PalmX 2025: The First Shared Task on Benchmarking LLMs on Arabic and Islamic Culture

Fakhraddin Alwajih[λ] **Abdellah El Mekki**[λ] **Hamdy Mubarak**[γ]
Majd Hawasly[γ] **Abubakr Mohamed**[γ] **Muhammad Abdul-Mageed**[λ]

[λ]The University of British Columbia [γ]Qatar Computing Research Institute

{fakhr.alwajih,abdellah.elmekki,muhammad.mageed}@ubc.ca
{hmubarak,mhawasly,abumohamed}@hbku.edu.qa

Figure 1: An Overview of the PalmX 2025 Shared Task

Abstract

Large Language Models (LLMs) inherently reflect the vast data distributions they encounter during their pre-training phase. As this data is predominantly sourced from the web, there is a high chance it will be skewed towards high-resourced languages and cultures, such as those of the West. Consequently, LLMs often exhibit a diminished understanding of certain communities, a gap that is particularly evident in their knowledge of Arabic and Islamic cultures. This issue becomes even more pronounced with increasingly under-represented topics. To address this critical challenge, we introduce PalmX 2025, the first shared task designed to benchmark the cultural competence of LLMs in these specific domains. The task is composed of two subtasks featuring multiple-choice questions (MCQs) in Modern Standard Arabic (MSA): General Arabic Culture and General Islamic Culture. These subtasks cover a wide range of topics, including traditions, food, history, religious practices, and language expressions from across 22 Arab countries. The initiative drew considerable interest, with 26 teams registering for Subtask 1 and 19 for Subtask 2, culminating in nine and six valid submissions, respectively. Our findings reveal that task-specific fine-tuning substantially boosts performance over baseline models. The top-performing systems achieved an accuracy of 72.15% on cultural questions and 84.22% on Islamic knowledge. Parameter-efficient fine-tuning emerged as the predominant and most effective approach among participants, while the utility of data augmentation was found to be domain-dependent. Ultimately, this benchmark provides a crucial, standardized framework to guide the development of more culturally grounded and competent Arabic LLMs. Results of the shared task demonstrate that general cultural and general religious knowledge remain challenging to LLMs, motivating us to continue to offer the shared task in the future.

1 Introduction

Despite their impressive capabilities, LLMs often display systematic Western- and Anglocentric biases, mirroring the over-representation of these perspectives in their training data (Adilazuarda et al., 2024; Pawar et al., 2025). This lack of cultural diversity can lead to outputs that are not only inappropriate but also harmful. For instance, an Arabic LLM trained on translated English data once suggested having a beer after prayer, a recommen-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 774–789
November 8-9, 2025

dation that fundamentally misunderstands (and indeed disrespects) core Arab cultural and religious norms (Naous et al., 2023). Incidents such as this underscore a critical distinction in LLM development between *cultural awareness*, which refers to the understanding of a culture's norms and values and *cultural alignment*, which is focused on the adaptation of actions to respect and reflect these norms and values (AlKhamissi et al., 2024). True progress requires models that are not just culturally aware, but culturally aligned as well.

The need for culturally aligned models is particularly acute in the Arab world, a region of over 450 million people spread across 22 countries. The Arab world comprises immense diversity in customs and traditions, as well as dialectal richness. While recent efforts have produced relatively fluent Arabic LLMs (Bari et al., 2024; Sengupta et al., 2023; Huang et al., 2024), many are trained on machine-translated datasets and evaluated on general NLP tasks in ways that largely overlook country-specific cultural competence. Foundational work on datasets like *Palm* (Alwajih et al., 2025a) has begun to address this by providing culturally inclusive, human-created Arabic instructions covering all 22 Arab countries. However, a standardized benchmark is still needed to systematically measure and compare the cultural understanding of different models.

To bridge this evaluation gap, we introduce the *PalmX 2025* Shared Task, the first benchmarking effort focused specifically on the cultural competence of LLMs in the Arabic context. In this task, we define culture as the collection of knowledge, beliefs, and behaviors encompassing the traditions, social etiquette, cuisine, history, arts, dialectal expressions, and religious practices that characterize communities across the Arab world. *PalmX* challenges models with multiple-choice questions designed to test deep cultural knowledge, not superficial pattern matching. The task is divided into two subtasks: one on *General Arabic Culture* and another on *General Islamic Culture*, reflecting the cornerstones of identity in the region. By providing a standardized evaluation framework, *PalmX* aims to drive the development of LLMs that are not only linguistically fluent but also culturally grounded and respectful.

This paper is organized as follows: Section 2 describes the *PalmX 2025* shared task, including data collection and annotation for both subtasks. Section 3 outlines the participation rules and evaluation methodology. Section 4 presents the participating teams and their results. Section 5 discusses the findings and provides analysis of the methodological approaches for the participating teams. Section 6 concludes with key insights and future directions. Appendix A provides a literature review of related work, and Appendix B presents detailed data analysis including country and topic distributions for datasets of both subtasks.

2 Task Description: PalmX 2025

The objective of the *PalmX 2025* Shared Task[1] is to enable evaluation of the competence of LLMs on Arabic and Islamic cultures through two independent subtasks: *general Arabic culture* and *general Islamic culture*. Each subtask is designed as a set of MCQs in MSA, each with four options (A-D) and a single correct answer; the questions target grounded knowledge. The distractors for each MCQ question are designed to plausible but incorrect, often sharing surface cues to minimize the chance of correct guesses. For each subtask, we provide training, development (dev), and test splits. The training split is provided to participants to support system development, allowing for various approaches such as fine-tuning. Additionally, the dev split is shared with participants to facilitate hyperparameter tuning and local evaluation of their systems before the test phase. The test split is kept private during the competition and is released publicly after the competition concludes. We apply basic quality filters to ensure clarity, a single unambiguous answer, and cultural correctness. This process involves removing off-topic questions unrelated to culture, those with multiple correct answers, biased content, and items with grammatical errors. Accuracy is the primary evaluation metric.

