# Commonsense Reasoning in Arab Culture

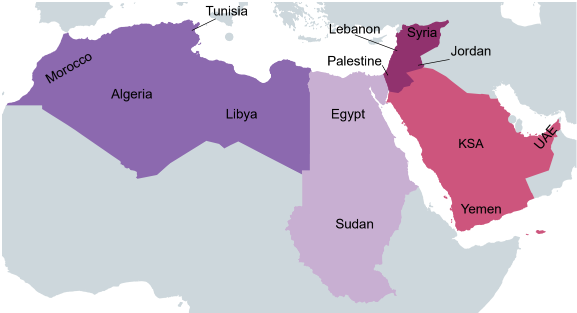
**Abdelrahman Sadallah**[1] **Junior Cedric Tonga**[1] **Khalid Almubarak**[2,5]
**Saeed Almheiri**[1] **Farah Atif**[1] **Chatrine Qwaider**[1] **Karima Kadaoui**[1]
**Sara Shatnawi**[3] **Yaser Alesh**[4] **Fajri Koto**[1]
[1]Mohamed bin Zayed University of Artificial Intelligence
[2]SDAIA [3]Al-Balqa Applied University
[4]Khalifa University [5]HUMAIN, Data and AI Models
{abdelrahman.sadallah,fajri.koto}@mbzuai.ac.ae

## Abstract

Despite progress in Arabic large language models, such as Jais and AceGPT, their evaluation on commonsense reasoning has largely relied on machine-translated datasets, which lack cultural depth and may introduce Anglocentric biases. Commonsense reasoning is shaped by geographical and cultural contexts, and existing English datasets fail to capture the diversity of the Arab world. To address this, we introduce ArabCulture, a commonsense reasoning dataset in Modern Standard Arabic (MSA), covering cultures of 13 countries across the Gulf, Levant, North Africa, and the Nile Valley. The dataset was built from scratch by engaging native speakers to write and validate culturally relevant questions for their respective countries. ArabCulture spans 12 daily life domains with 54 fine-grained subtopics, reflecting various aspects of social norms, traditions, and everyday experiences. Zero-shot evaluations show that open-weight language models with up to 32B parameters struggle to comprehend diverse Arab cultures, with performance varying across regions. These findings highlight the need for more culturally aware models and datasets tailored to the Arabic-speaking world.[1]

## 1 Introduction

Commonsense reasoning is the ability to make judgments and inferences based on everyday human knowledge and experiences (Sap et al., 2020). It is a fundamental aspect of human cognition and has been extensively studied in the context of large language models (LLMs) (OpenAI et al., 2024b; Grattafiori et al., 2024; Liu et al., 2023). However, commonsense reasoning is not universal—it is shaped by culture, which encompasses the shared knowledge, values, customs, and behaviors that define a society (Macionis, 2012; Giddens and Sutton, 2014).



| | |
|---|---|
| Location: **Algeria**, North Africa<br>Karim eats a dish of couscous for lunch<br>A. Karim eats couscous with kesra<br>**B. Karim eats couscous with meat sauce**<br>C. Karim eats couscous with soup | الموقع: **الجزائر**, شمال أفريقيا<br>يتناول كريم على الغداء طبق الكسكسي<br>أ.يتناول كريم الكسكسي مع الكسرة<br>**ب. يتناول كريم الكسكسي مع صلصة باللحم**<br>ج. يتناول كريم الكسكسي مع الشربة |
| Location: **Egypt**, Nile Valley<br>Asmaa prepares koshari for lunch<br>**A. Hot sauce and tomato sauce are added to the koshari**<br>B. Taro is added to the koshari<br>C. Peas are added to the koshari | الموقع: **مصر**, وادي النيل<br>تعد أسماء الكشري للغذاء<br>**أ.تضاف الشطة والصلصة إلي الكشري**<br>ب. يتم إضافة القلقاس إلي الكشري<br>ج. يتم إضافة البازلاء إلي الكشري |
| Location: **Syria**, the Levant<br>Mohammad eats fried kibbeh for lunch<br>A. The kibbeh is stuffed with a cheese filling<br>**B. The kibbeh is stuffed with a filling of meat and onions**<br>C. The kibbeh is stuffed with a variety of vegetables | الموقع: **سوريا**, الشام<br>ياكل محمد الكبة المقلية على الغداء<br>أ.يتم ملئ الكبة بحشوة من الأجبان<br>**ب. يتم ملئ الكبة بحشوة من اللحم و البصل**<br>ج. يتم ملئ الكبة بحشوة من الخضار المتنوعة |
| Location: **KSA**, the Gulf<br>Jareesh is one of the meals served in winter<br>A. Jareesh is served with fried garlic<br>B. Jareesh is served with bread<br>**C. Jareesh is served with fried onions** | الموقع: **السعودية**, الخليج<br>الجريش من الوجبات التي تقدم في فصل الشتاء<br>أ. الجريش يقدم مع الثوم المقلي<br>ب. الجريش يقدم مع الخبز<br>**ج. الجريش يقدم مع البصل المقلي** |

Figure 1: ArabCulture covers four regions across the Middle East and North Africa, spanning 13 countries. The highlighted areas on the map represent the regions included in ArabCulture. We present example questions for the *Lunch* category from Algeria, Egypt, Syria, and KSA, each with three answer choices, with the correct answer marked in bold. English translations are provided for illustration.

The Arab world, home to approximately 456 million people (Diab et al., 2017; Shoufan and Alameri, 2015), is characterized by its linguistic unity through Modern Standard Arabic (MSA) while encompassing diverse traditions, religions, and customs (Mirkin, 2010). This cultural diver-

[1] ArabCulture can be accessed at https://huggingface.co/datasets/MBZUAI/ArabCulture

*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7695–7710
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

sity influences not only social interactions but also reasoning patterns, making it crucial to develop evaluation benchmarks that reflect these variations. However, most existing commonsense reasoning datasets are developed with Western-centric assumptions, limiting their applicability to Arabic-speaking societies.

Despite recent advancements in Arabic LLMs (Sengupta et al., 2023; Huang et al., 2024; Team, 2024; Bari et al., 2024b), their evaluation has largely relied on machine-translated datasets originally created in English. Many commonsense reasoning benchmarks (Bisk et al., 2019; Sap et al., 2019a) fail to capture Arab cultural perspectives, as simple translations do not account for region-specific knowledge, potentially introducing bias. Given the significant cultural variations across the Arab world, these datasets do not provide a holistic measure of Arabic LLMs' ability to reason within culturally specific contexts. This raises an important question: *To what extent can existing LLMs accurately reason about commonsense knowledge in diverse cultural settings, particularly in the Arab world?*

To address this gap, we introduce ArabCulture, a commonsense reasoning dataset specifically designed to assess the cultural knowledge of Arabic LLMs. The dataset consists of 3,482 questions written in Modern Standard Arabic (MSA), covering 13 countries across the Gulf, Levant, North Africa, and the Nile Valley (see Figure 1). It spans 12 major domains and 54 fine-grained subtopics, reflecting various aspects of social norms, traditions, and daily life in the Arab world. Unlike existing benchmarks, which often rely on translated datasets, ArabCulture was built from scratch by directly engaging native speakers to write culturally relevant questions for their respective countries. We carefully implemented quality control measures throughout the dataset creation process, including rigorous validation steps to maintain accuracy, relevance, and cultural sensitivity.

We evaluate a range of closed-weight and open-weight Arabic and multilingual LLMs in a zero-shot setting to assess their cultural commonsense reasoning capabilities. Inspired by Koto et al. (2024b), we frame the task in two ways: multiple-choice questions (MCQ) and completion tasks. Additionally, we introduce three levels of location-based contextual grounding: (1) no additional location information, (2) specifying the broader region (e.g., Gulf or Levant), and (3) specifying the exact country along with its regional classification. This setup allows us to analyze how effectively LLMs incorporate geographical and cultural cues in their reasoning.

Our results show that even LLMs with up to 32B parameters struggle with cultural commonsense reasoning, with performance varying significantly across regions. We conduct a detailed analysis of the best-performing models, identifying strengths and weaknesses across different cultural contexts. We also conducted a small manual experiment to test the models' ability to explain their chosen answers. Additionally, we explore whether enriching prompts with cultural facts improves performance in smaller language models, finding that while it helps in some cases, it does not provide a universal solution.

