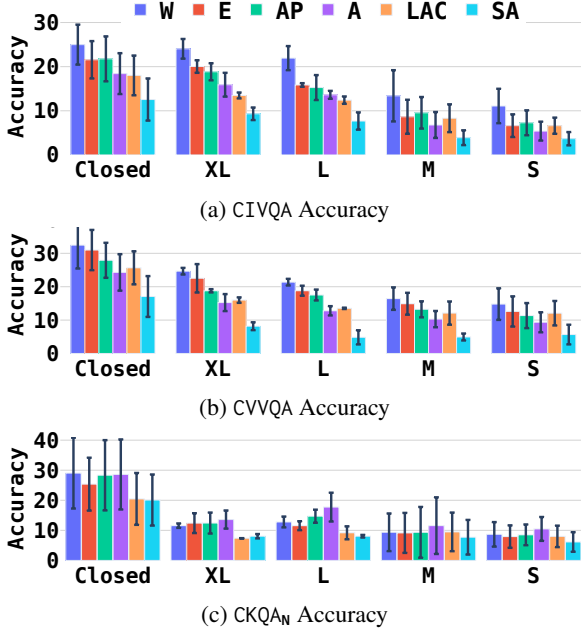


Figure 2: Aggregated results of the VQA tasks.

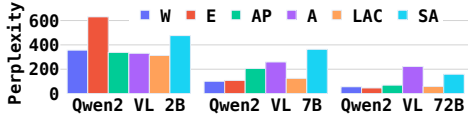


Figure 3: CIVQA ground-truth answer perplexity.

5.1 General Trends and Cultural Bias

We discuss general trends and investigate cultural bias across regions (Figures 2 and 3).

CIVQA & CVVQA. Figures 2a–c show clear regional performance disparities. Across all models—proprietary and open-weight, regardless of size—scores are highest for Western and Asian targets (■W, ■E, and ■AP) and lowest for ■SA. XL models, e.g., reach 24.04 on ■W and 9.32 on ■SA on average. ■A and ■LAC fall in between, with model performance varying by size. Since CIVQA is an open-answer task, often with rare culturally specific terms, we also evaluated the task with GPT-4o as LVLM-as-a-Judge to account for imperfect naming or spelling. While this method yields higher scores, it confirms the same trend: models exhibit a strong bias toward Western contexts. However, even the best model (GPT-4o) scores only 31.58% on ■W and 25.44% on average, highlighting GIMMICK as a challenging benchmark and the lack of fine-grained cultural knowledge in current models. We supplement our analysis with a more fine-grained investigation of how well models

“know” the cultural concepts discussed. Here, we focus on the QWENVL models on CIVQA and the compute perplexity of ground truth answers (conditioned on the input context) as a proxy of model cultural knowledge (details in §G.1.2). Figure 3 shows that for the 7B and 72B models, perplexity is consistently lower for ■W, ■E, and ■AP compared to ■A and ■SA, aligning with our performance findings. For the 2B model, however, ■E and ■SA yield the highest perplexities, which we attribute to the overall brittleness of the model. Moreover, we revisit the performance on questions about the prevalent cultural aspects in CIVQA (details in §G.1.2) and find that models perform notably better on tangible cultural aspects than on intangible ones. For instance, closed models achieve an accuracy of 30% for food-related questions and only 8% and 10% for questions concerning rituals or festivals. This highlights biases along the cultural dimension, which are particularly pronounced in non-Western contexts.

CKQA_N & CKQA_D. For CKQA_N, regional differences are minor, though proprietary models significantly outperform open-weight ones (see Figure 2c). The large error bars for closed models indicate inconsistent performance—particularly from GPT-4o MINI and GEMINI FLASH models, which perform similarly to large open-weight models. XL and L models perform worst on ■SA and ■LAC and best on ■A and ■AP with minor differences to ■W and ■E. For CKQA_D (Figure 6c), performance is 10~20% higher than on CKQA_N, likely because describing a CEF is easier than exactly naming it. However, regional biases are larger, with consistently higher scores on ■W than on ■SA, primarily for closed models like GPT-4o, which reaches 53.66 for ■W and 43.70 on ■SA.

COQA_C & COQA_R. Figure 6a shows minimal regional differences for COQA_C. Average accuracies range from close to or above 90% for closed, XL, and L models to 77.42% for S models. However, performance on COQA_R is lower than on COQA_C—85.02% vs. 81.17% on average over all models and regions—with models achieving the highest scores in ■AP. Notably, the regional ranking is mostly inverted compared to other tasks—■SA, ■A,

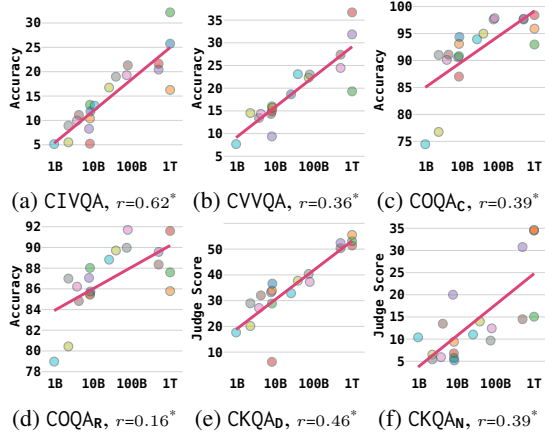


Figure 4: Model size vs. performance on GIMMICK tasks. The x-axis is in log scale. The trend line was computed using OLS regression. We report the Pearson correlation coefficient r (* indicates statistical significance).

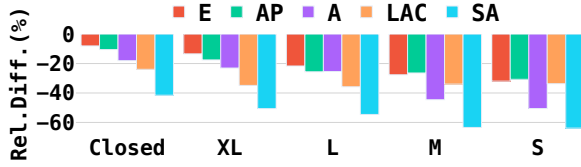


Figure 5: Relative Difference to **W** for CIVQA.

LAC, **E**, and **AP** score higher than **W**—suggesting more distinct visual and linguistic features in non-Western regions.

5.2 Influence of Model Size

We assess how model size impacts performance and whether it affects regions equally.

Figure 4 shows that model size¹² significantly influences performance, with moderate to strong Pearson correlations and steep regression lines across tasks except COQA_R, where the effect is minimal. Figure 5 shows that relative performance declines from the best-performing region (**W**) to others, particularly **SA**, varying by model size: the drops are −63.39 (S), −63.85 (M), −50.60 (L), −54.57 (XL), and −41.52 (Closed). We conclude that bigger sizes tend to result in smaller gaps without size presenting a strict ordering criterion.

5.3 Influence of Modalities

We explore how input modality—text-only, image-only, or text+image—affects perfor-

mance on COQA_C, COQA_R, and CKQA_D. Further, we compare LVLMS to their LLM backbones to assess potential losses in cultural knowledge during multimodal training.

Input Modalities. Figure 6 shows that text+image (I+T) inputs consistently yield the highest performance across all tasks, confirming that textual and visual data provide complementary cultural cues. The gap between I+T and text-only (T) is slightly more prominent for COQA_C than COQA_R, suggesting that visual information aids in inferring fine-grained, country-level details. In contrast, image-only (I) inputs perform poorly, indicating that textual information, such as CEF titles, carries more cultural context. The high variance in T results for the COQA tasks stems from the performance disparity between GEMINI PRO and CLAUDE 3.5 SONNET (e.g., 59.38 vs. 83.75 for **W**).

LVLMS vs. LLM-Backbone. Comparing LVLMS with their LLM backbones reveals that multimodal training can impair the acquisition of detailed cultural knowledge (notably in CKQA_D) while having minimal impact on coarse-grained cultural understanding (COQA). For large models, significant performance gaps—50.62 for QWEN2.5 72B vs. 40.02 for QWEN2VL 72B on **AP**—on the CKQA_D task between the LVLMS and their LLM backbones can be observed, whereas, for smaller models, the effect is subtle. Overall, our findings highlight that while images complement text for culturally grounded tasks, it is ultimately the synergy between both modalities that leads to robust and broad cultural understanding.

5.4 Influence of External Cues

We examine how external hints, i.e., informing a model about the country or region of a CEF, affect VQA performance. For CIVQA (Figure 7a), country hints consistently boost performance across model sizes and regions, while regional cues yield only modest—or even slightly adverse—effects in larger models. Gains from country hints are around 50% for most regions, but in **SA**, improvements nearly double (e.g., 97.48% for INTERNVL 2.5 78B and 97.13% for INTERNVL 2.5 38B). A similar pattern emerges for CVVQA (Figure 7b). Hints generally enhance performance across regions

¹²For closed source models, we manually set the number of parameters to 1T, except for Gemini Flash and GPT-4o mini, for which we set the number to 500B.

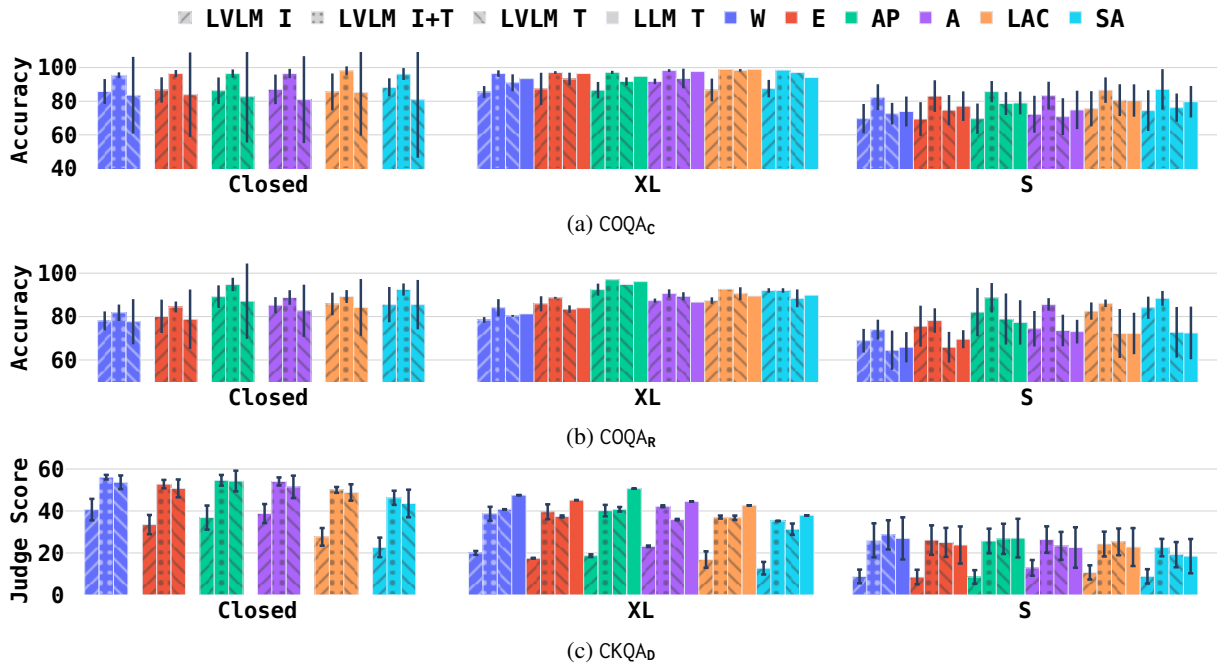


Figure 6: Aggregated results including multimodal input variations: Text-only, Image-only, Text+Image.

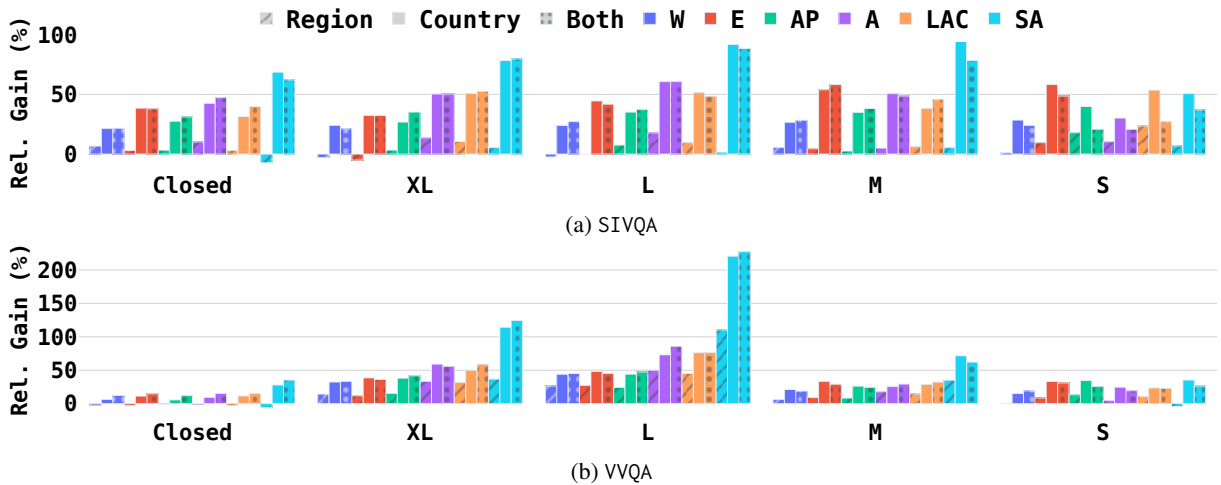


Figure 7: Relative gains on VQA tasks from providing external geographical hints.

and models, with **SA** showing the most significant gains. Proprietary and small models exhibit subtle improvements, whereas L and XL models see much higher relative gains—up to 240.7% for *INTERN VL 38B*. Notably, regional cues have a more positive impact on CVVQA than on CIVQA.

6 Conclusion

We introduce GIMMICK, a comprehensive benchmark to assess various aspects of cultural knowledge of current LVLMs and LLMs and introduce six tasks built upon three

novel datasets, which span 728 unique cultural events or facets (CEFs) from 144 countries grouped into six global macro-regions. Through extensive analyses, we study general cultural biases and the influence of model size, input modalities, and external cues. Our results consistently reveal a prominent bias toward Western cultures across all models. Interestingly, when only coarse cultural knowledge is required—such as regional origins—models performed remarkably better. Across all tasks, significant correlations between a model’s performance and its size are evident,

with a substantial gap between proprietary and open-weight models. Our analyses show that while models grasp broad cultural categories, they struggle with nuanced understanding. This suggests that GIMMICK poses a challenging benchmark and highlights the need for further advances in modeling broad cultural awareness.

Limitations

English-Only Benchmark Although we consider the performance on tasks requiring cultural understanding in English as an upper bound for the majority of models, it is yet to be tested if that hypothesis generally holds across tasks, model size, and model family. Especially for models like QWENVL and INTERNVL, which were pretrained on large portions of Chinese textual data, Chinese could be pivotal instead of English. Moreover, some cultural nuances might not be translatable to other languages.

Open-Ended VQA. CIVQA and CVVQA comprise open-ended answers to their questions, imposing challenges for adequate evaluation, especially when employing binary metrics like accuracy. This is especially true for rare, culturally specific answer terms, such as in our tasks, which are prone to spelling inaccuracies or might have different names in different cultures or languages. Although we alleviate this issue by computing scores using GPT-4o in an LVLM-as-a-Judge setting and thereby confirm our findings, this requires additional computational and financial resources. A typical solution for this is transforming the questions into multiple-choice questions, which, however, requires culturally expert annotators, who are challenging to find or train and expensive if hired via professional annotation companies.

Small Number of Samples. With a total of 7239 unique samples across all tasks in GIMMICK—2233 (CIVQA), 1809 (CVVQA), 982 (COQA_C), 759 (COQA_R), 728 (CKQA_D), and 728 (CKQA_M)—, the benchmark itself has the third most samples compared to other recent benchmarks. However, the per-task number falls relatively low, leading to even fewer counts per country or culture, making judgments about

single countries not informative.

To increase the number of samples, we consider two main options: 1) By expanding the number of annotations by employing expert annotators for an additional period of time and/or increasing the amount of silver data as described in §B.4, which would lead to an increase of samples for the CIVQA and CVVQA datasets. 2) By incorporating the newly released UNESCO data every year, as well as leveraging other high-quality sources such as UNESCO World Heritage¹³, the European Commission¹⁴, the Southeast Asian Cultural Heritage Alliance (SEACHA)¹⁵, the Journal of African Cultural Heritage¹⁶, or ICH Links¹⁷

Ethical Considerations

Country and Region Definitions. GIMMICK adopts the country and region classifications from the UNESCO ICH dataset. While these classifications are widely used, we recognize the potential for differing interpretations.

Potentially Offensive Questions. We employed semi-automatic data generation strategies to create the CIVQA dataset. Here, the silver data was generated using GPT-4o, which we showed displays significant cultural biases towards Western contexts. Although we provided the model with high-quality ground-truth information from the UNESCO ICH project and trained expert annotators with diverse cultural backgrounds to filter low-quality VQA samples, certain questions or their answers might still be offensive to people with certain cultural origins. Since this is subjective, we need to accept it as is for now. Nevertheless, we encourage contacting us if any offensive or otherwise harmful sample raises someone’s attention.

Acknowledgements

We thank our annotators for the CIVQA and CVVQA tasks with special thanks to Timm Dill, Narges Baba Ahmadi, Niloufar Baba Ahmadi,

¹³<https://www.unesco.org/world-heritage>

¹⁴<https://culture.ec.europa.eu/cultural-heritage>

¹⁵<https://seacha.org/>

¹⁶<https://jachs.org/>

¹⁷<https://www.ichlinks.com>

and Abdullah Abdelhafez for their extra efforts. The work of Carolin Holtermann and Anne Lauscher is funded by the Excellence Strategy of the German Federal Government and the Federal States.

References

- Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, and 68 others. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *CoRR*, abs/2404.14219.
- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards Measuring and Modeling “Culture” in LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15763–15784, Miami, Florida, USA.
- Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathé, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGEVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2598–2637, Mexico City, Mexico.
- Meta AI. 2024. Llama 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models.
- Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen R. McKeown. 2025. See It from My Perspective: Diagnosing the Western Cultural Bias of Large Vision-Language Models in Image Understanding. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, Singapore, Republic of Singapore.
- AI Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.
- Yujin Baek, ChaeHun Park, Jaeseok Kim, Yujung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Evaluating Visual and Cultural Interpretation: The K-Viscuit Benchmark with Human-VLM Collaboration. *CoRR*, abs/2406.16469.
- Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. 2024. From Local Concepts to Universals: Evaluating the Multicultural Understanding of Vision-Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6763–6782, Miami, FL, USA.
- Shaily Bhatt and Fernando Diaz. 2024. Extrinsic Evaluation of Cultural Competence in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 16055–16074, Miami, FL, USA.
- Olena Burda-Lassen, Aman Chadha, Shashank Goswami, and Vinija Jain. 2025. How Culturally Aware Are Vision-Language Models? In *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*, volume CFP2540Z-ART, pages 1–6, Lyon, France.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 52 others. 2024. InternLM2 Technical Report. *CoRR*, abs/2403.17297.
- Yong Cao, Wenyan Li, Jiaang Li, Yifei Yuan, Antonia Karamolegkou, and Daniel Hershcovich. 2024. Exploring Visual Culture Awareness in GPT-4V: A Comprehensive Probing. *CoRR*, abs/2402.06015.
- Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. MaXM: Towards Multilingual Visual

- Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682, Singapore.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024a. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 21 others. 2024b. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *CoRR*, abs/2412.05271.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, Seattle, WA, USA.
- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. CulturalBench: a Robust, Diverse and Challenging Benchmark on Measuring the (Lack of) Cultural Knowledge of LLMs. *CoRR*, abs/2410.02677.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya Expansive: Combining Research Breakthroughs for a New Multilingual Frontier. *CoRR*, abs/2412.04261.
- Tri Dao. 2024. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria.
- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA.
- Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2024. Vision Transformers Need Registers. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. VLMEvalKit: An Open-Source ToolKit for Evaluating Large Multi-Modality Models. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 11198–11201, Melbourne, VIC, Australia.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *CoRR*, abs/2306.13394.
- Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively Multicultural Knowledge Acquisition & LM Benchmarking. *CoRR*, abs/2402.09369.
- Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437.

- Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavaš. 2025. Centurio: On Drivers of Multilingual Ability of Large Vision-Language Model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Vienna, Austria.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. GPT-4o System Card. *CoRR*, abs/2410.21276.
- Antonia Karamolegkou, Phillip Rust, Yong Cao, Ruixiang Cui, Anders Søgaard, and Daniel Hershcovich. 2024. Vision-Language Models under Cultural and Inclusive Considerations. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 53–66, Bangkok, Thailand.
- Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024. FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 19077–19095, Miami, Florida, USA.
- Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10467–10485, Punta Cana, Dominican Republic.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 34892–34916, New Orleans, LA, USA.
- Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural Conditioning or Placebo? On the Effectiveness of Socio-Demographic Prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15811–15837, Miami, FL, USA.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. BLEND: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:78104–78146.
- Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 16366–16393, Bangkok, Thailand.
- Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking Vision Language Models for Cultural Understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5769–5790, Miami, FL, USA.
- Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas, and Alessandro Suglia. 2025. CROPE: Evaluating In-Context Adaptation of Vision and Language Models to Culture-Specific Concepts. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pages 7917–7936, Albuquerque, New Mexico.
- OpenAI. 2023. GPT-4 Vision System Card.

- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.
- David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesús-Germán Ortiz-Barajas, Emilio Villa-Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David Le Meur, Joan Nwatu, Ganzorig Batnasan, and 57 others. 2024. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada.
- Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *CoRR*, abs/2402.07927.
- Florian Schneider and Sunayana Sitaram. 2024. M5 - A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks. In *Findings of the Association for Computational Linguistics: (EMNLP)*, pages 4309–4345, Miami, Florida, USA.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context. *arXiv preprint arXiv:2403.05530*.
- Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 715–729, Abu Dhabi, United Arab Emirates.
- Norawit Uraileertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. SEA-VQA: Southeast Asian Cultural Context Dataset For Visual Question Answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185, Bangkok, Thailand.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution. *CoRR*, abs/2409.12191.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA.
- Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzaev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Ching Lam Cheng, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. WorldCuisines: A Massive-Scale Benchmark for Multilingual and Multicultural Visual Question Answering on Global Cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pages 3242–3264, Albuquerque, New Mexico.

- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45, Online.
- Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanguan Gu, Heng Huang, and Chunyuan Li. 2024. LLaVA-Critic: Learning to Evaluate Multimodal Models. *CoRR*, abs/2410.02712.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *CoRR*, abs/2408.01800.
- Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, Seattle, WA, USA.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, Kigali, Rwanda.
- Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA.

Appendix Overview

Due to the number of experiments, the general density of our work, and our aim to be as transparent as possible in the sense of open science, the following appendix is extensive. Hence, we provide a brief outline of its content to ease navigation and to get an overview quickly.

A GIMMICK Benchmark Details

Details on license, examples, regions, models.

B CIVQA Details

Details on examples, synthetic data generation, and the annotation project.

C VVQA Details

Details on examples and the annotation project.

D COQA Details

Details on prompts, and examples.

E CKQA Details

Details on prompts.

F Experimental Setup

Details on prompts hyperparameters.

G Results and Analyses

Details on complete results of all models and datasets and additional analyses.

A GIMMICK Benchmark Details

A.1 Data License

GIMMICK is built upon the open-access data from the UNESCO Intangible Cultural Heritage (ICH) project, which is organized as a knowledge graph. The graph can be downloaded in English, French, and Spanish on the ICH project website: <https://ich.unesco.org/en/open-access-to-dive-data-01218>, with details about its structure and subsets also provided. In GIMMICK, we work with the English graph only. The open-access license of the knowledge graph is defined on the UNESCO website¹⁸ as follows:

By 'open access' to the literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself.