All the resources of *PalmX 2025* shared task are publicly available, including data and evaluation code.[2]

2.1 Subtask 1: General Arabic Culture

The goal of this subtask is to encourage development of methods for incorporating Arabic general culture in LLMs, allowing them to comprehend and reason about diverse aspects of general Arabic culture. These aspects are coming from different cultural categories including *traditional customs*, *local etiquette*, *cuisine*, *historical events*, *famous*

[1] https://palmx.dlnlp.ai/

[2] https://github.com/UBC-NLP/palmx_2025

figures, *geography*, *local languages (dialects)*, and *arts*.

2.1.1 Data Collection and Annotation

The data for this subtask cover a number of cultural topics. To ensure this wide coverage, we follow two complementary data collection strategies, as described below.

Method 1: We source the data from *Palm* (Alwajih et al., 2025a) training split, which we convert into an MCQ format using Qwen3 30B (Yang et al., 2025). Using this method, we acquire $4,000$ samples.

Method 2: We crawl web pages from diverse online resources covering cultural knowledge, customs, etiquette, values, and practices across all Arab countries. Representative sources include *Cultural Crossing*,[3] *Commisceo*[4], *Cultural Atlas*,[5] and *Expatica*.[6] We then segment the collected pages into sections and subsections, and employ GPT-4o-mini to generate culturally relevant MCQs in both Arabic and English. We acquire $1,000$ samples using this method.

For both methods, two professional linguists independently reviewed the data for correctness, removal of low-quality or trivial questions, and acquisition of proper formatting. All discrepancies were reviewed in consolidation sessions. Finally, we shuffle answer options to minimize positional bias.

The final data for this subtask consists of $2,000$, 500, and $2,000$ questions for the training, dev, and test splits, respectively. The domain and country balance in the test set approximates that of the training data but includes some new entities and less frequent cultural items to test generalization. Samples from Subtask 1 are presented in Table 1.

2.2 Subtask 2: General Islamic Culture

This subtask aims to assess the capacity of LLMs to capture and understand the Islamic culture, which plays a foundational role in Arabic societies. It covers topics such as *Islamic rituals and practices (e.g., prayers and fasting)*, *Quranic knowledge*, *Hadith literature*, *historical developments in Islam*, and *religious holidays*.

[3] https://guide.culturecrossing.net/basics_business_student_details.php

[4] https://www.commisceo-global.com/resources/country-guides/

[5] https://culturalatlas.sbs.com.au/countries

[6] https://www.expatica.com/

2.2.1 Data Collection and Annotation

To enhance topical diversity, we employ two complementary methods to collect Islamic MCQs, yielding a nearly balanced distribution across sources.

Method 1: We create the data based on public Islamic competitions and general questions about Islamic culture using a university book [7]. We acquire 900 samples using this method.

Method 2: We crawl all Islamic articles from *Mawdoo3*, [8] one of the most reputable Arabic content platforms (category: Islam). From this corpus, we randomly select 200 pages and employ GPT-4o-mini to generate diverse MCQs per page. All generated Arabic items are independently reviewed by two professional linguists to verify correctness, eliminate low-quality or trivial content, and ensure proper formatting. Again, all discrepancies are reviewed in consolidation sessions and answer options are subsequently shuffled to reduce positional bias. We acquire $1,000$ samples using this method.

The final data for this subtask consists of 600, 300, and $1,000$ questions for the training, dev, and test splits, respectively.

Samples from Subtask 2 are presented in Table 2.

3 Rules and Evaluation

This section outlines the rules we establish for participation and the methods we employ for the evaluation of submissions. We design the framework to rigorously and fairly assess the intrinsic cultural and Islamic knowledge of the submitted language models.

Reproducibility Teams are instructed to document their data preprocessing, model architecture, external resources, prompt templates, and inference-time strategies.

3.1 Participation and Submission Guidelines

The primary objective of the shared task is to assess the internalized knowledge of LLMs. To ensure the evaluation focuses on the models' core understanding rather than their ability to query external information sources, we established two fundamental rules.

First, the use of systems with real-time data retrieval capabilities, such as retrieval-augmented generation (RAG) or live internet access, is strictly

[7] The Question Bank for Islamic Culture form Al-Balqa Applied University (BAU)

[8] https://mawdoo3.com

Split	Answer	D	C	B	A	Question
train	D	٢٣ سبتمبر	١ يناير	١٤ فبراير	٣٠ نوفمبر	متى يحتفل السعوديون باليوم الوطني؟
train	D	23 September	1 January	14 February	30 November	When do Saudis celebrate National Day?
train	B	الصداقة	الحب	الحزن	الفرح	ماذا ترمز زهور البنفسج في الثقافة الجزائرية؟
train	B	Friendship	Love	Sadness	Joy	What do violets symbolize in Algerian culture?
dev	D	طقس مهم بعد الزفاف	عملية خاصة بالعروس	نوع من الطعام	مباراة تقليدية	ما هو الجرتق في الزواج السوداني؟
dev	D	An important post-wedding ritual	A special process for the bride	A type of food	A traditional contest	What is "Jertiq" in Sudanese weddings?
test	A	الفرنسية	البرتغالية	الإيطالية	الأمازيغية	ما هي اللغة الأم لبعض المغاربة بجانب العربية؟
test	A	French	Portuguese	Italian	Amazigh	What is the mother tongue of some Moroccans besides Arabic?
test	A	المجبوس	الثريد	الكسكس	الكسرة	ما هو الطبق الموريتاني الأكثر شيوعاً في العالم العربي؟
test	A	Majboos	Thareed	Couscous	Kesra	What is the most common Mauritanian dish in the Arab world?

Table 1: Sample questions with their splits, correct answers, and options (A–D) for Subtask 1.