## 2 Related Work

### 2.1 Commonsense Reasoning in English

Early research on commonsense reasoning primarily focused on linguistic reasoning, as seen in the Winograd Schema Challenge (Levesque et al., 2012) and Winogrande (Sakaguchi et al., 2021), which evaluate pronoun coreference resolution within social and linguistic contexts. Other works have explored physical commonsense reasoning, assessing model's understanding of real-world properties and relationships (Bisk et al., 2019), as well as social reasoning, where models are tested on their ability to interpret human emotions, actions, and social norms (Sap et al., 2019b). Research has also expanded into numerical (Lin et al., 2020; Akhtar et al., 2023), temporal (Tan et al., 2023), and causal reasoning (Roemmele et al., 2011; Du et al., 2022), broadening the scope of commonsense evaluation. However, these benchmarks are all developed in English and shaped by Western cultural assumptions, limiting their applicability to the Arab world.

### 2.2 Arabic Large Language Models and Their Evaluation on Commonsense Reasoning

A limited number of Arabic language models have been developed with more than 7B parameters, all of which are decoder-only architectures. These include JAIS (Sengupta et al., 2023), Fanar (Fanar-Team, 2024), AceGPT (Huang et al., 2024), and ALLAM (Bari et al., 2024b). Their evaluation of commonsense reasoning has been primarily based on machine-translated datasets from English to

| **Dataset** | **Size** | **Data Construction Method** | **Cultural?** | **Location?** | **#Topic** | **#Country** | **Reasoning?** |
|---|---|---|---|---|---|---|---|
| ArabCulture (**Ours**) | 3,482 | Manually built, validated by native | ✓ | ✓ | 54 | 13 | ✓ |
| AraDiCE-Culture (Mousi et al., 2025) | 180 | Manually built, validated by native | ✓ | ✓ | 9 | 1 | – |
| AraDiCE-WinoGrande (Mousi et al., 2025) | 1,267 | Machine-translated, post-edited | – | – | – | – | ✓ |
| AraDiCE-PIQA (Mousi et al., 2025) | 1,838 | Machine-translated, post-edited | – | – | – | – | ✓ |
| AraDiCE-OpenBookQA (Mousi et al., 2025) | 500 | Machine-translated, post-edited | – | – | – | – | ✓ |
| AlGhafa (COPA Ar) (Almazrouei et al., 2023) | 89 | Machine-translated, verified by humans | – | – | – | – | ✓ |
| ACVA (**?**) | 2,486 | ChatGPT generated, verified by humans | ✓ | – | 50 | – | – |

Table 1: Comparison of our dataset with other Arabic cultural commonsense reasoning datasets. The metadata includes **Size** (number of Arabic instances), **Cultural?** (whether the data considers cultural nuances), **Location?** ( whether the data includes fine-grained location information, such as regions and countries per region), **#Topic** (number of fine-grained topics covered), **#Country** (total number of countries across all regions) and **Reasoning?** (whether the data emphasizes commonsense reasoning or not).

Arabic (e.g., Tawalbeh and Al-Smadi (2020); Al-Bashabsheh et al. (2021)). Although this approach provides a useful reference, it does not offer a comprehensive assessment of how well these models capture culturally grounded commonsense knowledge.

Much of the recent Arabic-centric benchmarks have focused on classic NLP tasks, including syntax, semantics, and question answering, as seen in LaraBench (Abdelali et al., 2024), natural language generation in DOLPHIN (Elmadany et al., 2023a), and natural language understanding in ORCA (Elmadany et al., 2023b). Only a few studies have shifted their focus toward knowledge-intensive tasks and reasoning abilities. Among them, ArabicMMLU (Koto et al., 2024a) compiles exam questions from different education levels across Arabic-speaking countries, offering a broad knowledge assessment but placing less emphasis on cultural reasoning.

Table 1 compares ArabCulture with related Arabic datasets. Most existing datasets are derived from machine translation with post-editing, prioritizing linguistic accuracy over cultural relevance. While some assess reasoning, they often lack cultural grounding, location metadata, and fine-grained topic categorization. ACVA (Huang et al., 2024), generated using ChatGPT (Ouyang et al., 2022), is not designed for reasoning evaluation. Similarly, AraDice-Culture (Mousi et al., 2025) consists of only 180 samples and focuses on open-ended cultural knowledge rather than structured reasoning tasks. These gaps highlight the need for larger, more diverse benchmarks that better capture Arabic cultural contexts and reasoning abilities.

## 3 ArabCulture Dataset

ArabCulture is a sentence completion task in MSA, comprising 3,482 unique instances. Each question consists of a one-sentence premise and three answer choices that are both logically and syntactically valid. As illustrated in Figure 1, instances are drawn from various Arab regions.

Solving these questions requires cultural knowledge specific to the country referenced, as the correct answer aligns with culturally relevant context. This makes ArabCulture a valuable benchmark for assessing an LLM's ability to incorporate cultural understanding and knowledge in Arabic-language tasks.

### 3.1 Dataset Construction

ArabCulture is built from scratch without relying on web-scraped text, minimizing the risk of training data leakage when evaluating LLMs. It is manually created and validated by native speakers from 13 Arabic-speaking countries. To further ensure quality, the authors conduct rigorous manual checks for lexical accuracy, semantic coherence, and contextual relevance.[2]

**Worker requirements** We hired 26 expert workers from 13 Arab countries, with two workers per country, that fit defined eligibility criteria: (1) The worker must be a native Arabic speaker; (2) They must have lived in the country for at least 10 years; (3) They must possess a strong understanding of local culture and traditions; (4) Their parents must also be from the country and reside there; (5) They must have at least a high school diploma, while higher degrees were considered an advantage.

[2] The authors of this paper represent most of the studied countries, contributing diverse regional perspectives to the dataset.

Among the 26 workers, 14 hold a Bachelor's degree, including seven with a Master's degree, two with a PhD, and three with a high school diploma.

All workers were required to attend a one-hour online workshop or watch a recorded session. The workshop introduced the project concept, explained task guidelines, and addressed any potential questions. To ensure a clear understanding of the task, we conducted a pilot study before the main annotation phase.

Each worker was assigned two tasks: (1) Writing instances and (2) Reviewing and verifying the work of their peer from the same country. The payment was determined based on the minimum monthly salary for data entry jobs in each worker's country, and each worker was compensated for the equivalent of four full-time working days.

**Country Selection** We selected countries that ensured broad geographic coverage of the Arab world while remaining within budget constraints. Priority was given to countries with larger populations and land areas, resulting in a selection of 13 countries across four regions, representing approximately 82% of the total Arab world population. These include: (1) The Gulf: Saudi Arabia, Yemen, UAE; (2) The Levant: Syria, Jordan, Palestine, Lebanon; (3) North Africa: Morocco, Algeria, Tunisia, Libya; and (4) The Nile Valley: Egypt and Sudan.

**Topic taxonomy** We define 12 daily life topics with 54 fine-grained subtopics to build ArabCulture. The topic selection is based on Koto et al. (2024b) and adapted to reflect Arab regional culture. Native Arabic speakers from nine countries contributed to determining these topics, ensuring cultural relevance, diversity, and regional representation (e.g., *Ramadan* traditions). Additionally, we carefully balanced the dataset by assigning an appropriate number of examples to each topic. Table 8 in the Appendix provides an overview of the topics and their subtopics, which include food, weddings, holiday activities, daily activities, habits, traditional games, death, art, parenting, agriculture, family relationships, and idioms.

**Instance Writing** In the first stage, each worker was tasked with writing short two-sentence stories. For each entry, they were provided with a predefined topic and instructed to write a one-sentence premise followed by three candidate completions for the second sentence. The completions had to adhere to the following rules: (1) all had to be valid syntactic continuations of the premise, (2) none could introduce logical contradictions (e.g., ensuring consistency in topic and narrative), and (3) only one of the three sentences should be culturally accurate for the specified country. This design ensures that model predictions are influenced by cultural knowledge rather than grammatical or logical inconsistencies. Each worker was required to write 150 instances with two workers assigned per country.