The images and videos within the data are shared via URLs and hosted by UNESCO or on YouTube, respectively. Further, each image and video node in the knowledge graph has individual copyright information attached. However, the licenses themselves are not discussed, and merely the name of the photographer or institution or UNESCO itself is stated. Unfortunately, we did not receive an answer to multiple emails in which we asked for clarification. Hence, we assume that the image and video content also fall under the definition of "open access". If you are a copyright holder of any of the images or videos and do not want your intellectual property to be used or shared by us, please reach out via email: florian.schneider-1@uni-hamburg.de.

A.2 Cultural Event or Facets (CEFs)

A.2.1 Examples

In the following, we provide one example of CEFs per region from the UNESCO ICH project. We also use the same information for the CKQA_N and CKQA_D tasks.

¹⁸<https://www.unesco.org/en/open-access>

Western Europe (■W)

Title: The skills related to perfume in Pays de Grasse: the cultivation of perfume plants, the knowledge and processing of natural raw materials, and the art of perfume composition

Countries: France

Regions: Western European and North American States

Description:

The skills related to perfume in Pays de Grasse cover three different aspects: the cultivation of perfume plants; the knowledge and processing of natural raw materials; and the art of perfume composition. The practice involves a wide range of communities and groups, brought together under the Association du Patrimoine Vivant du Pays de Grasse (Living Heritage Association of the Region of Grasse). Since at least the sixteenth century, the practices of growing and processing perfume plants and creating fragrant blends have been developed in Pays de Grasse, in a craft industry long dominated by leather tanning. Perfume plant cultivation involves a wide range of skills and knowledge, for instance pertaining to nature, soil, weather, biology, plant physiology and horticultural practices, as well as specific techniques such as extraction and hydraulic distillation methods. The inhabitants of Grasse have made these techniques their own and helped improve them. In addition to technical skills, however, the art also calls for imagination, memory and creativity. Perfume forges social bonds and provides an important source of seasonal labour. Related knowledge is mostly transmitted informally through a long learning process that still takes place primarily in perfumeries. In recent decades, however, there has been a growing interest in standardizing learning through formalized teaching.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/the-skills-related-to-perfume-i...>



Copyright: JM. Ghibaudo
APVPG 2011



Copyright: Musées de Grasse
2011



Copyright: N. Bédar APVPG
2015



Copyright: C. Barbiero/Musées
de Grasse 2010



Copyright: Daniel, Serre, M.
Roudnitska APVPG 2014



Copyright: Musées de Grasse
2012



Copyright: G. Voinot/Université
Sophia Antipolis 2011



Copyright: Esat Les Restanques
2013



Copyright: Forum des Associa-
tions Pays de Grasse 2014



Copyright: PH. Massé APVPG
2014

Eastern Europe (■E)

Title: Cultural Heritage of Boka Navy Kotor: a festive representation of a memory and cultural identity

Countries: Montenegro

Regions: Eastern European States

Description:

Boka Navy is a traditional, non-governmental maritime organization founded in Kotor, Montenegro in 809. Its origin is linked to the arrival of the relics of St. Tryphon, the patron saint of the city of Kotor. Comprised of a community of seafarers with military, economic, educational and humanitarian functions, Boka Navy has played a memorial role for two centuries, preserving and promoting maritime history and tradition. Membership is voluntary and open to men, women and children of all ages. The organization is founded on the respect of human rights and of religious, national and cultural diversity. During formal celebrations, members wear colourful traditional uniforms, carry historic weapons and perform the traditional circle kolo dance. Boka Navy is the backbone of the annual St. Tryphon festivities, which take place from 13 January through 3 February and include a procession and a series of rituals in the cathedral. The external festivities begin with the Boka Navy's traditional kolo circle dance and are followed by a procession carrying the relics of St. Tryphon through the main town squares and streets. Thousands of spectators attend the processions in the historic centre and observe the festive events. Hundreds of women, men and children also participate in preparations of the activities.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/cultural-heritage-of-boka-navy-...>



Copyright: Ministry of Culture of Montenegro



Copyright: Ministry of Culture of Montenegro



Copyright: Ministry of Culture of Montenegro



Copyright: Ministry of Culture of Montenegro



Copyright: Ministry of Culture of Montenegro



Copyright: Ministry of Culture of Montenegro



Copyright: Ministry of Culture of Montenegro



Copyright: Ministry of Culture of Montenegro



Copyright: Ministry of Culture of Montenegro



Copyright: Ministry of Culture of Montenegro

Title: Arts, skills and practices associated with engraving on metals (gold, silver and copper)

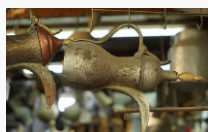
Countries: Algeria, Saudi Arabia, Egypt, Iraq, Morocco, Mauritania, Palestine, Sudan, Tunisia, Yemen

Regions: Arab States

Description:

Engraving on metals such as gold, silver and copper is a centuries-old practice that entails manually cutting words, symbols or patterns into the surfaces of decorative, utilitarian, religious or ceremonial objects. The craftsperson uses different tools to manually cut symbols, names, Quran verses, prayers and geometric patterns into the objects. Engravings can be concave (recessed) or convex (elevated), or the result of a combination of different types of metals, such as gold and silver. Their social and symbolic meanings and functions vary according to the communities concerned. Engraved objects, such as jewelry or household objects, are often presented as traditional gifts for weddings or used in religious rituals and alternative medicine. For instance, certain types of metals are believed to have healing properties. Engraving on metals is transmitted within families, through observation and hands-on practice. It is also transmitted through workshops organized by training centres, organizations and universities, among others. Publications, cultural events and social media further contribute to the transmission of the related knowledge and skills. Practised by people of all ages and genders, metal engraving and the use of engraved objects are means of expressing the cultural, religious and geographical identity and the socioeconomic status of the communities concerned.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/arts-skills-and-practices-assoc...>



Copyright: Huzaifa Ayad Bahaa El Din, Iraq, 2021



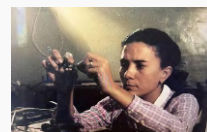
Copyright: Huzaifa Ayad Bahaa El Din, Iraq, 2021



Copyright: Huzaifa Ayad Bahaa El Din, Iraq, 2021



Copyright: Zahia Benabdallah, Algeria, 2021



Copyright: Azza Fahmy, Egypt, 2021



Copyright: Mustafa Kamil, Egypt, 2021



Copyright: National Heritage Preservation, Ministry of Culture, Youth and Sport and Relations with the Parliament, Egypt, 2022



Copyright: Direction du Patrimoine Culturel, Morocco, 2021



Copyright: Direction du Patrimoine Culturel, Morocco, 2021



Copyright: Ministry of Culture, Palestine, 2021

Asia and Pacific (■AP)

Title: Tugging rituals and games

Countries: Cambodia, Korea, Philippines, Vietnam

Regions: Asian and Pacific States

Description:

Tugging rituals and games in the rice-farming cultures of East Asia and Southeast Asia are enacted among communities to ensure abundant harvests and prosperity. They promote social solidarity, provide entertainment and mark the start of a new agricultural cycle. Many tugging rituals and games also have profound religious significance. Most variations include two teams, each of which pulls one end of a rope attempting to tug it from the other. The intentionally uncompetitive nature of the event removes the emphasis on winning or losing, affirming that these traditions are performed to promote the well-being of the community, and reminding members of the importance of cooperation. Many tugging games bear the traces of agricultural rituals, symbolizing the strength of natural forces, such as the sun and rain while also incorporating mythological elements or purification rites. Tugging rituals and games are often organized in front of a village's communal house or shrine, preceded by commemorative rites to local protective deities. Village elders play active roles in leading and organizing younger people in playing the game and holding accompanying rituals. Tugging rituals and games also serve to strengthen unity and solidarity and sense of belonging and identity among community members.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/tugging-rituals-and-games-01080...>



Copyright: Siyonn Sophearith, 2013



Copyright: Siyonn Sophearith, 2013



Copyright: Siyonn Sophearith, 2013



Copyright: Renato S. Rastrollo, NCCA



Copyright: Renato S. Rastrollo, NCCA



Copyright: Vietnam Institute of Culture and Arts Studies, 2013



Copyright: Vietnam Institute of Culture and Arts Studies, 2013



Copyright: Joo Byung Soo, 2006



Copyright: Joo Byung Soo, 2006

Title: Ancestral system of knowledge of the four indigenous peoples, Arhuaco, Kankuamo, Kogui and Wiwa of the Sierra Nevada de Santa Marta

Countries: Colombia

Regions: Latin-American and Caribbean States

Description:

The Ancestral System of Knowledge of the Arhuaco, Kankuamo, Kogui and Wiwa peoples of the Sierra Nevada de Santa Marta is comprised of sacred mandates that keep the existence of the four peoples in harmony with the physical and spiritual universe. Through many years of dedication, the knowledgeable men (Mamos) and women (Sagas) acquire the necessary skills and sensitivity to communicate with the snow-capped peaks, connect with the knowledge of the rivers and decipher the messages of nature. Based on the Law of Origin, a philosophy that governs human relationships to nature and the universe, the Ancestral System of Knowledge entails caring for sacred sites and partaking in baptism rituals, marriage rites, traditional dances and songs, and retributions or offerings to spiritual powers. This ancestral wisdom is believed to play a fundamental role in protecting the Sierra Nevada ecosystem and avoiding the loss of the cultural identity of the four peoples of the region. The Ancestral System of Knowledge is transmitted from generation to generation through cultural practice, community activities, the use of the indigenous language and the implementation of the sacred mandates. The transmission process includes the understanding of physical and spiritual relationships with Mother Nature and sacred sites.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/ancestral-system-of-knowledge-o...>



Copyright: William Diaz, 2021



Copyright: Jorge Mario Suarez/Government of Magdalena, 2017



Copyright: Jorge Mario Suarez/Government of Magdalena, 2017



Copyright: William Diaz, 2021



Copyright: Jorge Mario Suarez/Government of Magdalena, 2017



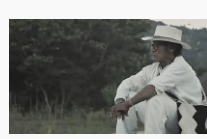
Copyright: Jorge Mario Suarez/Government of Magdalena, 2017



Copyright: Jorge Mario Suarez/Government of Magdalena, 2017



Copyright: Jorge Mario Suarez/Government of Magdalena, 2017



Copyright: Jorge Mario Suarez/Government of Magdalena, 2017



Copyright: William Diaz, 2021

Subsaharian Africa (■SA)

Title: Gada system, an indigenous democratic socio-political system of the Oromo

Countries: Ethiopia

Regions: Subsaharian African States

Description:

Gada is a traditional system of governance used by the Oromo people in Ethiopia developed from knowledge gained by community experience over generations. The system regulates political, economic, social and religious activities of the community dealing with issues such as conflict resolution, reparation and protecting women's rights. It serves as a mechanism for enforcing moral conduct, building social cohesion, and expressing forms of community culture. Gada is organized into five classes with one of these functioning as the ruling class consisting of a chairperson, officials and an assembly. Each class progresses through a series of grades before it can function in authority with the leadership changing on a rotational basis every eight years. Class membership is open to men, whose fathers are already members, while women are consulted for decision-making on protecting women's rights. The classes are taught by oral historians covering history, laws, rituals, time reckoning, cosmology, myths, rules of conduct, and the function of the Gada system. Meetings and ceremonies take place under a sycamore tree (considered the Gada symbol) while major clans have established Gada centres and ceremonial spaces according to territory. Knowledge about the Gada system is transmitted to children in the home and at school.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/gada-system-an-indigenous-democ...>



Copyright: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014



Copyright: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014



Copyright: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014



Copyright: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014



Copyright: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014



Copyright: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014



Copyright: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014



Copyright: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014



Copyright: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014



Copyright: Authority for Research and Conservation of Cultural Heritage (ARCCH), Ethiopia, 2014

A.2.2 CEFs as Python a dataclass

Listing 1 presents a CEF implemented as a Python dataclass.

```
from dataclasses import dataclass
```

```
@dataclass
class CEF:
    title: str
    description: str
    countries: list[str]
    regions: list[str]
    images: list[str] # URLs
    videos: list[str] # URLs
```

Listing 1: Python pseudo-code for a dataclass representing a CEF.

Region	Abbrev.	Countries	Countries
Arab	 A	18	Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, United Arab Emirates, Yemen
Asia & Pacific	 AP	35	Lao People’s Democratic Republic, Afghanistan, Australia, Bangladesh, Bhutan, Cambodia, China, Cook Islands, Democratic People’s Republic of Korea, Fiji, India, Indonesia, Iran, Japan, Kazakhstan, Korea, Kyrgyzstan, Malaysia, Micronesia, Mongolia, Myanmar, Nepal, New Zealand, Pakistan, Papua New Guinea, Philippines, Samoa, Singapore, Sri Lanka, Thailand, Timor-Leste, Tonga, Turkmenistan, Vanuatu, Vietnam
Eastern Europe	 E	25	Albania, Armenia, Azerbaijan, Belarus, Bosnia and Herzegovina, Bulgaria, Croatia, Czechia, Estonia, Georgia, Hungary, Latvia, Lithuania, Moldova, Montenegro, North Macedonia, Poland, Romania, Russia, Serbia, Slovakia, Slovenia, Tajikistan, Ukraine, Uzbekistan
Latin-America & Caribbean	 LAC	28	Antigua and Barbuda, Argentina, Bahamas, Belize, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Curaçao, Dominican Republic, Ecuador, El Salvador, Grenada, Guatemala, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama, Paraguay, Peru, Saint Kitts and Nevis, Saint Vincent and the Grenadines, Uruguay, Venezuela
Subsaharian Africa	 SA	40	Côte d’Ivoire, Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Congo, Democratic Republic of the Congo, Djibouti, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Kenya, Lesotho, Madagascar, Malawi, Mali, Mauritius, Mozambique, Namibia, Niger, Nigeria, Rwanda, Senegal, Seychelles, Somalia, South Africa, South Sudan, Togo, Uganda, Zambia, Zimbabwe
Western Europe & North America	 W	23	Andorra, Austria, Belgium, Canada, Cyprus, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Malta, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Türkiye, United Kingdom of Great Britain and Northern Ireland

Table 4: Caption

A.3 Regions

A.3.1 Number of Samples per Task per Region

A.4 Models

We present the comprehensive list of all 31 models evaluated in GIMMICK in Table 6.

B CIVQA Details

REGION	CIVQA	CVVQA	COQA _R	COQA _C	CKQA _D	CKQA _N
■ A	375	296	71	127	71	71
■ A ■ AP	4	4	2	2	1	1
■ A ■ AP ■ E ■ W	5	5	0	36	2	2
■ A ■ E ■ W	1	0	3	7	1	1
■ A ■ SA	8	0	2	3	1	1
■ AP	444	407	211	222	211	211
■ AP ■ E	7	7	6	6	3	3
■ AP ■ E ■ LAC ■ SA ■ W	1	1	0	8	1	1
■ AP ■ E ■ W	10	7	21	35	7	7
■ AP ■ W	4	3	2	3	1	1
■ E	302	242	125	136	125	125
■ E ■ W	21	20	22	56	11	11
■ LAC	420	341	96	106	96	96
■ LAC ■ W	2	2	2	2	1	1
■ SA	388	299	71	80	71	71
■ W	241	175	125	153	125	125

Table 5: Number of samples per region(s) in GIMMICK tasks.

MODEL ID	PAPER NAME	OPEN-WEIGHT	SIZE GROUP	IMAGE INPUT	VIDEO INPUT	TEXT INPUT	LLM BACKBONE
claude-3-5-sonnet-20241022	Claude 3.5 Sonnet (Anthropic, 2024)	No	A	Yes	Yes	Yes	–
gemini-1.5-pro-002	Gemini Pro (Team et al., 2024)	No	A	Yes	Yes	Yes	–
gemini-1.5-flash-002	Gemini Flash (Team et al., 2024)	No	A	Yes	Yes	Yes	–
gpt-4o-2024-11-20	GPT-4o (Hurst et al., 2024)	No	A	Yes	Yes	Yes	–
gpt-4o-mini-2024-07-18	GPT-4o Mini (Hurst et al., 2024)	No	A	Yes	Yes	Yes	–
opengvlab/internvl2_5-78b	InternVL2.5 78B (Chen et al., 2024b)	Yes	XL	Yes	Yes	Yes	qwen/qwen2.5-72b-instruct
qwen/qwen2-vl-72b-instruct	Qwen2 VL 72B (Wang et al., 2024)	Yes	XL	Yes	Yes	Yes	qwen/qwen2.5-72b-instruct
opengvlab/internvl2_5-26b	InternVL2.5 26B (Chen et al., 2024b)	Yes	L	Yes	Yes	Yes	internlm/internlm2_5-20b-chat
opengvlab/internvl2_5-38b	InternVL2.5 38B (Chen et al., 2024b)	Yes	L	Yes	Yes	Yes	qwen/qwen2.5-32b-instruct
meta-llama/llama-3.2-11b-vision-instruct	Llama 3.2 11B Vision (AI, 2024)	Yes	M	Yes	Yes	Yes	–
qwen/qwen2-vl-7b-instruct	Qwen2 VL 7B (Wang et al., 2024)	Yes	M	Yes	Yes	Yes	qwen/qwen2.5-7b-instruct
openbmb/minicpm-v-2_6	MiniCPM V 2.6 (Yao et al., 2024)	Yes	M	Yes	Yes	Yes	–
wuenlp/centurio_aya	Centurio Aya (Geigle et al., 2025)	Yes	M	Yes	Yes	Yes	cohereforai/aya-expanse-8b
opengvlab/internvl2_5-8b	InternVL2.5 8B (Chen et al., 2024b)	Yes	M	Yes	Yes	Yes	internlm/internlm2_5-7b-chat
wuenlp/centurio_qwen	Centurio Qwen (Geigle et al., 2025)	Yes	M	Yes	Yes	Yes	qwen/qwen2.5-7b-instruct
qwen/qwen2-vl-2b-instruct	Qwen2 VL 2B (Wang et al., 2024)	Yes	S	Yes	Yes	Yes	qwen/qwen2.5-1.5b-instruct
microsoft/phi-3.5-vision-instruct	Phi 3.5 Vision (Abdin et al., 2024)	Yes	S	Yes	Yes	Yes	microsoft/phi-3.5-mini-instruct
opengvlab/internvl2_5-4b	InternVL2.5 4B (Chen et al., 2024b)	Yes	S	Yes	Yes	Yes	qwen/qwen2.5-3b-instruct
opengvlab/internvl2_5-1b	InternVL2.5 1B (Chen et al., 2024b)	Yes	S	Yes	Yes	Yes	qwen/qwen2.5-0.5b-instruct
opengvlab/internvl2_5-2b	InternVL2.5 2B (Chen et al., 2024b)	Yes	S	Yes	Yes	Yes	internlm/internlm2_5-1.8b-chat
qwen/qwen2.5-72b-instruct	Qwen2.5 72B (Yang et al., 2024)	Yes	XL	No	No	Yes	–
qwen/qwen2.5-32b-instruct	Qwen2.5 32B (Yang et al., 2024)	Yes	L	No	No	Yes	–
internlm/internlm2_5-20b-chat	InternLM2.5 20B (Cai et al., 2024)	Yes	L	No	No	Yes	–
cohereforai/aya-expanse-8b	Aya Expanse 8B (Dang et al., 2024)	Yes	M	No	No	Yes	–
internlm/internlm2_5-7b-chat	InternLM2.5 7B (Cai et al., 2024)	Yes	M	No	No	Yes	–
qwen/qwen2.5-7b-instruct	Qwen2.5 7B (Yang et al., 2024)	Yes	M	No	No	Yes	–
qwen/qwen2.5-0.5b-instruct	Qwen2.5 0.5B (Yang et al., 2024)	Yes	S	No	No	Yes	–
qwen/qwen2.5-3b-instruct	Qwen2.5 3B (Yang et al., 2024)	Yes	S	No	No	Yes	–
qwen/qwen2.5-1.5b-instruct	Qwen2.5 1.5B (Yang et al., 2024)	Yes	S	No	No	Yes	–
internlm/internlm2_5-1.8b-chat	InternLM2.5 1.8B (Cai et al., 2024)	Yes	S	No	No	Yes	–
microsoft/phi-3.5-mini-instruct	Phi 3.5 Mini (Abdin et al., 2024)	Yes	S	No	No	Yes	–

Table 6: Details about the models evaluated within the GIMMICK benchmark. The size “A” indicates that the model is a proprietary API model with unknown size.

B.1 Examples

In the following, we provide one random sample per region for the CIVQA task. Note that the lower part of the examples, where the related CEF is provided, is *not* part of the actual sample.

■ A



Copyright: Conseil municipal de Sefrou, 2010

Question: What title is given to the woman wearing the sash in the image?

Answer: Cherry Queen

Related Cultural Event or Facet

Title: Cherry festival in Sefrou

Countries: Morocco

Regions: Arab States

Description:

For three days in June each year, the local population of Sefrou celebrates the natural and cultural beauty of the region, symbolized by the cherry fruit and that year's newly chosen Cherry Queen, selected during a pageant that draws competitors from the region and entire country. The highlight of the festival is a parade with performing troupes, rural and urban music, majorettes and bands, and floats featuring local producers. At the centre is the Cherry Queen, who offers cherries to onlookers while dressed ornately and surrounded by attendants. The whole population contributes to the success of the festival: craftswomen make silk buttons for traditional dresses, fruit growers supply cherries, local sports clubs participate in competitions, and music and dancing troupes animate the entire festival. The cherry festival provides an opportunity for the entire city to present its activities and achievements. The younger generation are also integrated into festival activities to ensure their sustainability. The festival is a source of pride and belonging that enhances the self-esteem of the city and its people and constitutes a fundamental contribution to their local identity.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/cherry-festival-in-sefrou-00641...>



Copyright: 2010 by Centre for Research and Development of Culture, Indonesia

Question: What traditional dance are the performers engaging in, as seen in the image?

Answer: Saman dance

Related Cultural Event or Facet

Title: Saman dance

Countries: Indonesia

Regions: Asian and Pacific States

Description:

The Saman dance is part of the cultural heritage of the Gayo people of Aceh province in Sumatra. Boys and young men perform the Saman sitting on their heels or kneeling in tight rows. Each wears a black costume embroidered with colourful Gayo motifs symbolizing nature and noble values. The leader sits in the middle of the row and leads the singing of verses, mostly in the Gayo language. These offer guidance and can be religious, romantic or humorous in tone. Dancers clap their hands, slap their chests, thighs and the ground, click their fingers, and sway and twist their bodies and heads in time with the shifting rhythm – in unison or alternating with the moves of opposing dancers. These movements symbolize the daily lives of the Gayo people and their natural environment. The Saman is performed to celebrate national and religious holidays, cementing relationships between village groups who invite each other for performances. The frequency of Saman performances and its transmission are decreasing, however. Many leaders with knowledge of the Saman are now elderly and without successors. Other forms of entertainment and new games are replacing informal transmission, and many young people now emigrate to further their education. Lack of funds is also a constraint, as Saman costumes and performances involve considerable expense.

UNESCO ICH URL: <https://ich.unesco.org/en/USL/saman-dance-00509...>



Copyright: 2010 by M.Rahimov/Ministry of Culture and Tourism

Question: What is the name of the musical instrument observed by the man in the image?