Split	Answer	D	C	B	A	Question
dev	B	رحمة لا تتعلق بالله	رحمة محدودة	رحمة تشمل جميع المخلوقات	رحمة خاصة بالمؤمنين	أي من العبارات التالية تعبر عن معنى اسم الرحمن؟
dev	B	Mercy unrelated to God	Limited mercy	Mercy that includes all creatures	Mercy specific to believers	Which of the following phrases expresses the meaning of the name "Ar-Rahman"?
train	A	عثمان بن عفان	معاذ بن جبل	عبد الرحمن بن عوف	أبو عبيدة بن الجراح	من هو الصحابي الذي لُقّب بأمين هذه الأمة؟
train	A	Uthman ibn Affan	Muadh ibn Jabal	Abdur Rahman ibn Awf	Abu Ubaidah ibn al-Jarrah	Which companion was nicknamed "the trustworthy of this nation"?
test	D	رفيدة بنت سعد الأسلمية رضي الله عنها	حفصة بنت عمر رضي الله عنها	عائشة بنت أبي بكر رضي الله عنها	أم أيمن رضي الله عنها	من هي أول ممرضة في الإسلام؟
test	D	Rufaidah bint Sa'd al-Aslamiyyah (may Allah be pleased with her)	Hafsa bint Umar (may Allah be pleased with her)	Aisha bint Abu Bakr (may Allah be pleased with her)	Umm Ayman (may Allah be pleased with her)	Who was the first nurse in Islam?

Table 2: Sample questions with their splits, correct answers, and options (A–D) for Subtask 2.

prohibited. This ensures that the task does not become a trivial information retrieval challenge. Consequently, submissions are limited to the following format:

1. **Model Weights:** Participants are required to submit the fine-tuned weights of a decoder-only generative language model.
2. **Parameter Limit:** To maintain computational fairness across all participants, the submitted models are constrained to a maximum size of 13 billion (13B) parameters.
3. **Secure Submission:** For privacy and accessibility, participants are instructed to host their models in a private repository on Hugging Face. The final submission consists of the repository ID and a fine-grained access token that provided the organizers with read-only access to the model for evaluation.

Second, to ensure integrity of the results, the test set was held out and remained private to the organizers[9]. This blind evaluation protocol guarantees that no participant had prior access to the test data, enabling a realistic assessment of each model's generalization capabilities in the domain of Arabic cultural and Islamic awareness.

3.2 Evaluation Method

To evaluate the MCQs from our test set, we adopt the likelihood-based method commonly used in frameworks like the EleutherAI Language Model Evaluation Harness (Biderman et al., 2024). This approach assesses a model's understanding by measuring how likely it is to choose the correct answer label after being presented with the question and all possible choices, rather than relying on generative decoding. We develop an in-house script to implement this method and share it with participants during the development phase to ensure they understand how their submissions would be evaluated.

3.2.1 Likelihood-based MCQ Evaluation

For each MCQ item, we construct a prompt that includes the question followed by the list of choices, each prefixed with a letter (e.g., A, B, C, D). The prompt is structured as follows:

[9]Test data was shared only after the leaderboard announcement.

<Question>
A. <Choice 1>
B. <Choice 2>
C. <Choice 3>
D. <Choice 4>
Answer:

The model's task is to determine which choice label (A, B, C, or D) is the most probable continuation of the prompt. We calculate the likelihood of the model generating each choice label. This approach of scoring only the label, rather than the full text of the choice, ensures the evaluation is not biased by the length of the answer strings.

Specifically, for a given question prompt P and a set of possible choices $\{C_1, C_2, \ldots, C_n\}$, we create n distinct sequences. Each sequence is formed by concatenating the prompt P with the text corresponding to one of the choice labels (e.g., " A", " B", etc.).

Let the tokens for the choice label C_i be $c_{i,1}, c_{i,2}, \ldots, c_{i,k}$. The score for choice C_i is its log-likelihood, calculated as the sum of the conditional log-probabilities of its tokens given the prompt and the preceding tokens of the choice label:

$$\text{score}(C_i) = \log p(C_i|P) = \sum_{j=1}^{k} \log p(c_{i,j}|P, c_{i,1}, \ldots, c_{i,j-1}) \quad (1)$$

These log-likelihood scores are computed for all choices. To select the model's final answer, we normalize these scores into a probability distribution using the softmax function:

$$\text{P}(C_i) = \frac{e^{\text{score}(C_i)}}{\sum_{j=1}^{n} e^{\text{score}(C_j)}}$$

The choice with the highest resulting probability is selected as the model's prediction.

3.2.2 Evaluation Metric

The final performance is measured using **accuracy**. The model's predicted label is compared against the ground-truth label for each question. The overall accuracy is the percentage of questions the model answered correctly:

$$\text{Accuracy} = N_{correct}/N_{total}$$

Where $N_{correct}$ is number of correct predictions and N_{total} is total number of questions. This method provides a robust measure of a model's preference for the correct answer among the given options. The entire process, from prompt construction to likelihood calculation and accuracy scoring, was automated using the provided evaluation script.

4 Shared Task Teams & Results

4.1 Participating Teams

The *PalmX 2025* shared task attracted significant interest from the research community, with 26 teams registering for Subtask 1 (General Culture) and 19 teams registering for Subtask 2 (General Islamic). However, actual participation rates varied between the subtasks. For Subtask 1, eleven teams successfully submitted their models or systems. Among these submissions, two were subsequently rejected due to non-compliance with the established submission guidelines, resulting in nine valid submissions that were evaluated and ranked. For Subtask 2, six teams submitted their approaches, all of which met the submission requirements and were successfully evaluated. Notably, five teams participated in both subtasks, demonstrating their commitment to addressing both domains. This cross-participation allowed for interesting comparisons of team performance across different cultural contexts and question types. Table 3 provides a comprehensive overview of all participating teams, including their subtask involvement and institutional affiliations.

4.2 Baselines

We established baseline performance (accuracy) using the NileChat-3B model (Mekki et al., 2025) without any task-specific fine-tuning (zero-shot):

- **Subtask 1 (General Culture)**: 70.00% on dev and 67.55% on test.
- **Subtask 2 (General Islamic)**: 64.00% on dev and 75.12% on test.

4.3 Shared Task Results

The shared task attracted strong participation, with many teams significantly outperforming the baseline models. This outcome highlights the value of applying task-specific fine-tuning and data augmentation techniques.