**Two-stage of Quality Control** In stage 1, each of the 13 countries had a designated representative involved in dataset development, all of whom are also authors of this project. After workers completed their assigned instances, the respective country representatives manually reviewed their submissions to ensure adherence to the guidelines. Linguistic errors were corrected through manual editing, but if an instance did not meet the guidelines, workers were required to revise and resubmit it.

To further ensure quality, stage 2 involved a peer validation process, where each worker reviewed their colleague's work. The data was reformatted into multiple-choice questions, with the second sentence of each instance shuffled among three options. Workers were then asked to select the correct culturally appropriate completion and were allowed to consult external sources if unsure. If the worker selected the correct answer, it indicated agreement between annotators on the cultural validity of the instance. However, if the worker selected the wrong answer, the example was discarded, as it suggested ambiguity or cultural disagreement in the instance.

**Country-Specific Annotation** Beyond quality control, we also tasked the quality control workers with annotating whether the cultural context described in an instance could be relevant to other countries. The goal was to distinguish instances that are truly unique to the designated country from those that are shared across multiple countries. To ensure accuracy, the authors of this paper conducted a second round of annotation. If an instance was marked as culturally relevant to more than one country by at least one annotator, we flagged it as Not Country-Specific ($\neg$CS); otherwise, it was labeled as Country-Specific (CS). This categorization was used in our analysis experiments to better understand the distribution of culturally unique and widely shared knowledge.

| Region | #data | CS (%) | $\mu$(words) | $\mu$(chars) |
|---|---|---|---|---|
| **Gulf** | 817 | 49.1 | 34.6 | 188 |
| KSA | 261 | 36.4 | 33.2 | 185 |
| UAE | 283 | 35.3 | 38.1 | 205 |
| Yemen | 273 | 75.5 | 32.4 | 172 |
| **Levant** | 1,097 | 16.9 | 30.2 | 170 |
| Lebanon | 255 | 38.8 | 29.5 | 167 |
| Syria | 279 | 16.5 | 27.7 | 143 |
| Palestine | 273 | 8.4 | 31.7 | 177 |
| Jordan | 290 | 5.9 | 31.6 | 190 |
| **North Africa** | 1,047 | 32.4 | 32.4 | 179 |
| Tunisia | 261 | 31.8 | 28.1 | 154 |
| Algeria | 271 | 27.3 | 32.5 | 180 |
| Morocco | 276 | 37.3 | 28.6 | 161 |
| Libya | 239 | 33.1 | 41.4 | 226 |
| **Nile Valley** | 521 | 65.5 | 34.3 | 195 |
| Egypt | 265 | 74.3 | 32.4 | 178 |
| Sudan | 256 | 56.2 | 36.3 | 213 |
| **All** | 3,482 | 46.0 | 32.5 | 181 |

Table 2: Overall statistics of `ArabCulture`. CS samples represent the percentage of country-specific instances for each location. The last two columns include the average number of words and characters.

### 3.2 Data Statistics

During the instance writing phase, we initially aimed to collect 3,900 samples (26 workers × 150 samples each). However, the first quality-check round (§3.1) resulted in 3,606 samples. Following the second quality control, we discarded 124 samples, leaving a final dataset of 3,482 instances.

Table 2 shows the distribution of `ArabCulture` across regions and their respective countries. The overall proportion of country-specific instances is 46%, indicating notable cultural similarities among Arab countries. In terms of word and character count, the dataset shows consistent length across countries, with an average of 32.5 words and 181 characters per instance. However, Libyan examples tend to be longer, averaging 226 characters per instance. Furthermore, Figure 2 illustrates the total number of samples for each topic. Overall, the dataset covers a wide range of topics, with food, daily activities, and holiday activities being the most frequent, while parenting, family relationships, and agriculture are the least frequent.

## 4 Experiment

### 4.1 Experimental Setup

We conducted zero-shot experiments across 31 models, categorized into the following groups: (1) 20 multilingual models of various sizes, such as BLOOMZ (Muennighoff et al., 2023), mT0-xxl (Muennighoff et al., 2023), Llama–3 (Grattafiori et al., 2024), Aya-Expanse (Dang et al., 2024), Gemma-2 (Riviere et al., 2024), and Qwen2.5 (Qwen et al., 2025); (2) 10 Arabic-centric models of different sizes, including Jais (Sengupta et al., 2023), SILMA (Team, 2024), AceGPT-v2 (Huang et al., 2024), and ALLaM (Bari et al., 2024a); (3) 3 reasoning models that are distilled from the DeepSeek-R1 model (DeepSeek-AI et al., 2025). (4) 1 closed-weight model, GPT-4o (OpenAI et al., 2024a). All experiments are done using zero temperature (greedy sampling) to enforce the models to produce factual outputs.

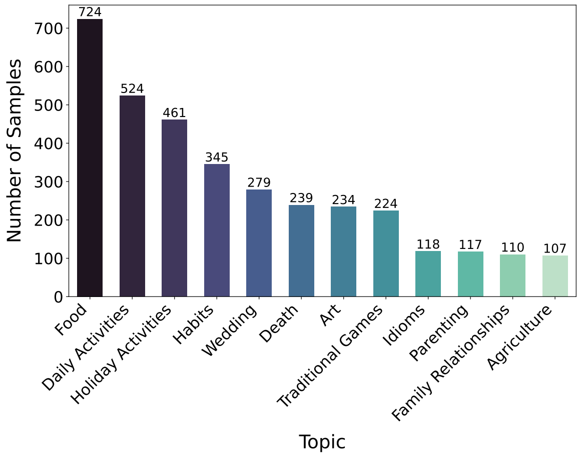


Figure 2: Total number of samples for each topic.

We conducted experiments using both Arabic and English prompts ( Figure 5) and evaluated language models using two strategies: (1) sentence completion and (2) MCQ. In the sentence completion approach, we concatenate the premise with each candidate's second sentence and select the one with the highest likelihood. For MCQ, we assign alphabetical labels to the answer choices (A, B, C for English, and أ, ب, ج for Arabic), and the selected answer corresponds to the option with the highest probability. We constructed these experiments using the LM-Evaluation-Harness Framework (Gao et al., 2024). Note that for the closed-weight model, we only perform MCQ-style evaluation, instructing the model to generate the answer as a JSON object containing only the answer character.

As discussed in Section 1, cultural knowledge varies across locations, and we hypothesize that providing geographical context can enhance a model's reasoning ability in cultural contexts. To test this, we complement our experiments with