Answer: Tar

Related Cultural Event or Facet

Title: Craftsmanship and performance art of the Tar, a long-necked string musical instrument

Countries: Azerbaijan

Regions: Eastern European States

Description:

The Tar is a long-necked plucked lute, traditionally crafted and performed in communities throughout Azerbaijan. Considered by many to be the country's leading musical instrument, it features alone or with other instruments in numerous traditional musical styles. Tar makers transmit their skills to apprentices, often within the family. Craftsmanship begins with careful selection of materials for the instrument: mulberry wood for the body, nut wood for the neck, and pear wood for the tuning pegs. Using various tools, crafters create a hollow body in the form of a figure eight, which is then covered with the thin pericardium of an ox. The fretted neck is affixed, metal strings are added and the body is inlaid with mother-of-pearl. Performers hold the instrument horizontally against the chest and pluck the strings with a plectrum, while using trills and a variety of techniques and strokes to add colour. Tar performance has an essential place in weddings and different social gatherings, festive events and public concerts. Players transmit their skills to young people within their community by word of mouth and demonstration, and at educational musical institutions. Craftsmanship and performance of the tar and the skills related to this tradition play a significant role in shaping the cultural identity of Azerbaijanis.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/craftsmanship-and-performance-a...>



Copyright: Py, 2019

Question: What traditional tool from the Guaraní culture is depicted in the image for drinking Terere?

Answer: Bombilla

Related Cultural Event or Facet

Title: Practices and traditional knowledge of Terere in the culture of Pohã Ñana, Guaraní ancestral drink in Paraguay

Countries: Paraguay

Regions: Latin-American and Caribbean States

Description:

The practices and traditional knowledge of Terere in the culture of Pohã Ñana, Guaraní ancestral drink in Paraguay, are widespread in the Paraguayan territory and involve a variety of bearers. Terere is a traditional drink prepared in a jug or thermos, in which cold water is mixed with Pohã Ñana crushed in a mortar. It is served in a glass pre-filled with yerba mate and sucked with a bombilla (metal or cane straw). Preparing the Terere is an intimate ritual involving a series of pre-established codes and each Pohã Ñana herb has health benefits linked to popular wisdom passed down through the generations. Terere practices in the culture of Pohã Ñana have been transmitted in Paraguayan families since approximately the sixteenth century. Traditional knowledge about the healing attributes of the medicinal herbs that make up the Pohã Ñana and their correct use are also transmitted spontaneously within the family. In recent years, the figure of apprentices has risen, but family transmission remains the main mode of transmission. The practice of the Terere in the culture of Pohã Ñana fosters social cohesion as the time and space dedicated to preparing and consuming the Terere promote inclusion, friendship, dialogue, respect and solidarity. The practice also strengthens new generations' appreciation of the rich cultural and botanical heritage of Guaraní origin.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/practices-and-traditional-knowl...>



Copyright: The Authority for Research and Conservation of Cultural Heritage (ARCCH), 2013

Question: What festival are the people in the image celebrating?
 Answer: Fichée-Chambalaalla

Related Cultural Event or Facet

Title: Fichée-Chambalaalla, New Year festival of the Sidama people

Countries: Ethiopia

Regions: Sub-Saharan African States

Description:

Fichée-Chambalaalla is a New Year festival celebrated among the Sidama people. According to the oral tradition, Fichée commemorates a Sidama woman who visited her parents and relatives once a year after her marriage, bringing "buurisame", a meal prepared from false banana, milk and butter, which was shared with neighbours. Fichée has since become a unifying symbol of the Sidama people. Each year, astrologers determine the correct date for the festival, which is then announced to the clans. Communal events take place throughout the festival, including traditional songs and dances. Every member participates irrespective of age, gender and social status. On the first day, children go from house to house to greet their neighbours, who serve them "buurisame". During the festival, clan leaders advise the Sidama people to work hard, respect and support the elders, and abstain from cutting down indigenous trees, begging, indolence, false testimony and theft. The festival therefore enhances equity, good governance, social cohesion, peaceful co-existence and integration among Sidama clans and the diverse ethnic groups in Ethiopia. Parents transmit the tradition to their children orally and through participation in events during the celebration. Women in particular, transfer knowledge and skills associated with hairdressing and preparation of "buurisame" to their daughters and other girls in their respective villages.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/fichée-chambalaalla-new-year-fe...>



Copyright: Município de Estremoz, 2015

Question: What specific region's attire is represented by the figures in the image?

Answer: Alentejo

Related Cultural Event or Facet

Title: Craftmanship of Estremoz clay figures

Countries: Portugal

Regions: Western European and North American States

Description:

The Craftmanship of Estremoz Clay Figures involves a production process lasting several days: the elements of the figures are assembled before being fired in an electric oven and then painted by the artisan and covered with a colourless varnish. The clay figures are dressed in the regional attires of Alentejo or the clothing of religious Christian iconography, and follow specific themes. The production of clay figures in Estremoz dates back to the seventeenth century, and the very characteristic aesthetic features of the figures make them immediately identifiable. The craft is strongly attached to the Alentejo region, since the vast majority of the figures depict natural elements, local trades and events, popular traditions and devotions. The viability and recognition of the craft are ensured through non-formal education workshops and pedagogical initiatives by the artisans, as well as by the Centre for the Appreciation and Safeguarding of the Estremoz Clay Figure. Fairs are organized at the local, national and international levels. Knowledge and skills are transmitted both in family workshops and professional contexts, and artisans teach the basics of their craft through non-formal training initiatives. Artisans are actively involved in awareness-raising activities organized in schools, museums, fairs and other events.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/craftmanship-of-estremoz-clay-f...>

B.2 Cultural Aspects

During the synthetic data generation phase of the CIVQA, we also obtained a “target aspect” per question (see §B.4 and §B.4.1). We report these aspects in the following.

B.3 External Hint Variations

For the CIVQA (and CVVQA) task, we ablate the effect of external cues or hints on the task performance of models. In the following, we provide the Python pseudo-code snippet to generate the prompt for a given sample.

B.4 Synthetic Data Generation

Aspect	Questions	Aspect	Questions	Aspect	Questions
traditions	390	education	3	jewelry	1
rituals	241	culture	3	objects	1
art	233	games	3	animal	1
music	210	performing arts	3	plants	1
craftsmanship	177	language	3	process	1
instruments	155	performance	3	agriculture	1
festivals	151	characters	2	celebrations	1
dance	150	practices	2	details	1
tools	108	skills	2	historical	1
food	96	origin	2	function or usage	1
clothing	93	cultural identity	2	symbolism	1
architecture	52	technology	1	healthcare	1
sports	38	people	1	knowledge	1
location	28	community	1	social status	1
symbols	19	identity	1	religion	1
drinks	14	environment	1	cultural space	1
customs	13	traditional medicine	1	social space	1
cultural significance	6	nature	1	cultural practice	1
theatre	4	communication	1	unknown	1

Table 7: Cultural aspects targeted by the questions within the CIVQA task.

Python Pseudo-Code for the external cue settings of the CIVQA and CVVQA tasks.

```
def apply_gimmick_prompt_template(
    sample: dict[str, Any],
    regions_hint: bool,
    countries_hint: bool,
) -> str:

    prompt_template = "{QUESTION}\n{HINTS}\n"
    hints = ""

    if regions_hint:
        hints += (
            "Hint: The question is related to a cultural event or facet from the following  

            ↳ region(s): "
            f"'{', '.join(sample['regions'])}'\n"
        )

    if countries_hint:
        hints += (
            "Hint: The question is related to a cultural event or facet from the following  

            ↳ country or countries: "
            f"'{', '.join(sample['countries'])}'\n"
        )

    return prompt_template.format(
        QUESTION=sample["prompt"],
        HINTS=hints,
    )
```

Figure 73: Python Pseudo-Code to generate the prompt for a given CIVQA (or CVVQA) sample for the external cues settings.

B.4.1 System Prompt

Your Role

You are a professional annotator specialized in creating VQA samples based on a provided
↪ intangible cultural heritage(ICH) item. You will be given the following information
↪ related to the item:

- Image: An image representing one aspect of the ICH item.
- Countries of Origin: The country or countries where this ICH is recognized.
- Regions of Origin: The country or countries where this ICH is recognized.
- Title: The official title of the ICH item.
- Description: A detailed description of the ICH item, including relevant details.

Your Task

Your task is it to generate high-quality question-answer pairs in a VQA style to assess the
↪ cultural knowledge of the intangible cultural heritage (ICH) item of state-of-the-art
↪ multimodal AI models. Be sure to follow the annotation guidelines provided below to ensure
↪ the quality and relevance of the question-answer pairs.

Annotation Guidelines

Question Requirements

Make sure the question meets all of the following requirements:

1. Clear and Concise
The question is clear and concise and no longer than a single sentence.
2. Directly related to the ICH item
The question is directly related to the ICH item.
3. Directly related to the visible content
The question is directly related to the visible content in the image and requires visual
↪ analysis to answer.
4. Does not (partially) contain the answer
The question does not contain any hints or clues to or parts of the answer that would make
↪ the answer obvious.
5. Does not contain subjective words
The question does not contain subjective words like 'likely', 'possibly', 'probably',
↪ 'eventually', 'might', 'could', 'should', etc., which could introduce ambiguity.
6. Requires both image and cultural knowledge to answer
The question requires both image and cultural knowledge to answer and is not answerable by
↪ looking only at the image or only knowing about the ICH item or reading the textual
↪ description.
7. (optional) Includes specific cultural terms
The answer includes specific cultural terms, names, or phrases related to the ICH item.
↪ E.g., particular names mentioned in the description or parts of the title.

Answer Requirements

Make sure the answer meets all of the following requirements:

1. Single Word or Multiword Expression
The answer is a single word or multiword expression.
2. Clear, Objective, and Correct
The answer is clear, objective, and unambiguously correct.
3. Directly Related to Visual Content
The answer is directly related to the visual content of the image.
4. No General or Abstract Words
The answer does not contain general, abstract, or non-depictable words like "Traditional",
↪ "Cooperation", "Gathering", "Solidarity", "Community", "Indoor", "Outdoor", "Urban",
↪ "Rural", etc.
5. Verifiable by Text and Image
The answer is unambiguously verifiable by reading the textual information and inspecting
↪ the image.
6. (optional) Includes specific cultural terms
The answer includes specific cultural terms, names, or phrases related to the ICH item.
↪ E.g., particular names mentioned in the description or parts of the title.

Question Characteristics

Target Aspects

Make sure the question targets different aspects of the ICH item, such as:

- Food
- Drinks
- Clothing
- Art
- Tools
- Sports
- Instruments
- Dance
- Music
- Rituals
- Traditions
- Festivals
- Customs
- Symbols
- Architecture
- Other

Question Categories

Make sure the question falls into different categories, such as:

- Identification
 - Questions that ask for the identification of objects, people, or elements in the image.
 - ↪ E.g.: What is the name of the instrument shown in the image?
- Origin
 - Questions that inquire about the origin or source of the CEF. E.g.: Which culture or
 - ↪ country does this artifact belong to?
- Cultural Significance
 - Questions that explore the cultural or religious significance of the depicted element. E.g.:
 - ↪ What cultural or religious significance does this item hold in its native context?
- Function or Usage
 - Questions that ask about the traditional or historical function or usage of the depicted
 - ↪ element. E.g.: What was this object traditionally used for?
- Material and Craftsmanship
 - Questions that focus on the materials used and the craftsmanship involved in creating the
 - ↪ depicted element. E.g.: What material is used to construct this artifact?
- Location
 - Questions that ask about the geographical location where the cultural event or facet takes
 - ↪ place. E.g.: In which place does this dance take place?
- Symbolism
 - Questions that delve into the symbolic meanings associated with the depicted element. E.g.:
 - ↪ What does the color red symbolize in this cultural context?
- Historical
 - Questions that relate to historical events or contexts depicted in the image. E.g.: What
 - ↪ historical event is depicted in this image?
- Details
 - Questions that ask for specific details about the formation, arrangement, or other aspects
 - ↪ of the depicted element. E.g.: What formation are the dancers in?
- Other
 - Questions that do not fall into the above categories but are relevant to the ICH item.

Task Strategy

Before generating a question-answer pair, first think step-by-step and analyse the image:

1. What is visible in the image? Generate a highly detailed description of the key elements,
 - ↪ objects, or people in the image. Take into account the textual description provided to
 - ↪ identify details.
2. How does the visible content relate to the intangible cultural heritage item? Identify the
 - ↪ connection between the contents of the image and the intangible cultural heritage item.

Then, think step-by-step about potential questions:

1. What can be asked about the image that is directly related to the visible content and the
↳ intangible cultural heritage item?
2. Can a concise and clear answer to the questions be inferred from the image and the provided
↳ information?

Finally, think step-by-step before generating the final question-answer pairs:

1. Does the question-answer pair strictly adhere to the guidelines provided above? Percisly
↳ check every part of the guidelines and drop the question-answer pair if it does not meet
↳ the criteria.
2. What aspect of the intangible cultural heritage item is targeted with the question?
3. What category does the question fall into?

Output Format

For each question-answer pair, provide the following information in the following format:

```
```xml
<vqa-task>
 <image-analysis>
 <description>
 <!-- PUT YOUR DETAILED DESCRIPTION OF THE IMAGE HERE -->
 </description>
 <cultural-relatetness>
 <!-- PUT YOUR ANALYSIS OF HOW THE CONTENTS OF THE IMAGE RELATE TO THE INTANGIBLE
 ↳ CULTURAL HERITAGE ITEM HERE -->
 </cultural-relatetness>
 </image-analysis>
 <potential-questions>
 <qa-candidate>
 <question>
 <!-- PUT YOUR QUESTION HERE -->
 </question>
 <answer>
 <!-- PUT YOUR ANSWER HERE -->
 </answer>
 <guideline-adherence>
 <question-requirments>
 <clear-and-concise>
 <!-- YES OR NO -->
 </clear-and-concise>
 <directly-related-to-ich>
 <!-- YES OR NO -->
 </directly-related-to-ich>
 <directly-related-to-visual-content>
 <!-- YES OR NO -->
 </directly-related-to-visual-content>
 <does-not-contain-answer>
 <!-- YES OR NO -->
 </does-not-contain-answer>
 <does-not-contain-subjective-words>
 <!-- YES OR NO -->
 </does-not-contain-subjective-words>
 <requires-both-image-and-cultural-knowledge>
 <!-- YES OR NO -->
 </requires-both-image-and-cultural-knowledge>
 <includes-specific-cultural-terms>
 <!-- YES OR NO -->
 </includes-specific-cultural-terms>
 </question-requirments>
 <answer-requirments>
 <single-word-or-multiword-expression>
 <!-- YES OR NO -->
 </single-word-or-multiword-expression>
 <clear-objective-and-correct>
 <!-- YES OR NO -->
 </clear-objective-and-correct>
 </answer-requirments>
 </guideline-adherence>
 </qa-candidate>
 </potential-questions>
</vqa-task>
```

```

 </clear-objective-and-correct>
 <directly-related-to-visual-content>
 <!-- YES OR NO -->
 </directly-related-to-visual-content>
 <no-general-or-abstract-words>
 <!-- YES OR NO -->
 </no-general-or-abstract-words>
 <verifiable-by-text-and-image>
 <!-- YES OR NO -->
 </verifiable-by-text-and-image>
 <includes-specific-cultural-terms>
 <!-- YES OR NO -->
 </includes-specific-cultural-terms>
 </answer-requirements>
 </guideline-adherence>
</qa-candidate>
...
</potential-questions>
<final-qa-pairs>
 <!-- PUT ALL QA PAIRS THAT MEET ALL MANDATORY REQUIREMENTS HERE -->
 <qa-pair>
 <meets-requirements>
 <!-- DOES YOUR QUESTION-ANSWER PAIR MEET ALL MANDATORY REQUIREMENTS? YES OR NO -->
 </meets-requirements>
 <final-result-json>
 <!-- PUT YOUR FINAL RESULT AS JSON HERE -->
 {
 "question": <insert question here>,
 "answer": <insert answer here>,
 "target_aspect": <insert target aspect here>
 "question_category": <insert question category here>
 }
 </final-result-json>
 </qa-pair>
 ...
</final-qa-pairs>
</vqa-task>
...

```

## B.4.2 User Prompt Template

```

Intangible Cultural Heritage Item

Image

{IMAGE_PLACEHOLDER}

Countries of Origin:

{LIST_OF_COUNTRIES}

Regions of Origin

{LIST_OF_REGIONS}

Title

{TITLE}

Description

{DESCRIPTION}

```

## B.5 Annotation Project Details

We first conducted several internal pilot studies to iteratively create a straightforward annotation task, guidelines, and an intuitive interface for the final annotation project. To find annotators, we advertised the task in our faculty research network, emphasizing our goal of creating a culturally diverse benchmark for assessing the cultural awareness of current AI models. Therefore, we targeted primarily individuals from non-Western cultural backgrounds. We found 18 volunteers who have spent most of their lives in 10 different countries from all six regions and thus cover diverse cultural backgrounds (see Table 8). To train the annotators, we provided detailed annotation guidelines, followed by an oral introduction to the task. For more details, refer to the (anonymized) original annotation guidelines we [shared here](#).

For the second annotation round, we hired 5 of the previous volunteering annotators (0, 1, 8, 15, 17) who assessed the kept samples from the first round to obtain two annotations (from distinct annotators) per sample. We paid the second-round annotators a salary of roughly 12.5€ per hour.

ID	AGE	PRONOUNS	EDUCATION	COUNTRY	REGION	ROUND(S)
0	23	she/her	Bachelor	Iran	■ AP	1, 2
1	23	she/her	Bachelor	Iran	■ AP	1, 2
2	28	she/her	PhD	Russia	■ E	1
3	35	he/him	Master	Germany	■ W	1
5	29	he/him	Bachelor	Guatemala	■ LAC	1
6	29	he/him	Master	Germany	■ W	1
7	42	he/him	PhD	Ethiopia	■ SA	1
8	23	he/him	Bachelor	Egypt	■ A	1, 2
9	33	she/her	Master	Iran	■ AP	1
10	29	she/her	Bachelor	Afghanistan	■ AP	1
11	23	she/her	Bachelor	India	■ AP	1
12	33	he/him	Bachelor	Germany	■ W	1
13	22	she/her	Bachelor	Pakistan	■ AP	1
14	27	he/him	Master	China	■ AP	1
15	29	she/her	High School	Germany	■ W	1, 2
16	22	she/her	Bachelor	China	■ AP	1
17	26	he/him	High School	Germany	■ W	1, 2, 3

Table 8: Demographics of the annotators who participated in our VQA annotation project. For the country, we asked the question, “Where did you spend most of your life?”. The Round(s) column indicates which annotation rounds the annotator participated in.


### B.5.1 CIVQA Annotation Interface

For the annotation project, we used a self-hosted Label Studio<sup>19</sup> instance with a custom

labeling interface (see Figure 74) for all annotation projects.

<sup>19</sup><https://labelstud.io/>

### Cultural Event or Facet (CEF)



**Title**  
Cultural practices and expressions linked to the M'Bolon, a traditional musical percussion instrument

**Regions**  
'Subsaharian African States'

**Countries**  
'Mali'

**Description**  
The M'Bolon is a musical instrument with a large calabash sound box covered with cowhide and a bow-shaped wooden neck with strings. To amplify the sound vibrations, the player often wears a belt-like device made of metal plates with small oval-shaped lobes. This device is fitted with small iron rings and attached to the player's hand by means of a pad with cords or an elastic band. The number of strings of the M'Bolon determines how it is used. Single-stringed and two-stringed M'Bolon are used for popular events and celebrations, as well as for rituals and religious ceremonies. Three-stringed and four-stringed M'Bolon are the most common. They are used to accompany the praising of traditional chiefs, celebrate the heroic deeds of kings, accompany farmers in the fields and rouse warriors. The M'Bolon can be played on its own or with other instruments, including the xylophone, talking drum and lutes. It is played in southern Mali by people of all ethnicities, genders and religions, and is taught through apprenticeship and by local associations. However, there is a limited number of initiatives, and the practice is threatened by factors such as urbanization, the introduction of religions that prohibit traditional initiatory rites and practices, and decreasing interest among youth.

### Generated Question-Answer Pair

**Question**  
What traditional musical instrument is the man holding in the field?

**Answer**  
M'Bolon

**Answer Sentences**  
The following list contains sentences from the description containing the answer. This should help you to quickly verify the answer.

- The **M'Bolon** is a musical instrument with a large calabash sound box covered with cowhide and a bow-shaped wooden neck with strings.
- The number of strings of the **M'Bolon** determines how it is used.
- Single-stringed and two-stringed **M'Bolon** are used for popular events and celebrations, as well as for rituals and religious ceremonies.
- Three-stringed and four-stringed **M'Bolon** are the most common.
- The **M'Bolon** can be played on its own or with other instruments, including the xylophone, talking drum and lutes.

### Questionnaire

**Question Requirements**  
Check all requirements that are met by the question.

> Show Description

- ☐ Clear and Concise<sup>[1]</sup>
- ☐ Directly related to the CEF<sup>[2]</sup>
- ☐ Directly related to the visible content<sup>[3]</sup>
- ☐ Does not (partially) contain the answer<sup>[4]</sup>
- ☐ Does not contain subjective words<sup>[5]</sup>
- ☐ Requires the image to answer<sup>[6]</sup>
- ☐ Requires cultural knowledge to answer<sup>[7]</sup>

**Answer Requirements**  
Check all requirements that are met by the answer.

> Show Description


- ☐ Single Word or Multiword Expression<sup>[8]</sup>
- ☐ Clear, Objective, and Correct<sup>[9]</sup>
- ☐ Directly Related to Visual Content<sup>[10]</sup>
- ☐ No General or Abstract Words<sup>[11]</sup>

**Keep or Reject**  
Decide whether to keep or reject the question-answer pair.

> Show Description

- ☐ Keep - All requirements are met<sup>[12]</sup>
- ☐ Reject - Not all requirements are met<sup>[13]</sup>

### Cultural Event or Facet (CEF)



**Title**  
Cherry festival in Sefrou

**Regions**  
'Arab States'

**Countries**  
'Morocco'

**Description**  
For three days in June each year, the local population of Sefrou celebrates the natural and cultural beauty of the region, symbolized by the cherry fruit and that year's newly chosen Cherry Queen, selected during a pageant that draws competitors from the region and entire country. The highlight of the festival is a parade with performing troupes, rural and urban music, magicians and bands, and floats featuring local products. At the centre is the Cherry Queen, who offers cherries to onlookers while dressed ornately and surrounded by attendants. The whole population contributes to the success of the festival: craftsmen make silk buttons for traditional dresses, fruit growers supply cherries, local sports clubs participate in competitions, and music and dancing troupes animate the entire festival. The cherry festival provides an opportunity for the entire city to present its activities and achievements. The younger generation are also integrated into festival activities to ensure their sustainability. The festival is a source of pride and belonging that enhances the self-esteem of the city and its people and constitutes a fundamental contribution to their local identity.

### Generated Question-Answer Pair

**Question**  
What title is given to the woman wearing the sash in the image?

**Answer**  
Cherry Queen

**Answer Sentences**  
The following list contains sentences from the description containing the answer. This should help you to quickly verify the answer.

- For three days in June each year, the local population of Sefrou celebrates the natural and cultural beauty of the region, symbolized by the cherry fruit and that year's newly chosen **Cherry Queen**, selected during a pageant that draws competitors from the region and entire country.
- At the centre is the **Cherry Queen**, who offers cherries to onlookers while dressed ornately and surrounded by attendants.

### Questionnaire

**Question Requirements**  
Check all requirements that are met by the question.

> Show Description

- ☐ Clear and Concise<sup>[1]</sup>
- ☐ Directly related to the CEF<sup>[2]</sup>
- ☐ Directly related to the visible content<sup>[3]</sup>
- ☐ Does not (partially) contain the answer<sup>[4]</sup>
- ☐ Does not contain subjective words<sup>[5]</sup>
- ☐ Requires the image to answer<sup>[6]</sup>
- ☐ Requires cultural knowledge to answer<sup>[7]</sup>

**Answer Requirements**  
Check all requirements that are met by the answer.

> Show Description


- ☐ Single Word or Multiword Expression<sup>[8]</sup>
- ☐ Clear, Objective, and Correct<sup>[9]</sup>
- ☐ Directly Related to Visual Content<sup>[10]</sup>
- ☐ No General or Abstract Words<sup>[11]</sup>

**Keep or Reject**  
Decide whether to keep or reject the question-answer pair.

> Show Description

- ☐ Keep - All requirements are met<sup>[12]</sup>
- ☐ Reject - Not all requirements are met<sup>[13]</sup>

### Cultural Event or Facet (CEF)



**Title**  
Fijri

**Regions**  
'Arab States'

**Countries**  
'Bahrain'

**Description**  
Fijri is a musical performance that commemorates the history of pearl diving in Bahrain. Dating back to the late nineteenth century, it was traditionally performed by pearl divers and pearling crews to express the hardships faced at sea. The performers sit in a circle, singing and playing different types of drums, finger chimes and a jahl, a clay pot used as an instrument. The centre of the circle is occupied by the dancers and the lead singer, who is in charge of conducting the performance. Fijri originated on the Island of Muharraq, where, up until the mid-twentieth century, most of the population formed part of the pearling community. However, today the practice has reached a wider audience through performances in festivals across all regions of Bahrain. It is now well-known across the country and is viewed as a means of expressing the connection between the Bahraini people and the sea. Fijri is usually performed in cultural spaces called durs by descendants of pearl divers and pearling crews and by other individuals. Although it is performed by all-male groups, Fijri is enjoyed by all members of the community. The words, rhythms and instruments are used to convey the values of perseverance, strength and resourcefulness.

### Generated Question-Answer Pair

**Question**  
What is the name of the cultural performance shown in the image?

**Answer**  
Fijri

**Answer Sentences**  
The following list contains sentences from the description containing the answer. This should help you to quickly verify the answer.

- Fijri** is a musical performance that commemorates the history of pearl diving in Bahrain.
- Fijri** originated on the Island of Muharraq, where, up until the mid-twentieth century, most of the population formed part of the pearling community.
- Fijri** is usually performed in cultural spaces called durs by descendants of pearl divers and pearling crews and by other individuals.
- Although it is performed by all-male groups, **Fijri** is enjoyed by all members of the community.

### Questionnaire

**Question Requirements**  
Check all requirements that are met by the question.

> Show Description

- ☐ Clear and Concise<sup>[1]</sup>
- ☐ Directly related to the CEF<sup>[2]</sup>
- ☐ Directly related to the visible content<sup>[3]</sup>
- ☐ Does not (partially) contain the answer<sup>[4]</sup>
- ☐ Does not contain subjective words<sup>[5]</sup>
- ☐ Requires the image to answer<sup>[6]</sup>
- ☐ Requires cultural knowledge to answer<sup>[7]</sup>

**Answer Requirements**  
Check all requirements that are met by the answer.

> Show Description

- ☐ Single Word or Multiword Expression<sup>[8]</sup>
- ☐ Clear, Objective, and Correct<sup>[9]</sup>
- ☐ Directly Related to Visual Content<sup>[10]</sup>
- ☐ No General or Abstract Words<sup>[11]</sup>

**Keep or Reject**  
Decide whether to keep or reject the question-answer pair.

> Show Description

- ☐ Keep - All requirements are met<sup>[12]</sup>
- ☐ Reject - Not all requirements are met<sup>[13]</sup>

Figure 74: Three screenshots showing examples of the Label Studio interface used in our CIVQA annotation tasks.

## B.5.2 First Annotation Round Statistics

Country	Count	Country	Count
United Arab Emirates	101	Nicaragua	18
China	98	Chile	17
Oman	91	Serbia	17
Saudi Arabia	87	Cambodia	17
France	86	Bangladesh	17
Croatia	84	Bulgaria	17
Algeria	82	Qatar	17
Morocco	81	Ireland	17
Türkiye	78	Panama	16
Peru	75	Ukraine	16
Spain	74	Malaysia	16
Azerbaijan	69	Namibia	16
Colombia	68	Philippines	15
Islamic Republic of Iran	66	Bosnia and Herzegovina	15
Mali	65	Niger	15
Mexico	64	Estonia	14
Republic of Korea	62	Netherlands	14
Egypt	62	Zimbabwe	14
Tunisia	56	Senegal	14
Iraq	54	Madagascar	14
Japan	52	Belarus	13
Brazil	50	Luxembourg	13
Italy	50	Togo	12
Belgium	50	Burundi	12
Plurinational State of Bolivia	49	Dominican Republic	12
Mauritania	49	Congo	11
Bolivarian Republic of Venezuela	47	Democratic Republic of the Congo	11
Nigeria	46	Benin	11
India	45	Finland	11
Malawi	43	Angola	10
Palestine	40	Afghanistan	10
Greece	38	Seychelles	10
Uzbekistan	37	Democratic People's Republic of Korea	10
Kuwait	37	Norway	9
Kyrgyzstan	36	Lao Peoples Democratic Republic	9
Cuba	35	Burkina Faso	9
Mauritius	34	Sweden	9
Mongolia	34	Bahamas	9
Czechia	34	Georgia	9
Jordan	32	Albania	9
Zambia	31	Republic of Moldova	9
Côte d'Ivoire	31	Cabo Verde	8
Syrian Arab Republic	31	North Macedonia	8
Kazakhstan	30	Jamaica	8
Portugal	29	Honduras	7
Switzerland	29	Latvia	7
Uganda	29	Denmark	7
Ethiopia	29	Pakistan	7
Botswana	28	Belize	7
Viet Nam	28	Uruguay	7
Argentina	28	Timor-Leste	6
Armenia	28	Montenegro	6
Yemen	28	Sri Lanka	6
Turkmenistan	26	Thailand	6
Sudan	26	Guinea	6
Bahrain	26	Malta	5
Indonesia	26	Andorra	5
Ecuador	25	Russian Federation	5
Mozambique	25	Lithuania	5
Tajikistan	25	Tonga	4
Austria	24	Costa Rica	4
Hungary	24	Cameroon	4
Slovakia	23	Vanuatu	3
Lebanon	23	Singapore	3
Cyprus	22	Gambia	3
Slovenia	22	Iceland	3
Paraguay	21	Federated States of Micronesia	2
Germany	21	Grenada	2
Romania	21	Samoa	2
Guatemala	20	Bhutan	1
Kenya	20	Djibouti	1
Poland	20	Central African Republic	1

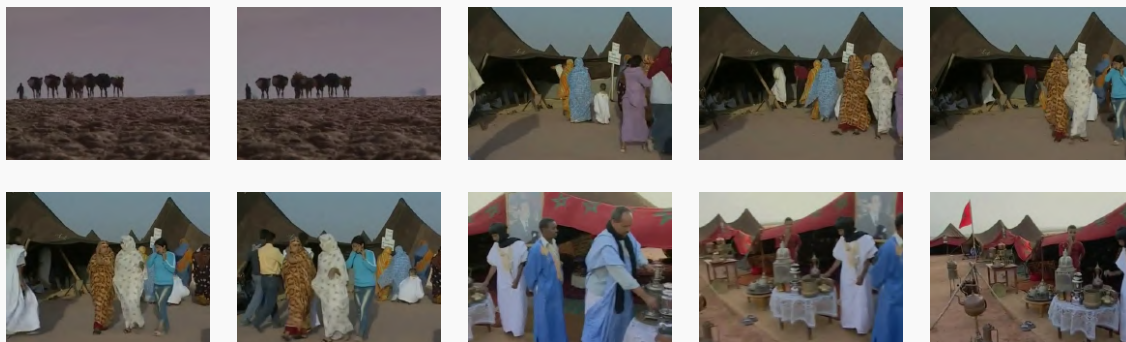
Table 9: The number of countries related to the QA pairs collected in the first annotation round for CIVQA.

## C VVQA Details

### C.1 Examples

In the following, we provide one random sample per region for the CVVQA task. Note that the lower part of the examples, where the related CEF is provided, is *not* part of the actual sample.

■ A



Question: What event are the women in the video participating in?

Answer: Moussem of Tan-Tan

#### Related Cultural Event or Facet

Title: Moussem of Tan-Tan

Countries: Morocco

Regions: Arab States

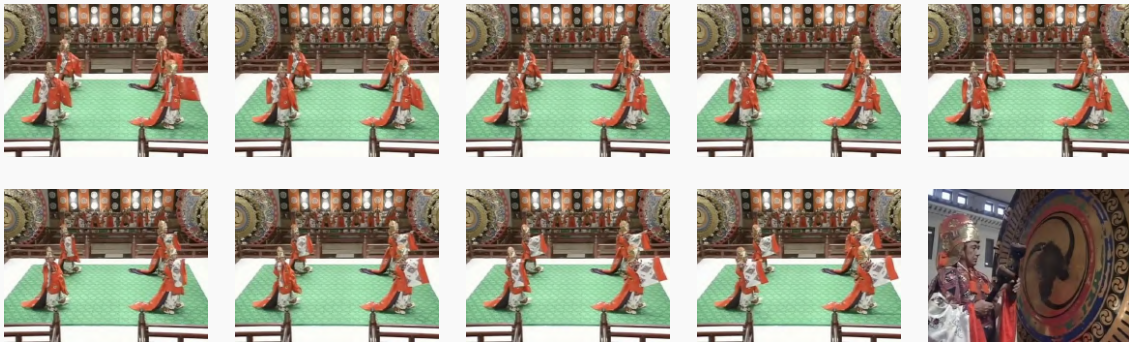
Description:

The Moussem of Tan-Tan in southwest Morocco is an annual gathering of nomadic peoples of the Sahara that brings together more than thirty tribes from southern Morocco and other parts of northwest Africa. Originally this was an annual event around the month of May. Part of the agricultural and herding calendar of the nomads, these gatherings were an opportunity to group together, buy, sell and exchange foodstuffs and other products, organize camel and horse-breeding competitions, celebrate weddings and consult herbalists. The Moussem also included a range of cultural expressions such as musical performances, popular chanting, games, poetry contests and other Hassanie oral traditions.

These gatherings took the form of a Moussem (a type of annual fair with economic, cultural and social functions) in 1963 when the first Moussem of Tan-Tan was organized to promote local traditions and provide a place for exchange, meeting and celebration. The Moussem is said to have been initially associated with Mohamed Laghdaf, who resisted the Franco-Spanish occupation. He died in 1960, and his tomb lies near the town. However, between 1979 and 2004 it was not possible to hold the Moussem because of security problems in the region.

Today, the nomadic populations are particularly concerned to protect their way of life. Economic and technical upheavals in the region have profoundly altered the lifestyle of the nomadic Bedouin communities, forcing many of them to settle. Moreover, urbanization and rural exodus have contributed to the loss of many aspects of the traditional culture of these populations, such as crafts and poetry. Because of these risks, Bedouin communities rely strongly on the renewed Moussem of Tan-Tan to assist them in ensuring the survival of their know-how and traditions.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/moussem-of-tan-tan-00168...>



Question: What traditional Japanese performance art is depicted by the performers in the video?

Answer: Gagaku

## Related Cultural Event or Facet

Title: Gagaku

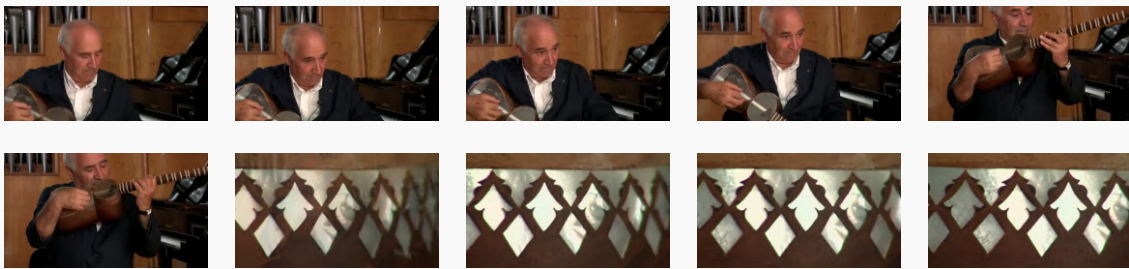
Countries: Japan

Regions: Asian and Pacific States

Description:

Gagaku, characterized by long, slow songs and dance-like movements, is the oldest of the Japanese traditional performing arts. It is performed at banquets and ceremonies in the Imperial Palace and in theatres throughout the country, and encompasses three distinct arts. The first, Kuniburi no Utamai, features ancient Japanese songs, partial accompaniment by harp and flute and simple choreography. The second consists of instrumental music (especially wind instruments) and a ceremonial dance developed on the Asian continent and subsequently adapted by Japanese artists. The third, Utamono, is danced to vocal music whose texts include Japanese folk songs and Chinese poems. Influenced by the politics and culture of different periods over its long evolution, Gagaku continues to be transmitted to apprentices by masters in the Music Department of the Imperial Household Agency, many of whom are the descendants of families with deep roots in the art. It is not only an important cultural tool in confirming Japanese identity and a crystallization of the history of Japanese society, but also a demonstration of how multiple cultural traditions can be fused into a unique heritage through constant recreation over time.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/gagaku-00265...>



Question: What instrument is the individual playing in the video?  
 Answer: Tar

## Related Cultural Event or Facet

Title: Craftsmanship and performance art of the Tar, a long-necked string musical instrument

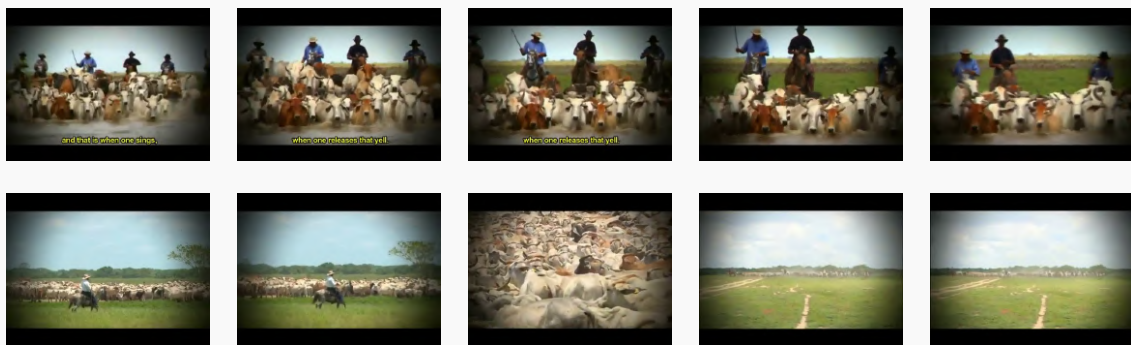
Countries: Azerbaijan

Regions: Eastern European States

Description:

The Tar is a long-necked plucked lute, traditionally crafted and performed in communities throughout Azerbaijan. Considered by many to be the country's leading musical instrument, it features alone or with other instruments in numerous traditional musical styles. Tar makers transmit their skills to apprentices, often within the family. Craftsmanship begins with careful selection of materials for the instrument: mulberry wood for the body, nut wood for the neck, and pear wood for the tuning pegs. Using various tools, crafters create a hollow body in the form of a figure eight, which is then covered with the thin pericardium of an ox. The fretted neck is affixed, metal strings are added and the body is inlaid with mother-of-pearl. Performers hold the instrument horizontally against the chest and pluck the strings with a plectrum, while using trills and a variety of techniques and strokes to add colour. Tar performance has an essential place in weddings and different social gatherings, festive events and public concerts. Players transmit their skills to young people within their community by word of mouth and demonstration, and at educational musical institutions. Craftsmanship and performance of the tar and the skills related to this tradition play a significant role in shaping the cultural identity of Azerbaijanis.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/craftsmanship-and-performance-a...>



**Question:** In which environment do the cultural practices depicted in the video typically occur?

**Answer:** Llanos

## Related Cultural Event or Facet

**Title:** Colombian-Venezuelan llano work songs

**Countries:** Colombia, Venezuela (Bolivarian Republic of)

**Regions:** Latin-American and Caribbean States

**Description:**

Colombian-Venezuelan llano work songs are a practice of vocal communication consisting of tunes sung individually, a capella, on the themes of herding and milking. The practice emerged from the close relationship between human communities and cattle and horses and is in harmony with the environmental conditions and the dynamics of nature, forming part of the traditional animal husbandry system of the Llanos. Transmitted orally from childhood, the songs are repositories of the individual and collective stories of the llaneros. Llano work songs have been gradually affected by economic, political and social processes that, modifying the llanero cultural universe, have significantly weakened the practice. For example, ambitious government plans conceived from a developmental perspective have led to profound changes in the use of the land and in ownership systems, and the modification of the social, cultural and natural sites of the songs have resulted in a loss of interest in the values and techniques of llano work. Llanero work songs thus face various threats to their viability. Efforts to safeguard the element are nonetheless widespread, including a pedagogical strategy involving more than twenty meetings for bearers and young people in the region, training projects for schoolteachers and a proliferation of festivals.

UNESCO ICH URL: <https://ich.unesco.org/en/USL/colombian-venezuelan-llano-wor...>



Question: What type of theatre is depicted in the video, known for using elaborate costumes and performances?

Answer: Kwagh-Hir

### Related Cultural Event or Facet

Title: Kwagh-Hir theatrical performance

Countries: Nigeria

Regions: Sub-Saharan African States

Description:

Kwagh-Hir theatrical performance is a composite art form encompassing a spectacle that is both visually stimulating and culturally edifying. Kwagh-hir has its roots in the story-telling tradition of the Tiv people called 'kwagh-alom', a practice where the family was treated to a storytelling session by creative storytellers, usually in the early hours of the night after the day's farming work. With time, creative storytellers began to dramatize these stories, culminating in the present stage and status of Kwagh-hir. The practice is a social performance with the potential to entertain and teach moral lessons through the dramatization and performance of past and current social realities. As a form of total theatre, Kwagh-hir incorporates puppetry, masquerading, poetry, music, dance and animated narratives in articulating the reality of the Tiv people. People's daily struggles, aspirations, successes and failures are all given expression through creative dramatization. Kwagh-hir theatre is owned by the community, with knowledge and skills being transmitted through apprenticeship. People who indicate an interest in the troupe's activities are trained and mentored until they reach a certain level of proficiency; they are then accepted into the troupe. Regular performances are held to ensure the art is kept alive and that the younger generation continues to identify with it.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/kwagh-hir-theatrical-performanc...>



Question: What traditional practice is depicted with the herders and sheep in the video?

Answer: Transhumance

### Related Cultural Event or Facet

Title: Transhumance, the seasonal droving of livestock

Countries: Albania, Andorra, Austria, Croatia, Spain, France, Greece, Italy, Luxembourg, Romania

Regions: Western European and North American States, Eastern European States

Description:

Transhumance refers to the seasonal movement of people with their livestock between geographical or climatic regions. Each year, in spring and autumn, men and women herders organise the movement of thousands of animals along traditional pastoral paths. They move on foot or horseback, leading with their dogs and sometimes accompanied by their families. An ancestral practice, transhumance stems from a deep knowledge about the environment and entails social practices and rituals related to the care, breeding and training of animals and the management of natural resources. An entire socio-economic system has been developed around transhumance, from gastronomy to local handicrafts and festivities marking the beginning and end of a season. Families have been enacting and transmitting transhumance through observation and practice for many generations. Communities living along transhumance routes also play an important role in its transmission, such as by celebrating herd crossings and organising festivals. The practice is also transmitted through workshops organised by local communities, associations and networks of herders and farmers, as well as through universities and research institutes. Transhumance thus contributes to social inclusion, strengthening cultural identity and ties between families, communities and territories while counteracting the effects of rural depopulation.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/transhumance-the-seasonal-drovi...>

## C.2 Annotation Project Details


The expert who annotated the samples was Annotator 17 from Table 8. As for the CIVQA task, we used a self-hosted Label Studio instance with a custom labeling interface. The UI is depicted in Figure 75.

## D COQA Details

### D.1 Prompts

In the following, the prompts for the COQA<sub>R</sub> and COQA<sub>C</sub> tasks are provided. For the variations involving images, the image placeholder gets replaced  $N$  times, where  $N$  is the number of images related to the target CEF.

