Cultura-Arabica: Probing and Enhancing Arabic Cultural Awareness in Large Language Models via LoRA[*]

Pulkit Chatwal and **Santosh Kumar Mishra**
Rajiv Gandhi Institute of Petroleum Technology, Jais, India
pulkitchatwal@gmail.com and satosh.mishra@rgipt.ac.in

Abstract

Large Language Models (LLMs) have demonstrated impressive multilingual capabilities; however, their reasoning often reflects English-centric perspectives, which can limit accuracy in culture-specific contexts. Arabic, with its diverse dialects, rich historical heritage, and complex socio-cultural norms, presents a particularly challenging setting for such evaluation. To address this gap, we participated in the PalmX 2025 shared task, which benchmarks cultural reasoning in Arabic through multiple-choice questions covering traditions, social norms, history, geography, arts, and dialectal expressions. By applying parameter-efficient adaptation and culturally informed prompt formatting, we aligned model outputs with both linguistic correctness and cultural relevance. Our approach achieved an accuracy of **71.65%**, securing **second place** overall and closely matching the top system. These results demonstrate that targeted adaptation can significantly enhance cultural reasoning in LLMs, paving the way for more culturally aware Artificial Intelligence.

1 Introduction

Large Language Models (LLMs) have transformed natural language processing, excelling in multilingual understanding, reasoning, and text generation (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023). Yet, their reasoning often reflects predominantly English-centric worldviews (Bang et al., 2023; Piqueras and Søgaard, 2022), leading to gaps in interpreting culture-specific knowledge, norms, and perspectives. Cultural reasoning—integrating linguistic comprehension with contextual understanding of traditions, values, and social practices—is essential for fair, contextually appropriate AI systems (Tao et al., 2024).

Arabic, with its diverse dialects, historical depth, and socio-cultural richness, is a particularly challenging testbed. Despite the growth of Arabic NLP resources, most models remain optimized for syntactic and semantic accuracy rather than capturing the implicit socio-cultural knowledge needed to interpret idioms, customs, and worldview-specific references. As Marcus and Davis emphasize, LLMs are powerful pattern recognizers but lack genuine understanding and grounded reasoning, often reproducing correlations without true comprehension (Marcus and Davis, 2019). Recent work also shows that "models tend to exhibit Western bias even when prompted in non-English languages like Arabic" (Naous et al., 2023), underscoring persistent cultural blind spots.

Addressing this challenge requires moving beyond language correctness toward genuine cultural alignment—where models reason in ways consistent with the target community's norms and context. This work examines whether parameter-efficient adaptation can improve the cultural reasoning capabilities of Arabic LLMs, bridging the gap between linguistic competence and culturally grounded intelligence.

2 Related Work

Arabic Natural Language Processing (NLP) has advanced notably in recent years, driven by transformer-based architectures, culturally aligned datasets, and resource-efficient adaptation methods.

AraBERT (Antoun et al., 2020) pioneered Arabic-specific BERT pre-training, achieving state-of-the-art results in sentiment analysis, named entity recognition, and question answering. ARBERT and MARBERT (Abdul-Mageed et al., 2020) extended this to Modern Standard Arabic (MSA) and dialects, accompanied by ARLUE, a benchmark for multi-dialectal understanding. These works underscore the value of Arabic-specific pre-training.

[*] The final fine-tuned model is available at https://huggingface.co/Pulkit-28/PalmQA-3B-Arabic.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 837–842
November 8-9, 2025 ©2025 Association for Computational Linguistics

Culturally grounded datasets have emerged to address linguistic and cultural biases. CIDAR (Alyafeai et al., 2024) is the first open Arabic instruction-tuning dataset curated for cultural relevance, improving alignment of large language models (LLMs) with Arabic norms. Other domain-specific benchmarks include AraSTEM (Mustapha et al., 2024) for STEM knowledge and AlGhafa (Almazrouei et al., 2023) for diverse Arabic MCQs. Beyond Arabic, the *Survey of Cultural Awareness in Language Models* (Pawar et al., 2025) reviews methods for integrating cultural sensitivity into text and multimodal LLMs, with discussion of datasets, benchmarking, and ethics.

Resource-efficient fine-tuning has also gained traction. Low-Rank Adaptation (LoRA) (Hu et al., 2022) reduces trainable parameters while maintaining performance, and Quantized Low-Rank Adaptation (QLoRA)-based adaptation for Arabic (Aryan, 2024) achieves high-quality results with minimal hardware. Parameter-efficient methods have also been applied to dialect identification (Radhakrishnan et al., 2023) with competitive accuracy.

Large-scale Arabic foundation models like Jais and Jais-chat (Sengupta et al., 2023) set records in Arabic reasoning tasks, while LAraBench (Abdelali et al., 2023) offers a comprehensive benchmarking suite for Arabic NLP and speech, revealing gaps between general-purpose and specialized Arabic models. Beyond Arabic, *Beyond English-Centric LLMs* (Zhong et al., 2024) shows multilingual models may rely on multiple latent languages, stressing the need to study internal representation dynamics for better cultural adaptation.

In summary, advances in Arabic NLP arise from the synergy of specialized pre-training, culturally relevant datasets, efficient fine-tuning, and robust benchmarking—together enhancing accuracy, cultural sensitivity, and efficiency in Arabic-focused LLMs.