Subtask 1: General Arabic Culture

The general culture subtask was exceptionally competitive, with the top four teams finishing within a

Team Name	Affiliation	Subtask 1 (Arabic)	Subtask 2 (Islamic)
HAI (Hossain and Afli, 2025)	ADAPT, MTU	✓	✓
RGIPT (Chatwal and Mishra, 2025)	Rajiv Gandhi Inst. of Petroleum Tech.	✓	
AYA (Tajrin et al., 2025)	Qatar Computing Research Institute	✓	✓
Phoenix (Atou et al., 2025)	Mohammed VI Polytechnic University	✓	✓
CultranAI (Chatwal and Mishra, 2025)	Hamad Bin Khalifa University	✓	
ISL-NLP (Gomaa and Elmadany, 2025)	AAST	✓	
MarsadLab (Biswas et al., 2025)	Hamad Bin Khalifa University	✓	✓
Hamyaria (Al-Dhabyani and Alsayadi, 2025)	Hadhramout Univ., Cairo Univ.	✓	✓
Star (Elrefai et al., 2025)	Alexandria University	✓	
TarnishedLab*	UIR		✓

Table 3: Participating teams, their affiliations, and their subtasks in PalmX 2025. A checkmark (✓) indicates participation in the corresponding subtask. Teams marked with * did not submit their system description papers.

Rank	Name	Accuracy	Model	Size	Dataset(s)	Methodology (concise)
1st	**ADAPT-MTU HAI**	**72.15%**	*NileChat-3B*	**3B**	PalmX (train)	Full fine-tune (CLM); 3 ep; full-prompt supervision.
2nd	**RGIPT**	**71.65%**	*NileChat-3B*	**3B**	PalmX	LoRA (r=16, α=32); 3 ep; no external data.
3rd	**AYA**	**71.45%**	*Fanar-1-9B-Instruct*	**9B**	PalmX Cultural & Islamic (train)	LoRA fine-tune; 3 ep; paraphrase aug (no dev gain).
4th	**Phoenix**	**71.35%**	*Fanar-1-9B-Instruct*	**9B**	PalmX Cultural (train) + LLM aug	FT Fanar-9B with Gemini-based paraphrase/sample/dataset aug (∼18k added).
5th	**CultranAI**	**70.50%**	*Fanar-1-9B-Instruct*	**9B**	PalmX (train+dev), PalmX (test), NativQA MCQs (22k)	LoRA fine-tune; added 22k curated MCQs; train on combined set.
6th	**ISL**	**67.60%**	*NileChat-3B*	**3B**	PalmX Cultural (train)	Retrieval-augmented (Gemini) + PEFT; partial unfreeze of projections.
7th	**MarsadLabM**	**67.55%**	*Qwen2.5-7B-Instruct*	**7B**	PalmX Cultural (train)	LoRA on Qwen2.5-7B (r=16, α=32); 3 ep; 4-bit quantization.
–	**Baseline (ours)**	**67.55%**	*NileChat-3B*	**3B**	–	Zero-shot (no fine-tuning).
8th	**Hamyaria**	**65.90%**	*Qwen2.5-3B-Instruct*	**3B**	PalmX + shuffle/paraphrase aug	Augment (answer shuffle + Fanar-9B paraphrase) + FT Qwen2.5-3B; 5 ep.
9th	**Star**	**64.05%**	*Qwen3-4B*	**4B**	Arabic culture corpus (Wikipedia) + PalmX Cultural	Continual pretrain on Arabic culture corpus; SFT on PalmX with PEFT/LoRA.

Table 4: Approaches for Subtask 1: General Arabic Culture.

narrow 1% accuracy margin.

- **First Place:** The **ADAPT-MTU HAI Team** achieved the top score of **72.15%**. Their strategy involved a full fine-tuning of the NileChat-3B model using a causal language modeling (CLM) objective. They trained the model for three epochs, supervising it over the complete prompt to maximize learning.

- **Second Place:** The **RGIPT Team** secured second place with **71.65%** accuracy. They also used the NileChat-3B model but opted for a parameter-efficient Low-Rank Adaptation (LoRA) approach (r=16, alpha=32). Their model was trained for three epochs on prompt-response pairs derived solely from the provided training data.

- **Third Place:** The **AYA Team** finished third with **71.45%** accuracy. They utilized the larger Fanar-1-9B-Instruct model and experimented with data augmentation by paraphrasing questions with other LLMs. However, this augmentation did not lead to improved performance on the development set, so their final result was based on LoRA fine-tuning for three epochs with a maximum sequence length of 512.

Subtask 2: General Islamic Culture

In the Islamic knowledge subtask, the performance differences between teams were more distinct.

- **First Place:** The **AYA Team** ranked first with a commanding accuracy of **84.22%**, using the ALLaM-7B-Instruct model. Their success stemmed from a combination of effective data augmentation and efficient LoRA fine-tuning, a strategy that proved more successful in the Islamic domain than in the general culture subtask.

- **Second Place:** The **Phoenix Team** took second place with **83.82%** accuracy, also employing the ALLaM-7B-Instruct model. They developed "PhoenixIs" by focusing on paraphrasing for data augmentation and notably included the cultural PalmX dataset in their

Rank	Name	Accuracy	Model	Size	Dataset(s)	Methodology (concise)
1st	**AYA**	**84.22%**	*ALLaM-7B-Instruct*	**7B**	PalmX Islamic (train) + aug	LoRA fine-tune on ALLaM-7B with data augmentation.
2nd	**Phoenix**	**83.82%**	*ALLaM-7B-Instruct*	**7B**	PalmX Islamic (train) + aug + PalmX Cultural	FT ALLaM-7B; paraphrase-focused aug; +Cultural data (∼4.5k).
3rd	**ADAPT-MTU HAI**	**82.52%**	*ALLaM-7B-Instruct-preview*	**7B**	PalmX Cultural & Islamic (train)	LoRA (8-bit load); add CoT cue "Let's think step-by-step".
–	**Baseline (ours)**	**75.12%**	*NileChat-3B*	**3B**	–	Zero-shot (no fine-tuning).
4th	**MarsadLabM**	**74.13%**	*Qwen2.5-7B-Instruct*	**7B**	PalmX Cultural	LoRA on Qwen2.5-7B; 3 ep; 4-bit quantization.
5th	**Hamyaria**	**70.83%**	*Qwen2.5-3B-Instruct*	**3B**	PalmX (no aug)	Plain fine-tune on original set; 10 ep.
6th	**TarnishedLab**	**62.84%**	*Qwen2.5-3B-Instruct*	–	–	–

Table 5: Approaches for Subtask 2: General Islamic Culture.

fine-tuning mixture, which expanded their training data to 4,500 questions.