### Cultural Event or Facet (CEF)



**Title**

**Regions**  
Asian and Pacific States\*


**Countries**  
Mongolia†

**Description**

Tsuur music is based on a combination of instrumental and vocal performance – a blending of sounds created simultaneously by both the musical instrument and the human throat. Tsuur music has an inseparable connection to the Ulaanbaatar Mongolians of the Altai Region, and remains an integral part of their daily life. Its origins lie in an ancient practice of worshipping nature and its guardian spirits by emulating natural sounds. The Tsuur is a vertical pipe-shaped wooden wind instrument with three fingerholes. Simultaneously touching the mouthpieces of the pipe with one's front teeth and spilling one's breath produces a unique timbre comprising a clear and gentle whistling sound and a drone. The Tsuur is traditionally played to ensure success for hunts, for benign weather, as a benediction for safe journeys or for weddings and other festivities. The music reflects one's inner feelings when travelling alone, connects a human to nature, and serves as a performing art. The Tsuur tradition has faded over recent decades as a consequence of negligence and animosity toward folk customs and religious faith, leaving many locales with no Tsuur performer and no families possessing a Tsuur. The forty known pieces preserved among the Ulaanbaatar Mongolians are transmitted exclusively through the memory of successive generations – a feature making this art highly vulnerable to the risk of disappearing.

### Video CLIP

The following video clip shows a part of the Cultural Event or Facet (CEF).  
It must be possible to answer the question by watching the video clip.



132 of 239    00:05:11    00:09:23

### Questionnaire: Requirements

**Video Requirements**

Check all requirements that are met!

- ☐ The video contains one or more frames very similar to the image of the CEF.<sup>[1]</sup>
- ☒ It is possible to clearly answer the question by watching the video.<sup>[1]</sup>
- ☐ None of the above.<sup>[1]</sup>

**Video Subtitles**

- ☒ The video does not contain subtitles.<sup>[1]</sup>
- ☐ The video contains English subtitles.<sup>[1]</sup>
- ☒ The video contains subtitles in a non-English language.<sup>[1]</sup>
- ☐ The video contains subtitles in English and a non-English language.<sup>[1]</sup>


**Keep or Reject**

Decide whether to keep or reject the question-answer pair:

- ☒ Keep - All requirements are met!<sup>[1]</sup>
- ☐ Reject - Not all requirements are met!<sup>[1]</sup>

---

### Cultural Event or Facet (CEF)



**Title**

**Regions**  
Asian and Pacific States\*


**Countries**  
Bangladesh†

**Description**

Shital Pati is the traditional art of making a handcrafted mat by weaving together strips of a green cane known as "Murta". The mat is used by people all over Bangladesh as a sitting mat, bedspread or prayer mat. The main weavers and practitioners are women living mostly in the low-lying villages in the greater Sylhet region of Bangladesh, but there are also pockets of Shital Pati weavers in other areas of the country. Both men and women participate in collecting and processing Murta, with women being more involved in the weaving process. The craft is a major source of livelihood and a strong marker of identity; primarily a family-based craft, it helps to reinforce family bonding and creates a harmonious social atmosphere. Mastery of the technique commands social prestige, and the practice empowers underprivileged communities, including women. The government promotes awareness of the element through local and national craft fairs, and Shital Pati communities are increasingly being organized into cooperatives to ensure the efficient safeguarding and transmission of the craft and guarantee its profitability. Safeguarding efforts involve the direct participation of the communities concerned and the practice is primarily transmitted from generation to generation within the families of craftspeople.

### Video CLIP

The following video clip shows a part of the Cultural Event or Facet (CEF).  
It must be possible to answer the question by watching the video clip.



1 of 239    00:00:00    00:09:23

### Questionnaire: Requirements

**Video Requirements**

Check all requirements that are met!

- ☒ The video contains one or more frames very similar to the image of the CEF.<sup>[1]</sup>
- ☒ It is possible to clearly answer the question by watching the video.<sup>[1]</sup>
- ☐ None of the above.<sup>[1]</sup>

**Video Subtitles**

- ☒ The video does not contain subtitles.<sup>[1]</sup>
- ☐ The video contains English subtitles.<sup>[1]</sup>
- ☒ The video contains subtitles in a non-English language.<sup>[1]</sup>
- ☐ The video contains subtitles in English and a non-English language.<sup>[1]</sup>

**Keep or Reject**

Decide whether to keep or reject the question-answer pair:

- ☒ Keep - All requirements are met!<sup>[1]</sup>
- ☐ Reject - Not all requirements are met!<sup>[1]</sup>

### Generated Question-Answer Pair

**Question**

*What activity is the woman in the video engaged in?*

**Answer**

Weaving

**Answer Sentences**

The following list contains sentences from the description containing the answer.  
This should help you to quickly verify the answer:

- Shital Pati is the traditional art of making a handcrafted mat by weaving together strips of a green cane known as "Murta".
- Both men and women participate in collecting and processing Murta, with women being more involved in the weaving process.

Figure 75: Two screenshots showing examples of the Label Studio interface used in our VVQA annotation tasks.

Region — Text-Only

From which of the following regions does the cultural event or facet with the title ``{TITLE}`` ↪ originate?

Choose from the following options and output only the corresponding letter.

A. {REGION\_OPTION\_A}  
B. {REGION\_OPTION\_B}  
C. {REGION\_OPTION\_C}  
D. {REGION\_OPTION\_D}

Your answer letter:

Region — Image-Only

<IMAGE\_PLACEHOLDER>

From which of the following countries does the cultural event or facet shown in the images ↪ originate?

Choose from the following options and output only the corresponding letter.

A. {REGION\_OPTION\_A}  
B. {REGION\_OPTION\_B}  
C. {REGION\_OPTION\_C}  
D. {REGION\_OPTION\_D}

Your answer letter:

Region — Text-Image

<IMAGE\_PLACEHOLDER>

From which of the following regions does the cultural event or facet with the title ``{TITLE}`` ↪ shown in the images originate?

Choose from the following options and output only the corresponding letter.

A. {REGION\_OPTION\_A}  
B. {REGION\_OPTION\_B}  
C. {REGION\_OPTION\_C}  
D. {REGION\_OPTION\_D}

Your answer letter:

Figure 76: Prompts for the COQA<sub>R</sub> task.

Country — Text-Only

From which of the following countries does the cultural event or facet with the title ↪ ``{TITLE}`` originate?

Choose from the following options and output only the corresponding letter.

A. {COUNTRY\_OPTION\_A}  
B. {COUNTRY\_OPTION\_B}  
C. {COUNTRY\_OPTION\_C}  
D. {COUNTRY\_OPTION\_D}

Your answer letter:

Country — Image-Only

<IMAGE\_PLACEHOLDER>

From which of the following countries does the cultural event or facet with the title ↪ ``{TITLE}`` originate?

Choose from the following options and output only the corresponding letter.

A. {COUNTRY\_OPTION\_A}  
B. {COUNTRY\_OPTION\_B}  
C. {COUNTRY\_OPTION\_C}  
D. {COUNTRY\_OPTION\_D}

Your answer letter:

Country — Text-Image

<IMAGE\_PLACEHOLDER>

From which of the following countries does the cultural event or facet with the title ↪ ``{TITLE}`` shown in the images originate?

Choose from the following options and output only the corresponding letter.

A. {COUNTRY\_OPTION\_A}  
B. {COUNTRY\_OPTION\_B}  
C. {COUNTRY\_OPTION\_C}  
D. {COUNTRY\_OPTION\_D}

Your answer letter:

Figure 77: Prompts for the COQA<sub>c</sub> task.

## D.2 Examples

In the following, we provide one random sample per region for the COQA<sub>c</sub> task in the image-only setting. For the other settings and the COQA tasks, the same pattern applies using the respective prompts from above. Note that the lower part of the examples, where the related CEF is provided, is *not* part of the actual sample.



Copyright: Huzaifa Ayad Bahaa El Din, Iraq, 2021



Copyright: Huzaifa Ayad Bahaa El Din, Iraq, 2021



Copyright: Huzaifa Ayad Bahaa El Din, Iraq, 2021



Copyright: Zahia Benabdallah, Algeria, 2021



Copyright: Azza Fahmy, Egypt, 2021



Copyright: Mustafa Kamil, Egypt, 2021



Copyright: National Heritage Preservation, Ministry of Culture, Youth and Sport and Relations with the Parliament, Egypt, 2022



Copyright: Direction du Patrimoine Culturel, Morocco, 2021



Copyright: Direction du Patrimoine Culturel, Morocco, 2021



Copyright: Ministry of Culture, Palestine, 2021

**Question:** In which of the following countries does the event shown in the images take place? Choose from the following options and output only the corresponding letter.

- A. Kuwait
  - B. Jordan
  - C. Egypt
  - D. United Arab Emirates
- Your answer letter:  
Answer: C

## Related Cultural Event or Facet

**Title:** Arts, skills and practices associated with engraving on metals (gold, silver and copper)

**Countries:** Algeria, Saudi Arabia, Egypt, Iraq, Morocco, Mauritania, Palestine, Sudan, Tunisia, Yemen

**Regions:** Arab States

**Description:**

Engraving on metals such as gold, silver and copper is a centuries-old practice that entails manually cutting words, symbols or patterns into the surfaces of decorative, utilitarian, religious or ceremonial objects. The craftsperson uses different tools to manually cut symbols, names, Quran verses, prayers and geometric patterns into the objects. Engravings can be concave (recessed) or convex (elevated), or the result of a combination of different types of metals, such as gold and silver. Their social and symbolic meanings and functions vary according to the communities concerned. Engraved objects, such as jewelry or household objects, are often presented as traditional gifts for weddings or used in religious rituals and alternative medicine. For instance, certain types of metals are believed to have healing properties. Engraving on metals is transmitted within families, through observation and hands-on practice. It is also transmitted through workshops organized by training centres, organizations and universities, among others. Publications, cultural events and social media further contribute to the transmission of the related knowledge and skills. Practised by people of all ages and genders, metal engraving and the use of engraved objects are means of expressing the cultural, religious and geographical identity and the socioeconomic status of the communities concerned.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/arts-skills-and-practices-assoc...>



Copyright: Public Foundation  
'Min Kiyal', Kyrgyzstan, 2018



Copyright: Public Foundation  
'Min Kiyal', Kyrgyzstan, 2018



Copyright: Public Foundation  
'Min Kiyal', Kyrgyzstan, 2018



Copyright: Public Foundation  
'Min Kiyal', Kyrgyzstan, 2018



Copyright: Public Foundation  
'Min Kiyal', Kyrgyzstan, 2018



Copyright: Public Foundation  
'Min Kiyal', Kyrgyzstan, 2018



Copyright: Public Foundation  
'Min Kiyal', Kyrgyzstan, 2018



Copyright: Public Foundation  
'Min Kiyal', Kyrgyzstan, 2018



Copyright: Public Foundation  
'Min Kiyal', Kyrgyzstan, 2018



Copyright: Public Foundation  
'Min Kiyal', Kyrgyzstan, 2018

Question: In which of the following countries does the event shown in the images take place? Choose from the following options and output only the corresponding letter.

A. Kyrgyzstan

B. Timor-Leste

C. Thailand

D. Turkmenistan

Your answer letter:

Answer: A

## Related Cultural Event or Facet

Title: Ak-kalpak craftsmanship, traditional knowledge and skills in making and wearing Kyrgyz men's headwear

Countries: Kyrgyzstan

Regions: Asian and Pacific States

Description:

Ak-kalpak craftsmanship is a traditional Kyrgyz handicraft. The Ak-kalpak is a traditional male hat made with white felt, which bears deep sacral meanings. Ak-kalpak craftsmanship is a cumulative, ever-evolving body of knowledge and skills passed down by craftswomen in the communities concerned comprising felting, cutting and sewing and pattern embroidery. Related knowledge and skills are transmitted via oral coaching, hands-on training and joint making in workshops. More than eighty kinds of Ak-kalpak can be distinguished, decorated with various patterns bearing a sacred meaning and history. Environmentally friendly and comfortable, the Ak-kalpak resembles a snow peak, with four sides representing the four elements: air, water, fire and earth. The four edging lines symbolize life, with the tassels on the top symbolizing ancestors' posterity and memory, and the pattern symbolizing the family tree. Ak-kalpak unites different Kyrgyz tribes and communities and makes Kyrgyz people recognizable to other ethnic groups. It also fosters inclusivity when representatives of other ethnic groups wear it on holidays or days of mourning to express unity and sympathy. There are workshops all over the country where related knowledge and skills are passed down, and in 2013 a project entitled 'From generation to generation' was conducted on traditional Ak-kalpak-making techniques nationwide, resulting in an exhibition and published book.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/ak-kalpak-craftsmanship-traditi...>



Copyright: Lithuanian National Culture Centre, Archive, 2021



Copyright: Vilnius Ethnic Culture Centre, Archive, 2021



Copyright: Vilnius Ethnic Culture Centre, Archive, 2021



Copyright: Lithuanian National Culture Centre, Archive, 2021



Copyright: Vilnius Ethnic Culture Centre, Archive, 2021



Copyright: Vilnius Ethnic Culture Centre, Archive, 2021



Copyright: Vilnius Ethnic Culture Centre, Archive, 2021



Copyright: Lithuanian National Culture Centre, Archive, 2021



Copyright: Marija Liugienė, Archive, 2003



Copyright: Lithuanian National Culture Centre, Archive, 2021

**Question:** In which of the following countries does the event shown in the images take place? Choose from the following options and output only the corresponding letter.

- A. Lithuania
- B. Bosnia and Herzegovina
- C. Russia
- D. Poland

Your answer letter:

Answer: A

## Related Cultural Event or Facet

Title: Sodai straw garden making in Lithuania

Countries: Lithuania

Regions: Eastern European States

Description:

Sodai straw gardens are hanging ornaments made from the stalks of grains. This practice involves the cultivation of grain (typically rye), the treatment of straw and the creation of geometric structures of varying sizes. The structures are then decorated with details symbolizing fertility and prosperity. Sodai gardens are believed to reflect the pattern of the universe and are associated with well-being and spirituality. They are hung over the cradles of babies and over a wedding or family table to wish happiness to newborns, fertility to newlyweds or harmony to the family. Lithuanian homes are also frequently decorated with sodai gardens for Easter and Christmas. Some sodai-making families have been practising the tradition for generations. Although most of the practitioners are women, workshops exist and are open to people of all ages and genders. The practice is passed on informally within families or during events such as festivals, exhibitions, conferences and summer camps. An integral part of traditional wooden home interiors, sodai gardens are viewed as spiritual gifts. They provide a sense of shared cultural heritage and continuity to the practising communities while strengthening communal partnerships, intergenerational bonds and cultural diversity.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/sodai-straw-garden-making-in-li...>



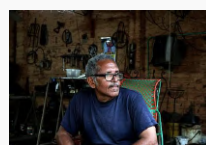
Copyright: Gerson Fonseca/Ministry of Culture of Colombia, 2018



Copyright: Gerson Fonseca/Ministry of Culture of Colombia, 2018



Copyright: Gerson Fonseca/Ministry of Culture of Colombia, 2018



Copyright: Gerson Fonseca/Ministry of Culture of Colombia, 2018



Copyright: Gerson Fonseca/Ministry of Culture of Colombia, 2018



Copyright: Gerson Fonseca/Ministry of Culture of Colombia, 2018



Copyright: Gerson Fonseca/Ministry of Culture of Colombia, 2018



Copyright: Gerson Fonseca/Ministry of Culture of Colombia, 2018



Copyright: Gerson Fonseca/Ministry of Culture of Colombia, 2018



Copyright: Gerson Fonseca/Ministry of Culture of Colombia, 2018

Question: In which of the following countries does the event shown in the images take place? Choose from the following options and output only the corresponding letter.

A. Dominican Republic

B. Chile

C. Colombia

D. Grenada

Your answer letter:

Answer: C

## Related Cultural Event or Facet

Title: Safeguarding strategy of traditional crafts for peace building

Countries: Colombia

Regions: Latin-American and Caribbean States

Description:

The safeguarding strategy of traditional crafts for peace building addresses the weakening of traditional crafts through a system of intergenerational transmission of knowledge between master and apprentice based on the non-formal 'learning by doing' method. The safeguarding strategy aims to train different sectors of the population, create labour connections and foster cultural entrepreneurship. It establishes a link between bearers of traditional crafts and skills who are recognized by their communities for their empirical knowledge of the peculiarities of their region and apprentices aged between fourteen and thirty-five who become builders of peace by learning a skill or craft, seeking to transform their situation of vulnerability. The safeguarding strategy is therefore geared at: allowing for the qualification of traditional crafts, thereby improving employment opportunities; implementing a Traditional Crafts Policy to guide and ensure continuity in the transmission and practice of these crafts; and enhancing the Workshop Schools Programme. Priority is accorded to young people who are exposed to the effects of armed conflict, a lack of opportunities, school desertion and unemployment. Training is also combined with work, guaranteeing apprentices' future employability. The strategy thus aims to foster the safeguarding of traditional crafts as a tool for social inclusion, employment and cultural entrepreneurship. In turn, the community can recognize the cultural and societal value of safeguarding different traditional skills and crafts.

UNESCO ICH URL: <https://ich.unesco.org/en/BSP/safeguarding-strategy-of-tradi...>



Copyright: Etienne Kokolo, Kinshasa, République du Congo, 2018



Copyright: Etienne Kokolo, Kinshasa, République du Congo, 2019



Copyright: Etienne Kokolo, Kinshasa, République du Congo, 2018



Copyright: Etienne Kokolo, Kinshasa, République du Congo, 2018



Copyright: Etienne Kokolo, Kinshasa, République du Congo, 2018



Copyright: Etienne Kokolo, Kinshasa, République du Congo, 2017



Copyright: Etienne Kokolo, Kinshasa, République du Congo, 2018



Copyright: Etienne Kokolo, Kinshasa, République du Congo, 2020



Copyright: Etienne Kokolo, Kinshasa, République du Congo, 2017



Copyright: Etienne Kokolo, Kinshasa, République du Congo, 2020

**Question:** In which of the following countries does the event shown in the images take place? Choose from the following options and output only the corresponding letter.

- A. Congo
- B. Togo
- C. Namibia
- D. Nigeria

Your answer letter:

Answer: A

## Related Cultural Event or Facet

Title: Congolese rumba

Countries: Congo, Democratic Republic of the Congo

Regions: Sub-Saharan African States

Description:

Congolese rumba is a musical genre and a dance common in urban areas of the Democratic Republic of the Congo and the Republic of the Congo. Generally danced by a male-female couple, it is a multicultural form of expression originating from an ancient dance called nkumba (meaning 'waist' in Kikongo). The rumba is used for celebration and mourning, in private, public and religious spaces. It is performed by professional and amateur orchestras, choirs, dancers and individual musicians, and women have played a predominant role in the development of religious and romantic styles. The tradition of Congolese rumba is passed down to younger generations through neighbourhood clubs, formal training schools and community organisations. For instance, rumba musicians maintain clubs and apprentice artists to carry on the practice and the manufacture of instruments. The rumba also plays an important economic role, as orchestras are increasingly developing cultural entrepreneurship aimed at reducing poverty. The rumba is considered an essential and representative part of the identity of Congolese people and its diaspora. It is perceived as a means of conveying the social and cultural values of the region and of promoting intergenerational and social cohesion and solidarity.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/congolese-rumba-01711...>



Copyright: Servicio de Patrimonio Histórico de la Región de Murcia, 2005



Copyright: Generalitat Valenciana, 2005



Copyright: Servicio de Patrimonio Histórico de la Región de Murcia, 2005



Copyright: Generalitat Valenciana, 2005



Copyright: Servicio de Patrimonio Histórico de la Región de Murcia, 2005



Copyright: Servicio de Patrimonio Histórico de la Región de Murcia, 2005



Copyright: Servicio de Patrimonio Histórico de la Región de Murcia, 2005



Copyright: Servicio de Patrimonio Histórico de la Región de Murcia, 2005



Copyright: Servicio de Patrimonio Histórico de la Región de Murcia, 2005



Copyright: Servicio de Patrimonio Histórico de la Región de Murcia, 2005

**Question:** In which of the following countries does the event shown in the images take place? Choose from the following options and output only the corresponding letter.

- A. Austria
- B. Spain
- C. Cyprus
- D. United Kingdom of Great Britain and Northern Ireland

Your answer letter:

Answer: B

## Related Cultural Event or Facet

**Title:** Irrigators' tribunals of the Spanish Mediterranean coast: the Council of Wise Men of the plain of Murcia and the Water Tribunal of the plain of Valencia  
**Countries:** Spain

**Regions:** Western European and North American States

**Description:**

The irrigators' tribunals of the Spanish Mediterranean coast are traditional law courts for water management that date back to the al-Andalus period (ninth to thirteenth centuries). The two main tribunals – the Council of Wise Men of the Plain of Murcia and the Water Tribunal of the Plain of Valencia – are recognized under Spanish law. Inspiring authority and respect, these two courts, whose members are elected democratically, settle disputes orally in a swift, transparent and impartial manner. The Council of Wise Men has seven geographically representative members, and has jurisdiction over a landowners' assembly of 23,313 members. The Water Tribunal comprises eight elected administrators representing a total of 11,691 members from nine communities. In addition to their legal role the irrigators' tribunals play a key part in the communities of which they are a visible symbol, as apparent from the rites performed when judgments are handed down and the fact that the tribunals often feature in local iconography. They provide cohesion among traditional communities and synergy between occupations (wardens, inspectors, pruners, etc.), contribute to the oral transmission of knowledge derived from centuries-old cultural exchanges, and have their own specialist vocabulary peppered with Arabic borrowings. In short, the courts are long-standing repositories of local and regional identity and are of special significance to local inhabitants.

UNESCO ICH URL: <https://ich.unesco.org/en/RL/irrigators-tribunals-of-the-spa...>

## E CKQA Details

### E.1 Prompts

In the following, the prompts for the CKQA<sub>N</sub> and CKQA<sub>D</sub> tasks are provided. For the variations involving images, the image placeholder gets replaced  $N$  times, where  $N$  is the number of images related to the target CEF. Examples without the respective prompts, i.e., only the related CEFs, are provided in §A.2.1.

**Naming — Image-Only**

Name the cultural event or facet depicted by the following images. Answer briefly and  
↔ concisely.

<IMAGE\_PLACEHOLDER>

Your answer:

Figure 138: Prompt for the CKQA<sub>N</sub> task.

**Describing — Text-Only**

Write a brief essay about the cultural event or facet with the title ``{TITLE}``.

Your answer:

**Describing — Image-Only**

Write a brief essay about the cultural event or facet depicted by the following images.

<IMAGE\_PLACEHOLDER>

Your answer:

**Describing — Text-Image**

Write a brief essay about the cultural event or facet depicted by the following images. It has  
↔ the title ``{TITLE}``.

<IMAGE\_PLACEHOLDER>

Your answer:

Figure 139: Prompts for the CKQA<sub>D</sub> task.

## F Experimental Setup

For inference, we load all models using the *transformers* library (v. 4.48.0) in 16-bit with Flash Attention 2 (Dao et al., 2022; Dao, 2024) (v. 2.7.3), PyTorch (v. 2.4.0), and CUDA (v12.1). We used A40 (46GB) GPUs for models up to 26B parameters, A100 (80GB) GPUs for models up to 38B parameters, and two H100 (96GB) GPUs for 70B+ models in a multi-GPU setup. To generate responses, we use greedy decoding, i.e., we use the following arguments for the generation method:

```
generation_kwargs = {
 "max_new_tokens": 512,
 "do_sample": False,
 "temperature": None,
 "top_p": None,
 "top_k": None,
}
```

More details and exact hyperparameters are documented in the code base: <https://github.com/floschne/gimmick>.

## G Results and Analyses

### G.1 CIVQA

#### G.1.1 Results

##### Relaxed Accuracy

Model	West EU & North America				Asia & Pacific				Subsaharian Africa				Arab				East EU				Latin-America & Caribbean				Average			
	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B
GPT-4o	31.58	34.39	41.05	40.70	29.89	31.37	36.63	37.68	17.38	17.63	32.49	31.74	25.70	30.53	39.19	39.95	26.80	32.56	42.94	41.79	23.17	25.77	30.26	32.39	25.44	28.17	36.59	37.08
Gemini Pro	27.02	30.53	31.23	32.28	22.53	26.11	31.16	29.68	16.84	14.61	26.20	24.18	19.85	22.39	28.50	28.50	25.07	25.94	31.12	32.56	22.46	19.86	24.35	27.19	21.50	22.84	28.30	28.30
GPT-4o Mini	23.86	25.26	30.18	29.82	21.05	21.89	26.74	26.53	9.32	10.58	16.12	15.37	17.30	19.85	25.45	25.95	19.02	19.02	28.53	28.24	16.31	17.73	23.64	22.93	17.38	18.54	24.81	24.59
Gemini Flash	22.81	25.96	27.02	24.91	18.95	20.21	26.11	25.89	12.91	10.83	20.40	18.89	15.27	17.56	20.36	20.61	20.17	19.31	24.78	24.50	14.66	16.55	22.22	20.57	16.85	18.00	23.29	22.44
InternVL2.5 78B	25.61	23.86	29.82	29.12	20.21	19.79	26.32	27.58	10.33	11.08	20.40	20.40	17.81	19.85	27.74	27.99	19.02	17.58	24.50	23.63	13.95	15.13	20.80	21.51	16.75	16.97	24.45	24.72
Qwen2 VL 72B	22.46	22.81	29.82	29.12	17.47	19.16	21.47	23.37	8.31	8.56	12.85	13.10	13.99	16.28	20.10	19.85	21.04	20.46	28.53	29.39	13.00	14.66	19.86	19.62	15.32	16.26	21.45	21.59
InternVL2.5 38B	23.86	23.16	28.77	29.82	17.26	17.89	22.32	23.16	9.07	8.82	17.88	16.62	14.25	17.30	23.16	22.65	16.14	17.29	24.78	23.92	11.82	12.29	17.97	17.49	14.55	15.41	21.99	21.63
Claude 3.5 Sonnet	19.65	17.19	22.11	24.21	16.42	12.84	18.11	22.95	6.30	4.53	10.58	11.59	13.99	11.20	17.81	20.61	16.71	14.12	21.90	21.61	13.48	13.00	17.97	22.93	14.02	11.64	17.60	20.24
InternVL2.5 26B	20.00	19.65	25.61	25.96	13.26	14.95	18.95	18.74	6.30	6.80	11.59	12.34	12.98	14.76	20.61	21.12	15.56	14.41	21.04	21.04	13.00	14.89	19.62	19.39	13.03	14.15	19.44	19.61
Llama 3.2 11B Vision	16.49	18.95	20.70	20.35	13.26	12.84	15.79	16.84	5.29	5.54	9.07	8.31	7.89	7.89	10.18	10.18	12.68	13.54	17.29	19.02	11.82	11.82	13.95	14.66	10.61	11.06	13.97	14.20
InternVL2.5 8B	19.30	17.89	23.16	23.51	11.79	12.00	16.42	16.84	5.04	6.30	10.58	9.57	9.41	9.67	14.50	14.25	9.80	9.80	15.27	15.56	9.46	9.69	13.71	14.89	10.34	10.39	15.41	15.41
Qwen2 VL 7B	17.19	17.19	20.35	18.95	9.47	9.47	12.00	11.37	5.79	6.30	8.56	8.31	8.91	9.92	11.45	11.45	10.95	12.10	15.56	14.41	9.69	11.11	13.00	13.24	9.63	10.26	12.76	12.36
Phi 3.5 Vision	14.39	12.63	20.00	18.95	8.84	10.74	13.89	13.47	6.05	6.05	8.31	8.82	6.62	8.14	9.41	9.92	8.93	8.65	14.70	14.70	8.27	9.93	13.71	12.77	8.55	9.18	12.99	12.85
MiniCPM V 2.6	12.98	14.39	14.39	17.19	10.74	10.32	13.68	14.74	2.52	3.27	6.55	6.05	6.36	6.36	9.67	9.67	10.09	9.80	13.26	14.70	9.46	9.22	12.77	13.24	8.11	8.15	11.60	11.96
InternVL2.5 4B	14.04	16.49	16.84	14.39	9.47	14.53	13.47	9.05	3.53	7.05	7.56	4.03	7.89	9.16	9.16	6.87	8.07	11.53	10.66	8.07	8.04	11.58	11.82	7.33	7.97	11.42	11.29	7.79
Qwen2 VL 2B	13.33	12.28	13.68	14.39	9.68	9.47	11.79	10.95	4.03	3.78	5.54	4.28	6.11	5.09	6.11	6.11	8.36	8.93	12.97	12.10	7.33	8.27	10.40	9.69	7.97	7.88	9.94	9.49
Centurio Qwen	11.23	9.12	14.39	14.39	9.05	8.42	10.32	9.68	3.02	1.76	6.05	6.05	6.87	5.34	9.92	8.91	6.34	5.48	11.24	11.24	6.62	5.67	9.22	9.46	6.81	5.69	9.85	9.76
InternVL2.5 2B	6.67	7.37	10.18	9.47	4.21	4.63	6.95	5.89	2.27	2.02	3.53	5.29	2.80	3.56	5.34	5.60	3.17	3.75	7.49	6.92	5.44	5.44	8.04	6.15	4.03	4.39	6.85	6.45
InternVL2.5 1B	7.02	7.37	10.53	11.58	4.21	3.58	4.84	4.63	2.52	0.76	2.77	2.77	3.56	3.82	5.09	4.07	4.61	3.46	6.63	7.49	4.02	5.67	6.86	6.15	4.03	4.03	5.96	5.87
Centurio Aya	3.16	7.37	8.77	8.77	2.95	5.68	9.05	9.68	1.76	1.51	4.79	3.53	1.27	3.56	5.60	6.11	2.02	3.46	7.20	7.20	2.84	5.44	6.38	7.09	2.24	4.39	6.99	7.17
Average X-Large	24.04	23.33	29.82	29.12	18.84	19.47	23.89	25.47	9.32	9.82	16.62	16.75	15.90	18.07	23.92	23.92	20.03	19.02	26.51	26.51	13.48	14.89	20.33	20.57	16.03	16.61	22.95	23.15
Average Large	21.93	21.40	27.19	27.89	15.26	16.42	20.63	20.95	7.68	7.81	14.74	14.48	13.61	16.03	21.88	21.88	15.85	15.85	22.91	22.48	12.41	13.59	18.79	18.44	13.79	14.78	20.71	20.62
Average Medium	13.39	14.15	16.96	17.19	9.54	9.79	12.88	13.19	3.90	4.11	7.60	6.97	6.79	7.12	10.22	10.09	8.65	9.03	13.30	13.69	8.31	8.83	11.51	12.10	7.96	8.32	11.76	11.81
Average Small	11.09	11.23	14.25	13.75	7.28	8.59	10.19	8.80	3.68	3.93	5.54	5.04	5.39	5.95	7.02	6.51	6.63	7.26	10.49	9.86	6.62	8.18	10.17	8.42	6.51	7.38	9.40	8.49
Average Open	15.18	15.37	19.13	19.06	10.79	11.56	14.48	14.40	5.05	5.31	9.07	8.63	8.45	9.38	12.54	12.32	10.45	10.68	15.41	15.29	8.98	10.06	13.21	12.84	9.33	9.97	13.66	13.39
Average Proprietary	24.98	26.67	30.32	30.39	21.77	22.48	27.75	28.55	12.55	11.64	21.16	20.35	18.42	20.31	26.26	27.12	21.56	22.19	29.86	29.74	18.01	18.58	23.69	25.20	19.04	19.84	26.12	26.53
Average	17.63	18.19	21.93	21.89	13.54	14.29	17.80	17.94	6.93	6.89	12.09	11.56	10.94	12.11	15.97	16.02	13.23	13.56	19.02	18.90	11.24	12.19	15.83	15.93	11.76	12.44	16.78	16.67

Table 10: Cultural Image Visual Question Answering (CIVQA) scores. The reported score is relaxed accuracy. The columns N, R, C, and B stand for the hints “None”, “Region”, “Country”, and “Both”, respectively.

### Judge Score

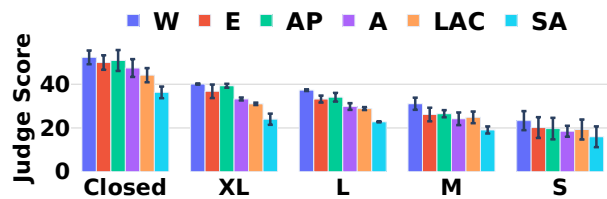


Figure 140: An overview of aggregated CIVQA Judge Score results.

Model	West EU & North America				Asia & Pacific				Subsaharian Africa				Arab				East EU				Latin-America & Caribbean				Average			
	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B
GPT-4o	55.86	55.46	64.33	64.49	56.91	56.42	63.09	63.20	39.13	35.97	48.90	47.93	51.89	56.53	65.71	67.59	54.26	55.62	67.56	67.33	48.32	46.69	53.05	54.09	51.06	51.12	60.44	60.77
Gemini Pro	55.44	56.30	63.15	62.06	54.23	54.15	59.84	59.14	38.74	39.51	47.63	46.74	49.83	52.96	58.10	57.65	51.67	52.23	62.48	63.77	46.54	45.32	51.95	52.03	49.41	50.08	57.19	56.90
Claude 3.5 Sonnet	50.65	51.35	65.33	64.14	50.11	51.07	63.70	63.59	35.06	39.90	57.56	56.52	48.99	55.34	67.49	67.86	50.42	51.80	70.12	70.20	41.49	46.10	57.85	59.53	46.12	49.26	63.68	63.64
Gemini Flash	50.49	49.93	56.29	56.11	48.19	47.42	53.20	53.11	35.60	31.32	39.58	38.54	44.04	44.25	50.85	49.03	46.66	47.34	57.24	55.22	43.59	41.58	47.74	46.88	44.76	43.64	50.82	49.81
GPT-4o Mini	48.98	48.47	53.10	54.28	44.83	44.24	49.96	49.05	32.78	35.06	35.49	35.98	42.24	43.83	47.98	48.58	46.57	43.40	55.06	53.76	40.74	38.24	44.21	44.19	42.69	42.21	47.63	47.64
Qwen2 VL 72B	40.28	40.05	48.04	46.90	38.65	39.25	43.02	44.15	25.79	26.30	31.52	30.57	32.77	35.27	42.89	41.64	38.95	39.44	50.55	49.74	30.64	33.06	39.66	40.27	34.51	35.56	42.61	42.21
InternVL2.5 78B	39.88	39.52	46.43	47.01	39.93	38.01	47.78	49.26	22.18	20.79	30.15	30.42	33.72	35.80	46.82	47.86	34.57	32.63	41.89	40.73	31.42	30.40	39.09	38.84	33.62	32.86	42.03	42.35
InternVL2.5 26B	37.00	34.65	39.75	41.00	32.64	32.97	39.10	39.47	22.63	21.71	29.40	27.22	30.89	31.39	38.10	38.81	34.34	32.38	41.14	41.53	29.34	29.69	37.05	37.78	31.14	30.47	37.42	37.64
InternVL2.5 38B	37.55	37.51	45.58	45.49	35.45	36.26	42.65	43.88	22.98	22.52	29.11	28.35	28.71	31.78	38.96	38.63	32.08	31.69	41.98	41.21	28.39	29.15	36.46	35.18	30.86	31.48	39.12	38.79
Qwen2 VL 7B	33.36	34.84	38.64	38.12	28.19	28.97	31.23	31.13	21.31	25.25	25.09	26.26	28.72	28.45	32.00	32.28	29.19	31.13	35.53	37.11	27.84	28.61	31.45	32.87	28.10	29.54	32.33	32.96
Llama 3.2 11B Vision	35.16	36.56	37.22	37.81	27.06	27.59	31.14	33.09	19.24	17.97	24.38	26.42	25.09	26.53	31.43	30.47	28.34	27.88	33.96	36.73	26.89	28.82	32.14	32.88	26.96	27.56	31.71	32.90
MiniCPM V 2.6	30.61	32.35	32.73	35.48	27.73	25.88	31.13	33.25	20.29	18.92	25.31	24.58	24.52	24.57	28.47	28.19	28.13	25.07	34.31	36.27	26.74	26.04	29.16	30.37	26.34	25.47	30.18	31.36
Qwen2 VL 2B	28.86	28.30	28.94	30.85	25.18	24.06	26.32	26.31	21.02	19.32	23.06	21.94	20.92	20.32	22.98	23.30	25.10	26.01	32.90	31.34	23.91	24.07	26.73	25.85	24.16	23.68	26.82	26.60
Centurio Aya	29.84	30.21	30.67	32.31	26.64	25.51	28.70	28.81	18.81	17.87	21.23	20.97	19.75	20.43	24.02	24.01	25.42	24.93	28.79	30.72	23.58	24.65	25.66	26.68	24.01	23.93	26.51	27.25
InternVL2.5 8B	30.12	32.19	35.35	36.47	23.62	23.93	29.75	29.92	16.70	17.20	20.54	21.81	23.61	23.46	30.65	29.92	24.94	24.67	32.73	33.66	22.80	22.13	26.78	27.99	23.63	23.93	29.30	29.96
Phi 3.5 Vision	24.84	26.93	33.43	33.45	23.46	25.36	29.18	29.02	21.28	21.65	23.63	25.92	21.06	23.26	26.18	26.48	24.70	24.88	31.32	31.82	24.47	25.73	29.67	30.56	23.30	24.64	28.90	29.54
Centurio Qwen	27.32	26.32	29.21	30.42	25.84	26.54	26.88	28.63	18.12	17.91	20.14	22.69	23.46	22.32	27.19	27.72	20.84	20.56	26.53	28.83	21.21	21.74	23.19	23.82	22.80	22.56	25.52	27.02
InternVL2.5 4B	25.18	26.06	26.71	29.67	20.67	22.04	26.29	27.53	12.32	14.45	14.99	17.91	18.42	21.62	24.43	25.80	20.22	22.07	26.43	28.43	17.56	20.93	23.45	24.29	19.06	21.19	23.72	25.61
InternVL2.5 1B	19.67	20.32	20.90	23.91	14.46	13.92	14.95	17.03	12.05	13.50	16.60	15.86	16.48	16.31	16.88	17.42	16.10	14.90	18.27	20.82	14.94	15.64	16.75	17.59	15.62	15.76	17.39	18.77
InternVL2.5 2B	18.19	19.35	20.25	21.95	14.75	15.99	16.42	18.36	13.14	10.52	12.88	14.96	15.55	14.08	16.69	18.03	14.77	13.73	17.43	18.32	15.57	15.86	18.04	18.77	15.33	14.92	16.95	18.40
Average X-Large	40.08	39.79	47.23	46.95	39.29	38.63	45.40	46.71	23.99	23.55	30.83	30.49	33.25	35.54	44.86	44.75	36.76	36.03	46.22	45.24	31.03	31.73	39.37	39.56	34.06	34.21	42.32	42.28
Average Large	37.28	36.08	42.67	43.25	34.05	34.61	40.88	41.68	22.81	22.12	29.25	27.79	29.80	31.58	38.53	38.72	33.21	32.04	41.56	41.37	28.86	29.42	36.75	36.48	31.00	30.98	38.27	38.22
Average Medium	31.07	32.08	33.97	35.10	26.51	26.40	29.80	30.80	19.08	19.19	22.78	23.79	24.19	24.29	28.96	28.77	26.14	25.71	31.98	33.89	24.84	25.33	28.06	29.10	25.31	25.50	29.26	30.24
Average Small	23.35	24.19	26.05	27.96	19.70	20.27	22.63	23.65	15.96	15.89	18.23	19.32	18.48	19.12	21.43	22.21	20.18	20.32	25.27	26.15	19.29	20.45	22.93	23.41	19.49	20.04	22.76	23.78
Average Open	30.52	31.01	34.26	35.39	26.95	27.08	30.97	31.99	19.19	19.06	23.20	23.73	24.24	25.04	29.85	30.04	26.51	26.13	32.92	33.82	24.35	25.10	29.02	29.58	25.30	25.57	30.03	30.76
Average Proprietary	52.28	52.30	60.44	60.22	50.85	50.66	57.96	57.62	36.26	36.35	45.83	45.14	47.40	50.58	58.03	58.14	49.92	50.08	62.49	62.05	44.14	43.59	50.96	51.34	46.81	47.26	55.95	55.75
Average	35.96	36.33	40.80	41.59	32.93	32.98	37.72	38.40	23.46	23.38	28.86	29.08	30.03	31.42	36.89	37.06	32.36	32.12	40.31	40.88	29.30	29.72	34.50	35.02	30.67	30.99	36.51	37.01

Table 11: Cultural Image Visual Question Answering (CIVQA) scores. The reported score is the average judge score. The columns **N**, **R**, **C**, and **B** stand for the hints “None”, “Region”, “Country”, and “Both”, respectively.

### G.1.2 Ground-Truth Answer Perplexity

The perplexity for every sample is computed as follows:

$$\text{PPL}(y \mid x) = \exp \left( -\frac{1}{N} \sum_{t=0}^N \log p(y_t \mid y_{t-1}, x) \right) \quad (1)$$

where  $x = \{s, v\}$  are the textual ( $s$ ) and visual ( $v$ ) prompt (prefix) tokens and  $y$  are the  $N$  ground-truth answer tokens.

### Results Per Cultural Aspect

We computed the average accuracy for questions targeting one of the ten most frequent cultural aspects (see §B.2), grouped by model size and region. For better interpretation, Table 12 reports the counts of questions associated with each cultural aspect per region. As shown in Table 13, our results reveal a consistent trend: models perform significantly better on tangible cultural aspects (e.g., food) than on intangible ones. For instance, across all regions, closed models achieve an average accuracy of 30% for food-related questions, compared to only 8% and 10% for questions concerning rituals and festivals, respectively. These findings highlight not only regional biases but also biases along the cultural dimension, the latter being particularly pronounced in non-Western contexts.

aspect	art	craftsmanship	dance	festivals	food	instruments	music	rituals	tools	traditions
<b>A</b>	45	32	20	6	33	20	37	32	30	76
<b>AP</b>	57	44	31	14	12	25	32	53	22	68
<b>E</b>	53	36	18	19	10	19	26	18	20	49
<b>LAC</b>	31	22	31	66	6	13	51	47	12	78
<b>SA</b>	14	16	40	16	22	64	41	73	7	70
<b>W</b>	33	27	10	30	13	14	23	18	17	49

Table 12: Number of questions targeting one of the top-10 cultural aspects per region in CIVQA.

## G.2 CVVQA

### G.2.1 Results

	AP					A					SA					W					E					LAC					OVERALL				
	A	XL	L	M	S	A	XL	L	M	S	A	XL	L	M	S	A	XL	L	M	S	A	XL	L	M	S	A	XL	L	M	S	A	XL	L	M	S
food	0.28	0.23	0.21	0.07	0.10	0.28	0.35	0.31	0.06	0.09	0.21	0.16	0.07	0.03	0.04	0.18	0.12	0.19	0.07	0.09	0.18	0.36	0.31	0.10	0.12	0.68	0.54	0.71	0.31	0.39	0.30	0.29	0.30	0.11	0.14
instruments	0.29	0.27	0.25	0.05	0.07	0.20	0.15	0.12	0.03	0.04	0.16	0.15	0.16	0.03	0.04	0.26	0.37	0.32	0.05	0.11	0.32	0.31	0.26	0.04	0.05	0.45	0.44	0.31	0.15	0.20	0.28	0.28	0.24	0.06	0.08
craftsmanship	0.15	0.17	0.15	0.04	0.07	0.18	0.24	0.22	0.08	0.08	0.11	0.09	0.08	0.04	0.02	0.14	0.19	0.12	0.10	0.08	0.26	0.28	0.27	0.13	0.17	0.12	0.12	0.12	0.06	0.07	0.16	0.18	0.16	0.08	0.08
music	0.20	0.32	0.26	0.07	0.09	0.10	0.09	0.12	0.03	0.04	0.13	0.11	0.13	0.03	0.03	0.25	0.27	0.28	0.10	0.13	0.10	0.10	0.05	0.02	0.02	0.19	0.22	0.22	0.08	0.12	0.16	0.19	0.18	0.06	0.07
tools	0.19	0.29	0.18	0.09	0.11	0.18	0.17	0.18	0.05	0.06	0.00	0.05	0.04	0.00	0.04	0.22	0.15	0.19	0.09	0.11	0.14	0.17	0.17	0.04	0.04	0.23	0.17	0.30	0.05	0.05	0.16	0.17	0.18	0.05	0.07
traditions	0.19	0.18	0.15	0.06	0.07	0.11	0.09	0.09	0.04	0.05	0.06	0.07	0.06	0.04	0.04	0.21	0.25	0.21	0.09	0.09	0.16	0.16	0.15	0.06	0.08	0.14	0.16	0.13	0.06	0.08	0.14	0.15	0.13	0.06	0.07
art	0.15	0.17	0.13	0.07	0.07	0.18	0.13	0.14	0.04	0.04	0.07	0.09	0.06	0.01	0.00	0.16	0.27	0.23	0.10	0.13	0.13	0.20	0.12	0.06	0.07	0.10	0.11	0.11	0.06	0.09	0.13	0.16	0.13	0.06	0.07
dance	0.07	0.07	0.04	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.05	0.04	0.02	0.01	0.01	0.26	0.39	0.35	0.17	0.11	0.22	0.23	0.17	0.04	0.03	0.14	0.11	0.08	0.04	0.03	0.13	0.14	0.11	0.05	0.03
festivals	0.18	0.20	0.18	0.09	0.08	0.02	0.06	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.04	0.06	0.13	0.11	0.05	0.02	0.01	0.13	0.11	0.11	0.04	0.04	0.10	0.10	0.07	0.03	0.03
rituals	0.09	0.08	0.09	0.02	0.03	0.06	0.06	0.05	0.02	0.03	0.06	0.07	0.04	0.02	0.04	0.12	0.10	0.08	0.02	0.02	0.13	0.16	0.08	0.01	0.01	0.04	0.06	0.03	0.01	0.00	0.08	0.09	0.06	0.02	0.02
Average	0.18	0.20	0.16	0.06	0.07	0.13	0.13	0.12	0.03	0.04	0.09	0.08	0.07	0.02	0.03	0.19	0.22	0.21	0.08	0.09	0.18	0.21	0.16	0.05	0.06	0.22	0.20	0.21	0.09	0.11	0.16	0.17	0.16	0.06	0.07

Table 13: The averaged accuracy per region per model size group (A, XL, L, M, S) per target cultural aspect for samples in the CIVQA task.

Model	West EU & North America				Asia & Pacific				Subsaharian Africa				Arab				East EU				Latin-America & Caribbean				Average			
	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B	N	R	C	B
GPT-4o	38.97	39.91	41.31	44.13	34.56	36.41	35.71	39.17	23.67	26.33	36.67	36.00	29.18	32.46	36.72	36.39	37.59	40.43	47.16	45.74	31.98	32.56	38.95	39.24	32.67	34.49	39.19	39.97
GPT-4o Mini	38.06	31.58	34.01	38.87	29.45	25.64	25.64	29.66	20.32	13.33	15.56	20.63	28.61	24.40	25.60	29.52	35.37	29.27	30.18	38.72	25.13	21.20	23.04	25.65	28.69	23.89	24.84	29.69
Gemini Pro	33.80	37.09	40.85	39.91	30.41	31.34	34.10	34.79	20.07	22.33	28.67	28.67	26.56	28.85	32.13	32.13	32.27	33.33	36.52	36.88	28.78	30.52	33.43	32.85	28.32	29.91	33.67	33.78
Gemini Flash	29.55	29.96	30.36	34.82	22.67	24.36	26.69	26.69	12.06	12.06	15.87	19.05	20.18	20.78	21.39	23.49	26.52	27.74	32.01	31.71	23.30	24.61	26.18	27.49	21.64	22.59	24.89	26.29
Claude 3.5 Sonnet	21.86	19.84	25.91	24.29	22.46	19.92	25.21	25.85	9.21	6.03	12.38	11.11	16.87	14.16	16.87	18.37	23.17	20.12	26.52	24.70	19.11	15.45	21.47	22.25	18.74	15.89	21.44	21.24
Qwen2 VL 72B	25.35	27.23	33.33	34.27	18.43	19.12	23.73	23.73	9.00	10.00	16.33	17.00	17.05	18.36	22.62	21.64	25.53	25.53	32.27	31.91	16.57	20.93	23.55	24.42	18.13	19.62	24.27	24.65
InternVL2.5 78B	23.94	29.11	31.92	31.46	19.12	24.19	28.11	29.49	7.33	12.33	18.67	19.67	13.44	22.30	25.90	25.90	19.50	24.82	30.14	29.43	15.41	21.22	24.42	26.45	15.75	21.56	25.98	26.70
InternVL2.5 38B	22.07	28.64	32.86	32.86	18.66	24.19	27.19	26.73	6.33	13.00	21.67	21.33	13.77	22.95	24.59	27.54	19.86	26.24	30.85	30.50	13.37	20.93	26.16	24.71	14.98	21.78	26.31	26.37
Qwen2 VL 2B	19.72	18.78	21.13	23.00	13.13	14.75	16.59	15.67	6.67	4.67	7.00	6.67	13.11	11.15	13.44	12.46	16.67	16.67	17.38	17.38	15.70	15.12	16.28	16.28	13.88	13.27	15.26	14.98
Qwen2 VL 7B	18.78	18.78	22.07	21.60	14.06	14.06	17.05	16.59	5.00	6.00	7.67	7.33	13.11	15.08	17.05	17.70	15.25	17.73	18.79	19.50	15.70	18.60	19.48	20.06	13.54	14.76	16.86	16.92
InternVL2.5 26B	20.66	25.35	28.64	29.11	16.36	19.35	23.27	24.88	3.33	7.33	9.33	10.33	11.80	15.41	19.67	20.00	17.73	21.63	24.82	24.11	13.66	18.31	21.51	22.97	13.32	17.30	20.78	21.61
