Pearl: A Multimodal Culturally-Aware Arabic Instruction Dataset

Fakhraddin Alwajih[1], Samar M. Magdy[1], Abdellah El Mekki[1], Omer Nacar[2,3], Youssef Nafea[4], Safaa Taher Abdelfadil[4], Abdulfattah Mohammed Yahya[5,6], Hamzah Luqman[7], Nada Almarwani[8], Samah Aloufi[8], Baraah Qawasmeh[9], Houdaifa Atou[10], Serry Sibaee[2], Hamzah A. Alsayadi[11], Walid Al-Dhabyani[12,5], Maged S. Al-shaibani[7], Aya El Aatar [13], Nour Qandos[14], Rahaf Alhamouri[15], Samar Ahmad[16], Mohammed Anwar Al-Ghrawi[17], Aminetou Yacoub[18], Ruwa AbuHweidi[19], Vatimetou Mohamed Lemin[18], Reem Abdel-Salam[12], Ahlam Bashiti [19], Aisha Alansari[7], Ahmed Ashraf[7], Nora Alturayeif[20], Alcides Alcoba Inciarte[1], Adel Ammar[2], Abdelrahim A. Elmadany[1], Mohamedou Cheikh Tourad[18], Ismail Berrada[10], Mustafa Jarrar[21,19], Shady Shehata[22], Muhammad Abdul-Mageed[1,22]

[1] The University of British Columbia, [2] Prince Sultan University, [3] Tuwaiq Academy, [4] Independent Researcher, [5] Hadhramout University, [6] Misr University for Science & Technology, [7] KFUPM, [8] Taibah University, [9] WMU, [10] UM6P, [11] IBB University, [12] Cairo University, [13] UCA, [14] Qafza Tech, [15] JUST, [16] KAUST, [17] Damascus University, [18] University of Nouakchott, [19] Birzeit University, [20] Imam Abdulrahman Bin Faisal University, [21] Hamad Bin Khalifa University, [22] Invertible AI

{fakhr.alwajih, muhammad.mageed}@ubc.ca

Figure 1: Overview of PEARL, covering all but three Arab countries (16 illustrated here) and presenting representative examples from 11 of our 13 challenging question categories. These questions require reasoning and deep cultural knowledge (English translations provided in Appendix D.1). Prompts span diverse cultural domains, with images sourced from Wikipedia and other publicly available resources.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 23048–23079
November 4-9, 2025

Abstract

Mainstream large vision-language models (LVLMs) inherently encode cultural biases, highlighting the need for diverse multimodal datasets. To address this gap, we introduce PEARL, a large-scale Arabic multimodal dataset and benchmark explicitly designed for cultural understanding. Constructed through advanced agentic workflows and extensive human-in-the-loop annotations by 37 annotators from across the Arab world, PEARL comprises over 309K multimodal examples spanning ten culturally significant domains covering all Arab countries. We further provide two robust evaluation benchmarks (PEARL and PEARL-LITE) along with a specialized subset (PEARL-X) explicitly developed to assess nuanced cultural variations. Comprehensive evaluations on state-of-the-art open and proprietary LVLMs demonstrate that reasoning-centric instruction alignment substantially improves models' cultural grounding compared to conventional scaling methods. PEARL establishes a foundational resource for advancing culturally-informed multimodal modeling research. All datasets and benchmarks are publicly available.[1]

1 Introduction

Mainstream large vision-language models (LVLMs) predominantly encode Western perspectives, often neglecting diverse cultural traditions. Recent efforts have partially addressed this issue through multilingual datasets (Romero et al., 2024; Liu et al., 2025b) and benchmarks like CultureVLM (Liu et al., 2025b), GIMMICK (Schneider et al., 2025), CVQA (Romero et al., 2024), and related initiatives (Vayani et al., 2024). However, some of these resources lack sufficient cultural depth, often relying on translated content, limited image sets, narrow topical coverage, and simplistic factual questions unsuitable for evaluating advanced LVLM capabilities. Furthermore, they frequently assume local cultural homogeneity, overlooking nuanced regional variations even within shared cultural artifacts. Thus, developing high-quality, culturally nuanced multimodal resources that adequately challenge contemporary LVLMs remains significantly underexplored, especially for many global regions, including the Arab world.

Addressing this gap, we introduce PEARL, a large-scale Arabic multimodal instruction dataset and benchmark explicitly designed for cultural understanding. To ground our work, we operationalize *culture* through the lens of *cultural heritage*. That is, the focus of the work is on the enduring traditions, values, and practices validated by native speakers (Pawar et al., 2025). This framing provides a clear scope that distinguishes deeply rooted customs from contemporary, globalized practices. Our approach aligns with recent efforts to systematically measure and model cultural dimensions in language models, which highlight the importance of heritage and regional specificity (Pawar et al., 2025; Liu et al., 2025a).

PEARL is the outcome of an extensive collaborative effort involving a diverse community of 37 contributors, all of whom are authors of this work, spanning the Arab region. By leveraging advanced agentic workflows alongside iterative human-in-the-loop refinement, our dataset integrates sophisticated LLM and LVLM outputs with the nuanced cultural expertise of native annotators. Covering ten culturally significant domains, PEARL authentically captures Arab cultural heritage. Moreover, we structure it around ten distinct question categories designed to test sophisticated LVLM capabilities, including *hypothesis formation, problem-solving, comparative analysis*, and *chronological sequencing*.

To specifically evaluate subtle cultural variations overlooked by existing benchmarks, we also introduce PEARL-X (Figure 3, a novel benchmark highlighting culturally shared yet visually distinct concepts (e.g., *coffee*) across different Arab contexts. Unique among benchmarks, PEARL-X incorporates both text-to-single-image and text-to-multiple-image pairings, enabling richer assessments of LVLM performance on complex multimodal tasks. We leverage our benchmarks to systematically evaluate a range of open and proprietary LVLMs across diverse sizes and capabilities. Our results demonstrate that models incorporating reasoning-based cultural alignment substantially outperform those relying solely on conventional scaling approaches.

The rest of this paper is organized as follows: Section 2 reviews related work on multilingual and cultural benchmarks. Section 3 details our data collection, annotation pipeline, and benchmark design. Section 5 describes the evaluation protocol and experimental setup. Section 6 presents our findings and analysis. Section 7 introduces the novel PEARL-X benchmark .Finally, Section 8 concludes the paper and outlines future directions.

[1] https://github.com/UBC-NLP/pearl

2 Related Work

Multilingual VQA Datasets. Multilingual VQA datasets are predominantly constructed by generating QA pairs in various target languages, often alongside English counterparts. The creation methodologies for these datasets include manual annotation (Romero et al., 2024; Liu et al., 2021; Pfeiffer et al., 2021; Das et al., 2024), fully automatic generation (Becattini et al., 2023), or automatic translation followed by human verification (Changpinyo et al., 2022; Vayani et al., 2024; Wang et al., 2024a; Das et al., 2024; Liu et al., 2025b). Notable examples include VISCOUNTH (Becattini et al., 2023), a large-scale Italian-English dataset featuring 500K images and 6.5M QA pairs that were semi-automatically generated using an existing ontology-based knowledge graph. Another significant resource is the M5 benchmark (Schneider and Sitaram, 2024), which encompasses eight datasets across five vision tasks (introducing M5-VGR and M5-VLOD). While M5 spans 13 languages, its collection of 79,470 images presents a limitation considering its broad linguistic coverage.

Multi-Cultural VQA Datasets. Relatively few VQA datasets prioritize cultural diversity in their image curation and QA generation processes. For instance, CulturalVQA (Nayak et al., 2024) incorporates images representing cultural concepts (2,378 image-QA pairs, 11 countries) but is limited to English. The ALM-bench (Vayani et al., 2024) benchmark covers cultural aspects from 73 countries and 100 languages (2,929 images, 22,763 QA pairs, 13 domains). Its QA pairs were automatically generated and translated using GPT-4o, followed by human verification. MaXM (Changpinyo et al., 2022) is a test-only VQA dataset in seven languages (five scripts), featuring 2,142 auto-generated and human-verified QA pairs. CVQA (Romero et al., 2024) provides QA pairs in the local languages of 30 countries, alongside English translations for 5,239 images (10,374 questions). More extensively, CultureVerse (Liu et al., 2025b) contains 74,959 images covering 19,682 cultural concepts from 188 countries, with its GPT-4o- generated QA pairs validated through automated and manual checks.

Arabic Cultural VQA. Despite the growing interest in VQA, Arabic cultural representation in existing datasets remains sparse. The Henna dataset (Alwajih et al., 2024) is specifically dedicated to Arabic culture, comprising 1,132 images from 11 Arabic countries. Other datasets offer minimal Arabic content: ALM-bench (Vayani et al., 2024) includes 168 images from Saudi Arabia, UAE, and Egypt; CVQA (Romero et al., 2024) contains approximately 200 Egyptian images; and CultureVerse (Liu et al., 2025b) features only seven Libyan and 272 Egyptian images. More recently, JEEM (Kadaoui et al., 2025) introduced a benchmark for five Arabic dialects, consisting of 10.89K QA pairs and 2,178 images across 13 cultural domains. While JEEM improves Arabic VQA coverage, it is limited to four question categories and highlights the ongoing need for broader and deeper cultural representation.

To the best of our knowledge, PEARL is the first large-scale, culturally diverse Arabic multimodal benchmark carefully constructed through extensive human supervision, covering a wide range of cultural domains and challenging question types requiring reasoning and deep cultural knowledge. A comparative analysis with existing datasets is presented in Table 1. Additional discussion on cultural biases in LVLMs and existing monolingual VQA datasets is provided in Appendix A.

3 Methodology

3.1 Annotation Process

Annotation team. Our pipeline for PEARL involves the use of an agentic workflow intertwined with human effort by our annotation team. The team comprises 37 local members from nine Arab countries[2] all native Arabic speakers, each holding at least a bachelor's degree. We cover all Arab countries except three[3] For each of these countries, we assigned at least two members either from the same country or a neighboring country (to ensure familiarity with local culture).

Annotation Guidelines. Over a period of six months, we developed an extensive set of annotation guidelines covering the different stages of the project. This includes a number of dimensions: (i) introduction of cultural domains and criteria of data selection based on uniqueness to culture and relevance of images to the articles we collect, (ii) illustrated definitions of question types and criteria for characterizing high-quality questions and answers and how to improve these, and (iii) illustrative screenshots of the annotation platform

[2]Egypt, Palestine, Yemen, Morocco, Tunisia, Syria, Jordan, Saudi Arabia, Mauritania

[3]These are Comoros, Djibouti, and Somalia.

Category	Dataset	Lang.	Domains	Images	AraQA/Total	Q-Type	Q-Form	Ann.	CC	BC
Multilingual	CVQA* (Romero et al., 2024)	30	10	5,239	200/10.4K	MCQ	Fixed	M	✓	✗
	MMBench (Liu et al., 2025c)	2	20	2974	00/3.2K	MCQ	Fixed	A + M	✗	✓
	EXAMS-V (Das et al., 2024)	11	20	20,932	00/20.9K	MCQ	Diverse	M	✗	✓
	MaRVL (Liu et al., 2021)	5	1	5464	00/5,464	TF	Fixed	M	✓	✓
	M3Exam (Zhang et al., 2023)	9	3	2,816	00/12.3K	MCQ	Diverse	A + M	✗	✓
	MaXM (Changpinyo et al., 2022)	7	-	700	00/2.1K	SVQA	Fixed	A + M	✓	✓
	xGQA (Pfeiffer et al., 2021)	8	-	459	00/14.4K	Y/N, SVQA	Fixed	M	✗	✗
	M4U (Wang et al., 2024a)	3	64	8,931	00/8.9K	MCQ	Fixed	A + M	✗	✓
	CultureVerse* (Liu et al., 2025b)	188	15	74,959	279/196.7K	MCQ, SVQA, LVQA	Diverse	A + M	✗	✗
	ALM-Bench* (Vayani et al., 2024)	100	19	2,929	1,008/22.7K	MCQ, SVQA, LVQA, TF	Diverse	A + M	✓	✓
	WorldCusine(Winata et al., 2024)	30	-	6,045	00/1.2M	MCQ, open-ended	-	A	✗	✗
X-Specific	MMMU (Yue et al., 2024)	English	30	11,550		MCQ, SVQA	Fixed	M	✗	✗
	ViTextVQA (Nguyen et al., 2024)	Vietnamese	-	16,762	00/50.3K	LVQA	Fixed	M	✗	✗
	MMT-Bench (Ying et al., 2024)	English	-	-	00/31.3K	MCQ	Fixed	A + M	✗	✗
	JMMMU (Onohara et al., 2024)	Japanese	-	1,118	00/1.3K	MCQ	Fixed	A + M	✓	✗
	HaVQA (Parida et al., 2023)	Hausa	-	1,555	00/6K	LVQA	Fixed	M	✗	✗
	CULTURALVQA (Nayak et al., 2024)	English	-	2328	2378	open-ended	Diverse	M	✓	✓
	VLBiasBench(Wang et al., 2024c)	English	11	46,848	00/128.3K	LVQA, MCQ	-	A + M	✗	✓
Arabic	Henna (Alwajih et al., 2024)	Arabic	5	120	1,132	SVQA	Diverse	A + M	✓	✗
	CAMEL-Bench (Ghaboura et al., 2024)	Arabic	8	-	29K/29K	MCQ	Fixed	A + M	✗	✗
	JEMM (Kadaoui et al., 2025)	Arabic	13	2,178	10,890K	open-ended	Diverse	A	✓	✓
	Pearl (ours)	Arabic	10	12k	309k/309k	13-Q-Type**	Diverse	A + M	✓	✓

Table 1: Comparison of related visual datasets, covering multilingual, cross-specific, and Arabic resources. **Lang.**: number of languages for multilingual datasets or the language name for cross-specific language datasets. **AraQA/Total**: number of Arabic questions compared to the total number of covered questions. **Q-Type**: types of questions, for instance, long VQA (LVQA), short VQA (SVQA), and True False (TF) questions. **Q-Form**: question phrasing of each image. **Ann.**: annotation method used while creating the datasets ("M:" manual data collection, filtering, and annotation; "A:" automatic). **CC**: inclusion of cultural content. **BC**: use of bias correction. *The CVQA contains Arabic samples. **PEARL has 13 different Q-types, as described in Table C.1.

itself. Our full annotation guidelines are available at https://github.com/UBC-NLP/pearl

Annotation Platform and Communication. We utilized Label Studio (Tkachenko et al., 2020) as our primary annotation platform, organizing annotators into country-specific teams. Each annotator carefully reviewed content relevant to their respective country, closely following our detailed guidelines. To ensure effective coordination, we maintained a dedicated Slack channel for real-time communication, feedback, and progress updates. Additionally, we conducted weekly full-team meetings, recorded and distributed to all team members, complemented by smaller team meetings scheduled as needed. Annotators received recorded video tutorials demonstrating practical annotation examples and addressing common challenges. Further details regarding the annotation process and platform setup can be found in Appendix B.