| **Model (#parameter)** | **Completion** | | | **MCQ** | | |
|---|---|---|---|---|---|---|
| | $\ell = \text{None}$ | $\ell = \text{R}$ | $\ell = \text{R} + \text{C}$ | $\ell = \text{None}$ | $\ell = \text{R}$ | $\ell = \text{R} + \text{C}$ |
| Human | – | – | 100.0 | – | – | 100.0 |
| Random | 33.3 | 33.3 | 33.3 | 33.3 | 33.3 | 33.3 |
| BLOOMZ (7B) | 31.6 | 31.4 | 31.7 | 57.9 | 57.8 | 58.5 |
| mT0$_{\text{xxl}}$ (14B) | 27.3 | 27.6 | 27.4 | 65.9 | 66.3 | 67.0 |
| Llama-3.1 (8B) | 29.7 | 29.7 | 29.6 | 35.0 | 34.7 | 34.9 |
| Llama-3.1 Instruct (8B) | 30.9 | 31.0 | 31.2 | 47.5 | 47.0 | 49.1 |
| Llama-3 Instruct (70B) | 39.1 | 39.5 | 39.4 | 34.3 | 34.3 | 34.3 |
| Llama-3.3 Instruct (70B) | 39.9 | **40.6** | **41.1** | 75.4 | 74.0 | 71.2 |
| Aya-Expanse (8B) | 33.7 | 37.2 | 38.2 | 39.6 | 40.7 | 41.8 |
| Aya-Expanse (32B) | 37.9 | 37.9 | 39.5 | 52.6 | 49.5 | 49.5 |
| Gemma-2 (9B) | 31.8 | 31.7 | 31.8 | 35.2 | 34.5 | 34.5 |
| Gemma-2 Instruct (9B) | 33.5 | 33.8 | 33.9 | 58.7 | 55.3 | 57.0 |
| Gemma-2 (27B) | 32.6 | 33.0 | 33.2 | 34.3 | 34.3 | 34.3 |
| Gemma-2 Instruct (27B) | 38.0 | 38.9 | 39.8 | 61.6 | 64.7 | 64.2 |
| Qwen2.5 (7B) | 29.0 | 31.5 | 31.8 | 52.1 | 48.1 | 49.0 |
| Qwen2.5 Instruct (7B) | 33.2 | 33.5 | 33.6 | 53.1 | 47.9 | 48.8 |
| Qwen2.5 (14B) | 33.6 | 34.5 | 35.4 | 55.2 | 62.5 | 61.6 |
| Qwen2.5 Instruct (14B) | 36.5 | 35.9 | 37.0 | 67.7 | 67.9 | 69.3 |
| Qwen2.5 (32B) | 34.9 | 35.6 | 35.9 | 51.6 | 56.6 | 53.3 |
| Qwen2.5 Instruct (32B) | 37.6 | 37.3 | 38.6 | 75.2 | 75.8 | 76.5 |
| Qwen2.5 (72B) | 35.5 | 36.7 | 37.4 | 56.1 | 51.6 | 51.8 |
| Qwen2.5 Instruct (72B) | **40.1** | 40.2 | 40.3 | **80.1** | **79.8** | **80.0** |
| DeepSeek-R1-Distill-Llama (70B) | 37.4 | 37.6 | 38.4 | 34.3 | 35.2 | 34.5 |
| DeepSeek-R1-Distill-Qwen (32B) | 34.8 | 34.7 | 35.0 | 34.3 | 34.3 | 34.3 |
| QwQ (32B) | 34.4 | 35.7 | 36.4 | 36.7 | 35.1 | 35.6 |
| Jais (13B) | 39.3 | 39.0 | 39.3 | 34.1 | 34.9 | 34.8 |
| Jais Chat (13B) | 40.8 | 40.8 | 41.9 | 58.3 | 54.1 | 54.4 |
| Jais-v3 (30B) | 39.4 | 38.4 | 39.1 | 34.3 | 34.3 | 34.3 |
| Jais-v3 Chat (30B) | 33.3 | 33.6 | 33.4 | 60.1 | 56.2 | 54.0 |
| SILMA Instruct (9B) | 32.7 | 33.0 | 33.2 | 71.5 | 71.0 | 72.0 |
| AceGPT-v2 (8B) | 32.4 | 34.0 | 34.6 | 35.1 | 35.0 | 35.1 |
| AceGPT-v2 Chat (8B) | 36.0 | 36.4 | 37.3 | 43.1 | 39.7 | 39.3 |
| AceGPT-v2 Chat (32B) | 38.5 | 39.2 | 40.0 | **79.7** | **79.1** | **79.6** |
| AceGPT-v2 Chat (70B) | **43.2** | **44.3** | **44.5** | 61.9 | 61.7 | 62.4 |
| ALLaM-Instruct-preview (7B) | 37.7 | 38.4 | 39.2 | 67.4 | 72.0 | 72.6 |
| GPT-4o | – | – | – | **88.5** | **89.6** | **90.0** |

Table 3: Zero-shot accuracy results for the English prompt across various models and settings. ”MCQ“ refers to the multiple-choice question evaluation method, and $\ell$ represents the inclusion of location context (”R“ indicates the region, and ”C“ denotes the corresponding country). Bolded numbers highlight the highest score within each model group

three different levels of location context $\ell \in \{\text{none}, \text{region}, \text{region} + \text{country}\}$.

## 4.2 Zero-shot Experiments

Our observations show that evaluation with English prompts outperforms Arabic prompts, consistent with findings from Koto et al. (2024a); Kmainasi et al. (2024). We speculate that this is due to the dominance of English instruction-tuning datasets in LLM development. Therefore, we present the results using English prompts in the main text and include the Arabic results in Appendix D.

**Overall Observation** The overall results presented in Table 3 reveal notable performance differences between open-weight and closed-weight models in understanding Arabic culture and norms. While some large-scale open-weight models achieve relatively high accuracy—such as Qwen-2.5-Instruct (72B) with 80%, LLaMA-3.3-Instruct (70B) with 75.4%, and AceGPT-v2-Chat (32B) with 79.7%—, closed-wight models represented by GPT-4o demonstrate significantly stronger performance. Arabic-centric models do not consistently outperform multilingual models. Some, such as Jais, struggle despite being tailored for Arabic, whereas certain multilingual models like Qwen2.5 and Llama-3.3 Instruct surpass them in accuracy. Within the same model family, performance generally improves with larger model sizes—except for Jais and AceGPT. However, across different model families, scaling does not guarantee better results. This indicates that factors beyond model size, such as pretraining data, architecture, and training recipes, significantly impact cultural comprehension. When comparing base models to instruction-tuned variants, we observe modest improvements in completion tasks but substantial gains in MCQ tasks.

While reasoning models have demonstrated impressive capabilities in domains such as mathematics and programming, they perform poorly on our cultural reasoning tasks. This discrepancy suggests that cultural reasoning involves fundamentally different challenges that remain underexplored and require dedicated attention.

Overall, these findings highlight the need for improvements in Arabic and multilingual models to enhance their comprehension of Arabic cultural contexts. Addressing this gap can also help mitigate cultural biases, which have been identified in recent studies, such as the work by Naous et al. (2024).

**Multiple-Choice (MCQ) Outperforms Completion** In Table 3, we observe that sentence completion is not as reliable as MCQ, despite being a more natural approach that aligns with the sentence completion framework of `ArabCulture`. Qwen-2.5 Instruct (32B), for example, achieves 75.2% accuracy in MCQ but drops significantly to 37.6% in sentence completion. Similar disparities are also evident in smaller models; for instance, BLOOMZ (7B) achieves 58.5% in MCQ but performs at random (31.7%) in sentence completion. Some models, such as Aya-Expanse (8B), show only a small gap between MCQ and sentence completion, likely due to pretraining or fine-tuning gaps that hinder their ability to leverage structured prompts effectively. Interestingly, the older version of Llama-3 Instruct (70B) does not benefit from the MCQ strategy, whereas the latest version (Llama-3.3 Instruct) shows a dramatic improvement, increasing from 41% to 71%. Given that MCQ yields the best results, we use it for further analysis.

**Impact of Location Granularity** Adding finer-grained location information in the prompt does not produce a consistently positive or negative effect on zero-shot performance, yielding mixed results. For instance, when region and country context are included, Jais-v3 Chat (30B) experiences a 6-point accuracy drop compared to the vanilla prompt ($\ell$ = None). In contrast, Qwen-2.5 (14B) shows a substantial improvement, increasing from 55.2% ($\ell$ = None) to 61.6% when provided with country-level context. These fluctuations suggest that in some models, location specificity does not necessarily enhance cultural understanding.

| Location | GPT-4o | | Qwen-2.5 | | AceGPT-v2 | |
|---|---|---|---|---|---|---|
| | CS | ¬CS | CS | ¬CS | CS | ¬CS |
| **Gulf** | 87.8 | 91.6 | 72.3 | 80.0 | 73.1 | 82.0 |
| KSA | 88.4 | 91.0 | 82.1 | 78.9 | 77.9 | 81.3 |
| UAE | 93.0 | 93.4 | 75.0 | 83.6 | 76.0 | 84.2 |
| Yemen | 85.0 | 88.1 | 66.5 | 73.1 | 69.4 | 77.6 |
| **Levant** | 83.8 | 93.2 | 69.9 | 86.7 | 69.4 | 85.6 |
| Lebanon | 78.8 | 82.1 | 61.6 | 68.6 | 63.6 | 66.7 |
| Syria | 85.0 | 93.7 | 75.0 | 85.4 | 67.5 | 83.3 |
| Palestine | 95.7 | 94.4 | 87.0 | 89.2 | 87.0 | 90.0 |
| Jordan | 100.0 | 97.8 | 90.9 | 95.7 | 90.9 | 94.3 |
| **Nile Valley** | 87.1 | 97.8 | 76.8 | 93.3 | 74.8 | 88.9 |
| Egypt | 87.3 | 97.1 | 78.2 | 91.2 | 73.1 | 83.8 |
| Sudan | 86.8 | 98.2 | 75.0 | 94.6 | 77.1 | 92.0 |
| **North Africa** | 86.1 | 86.9 | 70.8 | 81.1 | 73.7 | 79.9 |
| Tunisia | 75.9 | 80.9 | 61.4 | 69.7 | 62.7 | 72.5 |
| Algeria | 85.1 | 83.8 | 63.5 | 77.2 | 68.9 | 72.6 |
| Morocco | 91.3 | 94.2 | 75.7 | 93.1 | 80.6 | 93.6 |
| Libya | 91.1 | 89.4 | 81.0 | 85.6 | 81.0 | 82.5 |