- **Third Place:** The **ADAPT-MTU HAI Team** earned third place with **82.52%** accuracy using the ALLaM-7B-Instruct-preview model. They applied parameter-efficient fine-tuning (LoRA) to an 8-bit loaded version of the model and incorporated reasoning cues like"*Let's think step-by-step*" into their training instances to encourage more structured outputs.

Tables 4 and 5 display the full results for Subtasks 1 and 2, respectively, and briefly describe the system submissions provided by participants, including the backbone models used and their corresponding sizes.

5 Discussion

The results of this shared task provide valuable insights into the current state of Arabic cultural and Islamic knowledge Q&A, revealing several key findings about model performance, methodological approaches, and domain-specific challenges. We discuss a number of these insights here.

5.1 Performance Analysis

The competition demonstrated that task-specific fine-tuning significantly improves performance over baseline models. Most participating teams exceeded the NileChat-3B baseline (67.55% for culture, 75.12% for Islamic), with top performers achieving substantial improvements of 4.6% and 9.1% for Subtasks 1 and 2, respectively. Notably, the Islamic knowledge subtask showed higher overall accuracy scores, with the winning team reaching 84.22% compared to 72.15% for the cultural subtask. This performance difference suggests that Islamic knowledge questions may have more structured, canonical answers compared to the broader host of cultural domains.

5.2 Methodological Insights

Several key methodological trends emerged from the approaches employed by participating teams as we highlight next.

Model selection. Teams favored Arabic-centric models, with NileChat-3B, ALLaM-7B-Instruct, and Fanar-1-9B-Instruct being the most popular choices. Notably, larger models did not necessarily guarantee better performance. This is evidenced by the HAI and RGIPT teams winning the first and second place in subtask 1, respectively, using the smaller NileChat-3B model through effective (parameter-efficient) fine-tuning.

Parameter-efficient fine-tuning. LoRA emerged as the dominant fine-tuning strategy across teams, demonstrating its effectiveness. The success of LoRA-based approaches suggests that efficient adaptation methods can achieve competitive results while maintaining computational feasibility.

Data augmentation strategies. The impact of data augmentation varied significantly between subtasks. While the AYA Team's augmentation approach proved crucial for their success in the Islamic subtask, the same team reported that augmentation did not improve performance on the cultural development set. This suggests that augmentation effectiveness is highly domain- and data-dependent and requires careful study.

Cross-task learning. Teams participating in both subtasks showed varied success patterns. The ADAPT-MTU HAI Team achieved top performance in the cultural subtask but placed third in Islamic questions, while the AYA Team demonstrated the opposite pattern. This indicates that domain expertise and task-specific optimization

are crucial factors.

5.3 Domain-Specific Challenges

The performance gap between the two subtasks highlights distinct challenges in Arabic cultural versus Islamic knowledge representation, as follows:

Cultural Knowledge Complexity: The tighter competition in Subtask 1 (top four teams within 1%) suggests that cultural knowledge questions present more nuanced challenges. Cultural information spans diverse topics, regions, and interpretations, making it inherently more complex to model and evaluate.

Islamic Knowledge Structure: The higher accuracies and clearer performance hierarchy in Subtask 2 indicate that Islamic knowledge questions may be slightly less challenging due to being more structured and based on canonical sources and established scholarly consensus. This makes these questions more amenable to current language modeling approaches.

5.4 Technical Innovations

Several technical contributions stood out among the participating teams:

The ADAPT-MTU HAI Team's use of reasoning cues ("Let's think step-by-step") represents an interesting application of chain-of-thought prompting to Arabic cultural domains. The Phoenix team's comprehensive augmentation strategy, exploring paraphrasing, sample-based, and dataset-based approaches, provides valuable insights for future data augmentation research in Arabic NLP.

The ISL-Team's context-aware approach, combining external knowledge retrieval with instruction-based fine-tuning, demonstrates the potential of hybrid architectures for knowledge-intensive tasks in Arabic.

6 Conclusion

The PalmX 2025 Shared Task establishes the first standardized benchmark for evaluating Arabic and Islamic cultural competence in LLMs. Our evaluation framework revealed key insights: task-specific fine-tuning substantially improves performance over baselines, with parameter-efficient approaches (LoRA) emerging as the dominant methodology. The performance gap between cultural (72.15% best) and Islamic knowledge (84.22% best) subtasks suggests domain-specific challenges, with Islamic questions potentially benefiting from more structured canonical sources. Overall, models still struggle on both general cultural and general Islamic knowledge, motivating us to continue to offer the shared task in the future.

Strong community participation from diverse international teams demonstrates the critical need for culturally aligned Arabic LLMs. While participating teams achieved significant improvements over baselines, the modest absolute scores highlight substantial remaining challenges in achieving true cultural competence. PalmX 2025 benchmark provides a foundation for systematic progress tracking and comparison in Arabic cultural AI, driving development of more inclusive language technologies for Arabic-speaking communities worldwide.

Limitations

Several important limitations should be acknowledged:

- **Dataset Imbalances**: PalmX includes data from 22 Arab countries, but the distribution of questions is uneven. Countries like Iraq and Algeria are underrepresented, as shown in the appendix B, while others are overrepresented. This imbalance may bias the models toward frequently represented cultures and limit their generalization to underrepresented communities. Future releases should focus on targeted data collection to improve country-level representation.

- **Evaluation Constraints**: The benchmark is limited to multiple-choice questions in MSA. While this design ensures clarity, fairness, and reproducibility, it does not capture broader aspects of cultural and linguistic competence, such as open-ended reasoning, interactive dialogue, or sensitivity to dialectal variation.