3.2 Human-in-the-Loop Agentic Workflow

Our workflow begins by collecting image-article pairs from Wikipedia. Each image then undergoes a human review process to filter out irrelevant content (*Phase I*). Next, we employ specialized LLM agents to generate diverse categories of questions based on the selected images (*Phase II*). Finally, these image-question pairs undergo further human revision to ensure quality and cultural relevance (*Phase III*). This workflow is illustrated in Figure 2, and we describe each phase in greater detail below.

3.2.1 Phase I: Data Filtering

We initiate our data collection process by selecting ten culturally significant domains: *architecture, clothing, fauna, festivals and celebrations, flora, foods, geography, handcrafts, landmarks, and music*. These domains capture the diverse traditions and cultural identities prevalent across the Arab world, forming a robust foundation for curating a high-quality multimodal dataset comprising culturally contextualized text and images. Arabic *Wikipedia* serves as our primary source due to its extensive and accessible country-specific content available in the native language. We systematically gathered relevant articles along with their corresponding images, prioritizing those that distinctly highlight culturally meaningful topics from various Arab countries. Each article was carefully categorized into one of the predefined cultural domains, guided closely by *Wikipedia's* established internal taxonomy.

Due to the heterogeneous nature of *Wikipedia* content, initial retrieval yielded varied quality and cultural relevance. Hence, we implemented a hybrid human and automatic filtering strategy. First,

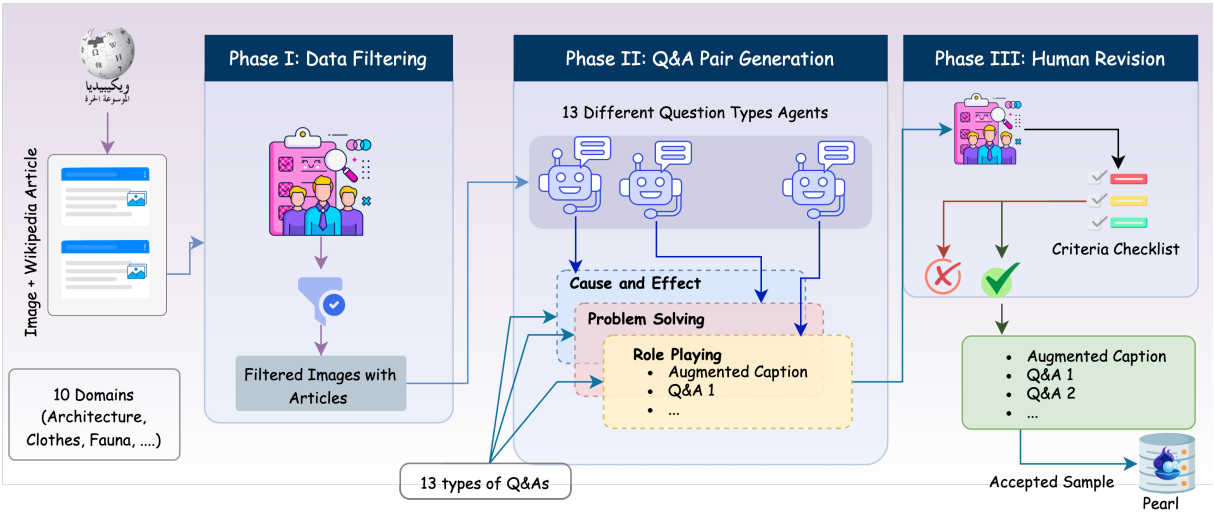

Figure 2: Illustrates the three main stages of our data annotation pipeline. (1) selecting culturally relevant images from *Wikipedia* articles across ten predefined domains (e.g., food, music, architecture). (2) Agents generate augmented captions and Q&A pairs across 13 question types. (3) Human reviewers apply a quality checklist to accept or reject samples for the final dataset.

an ***automated filtering*** evaluated each article and image based on basic metadata alignment criteria, such as consistency between image captions, articles, and categorical tags. This automated filter significantly reduced noise by discarding obviously irrelevant or misaligned images. After the initial automated filtering, human annotators carefully reviewed each item using a structured annotation interface on Label Studio. During this ***manual review*** phase, annotators ensured that images and articles were correctly aligned, culturally relevant, authentic, and factually accurate. Any image or article not meeting these standards of authenticity and relevance was excluded from further processing.

3.2.2 Phase II: Q&A Pair Generation

In this stage, filtered image-article pairs undergo an automated content generation process. We employ agents backboned with an LVLM — specifically `Qwen2.5-VL-72B-Instruct`—to create synthetic augmented captions and structured question–answer (Q&A) pairs for each image. We use a fixed schema of 13 predefined question types, each specifically tailored to assess different levels of understanding and reasoning skills. For example, we include questions involving *cause-and-effect, chronological ordering, comparative analysis, modern context interpretation, hypothesis formation*, and *scenario completion*. Table B.1 shows the full set of our question categories, and Figure 1 illustrates an example for each type.

For each filtered image-article pair, the LVLM generates an augmented caption that integrates visual details from the image with relevant cultural or historical context from the associated *Wikipedia* article (prompt details provided in Appendix D.1). Subsequently, the LVLM produces one augmented caption along with 2–4 structured Q&A pairs aligned explicitly with predefined reasoning categories (e.g., cause-and-effect reasoning). Each Q&A generation prompt clearly specifies the targeted reasoning type, ensuring the resulting questions are culturally meaningful and require synthesizing both visual and textual inputs. Detailed prompts used for generating Q&A pairs are provided in Appendices D.2 and D.3.

3.2.3 Phase III: Human Revision

We employ a *two-stage* human-in-the-loop quality control process. In *stage one*, annotators validate the augmented captions against the original image and article metadata based on four main criteria: (i) *cultural authenticity,* to ensure the captions genuinely reflect the cultural context depicted in the images; (ii) *visual relevance,* to confirm that the images clearly match the caption; (iii) *clarity and precision,* to guarantee question-answer pairs are understandable and grammatically correct; and (iv) *content accuracy,* to guarantee the question-answer pairs are consistent with the provided articles or other credible sources annotators can locate online. Figure D.5 illustrates caption-image validation and Figure D.6 illustrates the question-answer review process.

In *stage two* of human revision, carried out through a custom-built user interface, we sample a total of $11,000$ question-answer pairs from stage one and examine them to verify cultural relevance and ensure quality. All pairs not adhering to our revision criteria in *Phase III* are rejected. Figure B.1 shows a snapshot of the user interface we use for the second round of data revision.

4 Our Datasets

Apart from our main benchmarks (see Sections §4.1 and §7), our pipeline produces several high-quality datasets suitable for various purposes: (1) **Culturally-Relevant Images**, comprising $12,637$ manually selected images reflecting distinct country-level cultural nuances across ten domains; (2) **Augmented Captions**, a set of $135k$ carefully crafted captions via agents derived from Arabic Wikipedia articles. These enriched captions serve as a visual knowledge base, providing contextually rich and interconnected information beyond standard image descriptions, and are publicly released for community use; (3) **Automated Q&A Pairs**, including 309K question-answer pairs systematically generated by a suite of 13 specialized agents tailored to specific question categories; and (4) **Human-Revised Q&A Pairs**, a high-quality subset of 16K question-answer pairs reviewed by human annotators to ensure cultural accuracy and relevance. Collectively, these resources provide robust datasets for work involving VLMs. We now introduce our evaluation benchmarks.

4.1 PEARL Benchmarking Data

PEARL Benchmark. The PEARL benchmark is composed of $6,301$ high-quality Q&A pairs, carefully selected through a rigorous two-stage human evaluation process from an initial set of 16K pairs drawn from our larger corpus of 309K pairs. Specifically designed for model evaluation, PEARL includes $4,832$ *closed-form* multiple-choice and True/False questions focused on recognition and interpretation of culturally relevant visual information. This component evaluates LVLMs' fundamental accuracy and factual correctness, providing a reliable baseline to assess their foundational Arabic cultural knowledge. Additionally, the benchmark contains $1,469$ *open-ended* questions that require deeper engagement, such as *hypothesis testing, causal analysis, scenario completion*, and *role-playing*. These 11 different open-ended question types, as detailed in Table B.1, assess LVLMs' depth of cultural comprehension, requiring models not merely to recall cultural facts but also to generate nuanced explanations, reason through scenarios, and contextualize cultural elements within coherent, culturally informed narratives.

PEARL-LITE. A streamlined subset comprising 893 Q&A pairs (591 closed-form and 302 open-ended questions), PEARL-LITE is randomly sampled from Pearl while maintaining a balance of question types and countries. It is designed to facilitate efficient evaluation of proprietary models and minimize costly API usage.

5 Experimental Setup

5.1 Models

For evaluation of PEARL, we use both open and proprietary models (accessible via API calls). We use open models ranging in size between 3 billion to 72 billion parameters. To ensure a level playing field, all open models received the same input, sampling configurations, and temperature during the generation process. The specific models we evaluate are `Qwen2.5-VL` (Bai et al., 2025), `Aya-Vision` (Dash et al., 2025), `Gemma3` (Team et al., 2025), `Gemeni2.5 Pro` (DeepMind, 2025), `Claude Sonnet 4` (Anthropic, 2025), and `o3 (OpenAI)` (OpenAI, 2025). To promote reproducibility and future studies, we are making the complete inference logs for each evaluated model publicly available.

5.2 Evaluation Protocol

5.2.1 LVLM as Judge

We employ an automatic evaluation method using an *LVLM-as-judge* framework. For this role, we exclusively use `InternVL3.5-38B` (Wang et al., 2025), a model distinguished by its advanced reasoning capabilities. This model is part of the state-of-the-art InternVL3.5 family, which has demonstrated top-tier performance among open-source MLLMs across general multimodal, reasoning, text, and agentic tasks.

Metrics. Our evaluation protocol adopts two distinct scoring methods, each tailored to a specific question format. For ***closed-form*** questions (e.g., multiple-choice and True/False), we utilize a *relaxed-match accuracy* (ACC) metric. Here, the judge assesses semantic equivalence between candidate responses and gold-standard answers, permitting synonyms, paraphrases, or minor lexical variations. Each response is assigned a binary correctness score (1 for correct, 0 for incorrect),

aggregated into an overall accuracy. For ***open-ended*** questions—comprising 11 varied types such as cause-and-effect, comparative analysis, and scenario completion—the judge evaluates responses using a comprehensive structured rubric capturing four critical dimensions: *correctness*, *coherence*, *detail*, and *fluency*.[4] Each dimension is scored individually on a scale from 1 to 5, with an aggregated, weighted ***Overall Score*** calculated as follows:

$$\begin{aligned} \textit{Overall Score} = 0.4\ &\textit{Correctness} \\ &+ 0.2\ \textit{Coherence} \\ &+ 0.2\ \textit{Detail} \\ &+ 0.2\ \textit{Fluency} \end{aligned} \tag{1}$$

We also follow Burda-Lassen et al. (2025) in employing a *Cultural Awareness Score (CAS)*. CAS is a binary metric (0/1) indicating explicitly whether the candidate response mentions culturally-specific elements required by the reference answer, ensuring explicit cultural grounding in the evaluations.

5.2.2 Human Evaluation

We evaluate using a subset of PEARL-LITE open-ended dataset (N=70). To obtain human evaluations, we recruited four native Arabic speakers with deep familiarity in the relevant cultural contexts, who independently scored each sample according to the same evaluation rubric applied by our LVLM judges. The user interface utilized by human evaluators can be found in Appendix D.8.

To assess the reliability of our human annotations and LVLM judge, we conducted Intraclass Correlation Coefficient (ICC) (Shrout and Fleiss, 1979) analysis on the overall evaluation scores. First, we measured inter-annotator agreement among the four human evaluators, which revealed moderate reliability for individual raters ($ICC_{(3,1)} = 0.537, 95\%$ Confidence Interval (CI) $[0.42, 0.65]$) but good reliability when averaged across all annotators ($ICC_{(3,k)} = 0.823, 95\%$ CI $[0.74, 0.88]$). Subsequently, we evaluated the agreement between our LVLM judge and the consensus human ratings (averaged across the four annotators) to establish the reliability of our automated evaluation approach. The model demonstrated good agreement with human consensus ($ICC_{(3,1)} = 0.708, 95\%$ CI $[0.57, 0.81]$), demonstrating that our LVLM judge can produce evaluations that are reasonably consistent with human expert judgment.

[4] See Figure D.10 for detailed evaluation prompts defining these criteria explicitly.

6 Results and Discussion

6.1 Open Models

Table 2 presents the performance of nine openly-accessible LVLMs on the PEARL benchmark. We organize the discussion around two axes: (*i*) parameter scaling within the same family and (*ii*) cultural grounding and closed-form accuracy.