Table 4: Performance of `GPT-4o`, `Qwen-2.5-72B Instruct`, and `AceGPT-v2-32B Chat` across countries and regions. `CS` denotes country-specific examples, while ¬`CS` otherwise. Green cells indicate the top three scores, while red cells highlight the bottom three.

| Topic | GPT-4o | | Qwen-2.5 | | AceGPT-v2 | |
|---|---|---|---|---|---|---|
| | CS | ¬CS | CS | ¬CS | CS | ¬CS |
| Agriculture | 91.5 | 91.7 | 83.0 | 90.0 | 78.7 | 85.0 |
| Art | 87.7 | 92.4 | 74.8 | 84.8 | 73.5 | 84.8 |
| Daily Act. | 79.4 | 90.2 | 65.4 | 83.2 | 68.4 | 86.6 |
| Death | 84.4 | 91.3 | 78.1 | 83.1 | 84.4 | 81.2 |
| Family Rel. | 90.0 | 91.1 | 80.0 | 88.9 | 75.0 | 83.3 |
| Food | 87.0 | 90.4 | 71.5 | 79.4 | 70.9 | 79.4 |
| Habits | 81.8 | 91.0 | 68.8 | 86.2 | 71.4 | 82.1 |
| Holiday Act. | 89.2 | 94.4 | 70.7 | 89.1 | 66.9 | 90.1 |
| Idioms | 88.9 | 91.9 | 70.4 | 83.8 | 86.4 | 81.1 |
| Parenting | 76.2 | 93.8 | 66.7 | 87.5 | 66.7 | 84.4 |
| Trd. Games | 87.5 | 89.6 | 80.0 | 84.7 | 70.0 | 80.6 |
| Wedding | 89.4 | 89.1 | 78.0 | 79.6 | 81.8 | 78.9 |

Table 5: Performance of `GPT-4o`, `Qwen-2.5-72B Instruct`, and `AceGPT-v2-32B Chat` across topics. `CS` denotes country-specific examples, while ¬`CS` otherwise. Green cells indicate the top three scores, while red cells highlight the bottom three.

# 5 Analysis

## 5.1 Result by Categories

In this section, we expand on our findings based on the top three models from Table 3: (1) GPT-4o, (2) Qwen-2.5-72B-Instruct, and (3) AceGPT-v2-32B-Chat. We provide a detailed analysis across country and topic, with each sample categorized as either country-specific (CS) or non-country-specific (¬CS).

**Country** In Table 4, we observe significant variation in LLM performance across countries, emphasizing the need for country-specific adaptation when deploying models. Questions from Jordan are consistently predicted with high accuracy, exceeding 90% across all models. However, performance drops significantly for Lebanon and Tunisia, where models struggle to provide correct answers. Even the Arabic-centric AceGPT-v2 achieves only 63.6% accuracy for Lebanon and 62.7% for Tunisia. Across the four regions, we find that the Levant is the most challenging, underscoring the difficulty of achieving reliable performance across different cultural and linguistic contexts.

More interestingly, country-specific questions prove to be more challenging than non-country-specific ones across nearly all countries and models. For instance, in the Nile Valley region, GPT-4o experiences an accuracy drop of nearly 10 points when handling country-specific questions, while in the Levant region, Qwen-2.5 sees a 17-point decline. This suggests that when cultural knowledge is not shared across multiple countries or regions, it becomes more distinct and difficult for LLMs to capture accurately.

**Topic** Table 5 shows that LLMs encode cultural knowledge differently across various aspects of Arab culture. For example, GPT-4o performs best in agriculture and family relationships, while AceGPT excels in topics related to death and idioms. Meanwhile, Qwen achieves its highest accuracy in agriculture and traditional games. The accuracy gap between the highest- and lowest-performing topics across models ranges from 10 to 20 points, highlighting the difficulty of adapting cultural knowledge in LLMs. Additionally, we observe a consistent trend with Table 4, where country-specific samples are more challenging than non-country-specific ones.

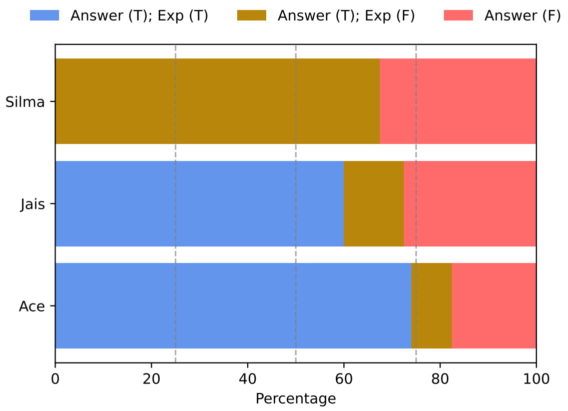


Figure 3: Performance comparison between jais-30b-chat-v3, AceGPT-v2-32B-Chat, and SILMA-9B-Instruct-v1.0 based on text generation output. “Answer (T)” indicates that the generated answer is true, while “Exp (F)” denotes that the answer explanation is false.

## 5.2 Can the Model Provide a Reasonable Explanation to Support the Answer?

We focus on Arabic-centric models—Jais, AceGPT, and Silma—to evaluate their actual generation capabilities. For 200 randomly selected samples, we generate responses by appending مع ذكر السبب ("with mentioning the reason") to the Arabic prompt (§C.1) to instruct the model to provide a brief explanation for its choice. We then manually assess the outputs to verify both the correctness of the answer and the validity of the explanation.

Figure 3 presents the results for each model, comparing their generation accuracy with their MCQ performance from Table 3. Jais demonstrated a significant improvement, increasing from 40% in MCQ to 72% in the manual evaluation. In contrast, Silma’s performance dropped from 73% in MCQ to 67% in the generation task. Notably, Silma often failed to generate explanations, instead providing only the answer key or, at times, just the answer text. Meanwhile, AceGPT maintained a consistent performance across both MCQ and generation tests, showing no significant change in accuracy.

## 5.3 Improving Small Language Model with Additional Context from GPT-4o

We evaluate six base models with $\leq$3B parameters to assess the impact of cultural context augmentation. Using GPT-4o, we generate five factual Arabic sentences conditioned on the premise, subtopic, and country, following the Arabic prompt in Figure 6. These sentences are then incorporated into

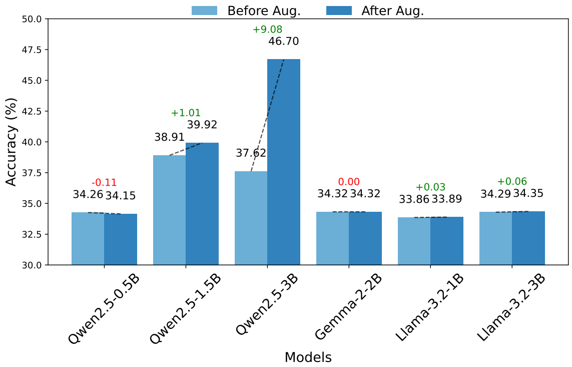


Figure 4: Accuracies per models before and after context augmentation. This experiment uses the Arabic prompt template for MCQ and location $\ell \in \{\text{region} + \text{country}\}$.

the Arabic multiple-choice question prompt (§C.2) with location context $\ell \in \{\text{region} + \text{country}\}$. Results (Figure 4) show accuracy gains for Qwen and Llama models, except for Qwen2.5-0.5B (decline) and Gemma-2-2B (unchanged). Qwen2.5-3B achieves the highest improvement across all countries. These variations likely stem from differences in training data: Qwen and Llama were trained on multilingual datasets, whereas Gemma-2, primarily trained in English, has limited multilingual support.