- **Language and Cultural Scope**: PalmX is designed with a focus on Arabic cultural and Islamic knowledge expressed in MSA. However, Arabic speaking communities are linguistically and culturally diverse, with extensive dialectal variation and localized traditions that MSA-based questions may not fully capture. Moreover, Islamic cultural practices extend far beyond the Arab world, but these dimensions are not addressed in a comprehensive way. Therefore, PalmX should be viewed as an initial step toward assessing

alignment with Arabic and Islamic cultural contexts, rather than as a complete evaluation of all cultural settings.

- **Quality and Methodology**: Although there were several levels of human review, covering all the dataset used, small sections of the dataset were generated or reformulated using LLMs (see Section 2.1.1), which could introduce subtle artifacts or stylistic biases. Furthermore, the topic classification used for dataset analysis (Appendix B) partially depended on automated methods that have imperfect accuracy. These factors may impact both the reliability of item difficulty and the interpretability of model performance.

Ethical Considerations

The development and evaluation of culturally-aware language models raises several ethical considerations that we have carefully addressed in PalmX 2025:

- **Cultural Representation and Bias**: While we strive for balanced representation across all 22 Arab countries, acknowledged geographical imbalances may inadvertently favor certain cultural perspectives over others. We mitigate this through transparent reporting of data distributions and encourage future work to address underrepresented regions.

- **Religious Sensitivity**: Questions involving Islamic knowledge require particular care to avoid misrepresentation or offense. All religious content was reviewed by qualified experts, and we acknowledge that legitimate scholarly disagreements exist on certain topics. The evaluation framework focuses on widely accepted knowledge rather than contentious interpretations.

- **Data Privacy and Consent**: All data sources used are publicly available or properly licensed. Web-crawled content was limited to public educational resources, and no personal information was collected or used in dataset construction.

- **Model Deployment Implications**: While this benchmark evaluates cultural competence, we emphasize that high performance does not guarantee appropriate real-world deployment. Cultural sensitivity extends beyond factual knowledge to include contextual appropriateness, respect for cultural values, and awareness of power dynamics.

- **Overfitting to Benchmarks**: The competitive nature of shared tasks may unintentionally promote overfitting to benchmark scores rather than fostering genuine cultural competence. As such, it is necessary to stress the importance of engaging with native speakers and experts in addition to the use of benchmarks.

- **Potential Misuse**: A benchmark that evaluates alignment to specific cultural and religious norms could be misapplied in harmful contexts. For instance, it could be used to justify censorship, surveillance, or exclusionary practices. The benchmark data and evaluation methods are designed for research purposes. We encourage responsible use and caution against deploying systems without adequate safeguards for cultural sensitivity and community feedback.

Acknowledgments

Muhammad Abdul-Mageed acknowledges 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,[10] and UBC Advanced Research Computing-Sockeye.[11]

References

Muhammad Abdul-Mageed, Abdelrahim Elmadany, Alcides Inciarte, Md Tawkat Islam Khondaker, and 1 others. 2023. Jasmine: Arabic gpt models for few-shot learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16721–16744.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit

[10] https://alliancecan.ca

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

Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Emran Al-Buraihy, Dan Wang, Tariq Hussain, Razaz Waheeb Attar, Ahmad Ali AlZubi, Khalid Zaman, and Zengkang Gan. 2025. Aratraditions10k bridging cultures with a comprehensive dataset for enhanced cross lingual image annotation retrieval and tagging. *Scientific Reports*, 15(1):19624.

Walid Al-Dhabyani and Hamzah A. Alsayadi. 2025. Hamyaria at PalmX2025: Leveraging Large Language Models to Improve Arabic Multiple-Choice Questions in Cultural and Islamic Domains. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, and 1 others. 2025a. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

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

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12753–12776, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Houdaifa Atou, Issam Ait Yahia, and Ismail Berrada. 2025. Phoenix at Palmx: Exploring Data Augmentation for Arabic Cultural Question Answering. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Lama Ayash, Hassan Alhuzali, Ashwag Alasmari, and Sultan Aloufi. 2025. Saudiculture: A benchmark for evaluating large language models' cultural competence within saudi arabia. *Journal of King Saud University Computer and Information Sciences*, 37(6):123.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

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. 2024. Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, and 11 others. 2024. Lessons from the trenches on reproducible evaluation of language models. *Preprint*, arXiv:2405.14782.

Md. Rafiul Biswas, Shimaa Ibrahim, Kais Attia, Mabrouka Bessghaier, Firoj Alam, and Wajdi Zaghouani. 2025. MarsadLab at PalmX Shared Task: An LLM Benchmark for Arabic Culture and Islamic Civilization. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Pulkit Chatwal and Santosh Kumar Mishra. 2025. Cultura-Arabica: Probing and Enhancing Arabic Cultural Awareness in Large Language Models via LORA. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Rochelle Choenni and Ekaterina Shutova. 2024. Self-alignment: Improving alignment of cultural values in llms via in-context learning. *arXiv preprint arXiv:2408.16482*.

Eman Elrefai, Esraa Khaled, and Alhassan Ehab. 2025. Star at PalmX 2025: Arabic Cultural Understanding via Targeted Pretraining and Lightweight Finetuning. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Sara Ghaboura, Ahmed Heakl, Omkar Thawakar, Ali Husain Salem Abdulla Alharthi, Ines Riahi, Abduljalil Radman, Jorma Laaksonen, Fahad Shahbaz Khan, Salman Khan, and Rao Muhammad Anwer. 2025. CAMEL-bench: A comprehensive Arabic LMM benchmark. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1970–1980, Albuquerque, New Mexico. Association for Computational Linguistics.

Mohamed Gomaa and Noureldin Elmadany. 2025. ISL-NLP at PalmX 2025: Retrieval-Augmented Fine-Tuning for Arabic Cultural Question Answering. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Shehenaz Hossain and Haithem Afli. 2025. ADAPT–MTU HAI at PalmX 2025: Leveraging Full and Parameter-Efficient LLM Fine-Tuning for Arabic Cultural QA. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

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.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

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

Zhaoming Liu. 2025. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, 3(2):224–244.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. JAWAHER: A multidialectal dataset of Arabic proverbs for LLM benchmarking. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12320–12341, Albuquerque, New Mexico. Association for Computational Linguistics.