Scaling within the same family. Within the `Qwen2.5-VL` line, performance steadily improves as model size grows: the overall score rises from 2.74 for the 3B-parameter instruct model to 3.16 for the 7B variant, reaching 3.46 for the 72B model. A similar upward trajectory is seen in the CAS, which climbs from 33.70% to 44.79% and then to 49.63%. Interestingly, the `Qwen2.5-VL-32B-Instruct`[5] reasoning model breaks this trend, outperforming even the 72B variant with an overall score of 3.77 and a CAS of 67.05%. This suggests that architectural or training enhancements in the reasoning variant provide stronger cultural grounding and open-ended response quality than parameter scaling alone.

Cross-family comparison highlights stylistic trade-offs. The `Gemma-3` series delivers balanced results despite lacking dedicated reasoning training. Its 12B and 27B variants achieve CAS values around $48 - 56$ higher than `Aya-Vision-8/32B`—and competitive Overall scores with other models. `Aya-Vision-32B`, on the other hand, favors fluent, stylistically polished answers (FLU 4.29) but lags in cultural specificity (CAS 51.2). These contrasts confirm that model design choices (pretraining corpora, vision encoder quality, alignment objectives) influence different quality axes in complementary ways.

Take-away. Among the *nine* open models, the clear front-runner is `Qwen2.5-VL-32B-Instruct`. It combines the highest Overall score (3.77) with the strongest closed-form accuracy (79.8%) and the best CAS (67.1%) of any system in Table 3. For downstream Arabic-cultural applications where access to proprietary systems is not feasible, `Qwen2.5-VL-32B` therefore offers the most reliable balance of factual correctness and cultural grounding. Looking ahead, further progress is likely to come from reasoning-centered alignment and

[5] The official release notes stating that Qwen2.5-VL-32B is specifically designed for enhanced reasoning and closer alignment with human preferences; see https://qwenlm.github.io/blog/qwen2.5-vl-32b/.

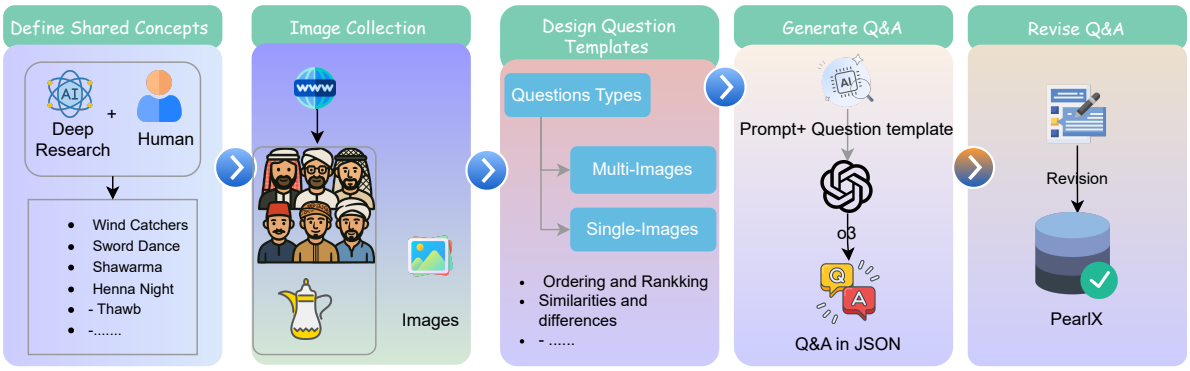

Figure 3: Workflow generation of PEARL-X.

Figure 4: PEARL-X exemplars for the Multiple-Image (top) and Single-Image (bottom) question formats. Each prompt is shown in Arabic with its English translation beneath. All depicted faces are synthetic and do not portray real individuals.

	Open-Ended						Closed
Model	**COR**	**COH**	**DET**	**FLU**	**OVR**	**CAS%**	**ACC%**
Qwen2.5-VL-3B-Instruct	2.48	3.07	2.05	3.60	2.74	33.70	71.15
gemma-3-4b-it	2.86	3.43	2.54	4.06	3.15	44.72	69.78
Qwen2.5-VL-7B-Instruct	2.89	3.55	2.45	4.02	3.16	44.79	73.01
aya-vision-8b	3.17	3.77	2.62	4.21	3.39	45.34	69.89
gemma-3-12b-it	2.98	3.54	2.54	4.10	3.23	48.13	75.83
gemma-3-27b-it	3.26	3.76	2.82	4.23	3.47	55.75	79.86
aya-vision-32b	3.36	3.91	2.75	4.29	3.53	51.19	79.12
Qwen2.5-VL-32B-Instruct ◆	3.61	3.99	3.30	4.32	3.77	67.05	79.78
Qwen2.5-VL-72B-Instruct	3.27	3.82	2.71	4.21	3.46	49.63	79.55

Table 2: Performance on the PEARL benchmark for open models. Open-ended metrics (COR, COH, DET, FLU, OVR) are averaged on a 1–5 scale, while CAS% and ACC% are reported as percentages. A ◇ marks models that use explicit reasoning techniques in this paper

culturally informed pre-training, rather than from simply adding more parameters.

	Open-Ended						Closed
Model	**COR**	**COH**	**DET**	**FLU**	**OVR**	**CAS%**	**ACC%**
Qwen2.5-VL-3B-Instruct	2.59	3.12	2.15	3.68	2.83	37.09	73.10
gemma-3-4b-it	2.97	3.56	2.62	4.12	3.25	47.68	70.73
Qwen2.5-VL-7B-Instruct	3.02	3.66	2.55	4.11	3.27	47.68	73.77
aya-vision-8b	3.23	3.85	2.67	4.21	3.44	46.69	70.56
gemma-3-12b-it	3.10	3.63	2.54	4.14	3.30	52.65	76.82
gemma-3-27b-it	3.38	3.81	2.95	4.26	3.55	60.93	80.88
aya-vision-32b	3.37	3.92	2.76	4.31	3.55	51.66	75.63
Qwen2.5-VL-32B-Instruct ◆	3.69	4.09	3.39	4.42	3.85	66.56	80.03
Qwen2.5-VL-72B-Instruct	3.36	3.91	2.76	4.25	3.53	55.63	79.36
claude-sonnet-4-20250514 ◆	3.77	4.03	3.71	4.43	3.94	76.49	79.53
gemini-2.5-pro ◆	4.36	4.48	4.45	4.77	4.48	83.11	89.00
o3-2025-04-16 ◆	4.39	4.44	4.52	4.71	4.49	87.09	86.97

Table 3: Results for the PEARL-LITE subset. A ◇ marks models that use explicit reasoning techniques in this paper.

6.2 Proprietary Models

Results on the PEARL-LITE benchmark (Table 3) clearly demonstrate the superiority of proprietary models (`Gemini Pro` and `OpenAI o3`) across all evaluation dimensions. In particular, `OpenAI o3`

Model	Accuracy %
Qwen2.5-VL-3B-Instruct	59.67
gemma-3-4b-it	64.58
Qwen2.5-VL-7B-Instruct	61.31
aya-vision-8b	64.03
gemma-3-12b-it	69.21
gemma-3-27b-it	69.21
aya-vision-32b	71.66
Qwen2.5-VL-32B-Instruct ◆	71.66
Qwen2.5-VL-72B-Instruct	73.57
gemini-2.5-pro-preview-05-06 ◆	77.93
o3-2025-04-16 ◆	78.75

Table 4: Accuracy on the PEARL-X shared-concepts benchmark.

achieves the highest overall score of 4.49 and the highest CAS of 87%. Furthermore, proprietary models consistently outperform open models on closed-form question accuracy, with `Gemini2.5 Pro` achieving 89%, surpassing the best-performing open model (`gemma-3-27b-it`) at 81%.

7 PEARL-X Benchmark

While working on PEARL, we observed existence of cultural element or practices prevalent across multiple regions or countries that, despite a fundamental similarity, exhibit subtle local variations in appearance, preparation, usage, or style. For instance, the traditional headband known as the *agal* "العقال", widely worn in countries such as the UAE, Saudi Arabia, and other Gulf nations, displays regional differences in shape, color, and style. Motivated by this cultural insight and supported iteratively by the `ChatGPT-o3` model,[6] we manually identified 61 such culturally shared concepts, each observed across at least two[7] Arab countries.[8] A complete list of the identified shared concepts is provided in Appendix D.2. We then manually located images from publicly accessible web resources. On average, we gathered approximately three representative images per concept, for a total of 347 images, ensuring a rich visual depiction of cultural diversity. We then developed *MCQ* and *True/False* questions exploiting the collected images. We include two categories of questions, differing based on whether we feed the model a single image or multiple images. *Single-image* questions focus on aspects specific to one country through an individual image, whereas *multiple-image* questions target variations among countries regarding the particular shared concept depicted across several images. The next step was to generate questions based on manually developed templates that we provide to `ChatGPT-o3` along with the images and the name of each shared concept, enabling it to generate relevant questions. The prompt we developed to generate the questions is shown in Appendix D.9. In total, we produce 367 questions, split into 177 single-image and 190 multiple-image questions. Finally, we conduct a thorough human review of all generated questions, ensuring that the questions and answers are accurate, meaningful, error-free, and diverse. Figure 3 demonstrates the workflow used to develop PEARL-X benchmark. Figure 4 shows examples of the shared concepts for both single and multiple image questions. Additionally, Appendix D.3 presents sample questions based on one-image and multi-image templates used for generating shared concept questions.

7.1 Evaluation on PEARL-X

As shown in Table 4, proprietary models (`Gemini 2.5 Pro` and `o3`) demonstrate superior performance, with `o3-2025-04-16` achieving the highest accuracy (78.75%). Among the open-source models, `Qwen2.5-VL-72B-Instruct` performs best, reaching an accuracy of 73.57%. Notably, reasoning-centric models generally outperform their counterparts, emphasizing the critical role of explicit reasoning alignment in culturally sensitive contexts.

8 Conclusion

In this paper, we presented PEARL, an Arabic multimodal instruction dataset and benchmarking suite tailored to enhance the cultural understanding capabilities of large vision-language models. PEARL addresses critical gaps in existing resources by encompassing diverse culturally-authentic materials across ten domains from the Arab world. Our rigorous annotation and agentic workflows, combined with the expertise of 37 local annotators, ensure high-quality, culturally-relevant content. Comprehensive evaluations confirm the superior performance of models explicitly optimized for reasoning tasks over parameter scaling, underscoring the importance of culturally-aware alignment methods. The specialized PEARL-X benchmark further allows nuanced assessment of cross-country cultural variations, setting the stage for more culturally sensitive model development.

[6] We used the model to generate the initial questions.

[7] For example, the traditional dish *kabsa* "كبسة" appears widely in seven Arab countries, with notable regional variations.

[8] Initially, we compiled a preliminary list of approximately 100 potential shared concepts. After thorough manual filtering to exclude irrelevant or inaccurately represented concepts, we finalized a refined set of 61 authentic cultural concepts.

Limitations

Despite its comprehensive scope, the PEARL dataset has a number of limitations. First, the dataset predominantly relies on publicly available resources like Wikipedia, potentially introducing biases towards topics and perspectives that are well-documented online. Second, while we involved annotators from diverse Arab countries, the coverage does not equally represent all regions. Additionally, although we employed rigorous human-in-the-loop annotation processes, subjective cultural interpretations may still influence data annotation consistency. Lastly, due to the focus on cultural specificity, generalizability to other non-Arabic cultures or languages may be limited, requiring additional datasets and evaluations tailored to different cultural contexts.

Ethics Statement

In developing PEARL, we emphasized cultural sensitivity, inclusivity, and ethical responsibility. All annotations were created by informed participants, each of whom is acknowledged and credited as a contributor. We adhered strictly to publicly available and reputable sources, refraining from using any private or sensitive data. Clear guidelines were provided to respect local norms, maintain data privacy, and secure participant consent. All images utilized in this dataset are sourced from Wikipedia under Creative Commons licenses, and any images originating outside Wikipedia have been masked or regenerated with alternative identities to ensure privacy and ethical compliance.

Although PEARL aims to mitigate biases in Arabic LVLMs, unintentional cultural bias may still occur—particularly in regions lacking direct local representation. We encourage ongoing community involvement to address these gaps, ensuring continual refinement and improvement of the dataset.

Reproducibility. Our test data, prompts, and code necessary to produce all results reported in this work are publicly available at `https://github.com/UBC-NLP/pearl`.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[9] and UBC Advanced Research Computing-Sockeye.[10] We also acknowledge the valuable contributions of Razan Khassib, Lina Hamad, Fatimah Alshamari, Cheikh Malainine, Doaa Qawasmeh, Tfeil Moilid, Sara Shatnawi, and Ahmed Aboeitta.

References

Manel Aloui, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. 101 billion arabic words dataset. *arXiv preprint arXiv:2405.01590*.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of arabic multimodal large language models and benchmarks. *arXiv preprint arXiv:2403.01031*.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran AQ Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, and 1 others. 2024. Cidar: Culturally relevant instruction dataset for arabic. *arXiv preprint arXiv:2402.03177*.

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

Anthropic. 2025. System card: Claude opus 4 & claude sonnet 4. System card, Anthropic.

Yujin Baek, ChaeHun Park, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vlm collaboration. *arXiv preprint arXiv:2406.16469*.

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

Federico Becattini, Pietro Bongini, Luana Bulla, Alberto Del Bimbo, Ludovica Marinucci, Misael Mongiovì, and Valentina Presutti. 2023. Viscounth: a large-scale multilingual visual question answering dataset for cultural heritage. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–20.

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. *arXiv preprint arXiv:2407.00263*.

[9] https://alliancecan.ca

[10] https://arc.ubc.ca/ubc-arc-sockeye

Olena Burda-Lassen, Aman Chadha, Shashank Goswami, and Vinija Jain. 2024. How culturally aware are vision-language models? *arXiv preprint arXiv:2405.17475*.

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)*, pages 1–6. IEEE.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2022. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*.

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

Google DeepMind. 2025. Gemini 2.5 pro. Vertex AI Documentation. Accessed May 18, 2025.