## 6 Conclusion

We introduced ArabCulture, a benchmark for evaluating cultural commonsense reasoning in the Arab world. The dataset comprises 3,482 questions across 13 countries, covering 12 daily life domains with 54 fine-grained subtopics, all authored and validated by native speakers. Evaluations on 31 LLMs show significant performance gaps, with open-weight models up to 32B struggling to capture Arab cultural contexts. Variability across countries, regions, and topics highlights the need for more culturally aware models and datasets tailored to the Arabic-speaking world.

## Limitations

**Culture is only one side of reasoning** While cultural knowledge plays a crucial role in shaping commonsense reasoning, it is only one of several dimensions that contribute to a model's overall reasoning capabilities (Plaat et al., 2024). Reasoning in LLMs encompasses a broad range of cognitive skills, including logical inference, numerical reasoning, and causal understanding, among others.

**The Influence of Dialects on Cultural Reasoning** The Arab world is characterized by rich dialects that vary not only across countries but also within different regions of the same country. These dialects significantly shape cultural expression, influencing language use in areas such as proverbs, humor, and everyday communication. This is particularly evident in topics like "idioms," where meaning and usage are deeply tied to specific dialects and local linguistic conventions.

However, to ensure that our evaluation isolates cultural commonsense reasoning rather than a model's proficiency in specific dialects, we constructed ArabCulture in Modern Standard Arabic (MSA). MSA serves as a unifying linguistic medium across Arabic-speaking countries, allowing us to control for dialectal variation while still capturing essential cultural knowledge. While this approach enhances comparability across regions, it also introduces a limitation: certain cultural concepts that are best expressed through dialect-specific phrasing or context may not be fully represented in our dataset.

**Location Leakage** Despite our efforts to systematically control the granularity of location information, some questions or corresponding multiple-choice completions inadvertently reveal the location through the inclusion of location cues (landmarks, national events, etc.) within prompts. Thus, resulting in unintended location leakage, where the model gains access to country-specific cues directly from text rather than controlled contexts, making it difficult to isolate the effect of the location granularity control.

**Coverage of All Arab Countries** While our study covers a significant portion of the Arab world, representing 82% of the total population, certain unique cultures remain underrepresented. Notably, countries such as Mauritania, Somalia, and Comoros were not included, despite their distinct cultural and linguistic characteristics. These nations, located in North Africa, the Horn of Africa, and the Indian Ocean, respectively, contribute to the broader diversity of the Arab world. Their exclusion was primarily due to the difficulty in sourcing human annotators from these regions.

## Acknowledgments

We are grateful to the supercomputing center at MBZUAI for providing resources for this research.

We also gratefully acknowledge the support of the university for providing funding to hire the human annotators.

## References

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, and 1 others. 2024. Larabench: Benchmarking arabic ai with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520.

Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15391–15405, Singapore. Association for Computational Linguistics.

Emran Al-Bashabsheh, Huthaifa Al-Khazaleh, Omar Elayan, and Rehab Duwairi. 2021. Commonsense validation for arabic sentences using deep learning. In *2021 22nd International Arab Conference on Information Technology (ACIT)*, pages 1–7. IEEE.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, Julien Launay, and Badreddine Noune. 2023. AlGhafa evaluation benchmark for Arabic language models. In *Proceedings of ArabicNLP 2023*, pages 244–275, Singapore (Hybrid). Association for Computational Linguistics.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2024a. Allam: Large language models for arabic and english. *Preprint*, arXiv:2407.15390.

M Saiful Bari, Yazeed Alnumay, Norah A Alzahrani, Nouf M Alotaibi, Hisham A Alyahya, Sultan AlRashed, Faisal A Mirza, Shaykhah Z Alsubaie, Hassan A Alahmed, Ghadah Alabduljabbar, and 1 others. 2024b. Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Mona Diab, Nizar Habash, and Imed Zitouni. 2017. NLP for Arabic and related languages. *Traitement Automatique des Langues*, 58(3):9–13.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Abdelrahim Elmadany, Ahmed El-Shangiti, Muhammad Abdul-Mageed, and 1 others. 2023a. Dolphin: A challenging and diverse benchmark for arabic nlg. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1404–1422.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023b. ORCA: A challenging benchmark for Arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

Fanar-Team. 2024. Fanar: An arabic-centric multimodal generative ai platform. https://fanar.qa/en. *arXiv preprint arXiv:2409.11404*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

A. Giddens and P.W. Sutton. 2014. *Essential Concepts in Sociology*. Polity Press.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu.

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 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Mohamed Bayan Kmainasi, Rakif Khan, Ali Ezzat Shahroor, Boushra Bendou, Maram Hasanain, and Firoj Alam. 2024. Native vs non-native language prompting: A comparative analysis. *Preprint*, arXiv:2409.07054.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024a. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024b. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, and 9 others. 2023. Llm360: Towards fully transparent open-source llms. *ArXiv*, abs/2312.06550.

J.J. Macionis. 2012. *Sociology: Fourteenth Edition*. Pearson.

Barry Mirkin. 2010. Population levels, trends and policies in the arab region: Challenges and opportunities.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *Preprint*, arXiv:2407.11511.

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

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 176 others. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI spring symposium series*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019a. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, and 1 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Abdulhadi Shoufan and Sumaya Alameri. 2015. Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Saja Khaled Tawalbeh and Mohammad Al-Smadi. 2020. Is this sentence valid? an arabic dataset for commonsense validation. *ArXiv*, abs/2008.10873.

Silma Team. 2024. Silma.

# A Dataset Statement for `ArabCulture`

## A.1 General Information

**Dataset title** `ArabCulture`
**Dataset version** 1.0 (Feb 2025)

## A.2 Executive Summary

`ArabCulture` is a cultural Arabic commonsense reasoning dataset covering 13 Arabic countries in the Gulf, Levant, North Africa, and the Nile Valley. The dataset spans 12 daily life domains and 54 fine-grained subtopics. It was created from scratch by native speakers who validated culturally relevant questions.

## A.3 Curation Rationale

`ArabCulture` serves as a cultural benchmark to assess large language models' ability to reason within culturally specific contexts. Built from scratch by native speakers, it avoids web-scraped text and undergoes rigorous quality control to ensure lexical accuracy, semantic coherence, and cultural sensitivity.

The dataset creation process involves:

1. **Coverage Determination:** Selecting relevant countries and topics.
2. **Annotator Selection:** Hiring qualified native speakers with deep cultural knowledge.
3. **Example Generation:** Each annotator produces 150 examples, with two annotators per country.
4. **Cross-Review:** Annotators validate each other's work by confirming correct answers.
5. **Final Review:** Unclear samples are revised or discarded based on comprehensive quality checks.

## A.4 Documentation for Source Datasets

`ArabCulture` is built entirely from scratch, without relying on web-scraped text, All data is manually created and validated by native speakers from 13 Arabic-speaking countries.

## A.5 Country and Regional Diversity

`ArabCulture` covers 13 Arab countries, ensuring rich cultural perspectives.

Our country selection was guided by the goal of broad geographic representation across the Arab world. These 13 countries span four key regions:

- **The Gulf:** Saudi Arabia, Yemen, UAE.
- **The Levant:** Syria, Jordan, Palestine, Lebanon.
- **North Africa:** Morocco, Algeria, Tunisia, Libya.
- **The Nile Valley:** Egypt, Sudan.

### A.6 Annotator Demographics

We recruited 26 expert annotators from 13 Arab countries, with two annotators representing each country. To ensure cultural authenticity and linguistic proficiency, we enforced the following eligibility criteria:

- Native Arabic speakers.
- Residency in the country for at least 10 years.
- Deep understanding of local culture and traditions.
- Both parents are native to and reside in the country.
- Minimum educational requirement of a high school diploma (higher degrees are preferred).

Among the 26 annotators, 14 hold a Bachelor's degree, seven have a Master's degree, two have a PhD, and three have a high school diploma. All annotators participated (attended or watched a record) in an initial online workshop to ensure a clear understanding of the project guidelines.