Abdellah El Mekki, Houdaifa Atou, Omer Nacar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. Nilechat: Towards linguistically diverse and culturally aware llms for local communities. *arXiv preprint arXiv:2505.18383*.

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.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

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.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. Commonsense reasoning in Arab culture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7695–7710, Vienna, Austria. 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*.

Jannatul Tajrin, Bir Ballav Roy, and Firoj Alam. 2025. AYA at PalmX 2025: Modeling Cultural and Islamic Knowledge in LLMs. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

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

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, and 1 others. 2025. Fanar:

An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Appendices

These appendices provide supplementary material supporting the main findings of this work. The content is organized as follows:

- **A: Literature Review**
  Reviews related work on cultural bias in LLMs, Arabic centric LLMs, and Arabic culturally-Aware datasets and benchmarks.

- **B: Data Analysis**
  This section presents the country-level and topical distributions of both subtasks' datasets.

A Literature Review

Our work is situated at the intersection of several active research areas: the evaluation of cultural biases in LLMs, the development of Arabic-centric models, and the creation of culturally grounded benchmarks.

1.1 Cultural Bias and Alignment in LLMs

The detection, mitigation, and control of cultural bias in LLMs is an expanding research area, seeking to produce generative models that are free of stereotypes and which align with a defined cultural perspective and value framework (Pawar et al., 2025).

Since many LLMs are trained primarily on widely available, high-quality English datasets, they inevitably reflect cultural elements present in those sources (Johnson et al., 2022). Techniques such as fine-tuning and reinforcement learning from human feedback (RLHF) are commonly employed to align such models with a desired value system (Bai et al., 2022; Li et al., 2024); however, this depends on the availability of high-quality instruction data that accurately reflects that system (Liu, 2025). Another approach is to use prompting and system roles to enforce a cultural identity (Tao et al., 2024; Choenni and Shutova, 2024).

1.2 Development of Arabic-Centric LLMs

To counter the dominance of English-centric models, significant efforts have been made to develop foundational LLMs for Arabic. Models like JAIS (Sengupta et al., 2023) pioneered a bilingual Arabic-English training strategy to leverage cross-lingual knowledge transfer. The Jasmine (Abdul-Mageed et al., 2023) suite of models was specifically designed to enhance few-shot learning capabilities in Arabic, while the AceGPT project (Huang et al., 2024) introduced a comprehensive localization pipeline, including pretraining, supervised fine-tuning (SFT), and reinforcement learning with a reward model sensitive to local values.

More recent models like ALLAM (Bari et al., 2024) and Fanar (Team et al., 2025) have further advanced Arabic capabilities. NileChat (Mekki et al., 2025), in particular, was developed as a linguistically diverse and culturally aware model specifically tailored for local communities. NileChat proved that it's possible to build a performant 3 billion parameters language model that can represent the Moroccan and Egyptian communities, including their dialects, cultural heritage, and values through controlled-generated synthetic data. While these models represent crucial advancements in Arabic linguistic competence, their evaluations have largely focused on standard NLP tasks (e.g., question answering, summarization) and general knowledge benchmarks like Arabic MMLU. They have not been systematically evaluated on their understanding of deep, country-specific cultural knowledge.

1.3 Arabic Culturally-Aware Datasets and Benchmarks

A growing body of work is dedicated to developing datasets and benchmarks that reflect Arab culture. One of the earliest benchmark efforts is the Arabic Cultural and Value Alignment dataset (Huang et al., 2024), comprising 8.7K yes–no questions synthetically generated by GPT-3.5 Turbo on various topics related to Arab values. AraDiCE-Culture (Mousi et al., 2025) is a fine-grained benchmark designed to assess cultural awareness across the Gulf, Egypt, and the Levant. Jawaher (Magdy et al., 2025) offers 10K multi-dialectal Arabic proverbs to evaluate understanding of cultural nuances through figurative language. ArabCulture (Sadallah et al., 2025) is a manually crafted dataset of 3.5K commonsense reasoning questions covering the cultures of 13 Arab countries across 54 subtopics.

On the other hand, instruction datasets aimed at embedding cultural understanding during model training include CIDAR (Alyafeai et al., 2024), a 10K culturally localized instruction dataset created

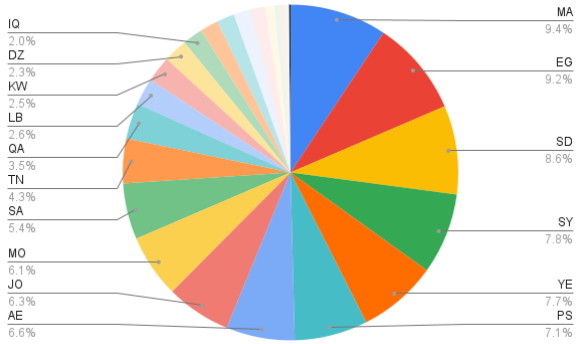

Figure 2: Country distribution of cultural questions in the **training** data.

via machine translation followed by human review, and Palm (Alwajih et al., 2025a), a 17K human-crafted instruction dataset spanning the cultures of the 22 Arab countries. Efforts to support local cultures also include datasets and models such as NileChat (Mekki et al., 2025) for Egyptian and Moroccan dialects, and benchmarks like SaudiCulture (Ayash et al., 2025).

More recently, a focus has emerged on culturally aware Arabic multimodal resources, including Peacock (Alwajih et al., 2024), CamelBench (Ghaboura et al., 2025), AraTraditions10K (Al-Buraihy et al., 2025), and Pearl (Alwajih et al., 2025b).

B Data Analysis

2.1 Subtask 1 Data Analysis

Country distributions of training, development, and test data are shown in Figures 2, 3, and 4. We use ISO 3166 Alpha-2 code for countries[12]. We note that certain countries, such as Iraq (IQ) and Algeria (DZ), are underrepresented across all data splits. In future releases of PalmX, we aim to ensure more balanced country distributions.