MiniCPM V 2.6	16.90	19.25	18.31	19.72	14.75	16.82	17.28	18.43	5.67	10.00	11.33	11.00	12.13	13.44	14.75	14.75	19.15	20.21	22.34	21.99	15.41	17.44	16.28	19.19	13.16	15.37	16.14	17.03
Phi 3.5 Vision	16.43	14.55	16.90	16.90	13.82	14.06	17.51	17.28	8.67	8.33	10.67	10.33	9.84	10.16	11.15	10.82	15.60	15.25	19.15	19.86	13.95	15.12	18.60	18.90	12.82	12.88	16.09	15.87
Centurio Qwen	20.19	17.84	23.00	21.13	15.67	15.44	18.43	17.74	6.00	6.33	7.33	7.33	9.51	10.82	10.16	10.49	14.89	15.96	22.34	20.92	11.63	11.92	15.70	14.53	12.38	12.55	15.70	15.15
InternVL2.5 8B	14.55	19.25	20.66	23.00	11.98	15.44	18.43	18.43	3.33	6.33	9.33	9.00	9.84	13.77	15.08	16.07	15.25	17.73	23.05	23.05	10.17	12.21	16.28	16.57	10.61	13.82	16.92	17.36
InternVL2.5 4B	14.55	15.96	18.78	18.31	12.67	14.29	17.74	16.59	5.67	6.33	9.00	9.00	8.52	9.84	13.11	12.46	11.35	15.25	19.50	18.09	11.34	14.24	15.70	15.12	10.45	12.38	15.70	14.70
Centurio Aya	11.74	12.21	15.49	12.21	9.68	9.91	12.21	11.06	4.67	4.67	6.67	5.33	6.89	7.54	7.54	7.54	9.93	9.57	12.77	10.64	7.56	9.01	10.17	9.59	8.46	8.96	10.95	9.62
InternVL2.5 1B	8.45	9.86	11.27	12.21	5.76	8.29	9.22	7.60	1.67	2.33	4.00	2.67	5.90	7.87	8.85	8.85	6.74	7.45	10.99	10.64	7.27	8.72	9.01	9.01	5.86	7.46	8.90	8.35
Average X-Large	24.65	28.17	32.63	32.86	18.78	21.66	25.92	26.61	8.17	11.17	17.50	18.33	15.25	20.33	24.26	23.77	22.52	25.18	31.21	30.67	15.99	21.08	23.98	25.44	16.94	20.59	25.12	25.68
Average Large	21.36	27.00	30.75	30.99	17.51	21.77	25.23	25.81	4.83	10.17	15.50	15.83	12.79	19.18	22.13	23.77	18.79	23.94	27.84	27.30	13.52	19.62	23.84	23.84	14.15	19.54	23.55	23.99
Average Medium	16.43	17.46	19.91	19.53	13.23	14.33	16.68	16.45	4.93	6.67	8.47	8.00	10.30	12.13	12.92	13.31	14.89	16.24	19.86	19.22	12.09	13.84	15.58	15.99	11.63	13.09	15.31	15.21
Average Small	14.79	14.79	17.02	17.61	11.35	12.85	15.26	14.29	5.67	5.42	7.67	7.17	9.34	9.75	11.64	11.15	12.59	13.65	16.76	16.49	12.06	13.30	14.90	14.83	10.75	11.50	13.99	13.47
Average Open	17.95	19.75	22.64	22.75	14.16	16.15	18.98	18.79	5.64	7.51	10.69	10.54	11.15	13.75	15.69	15.86	15.96	18.00	21.88	21.39	12.90	15.68	17.93	18.29	12.57	14.75	17.68	17.64
Average Proprietary	32.45	31.67	34.49	36.40	27.91	27.53	29.47	31.23	17.06	16.02	21.83	23.09	24.28	24.13	26.54	27.98	30.98	30.18	34.48	35.55	25.66	24.87	28.61	29.50	26.01	25.35	28.80	30.19
Average	21.98	23.07	25.93	26.54	17.98	19.31	21.90	22.24	8.81	9.88	13.79	14.03	14.80	16.63	18.70	19.23	20.13	21.39	25.38	25.32	16.45	18.23	20.90	21.40	16.30	17.69	20.77	21.13

Table 14: GIMMICK Video Visual Question Answering (VVQA) results. The reported score is relaxed accuracy. The columns N, R, C, and B stand for the hints “None”, “Region”, “Country”, and “Both”, respectively.

### G.3 COQA Details

#### G.3.1 Results

	WEST EU & NORTH AM.			EAST EU			ASIA & PACIFIC			LAT. AM. & CARIB.			ARAB			SUBS. AFRICA			AVERAGE			
	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	Avg.
GPT-4o	82.50	83.75	85.00	85.89	90.18	88.34	94.37	96.54	97.40	93.68	92.63	92.63	88.00	88.00	92.00	91.30	91.30	94.20	89.29	90.40	91.60	90.43
Claude 3.5 Sonnet	72.50	83.75	81.25	76.69	85.89	82.21	87.88	95.67	95.67	83.16	89.47	87.37	84.00	90.67	90.67	82.61	89.86	88.41	81.14	89.22	87.60	85.98
InternVL2.5 78B	77.50	80.00	86.88	83.44	82.21	88.96	94.37	94.81	96.97	88.42	92.63	92.63	88.00	90.67	92.00	92.75	91.30	92.75	87.41	88.60	91.70	89.24
Qwen2.5 72B	–	81.25	–	–	84.05	–	–	96.10	–	–	89.47	–	–	86.67	–	–	89.86	–	–	87.90	–	87.90
GPT-4o Mini	76.25	82.50	86.25	84.66	82.21	84.66	94.37	95.67	96.54	87.37	90.53	90.53	85.33	86.67	86.67	91.30	89.86	92.75	86.55	87.90	89.57	88.01
InternVL2.5 38B	81.25	81.25	84.38	85.89	84.66	85.28	90.04	95.24	92.64	86.32	86.32	92.63	89.33	90.67	92.00	89.86	86.96	91.30	87.11	87.51	89.70	88.11
Qwen2 VL 72B	79.38	80.62	81.25	88.34	84.66	88.34	90.48	94.81	96.97	86.32	88.42	92.63	86.67	88.00	89.33	91.30	85.51	91.30	87.08	87.00	89.97	88.02
Gemini Flash	82.50	78.75	78.13	85.28	80.37	84.66	87.01	91.34	94.81	85.11	87.37	90.53	89.19	86.67	90.67	89.86	91.30	91.30	86.49	85.97	88.35	86.94
Qwen2.5 32B	–	76.88	–	–	79.75	–	–	94.37	–	–	87.37	–	–	84.00	–	–	89.86	–	–	85.37	–	85.37
Qwen2 VL 7B	71.25	74.38	76.25	82.82	80.37	84.05	92.64	93.51	93.51	85.26	88.42	92.63	80.00	82.67	84.00	86.96	85.51	84.06	83.16	84.14	85.75	84.35
MiniCPM V 2.6	72.50	72.50	75.00	81.60	79.14	80.37	88.74	90.48	93.07	80.00	87.37	90.53	80.00	77.33	86.67	88.41	85.51	86.96	81.87	82.05	85.43	83.12
InternVL2.5 26B	77.50	74.38	80.62	87.12	75.46	87.12	91.77	91.77	96.54	88.42	84.21	93.68	84.00	85.33	88.00	91.30	79.71	86.96	86.69	81.81	88.82	85.77
Phi 3.5 Mini	–	74.38	–	–	72.39	–	–	88.31	–	–	83.16	–	–	81.33	–	–	86.96	–	–	81.09	–	81.09
InternLM2.5 7B	–	74.38	–	–	76.69	–	–	90.48	–	–	80.00	–	–	78.67	–	–	85.51	–	–	80.95	–	80.95
Centurio Qwen	75.63	74.38	80.00	79.75	76.69	82.82	86.58	92.64	92.21	83.16	86.32	89.47	78.67	77.33	88.00	86.96	76.81	89.86	81.79	80.69	87.06	83.18
InternLM2.5 20B	–	74.38	–	–	75.46	–	–	89.18	–	–	86.32	–	–	76.00	–	–	82.61	–	–	80.66	–	80.66
Qwen2.5 7B	–	71.88	–	–	72.39	–	–	93.51	–	–	85.26	–	–	77.33	–	–	81.16	–	–	80.26	–	80.26
Aya Expanse 8B	–	68.12	–	–	77.30	–	–	91.77	–	–	81.05	–	–	80.00	–	–	81.16	–	–	79.90	–	79.90
InternVL2.5 8B	68.12	72.50	75.63	83.44	76.07	83.44	87.88	89.61	94.37	84.21	83.16	92.63	84.00	73.33	89.33	88.41	81.16	92.75	82.68	79.31	88.03	83.34
Centurio Aya	80.62	68.12	78.75	82.21	75.46	80.37	90.91	85.71	92.21	84.21	82.11	85.26	81.33	82.67	85.33	85.51	81.16	91.30	84.13	79.21	85.54	82.96
Phi 3.5 Vision	65.62	72.50	75.63	69.94	70.55	76.69	89.18	91.34	95.24	80.00	81.05	86.32	72.00	80.00	86.67	85.51	79.71	88.41	77.04	79.19	84.82	80.35
InternVL2.5 4B	66.88	66.88	76.25	84.66	75.46	84.05	87.01	86.15	93.07	83.16	78.95	87.37	80.00	82.67	86.67	86.96	84.06	89.86	81.44	79.03	86.21	82.23
Qwen2 VL 2B	77.50	72.50	78.75	84.05	64.42	84.05	91.77	82.68	92.21	88.42	81.05	86.32	84.00	70.67	89.33	88.41	79.71	91.30	85.69	75.17	86.99	82.62
Qwen2.5 3B	–	68.75	–	–	73.01	–	–	83.12	–	–	73.68	–	–	74.67	–	–	75.36	–	–	74.76	–	74.76
Qwen2.5 1.5B	–	61.88	–	–	65.03	–	–	82.25	–	–	78.95	–	–	72.00	–	–	78.26	–	–	73.06	–	73.06
Qwen2.5 0.5B	–	68.12	–	–	72.39	–	–	67.53	–	–	65.26	–	–	70.67	–	–	55.07	–	–	66.51	–	66.51
InternLM2.5 1.8B	–	56.25	–	–	65.03	–	–	65.37	–	–	60.00	–	–	66.67	–	–	66.67	–	–	63.33	–	63.33
InternVL2.5 2B	70.62	51.88	72.50	76.69	58.28	71.78	77.92	72.29	82.68	83.16	62.11	83.16	73.33	66.67	82.67	84.06	60.87	89.86	77.63	62.02	80.44	73.36
InternVL2.5 1B	63.75	58.75	66.88	62.58	60.74	74.23	64.50	61.90	80.09	77.89	57.89	87.37	62.67	68.00	82.67	75.36	59.42	82.61	67.79	61.12	78.97	69.29
Gemini Pro	76.25	59.38	78.13	68.10	55.21	82.21	82.25	56.28	89.61	79.79	61.05	85.11	79.73	61.33	84.00	72.46	65.22	95.65	76.43	59.75	85.78	73.99
Average X-Large LVLs	78.44	80.31	84.06	85.89	83.44	88.65	92.42	94.81	96.97	87.37	90.53	92.63	87.33	89.33	90.67	92.03	88.41	92.03	87.24	87.80	90.84	88.63
Average Large LVLs	79.38	77.81	82.50	86.50	80.06	86.20	90.91	93.51	94.59	87.37	85.26	93.16	86.67	88.00	90.00	90.58	83.33	89.13	86.90	84.66	89.26	86.94
Average Medium LVLs	73.62	72.38	77.12	81.96	77.55	82.21	89.35	90.39	93.07	83.37	85.47	90.11	80.80	78.67	86.67	87.25	82.03	88.99	82.73	81.08	86.36	83.39
Average Small LVLs	68.88	64.50	74.00	75.58	65.89	78.16	82.08	78.87	88.66	82.53	72.21	86.11	74.40	73.60	85.60	84.06	72.75	88.41	77.92	71.31	83.49	77.57
Average LVLs	73.44	71.47	77.77	80.89	74.58	82.25	87.41	87.35	92.27	84.21	81.43	89.47	80.29	79.71	87.33	87.27	79.81	89.23	82.25	79.06	86.39	82.57
Average X-Large LLMs	–	81.25	–	–	84.05	–	–	96.10	–	–	89.47	–	–	86.67	–	–	89.86	–	–	87.90	–	87.90
Average Large LLMs	–	75.62	–	–	77.61	–	–	91.77	–	–	86.84	–	–	80.00	–	–	86.23	–	–	83.02	–	83.02
Average Medium LLMs	–	71.46	–	–	75.46	–	–	91.92	–	–	82.11	–	–	78.67	–	–	82.61	–	–	80.37	–	80.37
Average Small LLMs	–	65.88	–	–	69.57	–	–	77.32	–	–	72.21	–	–	73.07	–	–	72.46	–	–	71.75	–	71.75
Average LLMs	–	70.57	–	–	73.95	–	–	85.64	–	–	79.14	–	–	77.09	–	–	79.31	–	–	77.62	–	77.62
Average X-Large	78.44	80.62	84.06	85.89	83.64	88.65	92.42	95.24	96.97	87.37	90.18	92.63	87.33	88.44	90.67	92.03	88.89	92.03	87.24	87.83	90.84	88.39
Average Large	79.38	76.72	82.50	86.50	78.83	86.20	90.91	92.64	94.59	87.37	86.05	93.16	86.67	84.00	90.00	90.58	84.78	89.13	86.90	83.84	89.26	84.98
Average Medium	73.62	72.03	77.12	81.96	76.76	82.21	89.35	90.96	93.07	83.37	84.21	90.11	80.80	78.67	86.67	87.25	82.25	88.99	82.73	80.81	86.36	82.26
Average Small	68.88	65.19	74.00	75.58	67.73	78.16	82.08	78.10	88.66	82.53	72.21	86.11	74.40	73.33	85.60	84.06	72.61	88.41	77.92	71.53	83.49	74.66
Average Open	73.44	71.08	77.77	80.89	74.31	82.25	87.41	86.60	92.27	84.21	80.42	89.47	80.29	78.56	87.33	87.27	79.59	89.23	82.25	78.43	86.39	80.39
Average Proprietary	78.00	77.62	81.75	80.12	78.77	84.42	89.18	87.10	94.81	85.82	84.21	89.23	85.25	82.67	88.80	85.51	85.51	92.46	83.98	82.65	88.58	85.07
Average	74.64	72.17	78.82	80.69	75.05	82.82	87.88	86.68	92.94	84.63	81.05	89.41	81.59	79.24	87.72	86.80	80.58	90.08	82.71	79.13	86.96	81.17

Table 15: GIMMICK Cultural Origin Question Answering – Regions (COQA<sub>R</sub>) results. The reported score is relaxed accuracy. The columns **I** and **T** stand for **image-only** and **text-only** inputs to the model.

	West EU & North Am.			East EU			Asia & Pacific			Lat. Am. & Carib.			Arab			Subs. Africa			Average			
	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	I	T	I+T	Avg.
Claude 3.5 Sonnet	79.23	96.72	95.63	82.35	97.65	96.47	76.62	97.84	95.67	70.21	98.94	100.00	76.47	97.65	96.47	83.82	97.06	91.18	78.12	97.64	95.90	90.55
GPT-4o	93.44	95.08	96.17	94.71	98.24	98.24	93.51	97.40	98.27	97.87	98.94	98.94	95.29	95.29	98.82	95.59	97.06	100.00	95.07	97.00	98.41	96.83
InternVL2.5 78B	83.06	94.54	97.81	80.59	95.88	97.65	83.12	93.51	96.54	81.91	98.94	98.94	90.59	97.65	97.65	83.82	97.06	98.53	83.85	96.26	97.85	92.65
Qwen2.5 72B	–	93.44	–	–	96.47	–	–	94.81	–	–	98.94	–	–	97.65	–	–	94.12	–	–	95.90	–	95.90
GPT-4o Mini	89.07	93.99	95.63	90.00	95.29	97.65	90.48	93.51	96.97	90.43	95.74	100.00	94.12	88.24	97.65	91.18	95.59	98.53	90.88	93.73	97.74	94.11
Qwen2.5 32B	–	91.26	–	–	93.53	–	–	91.77	–	–	94.68	–	–	95.29	–	–	92.65	–	–	93.20	–	93.20
InternVL2.5 38B	78.69	91.80	92.35	77.06	91.18	92.94	77.49	93.07	93.94	79.79	95.74	96.81	84.71	94.12	95.29	88.24	92.65	98.53	80.99	93.09	94.98	89.69
Qwen2 VL 72B	87.98	87.43	95.08	94.12	90.59	96.47	90.04	90.04	97.84	91.49	97.87	98.94	92.94	89.41	98.82	91.18	97.06	98.53	91.29	92.07	97.61	93.66
Gemini Flash	90.56	89.01	97.27	90.59	88.82	97.06	91.77	90.48	98.70	90.43	93.62	97.87	90.59	88.24	97.65	88.24	95.59	97.06	90.36	90.96	97.60	92.97
InternVL2.5 26B	78.14	87.98	92.90	78.24	88.24	94.71	76.19	90.48	93.94	80.85	94.68	94.68	81.18	91.76	91.76	80.88	91.18	95.59	79.25	90.72	93.93	87.97
Qwen2.5 7B	–	86.34	–	–	88.24	–	–	85.28	–	–	95.74	–	–	90.59	–	–	94.12	–	–	90.05	–	90.05
Aya Expanse 8B	–	87.43	–	–	88.24	–	–	90.04	–	–	93.62	–	–	88.24	–	–	89.71	–	–	89.54	–	89.54
InternLM2.5 20B	–	86.89	–	–	87.06	–	–	90.91	–	–	90.43	–	–	85.88	–	–	89.71	–	–	88.48	–	88.48
MiniCPM V 2.6	81.97	84.70	90.16	81.18	86.47	92.94	78.79	87.45	92.21	86.17	87.23	92.55	82.35	89.41	96.47	88.24	92.65	94.12	83.12	87.98	93.08	88.06
Qwen2 VL 7B	87.43	83.61	90.71	82.35	85.29	92.94	87.01	84.85	94.37	91.49	88.30	94.68	84.71	88.24	96.47	92.65	94.12	97.06	87.61	87.40	94.37	89.79
Qwen2.5 3B	–	81.42	–	–	85.88	–	–	84.85	–	–	92.55	–	–	88.24	–	–	86.76	–	–	86.62	–	86.62
InternLM2.5 7B	–	83.61	–	–	85.88	–	–	85.71	–	–	90.43	–	–	77.65	–	–	88.24	–	–	85.25	–	85.25
Centurio Qwen	78.69	82.51	89.07	78.82	82.94	89.41	78.79	84.42	92.64	76.60	85.11	91.49	80.00	83.53	88.24	79.41	91.18	92.65	78.72	84.95	90.58	84.75
Centurio Aya	65.57	83.61	85.79	72.35	81.76	85.88	75.76	85.71	88.31	74.47	87.23	82.98	70.59	80.00	89.41	66.18	89.71	89.71	70.82	84.67	87.01	80.83
InternVL2.5 8B	68.31	82.51	88.52	70.59	84.71	90.00	75.32	86.58	91.34	75.53	87.23	94.68	76.47	83.53	90.59	82.35	82.35	89.71	74.76	84.49	90.81	83.35
Phi 3.5 Mini	–	80.87	–	–	82.94	–	–	83.98	–	–	84.04	–	–	82.35	–	–	88.24	–	–	83.74	–	83.74
InternVL2.5 4B	68.85	77.05	89.62	72.35	82.94	89.41	71.43	86.15	90.48	76.60	87.23	89.36	72.94	81.18	84.71	76.47	82.35	97.06	73.11	82.82	90.11	82.01
Phi 3.5 Vision	72.13	79.78	86.89	68.82	82.94	92.35	69.70	81.82	89.61	74.47	91.49	91.49	81.18	77.65	90.59	76.47	82.35	95.59	73.79	82.67	91.09	82.52
Qwen2.5 1.5B	–	78.69	–	–	81.18	–	–	82.68	–	–	82.98	–	–	75.29	–	–	80.88	–	–	80.28	–	80.28
Qwen2 VL 2B	83.06	74.32	87.43	84.71	77.06	87.65	83.55	80.95	90.48	92.55	81.91	94.68	83.53	76.47	91.76	89.71	80.88	94.12	86.18	78.60	91.02	85.27
Qwen2.5 0.5B	–	65.03	–	–	68.82	–	–	72.29	–	–	75.53	–	–	69.41	–	–	77.94	–	–	71.51	–	71.51
InternVL2.5 1B	61.20	66.12	73.77	59.41	65.88	73.53	62.34	75.76	77.06	67.02	74.47	76.60	56.47	63.53	75.29	55.88	72.06	70.59	60.39	69.64	74.47	68.17
InternLM2.5 1.8B	–	63.39	–	–	66.47	–	–	71.00	–	–	67.02	–	–	58.82	–	–	64.71	–	–	65.23	–	65.23
InternVL2.5 2B	62.84	65.57	74.32	61.76	64.71	72.35	61.04	68.40	80.95	67.02	68.09	80.85	67.06	55.29	74.12	73.53	63.24	77.94	65.54	64.22	76.76	68.84
Gemini Pro	76.67	43.17	92.70	75.88	39.41	92.94	78.79	34.20	92.64	78.72	39.36	93.62	78.82	35.29	91.76	82.35	19.12	94.12	78.54	35.09	92.96	68.86
Average X-Large LVLs	85.52	90.98	96.45	87.35	93.24	97.06	86.58	91.77	97.19	86.70	98.40	98.94	91.76	93.53	98.24	87.50	97.06	98.53	87.57	94.16	97.73	93.16
Average Large LVLs	78.42	89.89	92.62	77.65	89.71	93.82	76.84	91.77	93.94	80.32	95.21	95.74	82.94	92.94	93.53	84.56	91.91	97.06	80.12	91.90	94.46	88.82
Average Medium LVLs	76.39	83.39	88.85	77.06	84.24	90.24	79.13	85.80	91.77	80.85	87.02	91.28	78.82	84.94	92.24	81.76	90.00	92.65	79.01	85.90	91.17	85.36
Average Small LVLs	69.62	72.57	82.40	69.41	74.71	83.06	69.61	78.61	85.71	75.53	80.64	86.60	72.24	70.82	83.29	74.41	76.18	87.06	71.80	75.59	84.69	77.36
Average LVLs	75.57	81.54	88.17	75.88	82.90	89.16	76.47	84.94	90.69	79.71	87.54	91.34	78.91	82.27	90.08	80.36	86.34	92.12	77.82	84.26	90.26	84.11
Average X-Large LLMs	–	93.44	–	–	96.47	–	–	94.81	–	–	98.94	–	–	97.65	–	–	94.12	–	–	95.90	–	95.90
Average Large LLMs	–	89.07	–	–	90.29	–	–	91.34	–	–	92.55	–	–	90.59	–	–	91.18	–	–	90.84	–	90.84
Average Medium LLMs	–	85.79	–	–	87.45	–	–	87.01	–	–	93.26	–	–	85.49	–	–	90.69	–	–	88.28	–	88.28
Average Small LLMs	–	73.88	–	–	77.06	–	–	78.96	–	–	80.43	–	–	74.82	–	–	79.71	–	–	77.48	–	77.48
Average LLMs	–	81.67	–	–	84.06	–	–	84.85	–	–	87.81	–	–	82.67	–	–	86.10	–	–	84.53	–	84.53
Average X-Large	85.52	91.80	96.45	87.35	94.31	97.06	86.58	92.78	97.19	86.70	98.58	98.94	91.76	94.90	98.24	87.50	96.08	98.53	87.57	94.74	97.73	94.07
Average Large	78.42	89.48	92.62	77.65	90.00	93.82	76.84	91.56	93.94	80.32	93.88	95.74	82.94	91.76	93.53	84.56	91.54	97.06	80.12	91.37	94.46	89.83
Average Medium	76.39	84.29	88.85	77.06	85.44	90.24	79.13	86.26	91.77	80.85	89.36	91.28	78.82	85.15	92.24	81.76	90.26	92.65	79.01	86.79	91.17	86.45
Average Small	69.62	73.22	82.40	69.41	75.88	83.06	69.61	78.79	85.71	75.53	80.53	86.60	72.24	72.82	83.29	74.41	77.94	87.06	71.80	76.53	84.69	77.42
Average Open	75.57	81.60	88.17	75.88	83.41	89.16	76.47	84.90	90.69	79.71	87.66	91.34	78.91	82.45	90.08	80.36	86.24	92.12	77.82	84.38	90.26	84.29
Average Proprietary	85.79	83.59	95.48	86.71	83.88	96.47	86.23	82.68	96.45	85.53	85.32	98.09	87.06	80.94	96.47	88.24	80.88	96.18	86.59	82.88	96.52	88.66
Average	78.26	81.93	90.10	78.73	83.49	91.08	79.04	84.53	92.21	81.24	87.27	93.11	81.05	82.20	91.76	82.43	85.34	93.19	80.13	84.13	91.91	85.02

Table 16: GIMMICK Cultural Origin Question Answering – Country (COQA<sub>C</sub>) results. The reported score is relaxed accuracy. The columns **I** and **T** stand for **image-only** and **text-only** inputs to the model.

## G.4 CKQA

### G.4.1 LLM-as-a-Judge Evaluation

To evaluate the CKQA<sub>D</sub> and CKQA<sub>N</sub> tasks, we used GPT-4o (gpt-4o-2024-11-20) as a judge using the prompts shown in the next section. For each sample, we used the same system prompt and generated user prompts per sample individually.

## System Prompt

### # Your Role

You are an impartial judge who excels at critical and analytical thinking.

### # Your Task

Your task is it to thoroughly analyze and evaluate the correctness of a generated answer to a  
↪ Cultural Knowledge Test.

1. Carefully analyze the ground truth and the generated answer.
2. Provide a brief summary (1 - 3 sentences) of your analysis, covering the accuracy, relevance, and completeness of the generated answer.  
↪
3. Provide a one or two-sentence explanation justifying your final score. Ensure that your  
↪ explanation and score are consistent with each other and accurately reflect the quality of  
↪ the generated answer in relation to the ground truth.
4. Provide a single number from 0 to 100 representing the correctness of the generated answer,  
↪ where:  
0 = Completely incorrect or irrelevant.  
25 = Mostly incorrect or irrelevant.  
50 = Partially correct or relevant.  
75 = Mostly correct and relevant.  
100 = Perfectly correct and complete.

You may use any whole number within this range to reflect nuanced judgments.

### # Output Format

Provide your evaluation in the following format:

```
```xml
<evaluation>
<analysis>
<!-- Put your analysis summary here -->
</analysis>
<explanation>
<!-- Put your explanation here -->
</explanation>
<score>
<!-- Put your score here -->
</score>
</evaluation>
```
```

## User Prompt Template

Evaluate the correctness of the generated answer with respect to Ground Truth.

### # Ground Truth

```
```
{GROUND_TRUTH}
```
```

### # Generated Answer

```
```
{GENERATED_ANSWER}
```
```

### # Evaluation

## G.4.2 Results

|                      | WEST EU & NORTH AM. |       |       | EAST EU |       |       | ASIA & PACIFIC |       |       | LAT. AM. & CARIB. |       |       | ARAB  |       |       | SUBS. AFRICA |       |       | AVERAGE |       |       |       |
|----------------------|---------------------|-------|-------|---------|-------|-------|----------------|-------|-------|-------------------|-------|-------|-------|-------|-------|--------------|-------|-------|---------|-------|-------|-------|
|                      | I                   | T     | I+T   | I       | T     | I+T   | I              | T     | I+T   | I                 | T     | I+T   | I     | T     | I+T   | I            | T     | I+T   | I       | T     | I+T   | Avg.  |
| GPT-4o               | 46.98               | 56.78 | 57.21 | 38.20   | 54.30 | 54.87 | 44.71          | 59.20 | 58.56 | 34.08             | 51.53 | 52.45 | 44.41 | 57.76 | 56.91 | 29.04        | 50.68 | 51.37 | 39.57   | 55.04 | 55.23 | 49.95 |
| Claude 3.5 Sonnet    | 43.05               | 56.64 | 55.60 | 35.20   | 55.97 | 50.07 | 39.54          | 59.67 | 54.05 | 26.84             | 53.32 | 49.23 | 41.45 | 56.78 | 53.68 | 24.73        | 50.34 | 44.79 | 35.14   | 55.45 | 51.24 | 47.28 |
| Gemini Pro           | 42.28               | 53.29 | 57.21 | 36.80   | 50.07 | 53.67 | 37.47          | 52.94 | 55.18 | 29.18             | 49.44 | 50.00 | 38.68 | 48.68 | 54.08 | 22.05        | 41.23 | 46.23 | 34.41   | 49.28 | 52.73 | 45.47 |
| Qwen2.5 72B          | –                   | 47.55 | –     | –       | 45.17 | –     | –              | 50.62 | –     | –                 | 42.70 | –     | –     | 44.47 | –     | –            | 37.95 | –     | –       | 44.74 | –     | 44.74 |
| Qwen2.5 32B          | –                   | 47.89 | –     | –       | 43.73 | –     | –              | 48.45 | –     | –                 | 40.71 | –     | –     | 42.17 | –     | –            | 39.25 | –     | –       | 43.70 | –     | 43.70 |
| GPT-4o Mini          | 34.36               | 48.89 | 55.70 | 27.70   | 46.63 | 54.00 | 30.73          | 49.05 | 53.61 | 24.95             | 43.72 | 49.44 | 36.84 | 47.50 | 54.21 | 21.03        | 39.25 | 46.78 | 29.27   | 45.84 | 52.29 | 42.47 |
| Gemini Flash         | 36.54               | 52.75 | 54.70 | 29.67   | 46.87 | 51.30 | 31.78          | 50.40 | 51.62 | 23.20             | 46.07 | 49.07 | 32.43 | 46.64 | 51.45 | 16.44        | 36.37 | 42.12 | 28.34   | 46.52 | 50.04 | 41.63 |
| Phi 3.5 Mini         | –                   | 40.40 | –     | –       | 35.23 | –     | –              | 38.27 | –     | –                 | 34.80 | –     | –     | 34.87 | –     | –            | 30.00 | –     | –       | 35.60 | –     | 35.60 |
| Aya Expanse 8B       | –                   | 40.17 | –     | –       | 36.13 | –     | –              | 39.42 | –     | –                 | 34.18 | –     | –     | 36.32 | –     | –            | 26.71 | –     | –       | 35.49 | –     | 35.49 |
| Qwen2.5 7B           | –                   | 38.39 | –     | –       | 36.50 | –     | –              | 38.78 | –     | –                 | 34.23 | –     | –     | 34.01 | –     | –            | 29.04 | –     | –       | 35.16 | –     | 35.16 |
| InternLM2.5 20B      | –                   | 37.01 | –     | –       | 34.13 | –     | –              | 36.59 | –     | –                 | 31.17 | –     | –     | 32.83 | –     | –            | 27.53 | –     | –       | 33.21 | –     | 33.21 |
| Llama 3.2 11B Vision | –                   | 36.44 | –     | –       | 32.77 | –     | –              | 35.75 | –     | –                 | 30.00 | –     | –     | 33.68 | –     | –            | 27.40 | –     | –       | 32.67 | –     | 32.67 |
| InternVL2.5 38B      | 23.72               | 41.21 | 37.62 | 18.63   | 38.80 | 37.03 | 20.51          | 41.55 | 39.96 | 23.72             | 33.32 | 38.72 | 24.08 | 35.46 | 39.67 | 15.96        | 32.47 | 33.49 | 21.10   | 37.14 | 37.75 | 32.00 |
| InternVL2.5 78B      | 19.33               | 40.84 | 36.28 | 17.63   | 37.73 | 37.10 | 19.16          | 41.57 | 38.23 | 19.64             | 37.50 | 36.58 | 22.89 | 35.66 | 42.57 | 14.86        | 33.22 | 35.34 | 18.92   | 37.75 | 37.68 | 31.45 |
| Qwen2 VL 72B         | 20.67               | 40.81 | 41.01 | 17.37   | 37.03 | 42.13 | 18.23          | 40.02 | 42.10 | 14.08             | 36.02 | 37.60 | 23.42 | 36.12 | 41.91 | 10.62        | 29.38 | 35.14 | 17.40   | 36.56 | 39.98 | 31.31 |
| InternLM2.5 7B       | –                   | 34.33 | –     | –       | 32.30 | –     | –              | 34.62 | –     | –                 | 31.17 | –     | –     | 29.93 | –     | –            | 23.49 | –     | –       | 30.97 | –     | 30.97 |
| Qwen2.5 3B           | –                   | 32.75 | –     | –       | 28.90 | –     | –              | 33.05 | –     | –                 | 28.47 | –     | –     | 26.58 | –     | –            | 22.81 | –     | –       | 28.76 | –     | 28.76 |
| InternVL2.5 26B      | 11.91               | 39.97 | 34.43 | 12.63   | 36.10 | 34.07 | 13.76          | 38.92 | 34.00 | 13.11             | 34.18 | 28.98 | 15.33 | 34.01 | 36.38 | 9.38         | 29.59 | 27.95 | 12.69   | 35.46 | 32.64 | 26.93 |
| Qwen2 VL 7B          | 14.09               | 32.82 | 38.09 | 14.27   | 29.53 | 37.17 | 12.72          | 33.12 | 37.43 | 17.09             | 28.32 | 35.97 | 17.17 | 29.28 | 35.46 | 9.38         | 20.55 | 30.96 | 14.12   | 28.94 | 35.85 | 26.30 |
| MiniCPM V 2.6        | 18.49               | 34.70 | 36.28 | 13.60   | 30.47 | 34.60 | 15.88          | 33.76 | 34.96 | 18.67             | 29.03 | 34.34 | 17.50 | 30.20 | 36.84 | 9.93         | 18.42 | 24.52 | 15.68   | 29.43 | 33.59 | 26.23 |
| Centurio Qwen        | 14.87               | 31.38 | 32.05 | 14.47   | 29.23 | 29.97 | 15.07          | 31.57 | 34.76 | 16.38             | 27.40 | 35.77 | 18.62 | 27.30 | 36.32 | 13.01        | 20.41 | 32.60 | 15.40   | 27.88 | 33.58 | 25.62 |
| Phi 3.5 Vision       | 12.08               | 36.64 | 35.03 | 12.43   | 32.10 | 35.23 | 10.09          | 33.36 | 32.30 | 13.78             | 29.74 | 29.80 | 16.97 | 31.97 | 33.42 | 11.99        | 25.75 | 26.78 | 12.89   | 31.59 | 32.09 | 25.53 |
| InternVL2.5 8B       | 6.81                | 36.28 | 29.97 | 6.80    | 31.33 | 29.13 | 9.07           | 33.72 | 30.49 | 6.58              | 29.74 | 30.36 | 12.11 | 30.53 | 30.26 | 2.53         | 22.67 | 20.96 | 7.32    | 30.71 | 28.53 | 22.19 |
| InternVL2.5 4B       | 5.44                | 35.81 | 27.89 | 5.40    | 33.07 | 27.80 | 5.97           | 35.71 | 28.08 | 7.14              | 34.29 | 27.24 | 9.01  | 28.62 | 29.54 | 4.79         | 25.68 | 23.63 | 6.29    | 32.20 | 27.36 | 21.95 |
| Qwen2.5 1.5B         | –                   | 24.03 | –     | –       | 20.77 | –     | –              | 26.66 | –     | –                 | 20.87 | –     | –     | 21.45 | –     | –            | 16.23 | –     | –       | 21.67 | –     | 21.67 |
| InternLM2.5 1.8B     | –                   | 23.56 | –     | –       | 22.30 | –     | –              | 22.94 | –     | –                 | 18.52 | –     | –     | 21.78 | –     | –            | 14.66 | –     | –       | 20.63 | –     | 20.63 |
| Qwen2 VL 2B          | 11.41               | 23.29 | 31.95 | 11.57   | 20.03 | 28.90 | 12.48          | 20.97 | 29.00 | 14.18             | 20.51 | 28.16 | 16.32 | 18.29 | 29.47 | 11.92        | 13.63 | 25.96 | 12.98   | 19.45 | 28.91 | 20.45 |
| InternVL2.5 1B       | 9.87                | 24.09 | 16.51 | 8.50    | 20.43 | 16.83 | 9.78           | 21.28 | 18.50 | 11.22             | 20.41 | 16.07 | 12.83 | 16.51 | 18.75 | 9.93         | 14.93 | 17.60 | 10.35   | 19.61 | 17.38 | 15.78 |
| InternVL2.5 2B       | 5.30                | 23.26 | 18.72 | 4.80    | 19.50 | 21.90 | 4.47           | 22.26 | 20.58 | 7.30              | 21.48 | 19.95 | 9.34  | 21.38 | 20.72 | 5.68         | 15.96 | 18.77 | 6.15    | 20.64 | 20.11 | 15.63 |
| Centurio Aya         | 4.80                | 29.33 | 5.94  | 5.30    | 25.50 | 7.53  | 2.94           | 28.85 | 5.02  | 7.50              | 24.23 | 8.47  | 4.28  | 24.21 | 4.21  | 4.45         | 19.38 | 5.89  | 4.88    | 25.25 | 6.18  | 12.10 |
| Qwen2.5 0.5B         | –                   | 13.96 | –     | –       | 11.77 | –     | –              | 14.40 | –     | –                 | 11.43 | –     | –     | 8.29  | –     | –            | 8.70  | –     | –       | 11.42 | –     | 11.42 |
| Average X-Large LVLs | 20.00               | 40.83 | 38.64 | 17.50   | 37.38 | 39.62 | 18.70          | 40.80 | 40.16 | 16.86             | 36.76 | 37.09 | 23.16 | 35.89 | 42.24 | 12.74        | 31.30 | 35.24 | 18.16   | 37.16 | 38.83 | 31.38 |
| Average Large LVLs   | 17.81               | 40.59 | 36.02 | 15.63   | 37.45 | 35.55 | 17.14          | 40.24 | 36.98 | 18.42             | 33.75 | 33.85 | 19.70 | 34.74 | 38.03 | 12.67        | 31.03 | 30.72 | 16.90   | 36.30 | 35.20 | 29.46 |
| Average Medium LVLs  | 11.81               | 33.49 | 28.47 | 10.89   | 29.80 | 27.68 | 11.14          | 32.79 | 28.53 | 13.24             | 28.12 | 28.98 | 13.94 | 29.20 | 28.62 | 7.86         | 21.47 | 22.99 | 11.48   | 29.15 | 27.55 | 24.18 |
| Average Small LVLs   | 8.82                | 28.62 | 26.02 | 8.54    | 25.03 | 26.13 | 8.56           | 26.72 | 25.69 | 10.72             | 25.29 | 24.24 | 12.89 | 23.35 | 26.38 | 8.86         | 19.19 | 22.55 | 9.73    | 24.70 | 25.17 | 19.87 |
| Average LVLs         | 12.77               | 33.79 | 30.13 | 11.67   | 30.24 | 29.96 | 12.15          | 32.83 | 30.39 | 13.60             | 29.08 | 29.14 | 15.70 | 28.88 | 31.11 | 9.60         | 23.30 | 25.68 | 12.58   | 29.69 | 29.40 | 24.41 |
| Average X-Large LLMs | –                   | 47.55 | –     | –       | 45.17 | –     | –              | 50.62 | –     | –                 | 42.70 | –     | –     | 44.47 | –     | –            | 37.95 | –     | –       | 44.74 | –     | 44.74 |
| Average Large LLMs   | –                   | 42.45 | –     | –       | 38.93 | –     | –              | 42.52 | –     | –                 | 35.94 | –     | –     | 37.50 | –     | –            | 33.39 | –     | –       | 38.46 | –     | 38.46 |
| Average Medium LLMs  | –                   | 37.63 | –     | –       | 34.98 | –     | –              | 37.61 | –     | –                 | 33.19 | –     | –     | 33.42 | –     | –            | 26.41 | –     | –       | 33.87 | –     | 33.87 |
| Average Small LLMs   | –                   | 26.94 | –     | –       | 23.79 | –     | –              | 27.06 | –     | –                 | 22.82 | –     | –     | 22.59 | –     | –            | 18.48 | –     | –       | 23.62 | –     | 23.62 |
| Average LLMs         | –                   | 34.55 | –     | –       | 31.54 | –     | –              | 34.89 | –     | –                 | 29.84 | –     | –     | 30.25 | –     | –            | 25.12 | –     | –       | 31.03 | –     | 31.03 |
| Average X-Large      | 20.00               | 43.07 | 38.64 | 17.50   | 39.98 | 39.62 | 18.70          | 44.07 | 40.16 | 16.86             | 38.74 | 37.09 | 23.16 | 38.75 | 42.24 | 12.74        | 33.52 | 35.24 | 18.16   | 39.68 | 38.83 | 35.83 |
| Average Large        | 17.81               | 41.52 | 36.02 | 15.63   | 38.19 | 35.55 | 17.14          | 41.38 | 36.98 | 18.42             | 34.84 | 33.85 | 19.70 | 36.12 | 38.03 | 12.67        | 32.21 | 30.72 | 16.90   | 37.38 | 35.20 | 33.96 |
| Average Medium       | 11.81               | 34.87 | 28.47 | 10.89   | 31.53 | 27.68 | 11.14          | 34.40 | 28.53 | 13.24             | 29.81 | 28.98 | 13.94 | 30.61 | 28.62 | 7.86         | 23.12 | 22.99 | 11.48   | 30.72 | 27.55 | 27.41 |
| Average Small        | 8.82                | 27.78 | 26.02 | 8.54    | 24.41 | 26.13 | 8.56           | 26.89 | 25.69 | 10.72             | 24.05 | 24.24 | 12.89 | 22.97 | 26.38 | 8.86         | 18.84 | 22.55 | 9.73    | 24.16 | 25.17 | 21.74 |
| Average Open         | 12.77               | 34.11 | 30.13 | 11.67   | 30.79 | 29.96 | 12.15          | 33.70 | 30.39 | 13.60             | 29.40 | 29.14 | 15.70 | 29.46 | 31.11 | 9.60         | 24.07 | 25.68 | 12.58   | 30.26 | 29.40 | 27.21 |
| Average Proprietary  | 40.64               | 53.67 | 56.08 | 33.51   | 50.77 | 52.78 | 36.85          | 54.25 | 54.60 | 27.65             | 48.82 | 50.04 | 38.76 | 51.47 | 54.07 | 22.66        | 43.57 | 46.26 | 33.35   | 50.43 | 52.31 | 45.36 |
| Average              | 20.11               | 37.27 | 36.96 | 17.42   | 34.01 | 35.96 | 18.65          | 37.02 | 36.76 | 17.30             | 32.53 | 34.64 | 21.77 | 33.01 | 37.15 | 13.04        | 27.22 | 31.10 | 18.05   | 33.51 | 35.43 | 30.14 |

Table 17: Average Judge Score for the GIMMICK Cultural Knowledge Question Answering (CKQA) – Describing. The columns **I**, **T**, and **I+T** stand for **image-only**, **text-only**, and **image+text** input to the model.

|                       | WEST EU & NORTH AM. | EAST EU | ASIAN & PACIFIC | LATIN-AMERICA & CARIBBEAN | ARAB  | SUBSAHARIAN AFRICA | AVERAGE |
|-----------------------|---------------------|---------|-----------------|---------------------------|-------|--------------------|---------|
| GPT-4o                | 37.79               | 32.57   | 37.68           | 30.15                     | 38.03 | 28.42              | 34.11   |
| Claude 3.5 Sonnet     | 40.27               | 33.63   | 39.29           | 25.71                     | 38.16 | 24.25              | 33.55   |
| GPT-4o Mini           | 34.46               | 28.73   | 33.08           | 23.67                     | 34.87 | 25.89              | 30.12   |
| Centurio Qwen         | 18.69               | 19.10   | 21.97           | 18.67                     | 25.46 | 15.96              | 19.98   |
| Gemini Pro            | 16.91               | 15.60   | 16.71           | 11.13                     | 17.30 | 10.55              | 14.70   |
| Gemini Flash          | 15.77               | 16.27   | 14.87           | 11.60                     | 14.61 | 11.30              | 14.07   |
| InternVL2.5 38B       | 14.06               | 12.60   | 16.24           | 10.71                     | 21.12 | 8.36               | 13.85   |
| Phi 3.5 Vision        | 15.17               | 13.67   | 13.54           | 12.45                     | 14.28 | 10.75              | 13.31   |
| InternVL2.5 78B       | 12.08               | 14.73   | 14.89           | 7.35                      | 15.72 | 7.53               | 12.05   |
| InternVL2.5 26B       | 11.51               | 10.50   | 13.16           | 7.65                      | 14.34 | 7.74               | 10.82   |
| InternVL2.5 1B        | 10.20               | 9.43    | 10.42           | 10.71                     | 14.80 | 8.22               | 10.63   |
| Qwen2 VL 72B          | 11.04               | 10.07   | 9.96            | 7.40                      | 11.45 | 8.56               | 9.75    |
| MiniCPM V 2.6         | 8.89                | 8.60    | 11.42           | 4.74                      | 10.99 | 9.79               | 9.07    |
| Centurio Aya          | 6.95                | 6.57    | 6.06            | 8.78                      | 5.20  | 7.40               | 6.83    |
| InternVL2.5 2B        | 6.31                | 6.80    | 6.17            | 7.14                      | 8.49  | 3.08               | 6.33    |
| InternVL2.5 4B        | 6.28                | 5.47    | 5.07            | 6.02                      | 9.28  | 5.00               | 6.19    |
| InternVL2.5 8B        | 6.51                | 5.30    | 4.54            | 6.48                      | 9.28  | 3.77               | 5.98    |
| Qwen2 VL 2B           | 5.40                | 4.27    | 7.35            | 3.62                      | 5.53  | 3.63               | 4.97    |
| Qwen2 VL 7B           | 5.27                | 5.63    | 4.78            | 4.03                      | 6.32  | 3.70               | 4.96    |
| Average X-Large LVLMS | 11.56               | 12.40   | 12.42           | 7.38                      | 13.58 | 8.04               | 10.90   |
| Average Large LVLMS   | 12.78               | 11.55   | 14.70           | 9.18                      | 17.73 | 8.05               | 12.34   |
| Average Medium LVLMS  | 9.26                | 9.04    | 9.75            | 8.54                      | 11.45 | 8.12               | 9.36    |
| Average Small LVLMS   | 8.67                | 7.93    | 8.51            | 7.99                      | 10.48 | 6.14               | 8.29    |
| Average LVLMS         | 9.88                | 9.48    | 10.40           | 8.27                      | 12.30 | 7.39               | 9.62    |
| Average X-Large       | 11.56               | 12.40   | 12.42           | 7.38                      | 13.58 | 8.04               | 10.90   |
| Average Large         | 12.78               | 11.55   | 14.70           | 9.18                      | 17.73 | 8.05               | 12.34   |
| Average Medium        | 9.26                | 9.04    | 9.75            | 8.54                      | 11.45 | 8.12               | 9.36    |
| Average Small         | 8.67                | 7.93    | 8.51            | 7.99                      | 10.48 | 6.14               | 8.29    |
| Average Open          | 9.88                | 9.48    | 10.40           | 8.27                      | 12.30 | 7.39               | 9.62    |
| Average Proprietary   | 29.04               | 25.36   | 28.33           | 20.45                     | 28.59 | 20.08              | 25.31   |
| Average               | 14.92               | 13.66   | 15.12           | 11.47                     | 16.59 | 10.73              | 13.75   |