### A.7 Topic Diversity

`ArabCulture` features a carefully curated taxonomy of daily life topics. It includes 12 main topics with 54 fine-grained subtopics—covering areas such as food, weddings, holiday activities, daily routines, habits, traditional games, death, art, parenting, agriculture, family relationships, and idioms. This extensive range ensures that the dataset captures both common and unique cultural experiences across the Arab world.

## B Chosen Topics Distribution

Table 8 shows the distribution of topics and their corresponding subtopics.

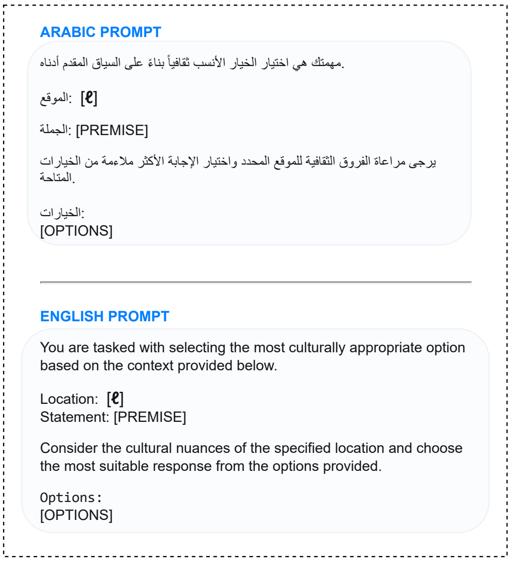
## C Prompts

### C.1 Zero-shot Experiment prompts

Figure 5 shows the prompts in Arabic and English that we used for the zero-shot experiments. The Arabic prompt is also used to generate responses for 5.2.

### C.2 Cultural Context prompts

Figure 6 displays the Arabic prompt used to generate five culturally grounded sentences for enhancing the small language model with additional context from GPT-4o. An English translation is provided to assist readers unfamiliar with Arabic in understanding the prompt.



ARABIC PROMPT

مهمتك هي اختيار الخيار الأنسب ثقافياً بناءً على السياق المقدم أدناه.

الموقع: [ℓ]

الجملة: [PREMISE]

يرجى مراعاة الفروق الثقافية للموقع المحدد واختيار الإجابة الأكثر ملاءمة من الخيارات المتاحة.

الخيارات:
[OPTIONS]

ENGLISH PROMPT

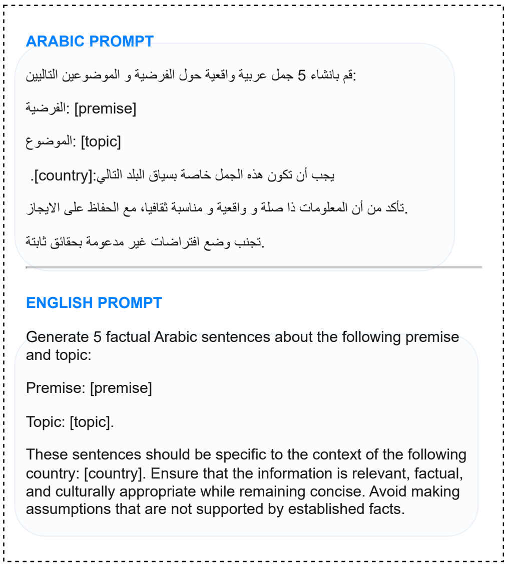
You are tasked with selecting the most culturally appropriate option based on the context provided below.

Location: [ℓ]
Statement: [PREMISE]

Consider the cultural nuances of the specified location and choose the most suitable response from the options provided.

Options:
[OPTIONS]

Figure 5: Templates for multiple-choice question prompts. Sentence completion prompt template is the same but without the options section. Location $\ell \in \{\text{none}, \text{region}, \text{region} + \text{country}\}$.



ARABIC PROMPT

قم بانشاء 5 جمل عربية واقعية حول الفرضية و الموضوعين التاليين:

الفرضية: [premise]

الموضوع: [topic]

يجب أن تكون هذه الجمل خاصة بسياق البلد التالي:[country].

تأكد من أن المعلومات ذا صلة و واقعية و مناسبة ثقافيا، مع الحفاظ على الايجاز.

تجنب وضع افتراضات غير مدعومة بحقائق ثابتة.

ENGLISH PROMPT

Generate 5 factual Arabic sentences about the following premise and topic:

Premise: [premise]

Topic: [topic].

These sentences should be specific to the context of the following country: [country]. Ensure that the information is relevant, factual, and culturally appropriate while remaining concise. Avoid making assumptions that are not supported by established facts.

Figure 6: Prompts to generate culturally relevant sentences for context augmentation. The Arabic prompt was used for the generation.

## D Results of the Arabic prompt in the zero-shot experiments

Table 9 presents the zero-shot experiment results using the Arabic prompt, which is illustrated in Figure 5.

### D.1 Results by Geographic Location - Arabic Prompt

| Topic | GPT-4o | | AceGPT-v2 | | Llama-3.3 | |
|---|---|---|---|---|---|---|
| | CS | ¬CS | CS | ¬CS | CS | ¬CS |
| **Gulf** | 81.0 | 92.5 | 76.3 | 85.3 | 69.1 | 81.5 |
| KSA | 89.5 | 93.4 | 77.9 | 84.3 | 75.8 | 80.1 |
| UAE | 90.0 | 93.4 | 75.0 | 90.2 | 66.0 | 87.4 |
| Yemen | 72.8 | 88.1 | 76.2 | 74.6 | 67.5 | 68.7 |
| **Levant** | 77.5 | 92.3 | 67.6 | 86 | 66.5 | 84.4 |
| Lebanon | 70.7 | 78.8 | 60.6 | 67.9 | 62.6 | 66.0 |
| Syria | 82.5 | 94.6 | 72.5 | 87.0 | 70.0 | 80.8 |
| Palestine | 87.0 | 94.4 | 78.3 | 87.2 | 65.2 | 88.8 |
| Jordan | 100.0 | 96.1 | 90.9 | 94.3 | 90.9 | 93.9 |
| **Nile Valley** | 86.8 | 96.1 | 70.7 | 86.1 | 76.0 | 85.0 |
| Egypt | 88.8 | 98.5 | 72.1 | 79.4 | 76.1 | 72.1 |
| Sudan | 84.0 | 94.6 | 68.8 | 90.2 | 75.7 | 92.9 |
| **North Africa** | 82.0 | 86.0 | 73.7 | 78.8 | 73.5 | 80.9 |
| Tunisia | 67.5 | 79.8 | 62.7 | 72.5 | 68.7 | 69.7 |
| Algeria | 83.8 | 84.3 | 68.9 | 73.1 | 63.5 | 80.2 |
| Morocco | 90.3 | 94.2 | 81.6 | 93.1 | 83.5 | 93.1 |
| Libya | 84.8 | 86.2 | 79.7 | 77.5 | 74.7 | 81.2 |

Table 6: Performance of the best three models using the Arabic prompt. The results for `GPT-4o` are using the Region prompt, while the results for `AceGPT-v2-32B-Chat` and `Llama-3.3-70B-Instruct` are using the Country_Region prompt. The first column includes the different locations (countries and regions) with the regions in **bold**. `CS` refers to the Country Specific examples, while ¬`CS` refers to the rest of the examples. The green and red cells indicate the top three and bottom three scores, respectively.

Table 6 presents the country-level breakdown analysis based on the Arabic prompt.