Sara Ghaboura, Ahmed Heakl, Omkar Thawakar, Ali Alharthi, Ines Riahi, Abduljalil Saif, Jorma Laaksonen, Fahad S. Khan, Salman Khan, and Rao M. Anwer. 2024. Camel-bench: A comprehensive arabic lmm benchmark. *Preprint*, arXiv:2410.18976.

Xiaotian Han, Quanzeng You, Yongfei Liu, Wentao Chen, Huangjie Zheng, Khalil Mrini, Xudong Lin, Yiqi Wang, Bohan Zhai, Jianbo Yuan, and 1 others. 2023. Infimm-eval: Complex open-ended reasoning evaluation for multi-modal large language models. *arXiv preprint arXiv:2311.11567*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, and 1 others. 2024. Acegpt, localizing large language models in arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8132–8156.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, and 1 others. 2023. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*.

Karima Kadaoui, Hanin Atwany, Hamdan Al-Ali, Abdelrahman Mohamed, Ali Mekky, Sergei Tilga, Natalia Fedorova, Ekaterina Artemova, Hanan Aldarmaki, and Yova Kementchedjhieva. 2025. Jeem: Vision-language understanding in four arabic dialects. *arXiv preprint arXiv:2503.21910*.

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

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

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2025c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Zheng Ma, Mianzhi Pan, Wenhan Wu, Kanzhi Cheng, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2023. Food-500 cap: A fine-grained food caption benchmark for evaluating vision-language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5674–5685.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Aishwarya Agrawal, and 1 others. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.

Quan Van Nguyen, Dan Quang Tran, Huy Quang Pham, Thang Kien-Bao Nguyen, Nghia Hieu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2024. Vitextvqa: A large-scale visual question answering dataset for evaluating vietnamese text comprehension in images. *Preprint*, arXiv:2404.10652.

Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2024. Jmmmu: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. *Preprint*, arXiv:2410.17250.

OpenAI. 2025. OpenAI o3 and o4-mini System Card. Technical report, OpenAI. Version 2 of the Preparedness Framework.

Shantipriya Parida, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, Aneesh Bose, Guneet Singh Kohli, Ibrahim Said Ahmad, Ketan Kotwal, Sayan Deb Sarkar, Ondřej Bojar, and Habeebah Kakudi. 2023. HaVQA: A dataset for visual question answering and multimodal research in Hausa language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10162–10183, Toronto, Canada. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

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

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2021. xgqa: Cross-lingual visual question answering. *arXiv preprint arXiv:2109.06082*.

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

Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025. Gimmick–globally inclusive multimodal multitask cultural knowledge benchmarking. *arXiv preprint arXiv:2502.13766*.

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. *arXiv preprint arXiv:2407.03791*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, and 1 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 8612–8642. Curran Associates, Inc.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Xiujie Song, Mengyue Wu, Kenny Q. Zhu, Chunhao Zhang, and Yanyi Chen. 2025. A cognitive evaluation benchmark of image reasoning and description for large vision-language models. *Preprint*, arXiv:2402.18409.

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

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label studio: Data labeling software. *Open source software available from https://github. com/heartexlabs/label-studio*, 2022.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, and 1 others. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *arXiv preprint arXiv:2411.16508*.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. 2024a. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv preprint arXiv:2405.15638*.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. 2024b. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421.

Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. 2024c. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. *Preprint*, arXiv:2406.14194.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan,

Quanzeng You, and Hongxia Yang. 2024d. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

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 Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Ching Lam Cheng, Daud Abolade, Emmanuele Chersoni, and 32 others. 2024. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *Preprint*, arXiv:2410.12705.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, and 3 others. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *Preprint*, arXiv:2404.16006.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, 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. *Preprint*, arXiv:2311.16502.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Appendices

This appendix provides supplementary material to support the main findings of this work. It is organized as follows:

- **§A: Literature Review**
  Reviews related work on cultural bias in LLMs and VLMs, multilingual and monolingual VQA datasets, and visual reasoning benchmarks.
- **§B: Annotation Guidelines**
  Describes the iterative development of annotation guidelines, including domain definitions, question types, and quality control processes.
- **§C: Data Statistics**
  Provides detailed statistics on annotated samples, cultural category distributions, and clean filtered samples per Arab country.
- **§D: Additional Technical Details**
  Includes prompt templates for generating augmented captions and Q&A pairs, user interface screenshots, human evaluation examples, and evaluation score breakdowns.
- **§E: Fine-Grained Models Performance Analysis on Pearl-Lite by Country and Question Type**
  Presents detailed heatmaps of model performance on the Pearl-Lite benchmark, with results broken down by country and question type.

Key Tables

- **Table B.1:** Taxonomy of 11 culturally-focused question types.
- **Table C.1:** Image distribution across countries and cultural domains.
- **Table D.1:** Evaluation scores from LVLM judges and human annotators.
- **Table D.2:** A complete list of 61 identified shared concepts is provided.
- **Table D.3:** Examples of question templates for both multiple and single image prompts in Arabic, covering diverse reasoning types and cultural features.

Key Figures

- **Figure C.1:** Distribution of the number of images by cultural category for each Arab country in our dataset.
- **Figure D.1:** Shows prompt template step 1 used to generate augmented captions.
- **Figure D.2:** Illustrates template step 2 (Part 1) to generate Q&A pairs.
- **Figure D.3:** Displays prompt template step 2 (Part 2) used to generate answers.
- **Figure D.4:** Combined prompt template including the two steps for generating Q&A pairs.
- **Figure D.5:** Label Studio platform interface used by annotators in step one of the human revision phase to revise the augmented caption.
- **Figure D.6:** Step two of the revision phase, showing the process of reviewing and editing questions and answers on the platform.
- **Figure D.7:** Shows an example illustrates the format of the augmented caption followed by multiple Q&A pairs.
- **Figure D.8:** Human Eval UI
- **Figure D.10:** Evaluation prompts examples
- **Figure D.9:** Illustrates the prompt used to generate Q&A pairs for *Pearl X* using `OpenAI o3`.
- **Figure E.1:** Heatmap of accuracy scores on closed-form questions by model and country.
- **Figure E.2:** Heatmap of the Overall Score for open-ended questions by model and country.
- **Figure E.3:** Heatmap of the CAS for open-ended questions by model and country.
- **Figure E.4:** Heatmap of the Overall Score for open-ended questions by model and question type.

A Literature Review

Culture Bias in LLMs and LVMs. Beyond language understanding capabilities, cultural awareness and sensibility are critical to ensure practical effectiveness while mitigating the potential stereotypes and biases of text-based and multimodal LLMs. Thus, designing culturally aware LLMs requires understanding the various perspectives of a given culture. This involves recognizing diverse cultural dimensions such as traditions, beliefs, and social practices as one dimension, social interaction as the second dimension, and materialized objects as the third dimension (Pawar et al., 2024). Hence, different studies have been conducted to evaluate and identify cultural gaps and ensure diversity and inclusion in current SOTA LLMs.

(Pawar et al., 2024) provided a comprehensive survey on recent works on cultural awareness, exploring dataset creation methodologies, benchmarking techniques, and the ethical implications. For example, one of the main methodologies for dataset creation is relying on automatic pipelines that leverage public corpora to generate large datasets (Sengupta et al., 2023; Huang et al., 2024; Aloui et al., 2024). However, incorporating information from a specific culture into a general-purpose LLM can

lead to misinformation, stereotyping, biases, and misrepresentation when used to represent other cultures (Pawar et al., 2024). Generally, this stems from the use of machine-translated data to create multilingual LLMs, a widely adopted practice in the field (Hendrycks et al., 2020; Singh et al., 2024). In particular, (Singh et al., 2024) highlighted the propagated emphasis of Western perspectives on the topics covered in the translated English dataset, such as the MMLU benchmarks (Hendrycks et al., 2020).

To create a culturally aligned dataset, other works employ humans either as part of semi-automatic methods or completely to manually create cultural datasets from scratch (Pawar et al., 2024; Ma et al., 2023; Alyafeai et al., 2024; Huang et al., 2023; Baek et al., 2024). For example, to ensure a multilingual culturally diverse dataset, (Singh et al., 2024) introduced Global-MMLU, an improved set validated by humans and covering culturally sensitive and agnostic sets of 42 languages.

Similarly, relying on English pre-trained LLMs to build VLMs is the core cause of inherently encoding Western cultural knowledge. Thus, most vision-language models exhibit cultural misrepresentation (Burda-Lassen et al., 2024; Ananthram et al., 2024). Images often convey rich cultural stories and heritages; however, English-based VLM-generated captions tend to fail to accurately narrate cultural stories (Burda-Lassen et al., 2024). To study this phoneme, several works have been conducted to assess the performance of VLMs in culturally specific information (Burda-Lassen et al., 2024; Ananthram et al., 2024; Bhatia et al., 2024).

For example, (Burda-Lassen et al., 2024) assessed the performance of four VLMs' capabilities to identify cultural information in images to generate aligned culturally sensitive captions. Most of the models struggle to perform the task, as the highest cultural awareness score of 35% was achieved by Gemini Pro Vision. Furthermore, (Bhatia et al., 2024) introduced GLOBALRG benchmark to evaluate multicultural understanding and inclusivity in VLMs across universal and local concepts. Although the results of evaluating several SOTA VLMs on the image retrieval task varied across cultures, most of the retrieved images contain Western-specific elements. To mitigate the interpretation of images from the Western perspective only, (Ananthram et al., 2024) utilized a mix of diverse language-based VLMs to improve the model's understanding ability of Chinese cultural images.

Monolingual VQA Datasets. Several studies have addressed the VQA problem by developing language-specific datasets. Datasets like MMT-Bench (Ying et al., 2024), CAMEL-Bench (Ghaboura et al., 2024), and JMMMU (Onohara et al., 2024) considered only the MCQ type for English, Arabic, and Japanese languages, respectively. Similarly, the HaVQA (Parida et al., 2023) and ViTextVQA (Nguyen et al., 2024) datasets use open-ended questions in Hausa and Vietnamese languages, respectively. VLBiasBench (Wang et al., 2024c) offers a large English dataset with both open and closed questions. A semi-automated method was used to create K-Viscuit (Baek et al., 2024), a Korean cultural dataset generated with GPT-4 and human verification. In Arabic VQA, CAMEL-Bench (Ghaboura et al., 2024) includes 29,036 curated MCQs across various domains. This dataset integrates Arabic and translated content from existing English LMM benchmarks. It mainly focuses on Modern Standard Arabic (MSA) with limited attention to the Arabic dialects.

Visual Reasoning. Reasoning represents one of the fundamental human cognitive abilities derived from commonsense understanding and world knowledge. The cognitive process involves interpreting information, analyzing given conditions, and subsequently drawing inferences, solving problems, or making predictions. Although Multimodal LLMs (MLLMs) have achieved remarkable advancements across various domains, their capacity for robust reasoning, particularly in processing and integrating information from diverse modalities, remains a prominent and challenging research frontier. Consequently, a growing body of research has focused on evaluating and improving the reasoning capabilities of MLLMs, resulting in the development of specialized datasets and benchmarks aimed at rigorously assessing models' reasoning capabilities.

Several benchmarks have been proposed to evaluate different facets of MLLMs reasoning. InfiMM-Eval benchmark (Han et al., 2023) was designed to evaluate the reasoning capabilities of MLLMs through open-ended and multi-step reasoning across three categories: deductive, abductive, and analogical. This benchmark consists of 279 manually curated, high-quality questions paired with 342 images, placing a strong emphasis on the multi-step reasoning process. Further contributing to the evaluation and development of reasoning in

MLLMs, the Large-scale Visual Chain-of-Thought (Visual CoT) dataset (Shao et al., 2024) aims to improve the inheritability and reasoning abilities of MLLMs. It contains 438K visual question-answer pairs annotated with intermediate bounding boxes that highlight critical image regions. Around 98K of these pairs are further enriched with explicit CoT annotations that provide detailed reasoning steps.

Focusing specifically on spatial understanding, the SpatialEval dataset (Wang et al., 2024b) covers four spatial reasoning tasks, such as spatial relationships, navigation, positional understanding, and counting, to evaluate the spatial reasoning capabilities of LLMs and VLMs. Focusing on the same facet of reasoning, the SpatialVLM framework (Chen et al., 2024) introduces an approach to enhance the spatial reasoning abilities of VLMs by generating large-scale 3D spatial reasoning data. This framework transforms 2D images into detailed metric-scale 3D point cloud, enabling the synthesis of approximately two billion spatial reasoning QA pairs. These pairs are designed to cover both qualitative and quantitative spatial reasoning tasks.

CogBench (Song et al., 2025) is a cognitive evaluation benchmark designed to assess the reasoning abilities of LVLMs using a unique set of images inspired by the "Cookie Theft" cognitive assessment task. The dataset consists of manually collected images, mostly painting-style from Pinterest, which then underwent detailed human annotation. Annotators identified key entities, constructed explicit Chain-of-Reasoning (CoR) annotations, and provided thorough descriptions of the image content. This results in 251 annotated images, structured across eight reasoning dimensions and includes two primary evaluation tasks: an image description task and a multiple-choice VQA task. Similarly, the MM-Vet benchmark (Yu et al., 2023) has been designed to evaluate LMMs on complex multimodal tasks requiring integrated vision language capabilities. The benchmark covers six core VL abilities, including recognition, OCR, knowledge, spatial awareness, language generation, and arithmetic capability.

Despite these advancements, the development of comprehensive evaluation datasets remains crucial. A systematic review by (Wang et al., 2024d) highlighted the current State-of-the-art reasoning capabilities within MLLMs and noted that while various datasets have been proposed as multimodal reasoning benchmarks, many of which lack comprehensive reasoning steps. This underscores the importance of developing robust benchmarks that can accurately measure and drive progress in the multimodal reasoning capabilities of current models.