Table 6 presents the 15 most frequent topics, which together account for 95% of all test questions, along with illustrative examples. The topics were initially classified using GPT-4o and subsequently consolidated and manually verified. To estimate classification quality, 200 random questions were sampled, yielding an accuracy of 85%. We observe that roughly one-third of the test questions pertain to historical events in Arab countries, such as the dates of revolutions, the founding of political parties, or the birthdates of notable writers.

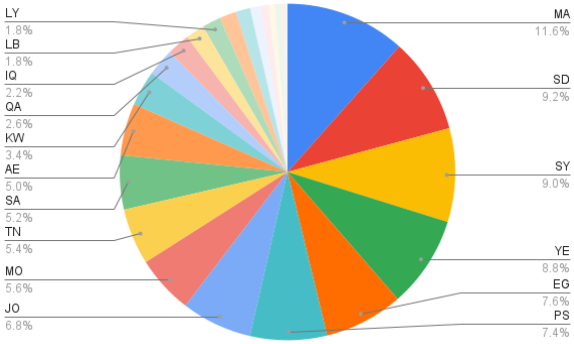

Figure 3: Country distribution of cultural questions in the **development** data.

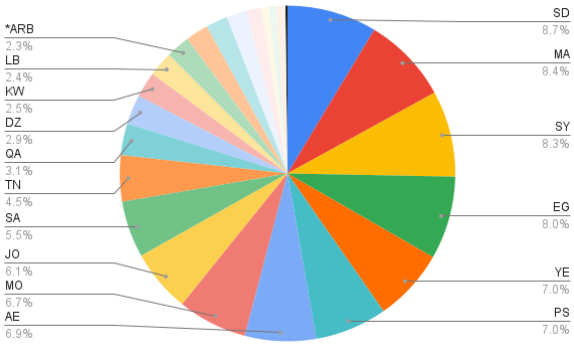

Figure 4: Country distribution of cultural questions in the **test** data. *ARB denotes questions related to Arab culture in general, rather than those tied to a specific country.

[12] https://www.iban.com/country-codes

Topic	Example	%
History	متى تم الاستقلال الجزائري؟ When did Algeria gain independence?	35.2
Geography/Environment	ما هو أكبر الأنهار في سوريا؟ What is the largest river in Syria?	10.0
Food	ما هو طبق البازين في ليبيا؟ What is the Bazin dish in Libya?	7.9
Customs	ما استعمال الحنة في الزواج السوداني؟ What is the use of henna in Sudanese marriage?	7.6
Arts	ما هي أهم فعالية سينمائية في تونس؟ What is the most important cinematic event in Tunisia?	6.0
Sports	ما هي الرياضة الأكثر شعبية في مصر؟ What is the most popular sport in Egypt?	5.1
Literature	متى بدأت الحركة الأدبية الحديثة في قطر؟ When did the modern literary movement begin in Qatar?	4.6
Economics	بماذا تشتهر مدينة بيت لحم في فلسطين من حيث الصناعات؟ What is Bethlehem, Palestine, famous for in terms of industries?	4.6
Religion	ما هو اليوم المقدس للمسلمين في الأسبوع؟ What is the holy day of the week for Muslims?	3.9
Language	ما معنى كلمة «صنطة» في اللهجة العراقية؟ What does the word 'santa' mean in the Iraqi dialect?	2.8
Clothing	ما هو الطربوش المغربي؟ What is the Moroccan fez?	2.4
Education	ما هي اللغة الثانية التي تُعتبر إلزامية في المدارس الكويتية؟ What second language is mandatory in Kuwaiti schools?	1.5
Politics	من يرأس حزب التجمع من أجل موريتانيا (تمام)؟ Who heads the Rally for Mauritania (RMA) party?	1.3
Tourism	ما يميز شاطئ أرتا في جيبوتي؟ What makes Arta Beach in Djibouti special?	1.3
Law	ما هو السن القانوني للتدخين في البحرين؟ What is the legal smoking age in Bahrain?	1.3
Other 10 topics	Technology, Architecture, Medecine, etc.	5.0

Table 6: Topic distribution of the cultural questions (translated to English) in the **test** set.

2.2 Subtask 2 Data Analysis

Table 7 presents the topic distribution along with examples from the test set. Topic labels were predicted using GPT-4o. To estimate accuracy, we sampled 200 questions and found a 91% agreement with manual annotations. Notably, about one-quarter of the questions concern historical events, such as battles, the birthplaces of scholars, or former names of places.

Topic	Example	%
History	أين وقعت معركة اليرموك؟ Where did the Battle of Yarmouk take place?	25.5
Worship	ما إحدى الفوائد المرتبطة بصلاة الفجر؟ What is one of the virtues of Fajr prayer?	18.2
Ethics	ما أحد مظاهر احترام الآخرين في الإسلام؟ What is one of the manifestations of respecting others in Islam?	12.4
Fiqh (Islamic Jurisprudence)	ما مقدار الزكاة الواجبة في المال؟ How much zakat is due on money?	12.3
Quranic Sciences	ما الآية التي تشير إلى انشقاق القمر؟ Which verse refers to the splitting of the moon?	10.3
Aqidah (Islamic theology)	كم عدد أركان الإيمان؟ How many pillars of faith?	9.4
Hadith Sciences	بماذا يتميز الحديث القدسي؟ What distinguishes the Hadith Qudsi?	3.5
Mu'amalat (Islamic Transactions)	ما الحكم العام للبيع بالتقسيط؟ What is the general ruling on installment sales?	2.4
Contemporary Issues	ما أحد مظاهر التطرف الديني؟ What is one manifestation of religious extremism?	2.1
Sirah (Biography of the Prophet)	من الذي صلى بالناس بعد أن اشتد مرض النبي؟ Who led the people in prayer after the Prophet's illness became severe?	2.0
Philosophy	ماذا يعني مفهوم عالميّة الإسلام؟ What does the concept of the universality of Islam mean?	2.0

Table 7: Topic distribution of the Islamic questions (translated to English) in the **test** set