### D.2 Results by Topic - Arabic Prompt

Table 7 presents the breakdown analysis by topic based on the Arabic prompt.

| Topic | GPT-4o | | AceGPT-v2 | | Llama-3.3 | |
|---|---|---|---|---|---|---|
| | CS | ¬CS | CS | ¬CS | CS | ¬CS |
| Agriculture | 87.2 | 95.0 | 76.6 | 85.0 | 78.7 | 83.3 |
| Art | 85.2 | 89.9 | 74.2 | 87.3 | 74.8 | 86.1 |
| Daily Activities | 76.5 | 89.9 | 67.6 | 83.2 | 69.9 | 84.3 |
| Death | 78.1 | 90.3 | 87.5 | 84.1 | 78.1 | 82.6 |
| Family Relationships | 85.0 | 91.1 | 80.0 | 85.6 | 60.0 | 85.6 |
| Food | 82.0 | 91.2 | 69.3 | 79.9 | 68.7 | 76.5 |
| Habits | 80.5 | 88.4 | 68.8 | 84.0 | 72.7 | 85.8 |
| Holiday Activities | 82.8 | 92.1 | 74.5 | 88.8 | 70.7 | 87.2 |
| Idioms | 84.0 | 97.3 | 80.2 | 78.4 | 70.4 | 75.7 |
| Parenting | 71.4 | 91.7 | 71.4 | 84.4 | 76.2 | 88.5 |
| Traditional Games | 86.2 | 89.6 | 65.0 | 77.8 | 72.5 | 77.8 |
| Wedding | 84.1 | 89.8 | 80.3 | 85.7 | 75.8 | 81.6 |

Table 7: Performance of the best three models using the Arabic prompt across the different Topics. The results for `GPT-4o` are using the Region prompt, while the results for `AceGPT-v2-32B-Chat` and `Llama-3.3-70B-Instruct` are using the Country_Region prompt. `CS` refers to the Country Specific examples, while ¬`CS` refers to the rest of the examples. The green and red cells indicate the top three and bottom three scores, respectively.

| **Topics** | **Sub-topics** | **#Samples** |
|---|---|---|
| Food | Breakfast (5), Lunch (5), Dinner (2), Sahoor (Ramadan) (5), Iftar (Ramadan) (5), Dessert (3), Fruits (3), Snacks (2) | 30 |
| Wedding | Wedding location (1), Wedding food (1), Wedding dowry (1), Wedding other logistics (2), Men ceremony vs. women ceremony (2), Songs and activities during the wedding (5) | 12 |
| Holiday Activities | Traditions before religious holidays (5), Traditions during religious holidays (10), Activities for non-religious holidays (5) | 20 |
| Daily Activities | Before going to work/college (4), While on the way to college/work (3), Things you do with colleagues/friends while at work/college (2), Things you do after coming back from work/uni (men) (2), Things you do after coming back from work/uni (women) (2), Household activities (groceries, fixing things, cleaning, etc.) (6), Things you do in your free time (indoors or outdoors) (5) | 24 |
| Habits | Eating habits (3), Stereotypes (5), Communication habits (3), Financial habits (1), Gift-giving practices (1), Cleanliness habits (2) | 14 |
| Traditional Games | Childhood games indoors (5), Childhood games outdoors (5) | 10 |
| Death | Before burying (3), Burying rituals (2), After burying ceremonies (4), Inheritance (1) | 10 |
| Art | Musical instruments (3), Local songs (3), Local dances (4) | 10 |
| Parenting | Parents-child actions (3), Grandparents-child actions (2) | 5 |
| Agriculture | What to plant (3), While planting (1), Harvest (1) | 5 |
| Idioms | Idioms in context (5) | 5 |
| Family Relationships | Between siblings (2), With cousins (1), Relationship of parents and child (2) | 5 |

Table 8: Overview of topics, sub-topics, and the sample counts for each topic. The number in parenthesis beside each subtopic represents the number of samples for each subtopic.

| **Model (#parameter)** | **Completion** | | | **MCQ** | | |
|---|---|---|---|---|---|---|
| | $\ell = \text{None}$ | $\ell = \text{R}$ | $\ell = \text{R + C}$ | $\ell = \text{None}$ | $\ell = \text{R}$ | $\ell = \text{R + C}$ |
| Human | – | – | 100.0 | – | – | 100.0 |
| Random | 33.3 | 33.3 | 33.3 | 33.3 | 33.3 | 33.3 |
| BLOOMZ (7B) | 30.1 | 30.7 | 30.9 | 50.6 | 52.2 | 52.7 |
| mT0$_{x}xl$ (14B) | 26.6 | 26.3 | 26.9 | 65.5 | 66.3 | 66.4 |
| Llama-3.1 (8B) | 27.7 | 27.5 | 27.6 | 34.3 | 34.1 | 34.2 |
| Llama-3.1 Instruct (8B) | 32.2 | 31.0 | 31.3 | 37.8 | 36.8 | 37.8 |
| Llama-3 Instruct (70B) | 36.6 | 37.5 | 38.5 | 47.1 | 37.0 | 38.7 |
| Llama-3.3 Instruct (70B) | 39.0 | 39.6 | 39.9 | **78.4** | **77.8** | **78.8** |
| Aya-Expanse (8B) | 34.7 | 35.8 | 36.9 | 36.3 | 38.1 | 39.3 |
| Aya-Expanse (32B) | 37.9 | 39.3 | 39.9 | 38.4 | 43.7 | 44.6 |
| Gemma-2 (9B) | 32.0 | 32.2 | 32.7 | 34.5 | 35.8 | 35.4 |
| Gemma-2 Instruct (9B) | 32.5 | 33.5 | 33.5 | 34.4 | 34.3 | 34.3 |
| Gemma-2 (27B) | 33.8 | 34.4 | 35.0 | 34.3 | 34.3 | 34.4 |
| Gemma-2 Instruct (27B) | 35.9 | 36.4 | 37.1 | 34.4 | 34.9 | 34.9 |
| Qwen2.5 (7B) | 30.0 | 30.4 | 30.4 | 47.9 | 47.0 | 47.7 |
| Qwen2.5 Instruct (7B) | 32.5 | 33.2 | 33.8 | 51.6 | 37.7 | 39.3 |
| Qwen2.5 (14B) | 32.4 | 33.1 | 33.5 | 46.5 | 57.8 | 57.4 |
| Qwen2.5 Instruct (14B) | 36.5 | 37.7 | 37.2 | 52.2 | 58.4 | 59.5 |
| Qwen2.5 (32B) | 33.4 | 34.2 | 34.5 | 42.0 | 43.7 | 42.7 |
| Qwen2.5 Instruct (32B) | 36.6 | 37.7 | 37.8 | 70.7 | 74.6 | 76.3 |
| Qwen2.5 (72B) | 35.4 | 36.2 | 36.4 | 48.0 | 52.2 | 57.3 |
| Qwen2.5 Instruct (72B) | **39.6** | **40.1** | **41.0** | 61.5 | 64.5 | 65.4 |
| DeepSeek-R1-Distill-Llama (70B) | 36.7 | 37.1 | 37.6 | 34.7 | 34.5 | 35.0 |
| DeepSeek-R1-Distill-Qwen (32B) | 33.6 | 34.4 | 35.0 | 34.3 | 34.3 | 34.3 |
| QwQ (32B) | 22.83 | 22.29 | 22.77 | 32.88 | 32.62 | 32.28 |
| Jais (13B) | 37.9 | 37.8 | 38.3 | 33.9 | 33.6 | 33.8 |
| Jais chat (13B) | 39.8 | 39.9 | 40.6 | 40.7 | 39.6 | 38.4 |
| Jais-v3 (30B) | 39.9 | 40.4 | 40.6 | 34.8 | 35.8 | 35.6 |
| Jais-v3 Chat (30B) | 34.0 | 34.0 | 34.5 | 35.2 | 36.2 | 39.7 |
| SILMA Instruct (9B) | 32.5 | 33.3 | 33.4 | 70.2 | 70.7 | 70.7 |
| AceGPT-v2 (8B) | 29.9 | 31.2 | 31.6 | 34.3 | 34.2 | 34.2 |
| AceGPT-v2 Chat (8B) | 34.6 | 35.1 | 35.9 | 44.8 | 45.0 | 45.4 |
| AceGPT-v2 Chat (32B) | 37.7 | 38.9 | 39.0 | **78.8** | **78.2** | **79.8** |
| AceGPT-v2 Chat (70B) | **42.9** | **44.5** | **45.1** | 73.6 | 73.0 | 74.4 |
| ALLaM-Instruct-preview (7B) | 36.5 | 37.2 | 37.9 | 70.0 | 71.9 | 74.4 |
| GPT-4o | | | | **88.5** | **91.9** | **91.1** |

Table 9: Zero-shot accuracy results for the Arabic prompt across various models and settings. ”MCQ“ refers to the multiple-choice question evaluation method, and $\ell$ represents the inclusion of location context (”R“ indicates the region, and ”C“ denotes the corresponding country). Bolded numbers highlight the highest score within each model group.