B Annotation Guidelines

We iteratively developed the annotation guidelines for PEARL over a period of six months, structured across two distinct phases. In *Phase I*, annotators focused on filtering *Wikipedia* articles based on two main criteria: cultural uniqueness and the relevance of images to their corresponding articles. Throughout this initial phase, annotators provided continuous feedback during weekly meetings, allowing us to promptly address and resolve any encountered issues. The annotation guidelines begin by outlining the primary goals of the project, followed by detailed descriptions of the ten domains, each supported by illustrative images and captions to ensure clarity. The second part of the guidelines explains each question type used in the project, providing clear definitions along with practical examples directly from the annotation platform to support annotators throughout the process.
Phase III involved reviewing automatically generated questions derived from the previously filtered data. Annotators were responsible for thoroughly validating these questions and ensuring the accuracy and appropriateness of question types and answers. Throughout this second phase, we further refined and enhanced the annotation guidelines based on annotators' experiences and feedback, ensuring consistency and effectiveness across all phases of the annotation process.

C Data Statistics

- We provide statistics on the number of images by cultural category for the Arab countries in PEARL in C.1.

D Additional Technical Details

Question Type	Count	Purpose
Cause and Effect	23,400	Ask about the reasons behind a cultural element and its consequences or impacts.
Chronological Sequence	22,614	Examine the historical development or timeline of a cultural element or practice.
Comparative Analysis	23,264	Compare or contrast cultural elements within or across contexts or regions.
General Q&A	35,034	Query a straightforward factual detail about the image or topic as stated in the article.
Hypothesis Formation	23,552	Pose a speculative why/how question about a cultural element, prompting an explanatory theory.
Modern Context	22,972	Connect a traditional cultural element to its relevance or adaptation in today's world.
Origin Identification	23,776	Inquire about the historical or geographical origin of a cultural element or practice.
Perspective Shifting	23,498	Explore different viewpoints or interpretations regarding the cultural element.
Problem Solving	23,764	Present a cultural challenge or issue and ask for a solution or mitigation approach.
Role Playing	23,184	Provide an answer from a specific role or persona related to the cultural context.
Scenario Completion	23,368	Present an incomplete scenario or sequence and ask to predict or complete the outcome.
Multiple Choice	30,654	Present several options for a question, requiring selection of the correct answer.
True/False	10,218	Present a statement and ask whether it is true or false, testing factual accuracy of cultural information.

Table B.1: Lists the 13 question types used in PEARL, with their counts in Pearl and intended purposes, each designed to elicit different forms of cultural reasoning.

Country	Architecture	Clothes	Fauna	Festivals & Celebrations	Flora	Food	Geography	Handicrafts	Landmarks	Music	Total
Algeria	0	76	0	0	0	0	0	1	32	143	252
Bahrain	0	0	1	0	0	16	143	1	0	17	178
Egypt	4,63	1	6	1	12	162	211	0	264	252	1,372
Iraq	170	3	14	0	2	62	74	0	0	85	410
Jordan	400	0	145	0	7	60	1,546	0	982	25	3,165
Kuwait	111	0	7	0	10	0	212	0	0	65	405
Lebanon	117	0	41	0	101	66	280	0	0	191	796
Libya	10	0	20	0	53	22	296	0	155	10	566
Mauritania	0	0	2	0	18	7	198	0	0	0	225
Morocco	223	105	52	1	203	235	310	133	123	256	1,641
Oman	141	12	55	0	36	4	331	0	0	5	584
Palestine	1,091	93	73	0	185	393	76	34	294	48	2,287
Qatar	0	0	13	0	0	1	144	0	0	4	162
Saudi Arabia	220	12	121	14	111	33	356	0	0	22	889
Sudan	0	0	1	0	0	16	339	0	113	0	469
Syria	877	11	8	0	134	61	253	2	0	7	1353
Tunisia	92	11	5	0	0	4	0	4	0	0	116
UAE	4	0	7	0	0	0	135	0	0	3	149
Yemen	385	2	0	0	42	63	551	0	0	23	1,066
Total	4,304	326	571	16	914	1,205	5,455	175	1,963	1,156	16,085

Table C.1: Distribution of the number of images by cultural category for each Arab country in PEARL (after revision).

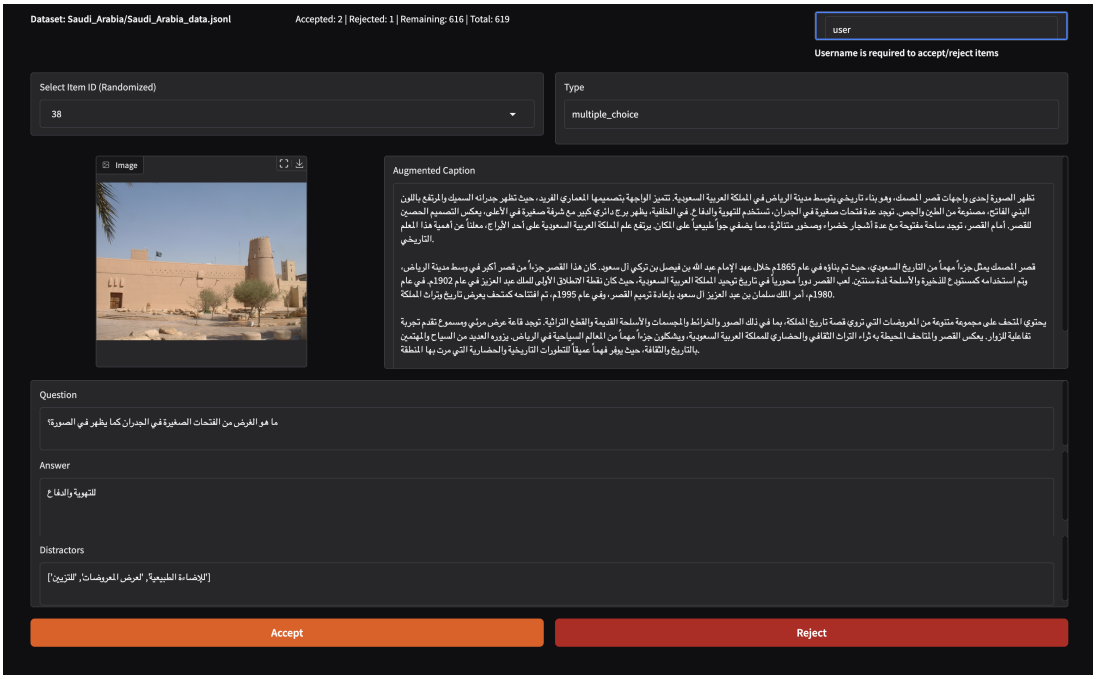

Figure B.1: User Interface used by annotators to revise 11K Q&A pairs for benchmarking. The revision process involved filtering samples based on cultural relevance and quality, enabling annotators to accept or reject entries accordingly.

Prompt

Step 1: Generate an Augmented Caption

Your first task is to create a detailed, extended description of the image based on the provided image caption and Wikipedia article. This augmented caption should:

1. START with تظهر الصورة (The image shows) followed by a comprehensive description of what is VISUALLY present in the image, using the provided image caption as a foundation.
2. Expand on the visual elements by adding relevant contextual information from the provided sources.
3. Focus particularly on details that would support {focus_area} question-answer pairs.
4. Describe the image's visual elements in detail including:
   - Specific objects, people, settings, activities, and artifacts visible
   - Spatial relationships between elements
   - Notable colors, textures, and visual features
   - Any text or inscriptions visible
   - Any actions or events being depicted

INFORMATION GUIDELINES:

1. Include information from the provided sources (category, title, country, image caption, Wikipedia article)
2. Present information as factual statements without attributing to the source
3. DO NOT use phrases like "according to the article," "as mentioned in the caption," or any reference to the sources
4. DO NOT mention Wikipedia, articles, captions, or sources in any way
5. Simply state the facts and information directly as established knowledge

Question Type-Specific Requirements for {question_type}:

{specific_requirements}

Input Information

- **Cultural Category:** {category}
- **Article Title:** {title}
- **Country:** {country}
- **Image Caption:** {image_caption}
- **Wikipedia Article:** {wikipedia_article}

REQUIRED STRUCTURE:

1. First paragraph: Begin with "تظهر الصورة" and describe what is VISUALLY present based on the image caption
2. Following paragraphs: Add relevant cultural context from the sources that directly relates to the visual elements
3. Final paragraph: Summarize elements specifically relevant to {question_type} questions

CRITICAL VERIFICATION STEP:

Before finalizing your augmented caption, verify that:

- The caption has sufficient detail to support {question_type} question-answer pairs
- You have NOT included any reference to articles, captions, or sources
- You have presented all information directly as factual statements

If the provided sources do not provide SUFFICIENT information to create a meaningful augmented caption for {question_type} question-answer pairs, return a JSON error message:

```
{"error": "Insufficient context in the provided sources to generate an augmented caption for {question_type} question-answer pairs."}
```

Figure D.1: Shows *prompt template step 1* used to generate augmented captions

Prompt (Step 2: Part 1)

Step 2: Generate Question-Answer Pairs

Using ONLY the augmented caption you created in Step 1, now generate {num_pairs} different question-answer pairs that {task_description}, without explicitly naming it.

All questions and answers MUST be written in Modern Standard Arabic only.

INFORMATION GUIDELINES:

1. Questions and answers MUST be based EXCLUSIVELY on information in the augmented caption
2. DO NOT introduce any new information not present in the augmented caption
3. DO NOT mention or reference any articles, captions, or sources
4. Present all information as direct factual statements without attribution
5. If the augmented caption lacks sufficient information for meaningful Q&A pairs, return an error

Question Requirements:

1. Your question MUST reference something clearly described in the augmented caption
2. Your question MUST use one of these exact phrases to refer to the element:
   - "الذي يظهر في الصورة" (that appears in the image)
   - "كما يظهر في الصورة" (as shown in the image)
   - "الظاهر في الصورة" (the visible element in the image)
3. NEVER mention the specific name of the element in the question
4. Each question should require both visual identification AND cultural knowledge
5. NEVER include any terms that could hint at the exact name of the object, tradition, landmark, or feature

Figure D.2: Illustrates template step 2 (Part 1) to generate Q&A pairs

Prompt (Step 2: Part 2)

Answer Requirements:

1. Base all answers EXCLUSIVELY on information in the augmented caption
2. NEVER add any details or context not in the augmented caption
3. Keep answers between 2-5 sentences in length
4. The answer MUST directly address and resolve the specific question being asked
5. Start your answer with a direct response to the question, then provide supporting details
6. DO explicitly name the object, tradition, or element in your answer
7. You SHOULD include the specific name of elements in the answer - unlike in the question
8. Use clear language and structured sentences that directly connect to the question
9. AVOID repeating the same information across multiple answers
10. Include the country/region name ONLY when it is relevant to the answer
11. NEVER mention or reference articles, captions, sources, Wikipedia, or any attribution phrases

Question Type-Specific Requirements for {question_type}:

{specific_requirements}

CRITICAL VERIFICATION STEP:

Before finalizing each Q&A pair, verify that:

- The answer contains information found ONLY in the augmented caption
- No new information has been introduced
- The question uses one of the required phrases
- The question doesn't name the specific element
- The answer DOES directly name the specific element
- The answer directly and clearly addresses the specific question being asked
- The answer follows the required length guidelines (2-5 sentences)
- NO reference is made to any sources such as articles or captions

If the augmented caption does not provide SUFFICIENT information to create meaningful question-answer pairs, return a JSON error message:

```
{"error": "Insufficient context in the augmented caption to generate meaningful question-answer pairs for this question type."}
```

Figure D.3: Displays prompt template step 2 (Part 2) used to generate answers

Prompt

Two-Step Process for Generating Question-Answer Pairs

{step1_prompt}

{step2_prompt}

{example}

Final Output Format Your final output must be a valid JSON object with the following structure:

```
{
 "augmented_caption": "The generated augmented caption based on the question-answer type.",
 "generated_QAs": [
{
 "question": "Your first question text in Modern Standard Arabic here",
 "answer": "Your first answer text in Modern Standard Arabic here"
 },
{
 "question": "Your second question text in Modern Standard Arabic here",
 "answer": "Your second answer text in Modern Standard Arabic here"
 }
 // Additional pairs as needed up to {num_pairs}...
 ]
 }
```

If there is insufficient context in the provided sources for either step, your output should be a JSON object with an error message:

```
{"error": "Specific error message explaining the lack of sufficient context."}
```

FINAL REMINDER - CRITICAL INSTRUCTIONS:

- The augmented caption MUST begin with "تظهر الصورة" followed by a detailed description of what is VISUALLY present
- ALL information must be based on the provided sources
- Present ALL information as direct factual statements
- DO NOT use phrases like "according to the article," "as mentioned," or similar attributions
- DO NOT reference Wikipedia, articles, captions, or any sources in the augmented caption, questions, or answers
- Questions should NOT mention the specific name of elements, but answers SHOULD explicitly name them
- Answers MUST directly respond to their corresponding questions and be between 2-5 sentences in length
- Each answer should begin with a sentence that directly addresses the question asked
- If you cannot generate meaningful content without inventing information, return an error message

Figure D.4: Combined prompt template including the two steps for generating Q&A pairs

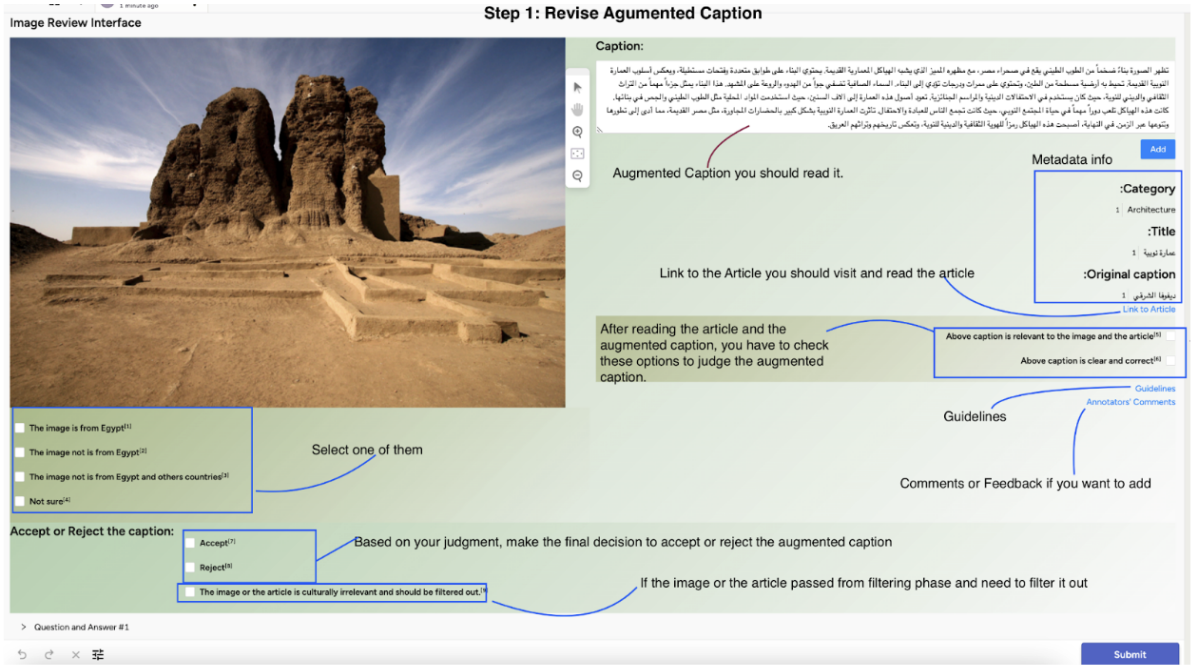

Figure D.5: Label Studio platform interface used by annotators in step one of the human revision phase to revise the augmented caption

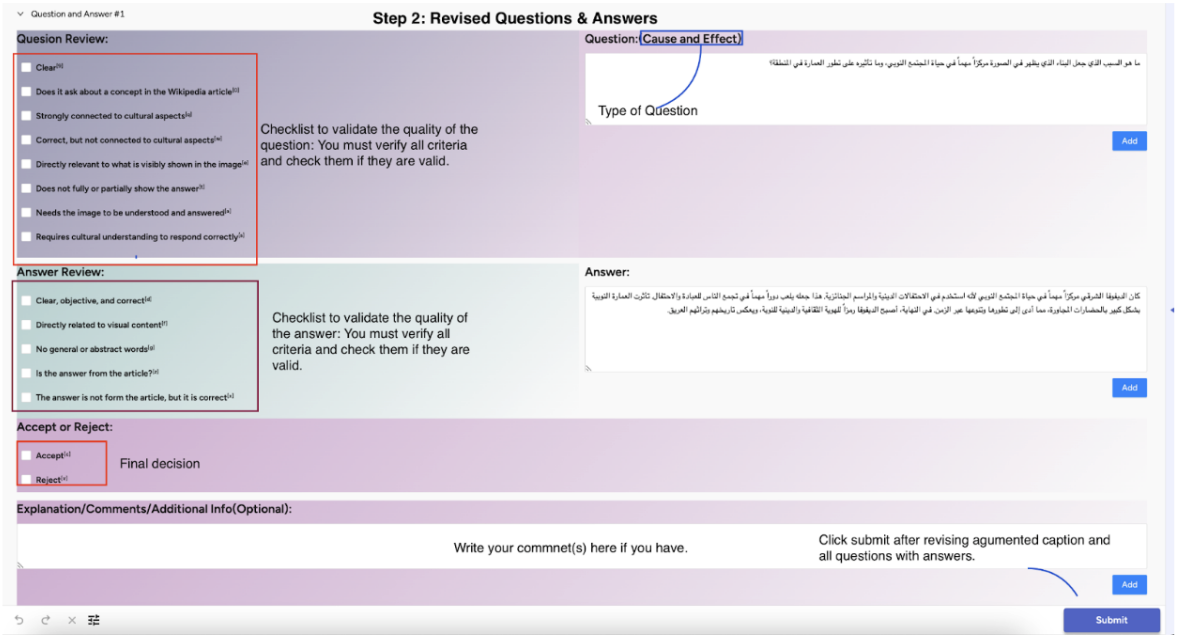

Figure D.6: Step two of the revision phase, showing the process of reviewing and editing questions and answers on the platform

Augmented Caption

تظهر الصورة حي البستكية التاريخي في إمارة دبي، حيث تظهر مباني تقليدية ذات تصميم معماري فريد مع أبراج الرياح التي تتميز بها العمارة التقليدية في المنطقة. في المقدمة، يظهر جمل مرتدياً زينة تقليدية يجلس على الرمال، مما يعكس التراث العربي القديم. في الخلفية، توجد خيمة تقليدية مزينة بالأعلام الوطنية، ورجلان يرتديان الزي التقليدي الإماراتي يقفان بالقرب من المباني. تحيط بالمنطقة أشجار نخيل ومباني حديثة، مما يبرز التوازن بين التراث والحداثة في دبي. ....

Modern Context Question	كيف يمكن أن يعكس الجمل الذي يظهر في الصورة التوازن بين التراث والحداثة في دبي؟
Answer	الجمل الذي يظهر في الصورة يعكس التوازن بين التراث والحداثة في دبي من خلال تمثيله للتراث العربي القديم في بيئة حضرية حديثة. يرمز الجمل إلى التراث الثقافي والحياة البدوية القديمة، بينما يظهر في خلفية حديثة تجمع بين المباني التقليدية والحديثة. هذا التوازن يعكس الهوية الثقافية الفريدة لدبي ودورها في الحفاظ على التراث العربي وتعزيزه في العصر الحديث.
General Q&A Question	ما هو الغرض من المبنى الذي يظهر في الصورة، والذي يتميز بتصميمه المعماري الفريد؟
Answer	المبنى هو برج هوائي تقليدي يستخدم لتبريد الهواء في المنازل والمباني. هذا النوع من التصميم كان شائعاً في المناطق الحارة لتحسين تدفق الهواء الطبيعي وخفض درجات الحرارة. يعكس هذا التصميم التراث المعماري الغني للإمارات العربية المتحدة.
General Q&A Question	ما هي الأهمية الثقافية للحي الذي يظهر في الصورة؟
Answer	الحي هو حي البستكية التاريخي في دبي، وهو يمثل التراث الثقافي والمعماري للإمارة. يحتفظ الحي بمبانيه التقليدية وتصميمه المعماري الفريد، مما يعكس تاريخ دبي وتطورها. يعد الحي معلماً سياحياً مهماً يجذب الزوار للاطلاع على التراث المحلي.
Problem Solving Question	ما هي الطرق الممكنة لتعزيز دور الخيمة التقليدية كما يظهر في الصورة في تعزيز الثقافة المحلية؟
Answer	يمكن تعزيز دور الخيمة التقليدية في حي البستكية من خلال تنظيم الفعاليات الثقافية والفنية التي تهدف إلى إبراز التراث المحلي. يجب تشجيع السكان والزوار على المشاركة في هذه الفعاليات لتعزيز التواصل الثقافي. كما يمكن استخدام الخيمة كمركز للحرف اليدوية التقليدية لدعم الحرفين المحليين ونقل مهاراتهم إلى الأجيال القادمة.

Figure D.7: Shows an example illustrates the format of the augmented caption followed by multiple Q&A pairs. Each question is linked to a specific reasoning type and is designed to reflect cultural understanding grounded in the image and caption.

Image	Question Type		Text	Translation
	Hypothesis formation	Caption	تظهر الصورة برج الرياح التقليدي في البحرين، وهو عنصر معماري تراثي يستخدم لتبريد المنازل قدمًا قبل ظهور التكييف الكهربائي	The picture shows a traditional wind tower in Bahrain—an architectural heritage feature that was once used to cool homes before the advent of electric air-conditioning.
		Question	لماذا تعتقد أن المعماريين في البحرين والخليج العربي اخترعوا البرج الذي يظهر في الصورة؟	Why do you think architects in Bahrain and the Gulf region invented the tower shown in the picture?
	Hypothesis formation	Caption	تظهر الصورة زوجا من حيوانات الفنك، يتميزان بفرائهما الرملي الكثيف وآذانهما الكبيرة البارزة، وهما مستلقيان	The image depicts a pair of fennec foxes, distinguished by their thick sandy fur and large, prominent ears, reclining side by side.
		Question	لماذا تعتقد أن الفنك أصبح رمزا ثقافيا مهما في الجزائر، ويستخدم في الرياضة؟	Why do you think the fennec fox has become an important cultural symbol in Algeria and is even used in sports?
	Role Playing	Caption	تظهر الصورة واحدة من أشهر وأندر الأشجار في العالم ، وهي شجرة دم الأخوين، التي تنمو في جزيرة سقطرى اليمنية.	The picture shows one of the world's most famous and rarest trees: the Dragon's Blood Tree, which grows on Yemen's Socotra Island.
		Question	تخيل أنك دليل سياحي في جزيرة سقطرى، كيف تشرح للزوار أهمية شجرة دم الأخوين؟	Imagine you are a tour guide on Socotra—how would you explain to visitors the significance of the Dragon's Blood Tree?
	Problem Solving	Caption	تظهر الصورة حي البستكية التاريخي في إمارة دبي، حيث تظهر مباني تقليدية ذات تصميم معماري فريد مع أبراج الرياح ...	The image shows Dubai's historic Al-Bastakiya quarter, where traditional buildings with a distinctive architectural style and wind towers can be seen. . .
		Question	كيف يمكن الحفاظ على المباني التاريخية التي تظهر في الصورة مع التطور الحضري السريع في المنطقة؟	How can the historic buildings shown in the picture be preserved amid the region's rapid urban development?
	Modern Context	Caption	تظهر الصورة رجل تونسي يرتدي لباسا تقليديا يعرف بالكدرون، وهو قطعة ملابس مصنوعة	The photo shows a Tunisian man wearing a traditional garment known as the qadrūn, a piece of clothing made of. . .
		Question	كيف يمكن دمج الزي التقليدي الذي يظهر في الصورة في الحياة اليومية الحديثة في تونس؟	How can the traditional attire shown in the picture be integrated into modern everyday life in Tunisia?
	Original Identification	Caption	تظهر الصورة حكواتيا معاصرا في أحد المقاهي الدمشقية التقليدية وهو يرتدي الطربوش واللباس الشعبي ويجلس	The picture shows a contemporary storyteller in a traditional Damascene cafe, wearing a fez and folk dress, seated . . .
		Question	ما هو الأصل التاريخي أو الجغرافي لشخصية الحكواتي الذي يظهر في الصورة؟	What is the historical or geographical origin of the hakawati figure shown in the picture?
	Scenario Completion	Caption	تظهر الصورة أحد المشاركين في رقصات الطرق الصوفية في السودان، وهو يرتدي زيا تقليديا مميزا غنيا بالألوان والرموز.....	The image depicts a participant in Sudanese Sufi order dances, dressed in a distinctive traditional costume rich in colours and symbols. . .
		Question	بدأ الرجل في الصورة يرقص مع جماعته في احتفال صوفي. اذا توقف توقف عن الرقص في الحفل برأيك، ماذا سيحدث بعد ذلك؟	"The man in the picture began dancing with his group at a Sufi celebration. If he were to stop dancing during the ceremony, what do you think would happen next?"
	General QA	Caption	تظهر الصورة المسجد النبوي في المدينة المنورة، وهو معلم ديني وثقافي بارز في المملكة العربية السعودية. يظهر ف.....	The picture shows the Prophet's Mosque in Medina, a major religious and cultural landmark in the Kingdom of Saudi Arabia. It features . . .
		Question	ما هي الميزة المعمارية البارزة الذي يظهر في الصورة والتي تغطي قبر النبي محمد صلى الله عليه وسلم ؟	What is the prominent architectural feature shown in the picture that covers the grave of the Prophet Muhammad (peace be upon him)?
	Modern Context	Caption	تظهر الصورة مجموعة من الدمى ترتدي الزي التقليدي للمرأة القطرية، وقد عرضت داخل أحد أروقة متحف الشيخ فيصل بن	The picture displays a collection of dolls wearing traditional Qatari women's attire, exhibited in one of the galleries of the Sheikh Faisal bin . . . Museum.
		Question	كيف يمكن استخدام الأزياء التقليدية القطرية المعروضة في الصورة في عالم اليوم؟	How can the traditional Qatari costumes displayed in the picture be used in today's world?
	Prespective Shifting	Caption	تظهر الصورة أطفال فلسطينيين يرتدون الزي الفلسطيني التقليدي.	The picture shows Palestinian children dressed in traditional Palestinian clothing.
		Question	ناقش كيف يختلف فهم الزي الفلسطيني التقليدي الذي في الصورة ودلالته عند النظر إليه من وجهة نظر فلسطيني يعيش في المهجر؟	Discuss how the understanding and symbolism of the traditional Palestinian attire in the picture differ when viewed from the perspective of a Palestinian living in the diaspora.
	Chronological Sequence	Caption	تظهر الصورة الخنجر العماني التقليدي، والمعروف أيضا باسم الجنبية في بعض المناطق. يتميز الخنجر ...	The picture depicts the traditional Omani dagger, also known as the janbiyyah in some regions. The dagger is distinguished by . . .
		Question	كيف تطور استخدام العنصر الذي يظهر في الصورة عبر الزمن؟	How has the use of the item shown in the picture evolved over time?
	Comparative analysis	Caption	تظهر الصورة مجموعة متنوعة من الأحذية التقليدية المغربية الملونة والمزخرفة، مرتبة بشكل منظم	The image shows a neatly arranged selection of colourful, ornamented traditional Moroccan shoes . . .
		Question	كيف يختلف الحذاء التقليدي الذي يظهر في الصورة عن الأحذية الحديثة من حيث التصميم والوظيفة؟	How does the traditional shoe shown in the picture differ from modern footwear in terms of design and function?
	Original Identification	Caption	تظهر الصورة أطلال معبد روماني ضخم في مدينة بعلبك اللبنانية، وهو أحد أبرز المعالم الأثرية في لبنان والشرق الأوسط.	The picture presents the ruins of a massive Roman temple in Baalbek, Lebanon—one of the most prominent archaeological landmarks in Lebanon and the Middle East.
		Question	ما هو الأصل التاريخي للعنصر المعماري الموجود في الصورة؟	What is the historical origin of the architectural element shown in the picture?
	Original Identification	Caption	تظهر الصورة طبقا شهيا من المجبوس أو الكبسة الخليجية، وهو أحد أشهر ال	The picture shows a mouth-watering dish of Gulf-style majboos (also called kabsa), one of the region's most celebrated rice dishes.
		Question	كيف يختلف طبق الذي في الصورة عن أطباق الأرز التقليدية في دول عربية أخرى مثل المنسف الأردني؟	How does the dish in the picture differ from other traditional rice dishes in Arab countries, such as Jordanian mansaf?
	Cause and Effect	Caption	تظهر الصورة مجموعة من العازفين العراقيين من فرقة فنون شعبية، يرتدون الزي التقليدي، ويؤدون عرضا موسيقيا .	The image features a group of Iraqi musicians from a folk-arts ensemble, dressed in traditional attire and giving a musical performance . . .
		Question	لماذا يستخدم المزمار في الموسيقى الشعبية العراقية؟	Why is the mizmar used in Iraqi folk music?
	Original Identification	Caption	الصورة تظهر مجموعة من فوانيس رمضان التقليدية المعلقة على أغصان شجرة، بألوان زاهية وزخارف مبهجة.....	The photo shows a collection of traditional Ramadan lanterns hanging from tree branches, glowing in vivid colours and cheerful designs . . .
		Question	ما هو الأصل التاريخي أو الجغرافي للعنصر الثقافي الموجود في الصورة؟	What is the historical or geographical origin of the cultural element shown in the picture?

Table D.1: Provides English version of the main figure, including the image, question types, and translations of the original Arabic texts

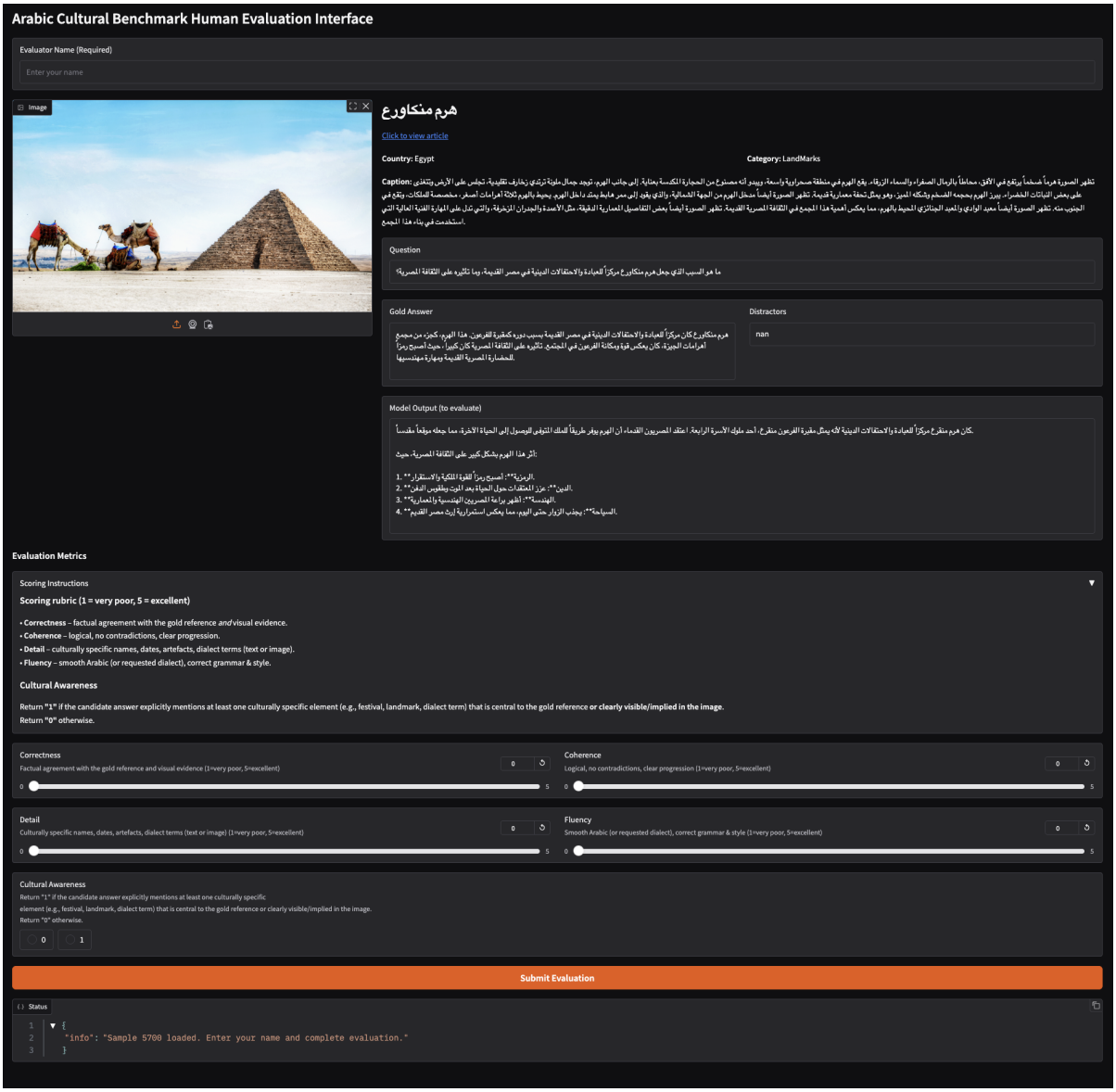

Figure D.8: Displays User interface employed by annotators during the human evaluation phase, where they rate the models' output based on Correctness, Coherence, Detail, and Fluency

Prompt

We are building Arabic culture multimodal datasets. In some cases, a single cultural concept is shared across multiple Arabic-speaking countries but manifests in distinct ways—visually, in preparation, or in usage. For instance, Kabsa is a traditional dish enjoyed in places like Saudi Arabia, Yemen, and Qatar, yet each locale has its own way of preparing and seasoning it. Likewise, "Agal" is a type of traditional head accessory worn across various Arab regions, though its style or method of wearing can differ from one country to another. These nuanced variations across shared concepts highlight the rich cultural diversity that exists within the broader Arabic-speaking world. We called this "Shared Concepts". We introduce Pearl-Shared Concepts as a specialized benchmark that evaluates the LVLM's ability to generalize and reason consistently about cultural concepts shared across multiple Arabic-speaking regions. Initially, we identified culturally common concepts and curated diverse multimodal examples (images and texts) reflecting variations across regions. We then developed template-based multiple-choice and true/false questions to systematically assess how robustly LVLMs generalize shared concepts despite variations in visual representation or cultural context. I will provide you with some images and templates of questions (MCQ and true and false), and I want you to generate questions based on the templates. Be creative and don't stick to these templates. Some of the templates require one image question and the others require multiple images. I will provide you with the name of the shared concept. The images are numbered in the format of number_name of country.extension. Also, I will provide you with a shared concept name. I want you to generate 2-3 single-image questions, 2-3 multiple-image questions.

The output format should be in the following json format:

```
{ "ID": 1,
"Question": Question? (In Arabic),
"Choices": {
"A": Choice 1 (In Arabic),
"B": Choice 2 (In Arabic),
"C": Choice 3 (In Arabic),
"D": Choice 3 (In Arabic)
},
"Answer": "A",
"Single_Multiple": "Multiple",
"Selected_images": [1, 2, 3, 4]
}
```

The concept name: Concept Name in Arabic: {concept_name}

```
{Template-Based Questions (In Arabic)
What is another name for [element X] in this country?
What is a key ingredient that distinguishes the preparation of [dish X] here?
On what occasion is the depicted [element X] worn or presented?
True or False: [Dish X] is usually served with Arabic bread in this country.
True or False: The visible color on [element X] indicates its Qatari origin.
Arrange the images according to the level of spiciness from mildest to spiciest.
Which of these images represents the oldest historical form of [element X]?
True or False: All the images show [element X] being worn in a similar way."}
```

Figure D.9: Illustrates the prompt used to generate Q&A pairs for PEARL-X using GPT-o3. in the prompt, we include few-shot examples based on templates for both single-image and multi-image question types.

Prompt

multiple_choice's promprt

```
You are an impartial evaluator.

TASK: Decide if the candidate's choice is correct **ignoring surface form** (letter, synonym,
capitalisation). Return ONLY "1" for correct, "0" for incorrect.

Question with options:
{question}
Gold correct answer:
{ground_truth}
Candidate's chosen answer:
{predicted_answer}

Reply with 1 or 0 - nothing else.
```

true_false's promprt

```
You are an impartial evaluator.

TASK: Compare the candidate's short answer with the gold answer. If they express the **same
fact** (allowing synonyms, paraphrase, spelling variants) return "1". Otherwise return "0".
Give no explanation.
Statement:
{question}
Gold label (True/False):
{ground_truth}
Candidate label:
{predicted_answer}

Reply with 1 or 0 - nothing else.
```

open-ended's promprt

```
You are an **impartial multimodal evaluator** for Arabic cultural benchmark.
Your task is to grade the candidate's Arabic answer to a culture-focused question, using
**both** the reference text and the attached image.
***Output ONE JSON object only*** in this schema:
{{ "correctness": <0-5>, "coherence": <0-5>, "detail": <0-5>, "fluency": <0-5> }}
### Scoring rubric (0= very poor,5= excellent)
• **Correctness 40% - factual agreement with the gold reference *and* visual evidence.
• **Coherence 20%** - logical, no contradictions, clear progression.
• **Detail 20%** - culturally specific names, dates, artefacts, dialect terms (text or image).
• **Fluency 20%** - smooth Arabic (or requested dialect), correct grammar & style.

Image description:
{image_description}
Question: {question}
Gold reference answer:
{ground_truth}
Candidate answer:
{predicted_answer}

Respond with the JSON object only— **no additional text**
```

Figure D.10: Evaluation prompts used for automatic judgment across different question formats in the Pearl benchmark. The `multiple_choice` prompt evaluates whether the predicted choice matches the correct answer, allowing for variation in surface form (e.g., casing or synonyms). The `true_false` prompt compares semantic equivalence between gold and predicted binary labels. The `short_answer` prompt assesses alignment between candidate and gold judgments on factual statements. All prompts instruct the model to return strictly binary outcomes (1 or 0) with no explanation. Detailed evaluation prompts are available on the project GitHub repository: https://github.com/UBC-NLP/pearl.

Architecture	Clothes	Fauna	Flora	Food	Geography	Handicrafts	Landmarks	Music
Courtyard Houses	Keffiyeh & Agal	Camel	Date Palm	Couscous	Nile River Valley	Al-Sadu Weaving	Historic Citadels and Forts	Oud
Souk Bazaars	Thawb/Dishdasha	Falconry	Olive Tree	Kabsa/Machbūs	Mediterranean Coast	Carpet and Rug Weaving	Grand Mosques of Early Islam	Dabke
Wind Catchers	Abaya	Saluki Dogs	Cedar Tree	Mandi	Atlas Mountains	Embroidery (Tatreez)	Historic City "Old Towns" (Medinas)	Mizmar
Mudbrick Architecture	Jellabiya	Arabian Oryx	Argan Tree	Falafel	Red Sea Coast	Khanjar	Caravanserais (Khans)	Sword Dance (Al-'Ardha)
Historic Citadels and Forts	Kaftan	Fennec Fox	Jasmine	Hummus	Desert Oases			
Grand Mosques of Early Islam	Keffiyeh			Ful Medames				
Historic City "Old Towns" (Medinas)	Djellaba			Stuffed Vegetables (Mahshi/Dolma)				
Caravanserais (Khans)	Bisht			Shawarma				
	Fez (Tarboosh)			Arabic Coffee				
				Harissa				
				Mulukhiyah				
				Shakshouka				
				Asida				
				Qatayef				

Table D.2: Shows *PearlX*'s comprehensive list of 61 culturally diverse concepts spanning various categories across Arab countries, which were used to generate Q&A pairs for benchmarking

Type	Template	Example	Q-Type
Multiple Image Templates	طابق كل صورة بالدولة الصحيحة استناداً إلى اختلاف نمط [العنصر].	طابق كل صورة بالعقال القطري، السعودي، العراقي. (القطري، السعودي، العراقي. القطري، السعودي، العراقي القطري، السعودي، العراقي.	Choose
	صنّف الصور إلى مجموعتين بحسب نوع البهارات المستخدمة في [الطبق].	صنّف صور الكبسة إلى مجموعة تستخدم الهيل وأخرى لا تستخدمه.	Categorieze
	اختر الصورة التي لا تنتمي إلى نفس المفهوم المشترك.	اختر الصورة التي لا تمثل طبق كبسة بين الصور الأربع.	Choose
	رتّب الصور حسب درجة حدة التوابل من الأخف إلى الأشد.	رتّب صور أطباق الكبسة السعودية، العمانية، واليمنية حسب حدة التوابل.	Reorder
	أي من هذه الصور تُمثّل أقدم شكل تاريخي لـ[العنصر]؟	أي من هذه الصور تمثّل الشكل التاريخي الأقدم للعقال؟	Idenify
	صحيح أم خطأ: كل الصور تُظهر [العنصر] يُرتدى بطريقة متشابهة.	صحيح أم خطأ: كل الصور تُظهر العقال يُرتدى بالطريقة نفسها.	T/F
	اختر الصورة التي توضّح خطوة الطهي الأولى في إعداد [الطبق].	اختر الصورة التي توضّح أول خطوة في طهي الكبسة بين الصور الثلاث.	MCQ
	أي من هذه الصور تُمثل نسخة حديثة مطوَّرة من [العنصر]؟	أي من هذه الصور تمثل نسخه حديثة مطورة للعقال باستخدام خامات صناعية؟	Choose
	طابق كل صورة بالدولة الصحيحة استناداً إلى اختلاف نمط [العنصر].	طابق صور العقال مع الدول: أ) قطر ب) السعودية ج) العراق	MCQ
Single Image Templates	ما الاسم الآخر الذي يُطلق على [العنصر] في دولة؟	ما الاسم الآخر الذي يُطلق على كبسة في اليمن؟	T/F
	اذكر المكوّن الرئيس الذي يميز [الطبق] في هذه الدولة عن نسخته في دولةٍ أخرى.	اذكر المكوّن الرئيس الذي يميز كبسة سلطنة عُمان عن كبسة السعودية.	T/F
	في أي مناسبة أو فصل عادةً ما يُستخدم [العنصر] هنا؟	في أي مناسبة عادةً ما يُرتدى هذا العقال في العراق؟	T/F
	ما اللون التي تظهر في [العنصر] وتدلّ على أصلِه من هذه الدولة؟	ما الزخرفة التي تظهر في هذا العقال وتدلّ على أصله القطري؟	T/F
	صحيح أم خطأ: يُقدَّم [الطبق] عادةً مع الخبز العربي في هذه الدولة.	صحيح أم خطأ: يُقدَّم الفول المدمس مع الخبز العربي في السودان.	T/F
	ما الاسم الآخر الذي يُطلق على [العنصر] في هذه الدولة؟	ما الاسم الآخر الذي يُطلق على كبسة في اليمن؟ أ) المندي ب) السلتة ج) الزربيان د) الصالونة	MCQ
	أي مكوّن رئيس يميّز إعداد [الطبق] هنا؟	أي مكوّن رئيس يميّز كبسة سلطنة عُمان؟ أ) الحلبة ب) الهيل ج) اللومي د) الكركم	MCQ
	في أي مناسبة يُرتدى [العنصر] الموضَّح؟	في أي مناسبة يُرتدى هذا العقال العراقي؟ أ) الأعراس ب) العمل اليومي ج) الصلاة د) الحج	MCQ
	أي حقبة تاريخية ارتبطت بظهور [العنصر] في هذه الدولة؟	أي حقبة تاريخية ارتبطت بظهور المجبوس في الكويت؟ أ) العباسيون ب) العثمانيون ج) الأندلسيون د) الفاطميون	MCQ
	ما أداة الطهي التقليدية المستخدمة لإعداد [الطبق] هنا؟	ما أداة الطهي التقليدية المستخدمة لإعداد الكبسة السعودية؟ أ) القدر الضغاط ب) التنور ج) القدر المقلوب د) المضغوط	MCQ

Table D.3: Examples of question templates for both multiple and single image prompts in Arabic that we used to generate True/False and MCQ questions using `ChatGPT-o3`

E Fine-Grained Performance Analysis on PEARL-LITE by Country and Question Type

In this section, we provide a more granular analysis of model performance on the PEARL-LITE benchmark. Figure E.1 presents a detailed breakdown of accuracy scores on closed-form questions, categorized by both model and country. The heatmap illustrates that while larger proprietary models like `Gemini 2.5 Pro` and the `o3` models consistently achieve high accuracy across nearly all countries, the performance of open-source models varies significantly depending on the geographic context. For instance, several models show lower performance on questions related to Lebanon and Jordan, indicating potential gaps in their regional cultural knowledge for fact-based retrieval tasks.

For the more challenging open-ended questions, we offer several detailed views. Figures E.2 and E.3 display the *Overall score* and *CAS score*, respectively, broken down by model and country. These results underscore the difficulty of generating culturally nuanced text. The proprietary models again lead in performance, but even they show variability, particularly in the CAS metric where scores for countries like Mauritania and Qatar are notably lower for many models. This highlights that a model's ability to be culturally aware is not uniform and can be highly dependent on the specific regional culture being evaluated.

To further dissect the reasoning capabilities of each model, Figure E.4 provides a performance breakdown of the Overall Score by the 11 different open-ended question types. This analysis reveals specific strengths and weaknesses in model reasoning. Complex tasks such as *Chronological Sequence* and *Origin identification* prove to be challenging for smaller models, which often score poorly. In contrast, more advanced models demonstrate stronger and more consistent performance across these sophisticated reasoning categories, emphasizing our finding that reasoning-centric alignment is crucial for achieving deep cultural comprehension.

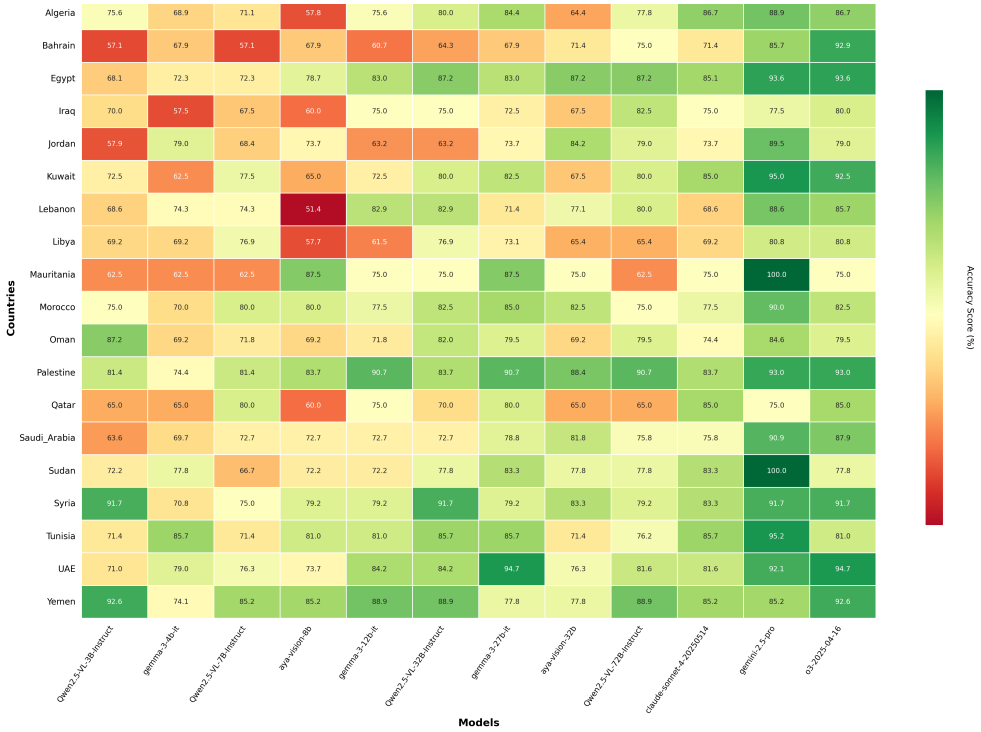

Figure E.1: Heatmap of accuracy scores (%) on closed-form questions, broken down by model and country.

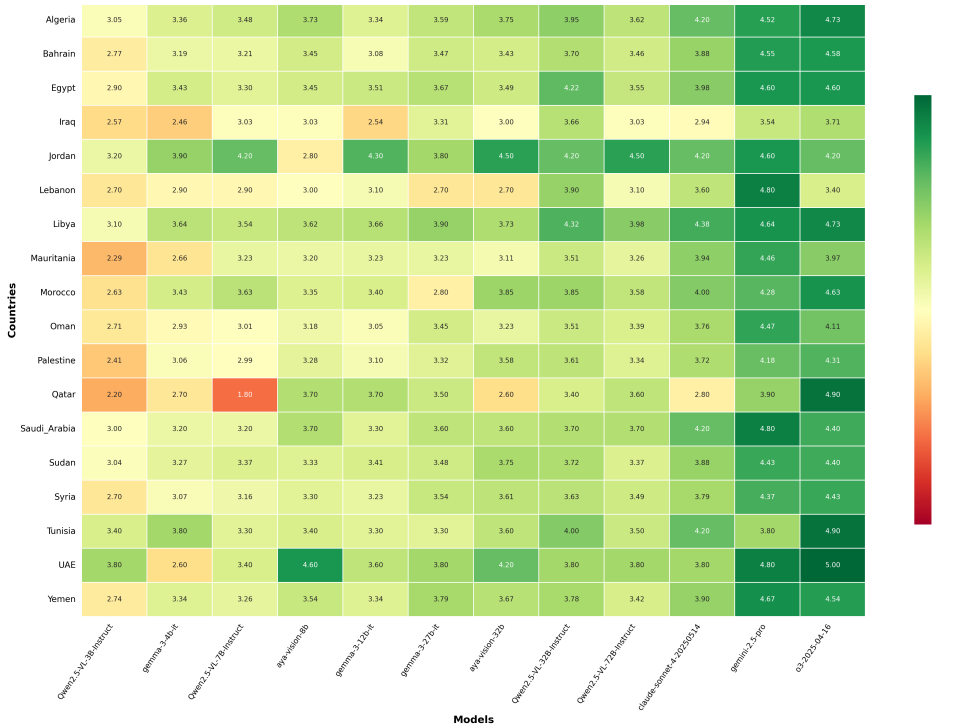

Figure E.2: Heatmap of the Overall Score (1-5) for open-ended questions, analyzed by model and country

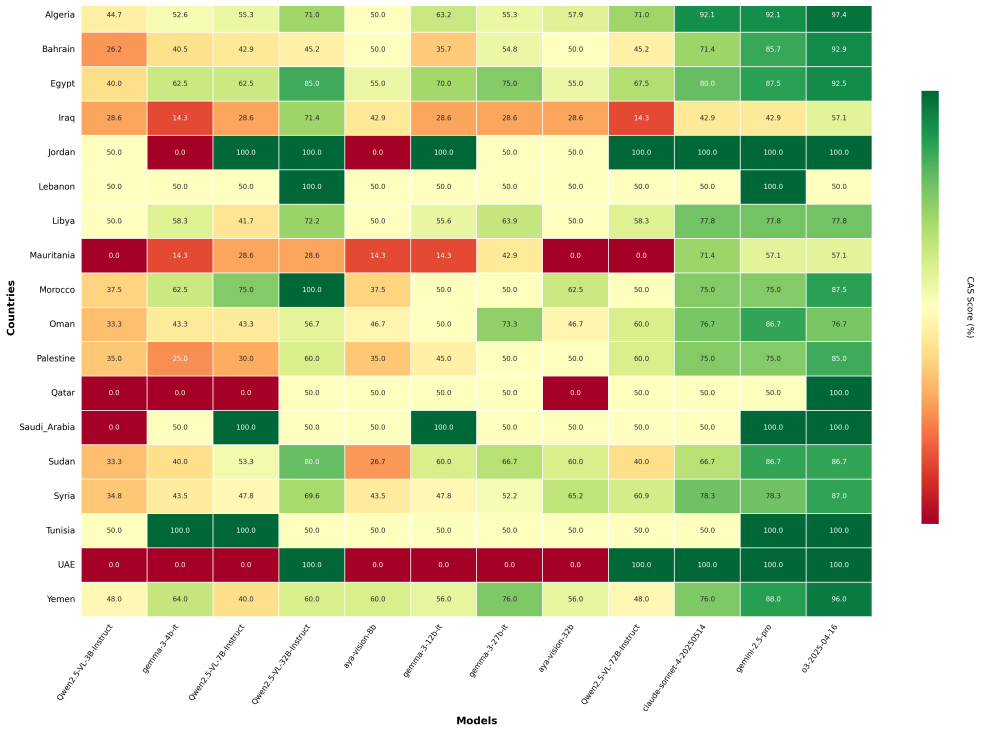

Figure E.3: Heatmap of the Cultural Awareness Score (CAS, in %) for open-ended questions, by model and country

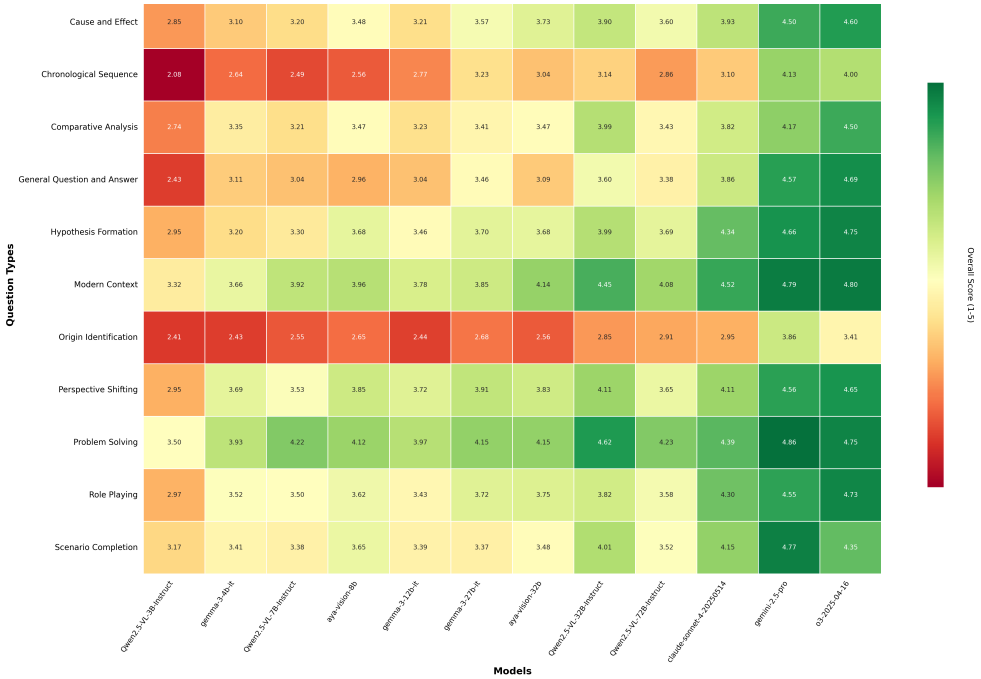

Figure E.4: Heatmap of the Overall Score (1-5) for open-ended questions, broken down by model and question type