3 Problem Statement

We participated in the PalmX 2025 shared task (Alwajih et al., 2025), which evaluates large language models (LLMs) on their ability to comprehend and reason about *Arabic general culture*—including traditions, social norms, history, geography, arts, and dialectal variations. Formally, let $\mathcal{Q} = \{q_1, \dots, q_n\}$ be a set of culturally grounded questions in Modern Standard Arabic, each with candidate answers $\mathcal{A}_i$, where exactly one a_i^* is correct. An LLM, modeled as $f_\theta : \mathcal{Q} \to \mathcal{A}$, aims to maximize:

$$\frac{1}{n}\sum_{i=1}^{n} \nVdash\{\hat{a}_i = a_i^*\}.$$

Unlike traditional benchmarks that focus on $P(\hat{a}_i = a_i^* \mid \text{linguistic knowledge})$, this task emphasizes $P(\hat{a}_i = a_i^* \mid \text{linguistic knowledge, cultural knowledge})$, ensuring models are both linguistically accurate and culturally grounded.

4 Dataset

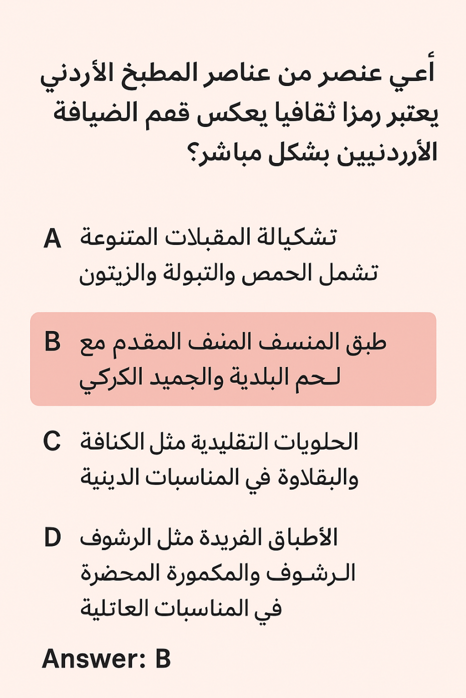
The dataset provided for the PalmX 2025 shared task (Alwajih et al., 2025) was specifically curated to evaluate cultural reasoning capabilities in Arabic LLMs. It consists of three partitions, each balanced across domains such as traditions, social norms, history, geography, arts, and dialectal expressions from diverse Arab countries. The statistics of the dataset are summarized in Table 1, and an example from the training set is shown in Figure 1.

Partition	**Number of MCQs**
Training set	2,000
Development set	500
Blind test set	2,000

Table 1: Summary statistics of the dataset provided for the PalmX 2025 shared task

5 Methodology

This section delineates the modeling framework employed to adapt a large Arabic language model for the PalmX 2025 cultural reasoning task. Recognizing that the task entails selecting the appropriate option from multiple culturally grounded choices, we formulate it as a causal language modeling problem augmented with structured prompts. This approach not only facilitates the model's acquisition of reasoning patterns encompassing both factual and cultural knowledge but also leverages the inherent generative capabilities of language models to handle nuanced, context-dependent queries effectively.



أعي عنصر من عناصر المطبخ الأردني
يعتبر رمزا ثقافيا يعكس قمم الضيافة
الأرردنيين بشكل مباشر؟

A تشكيالة المقبلات المتنوعة
تشمل الحمص والتبولة والزيتون

B طبق المنسف المنف المقدم مع
لـحم البلدية والجميد الكركي

C الحلويات التقليدية مثل الكنافة
والبقلاوة في المناسبات الدينية

D الأطباق الفريدة مثل الرشوف
الـرشـوف والمكمورة المحضرة
في المناسبات العاتلية

Answer: B

Figure 1: Sample culturally grounded MCQ from the training set.

5.1 Base Model

Our framework is built upon the `NileChat-3B` checkpoint (Mekki et al., 2025), a 3B-parameter decoder-only transformer specifically optimized for Arabic dialogue and general-purpose text generation. This model was selected due to its robust pre-training on a diverse corpus of Arabic text, which includes dialectal variations and cultural contexts, making it particularly suitable for tasks requiring deep linguistic and sociocultural understanding. The architecture adheres to an autoregressive GPT-style design, comprising 24 stacked multi-head self-attention layers interspersed with feed-forward blocks, all geared toward efficient left-to-right token prediction. The tokenizer, derived from the same checkpoint, utilizes byte-pair encoding (BPE) with a vocabulary size of 50,000 tokens to accommodate both Arabic and non-Arabic scripts, with the end-of-sequence (EOS) token repurposed as the padding token to ensure seamless compatibility with causal modeling paradigms.

5.2 Fine-Tuning Strategy

For efficient adaptation, we leverage Low-Rank Adaptation (LoRA) (Hu et al., 2022), a parameter-efficient fine-tuning technique that introduces trainable low-rank decomposition matrices into the transformer's projection layers while keeping the original weights frozen. This method allows us to fine-tune fewer than 1% of the total parameters, achieving an optimal trade-off between computational overhead and expressive capacity, which is especially beneficial for resource-constrained environments and multilingual models where full fine-tuning could lead to catastrophic forgetting of pre-trained knowledge. The specific LoRA configuration adopted in this study is as follows:

- Rank (r): 16
- Scaling factor (α): 32
- Target modules: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`
- Dropout rate: 0.05
- Bias: none

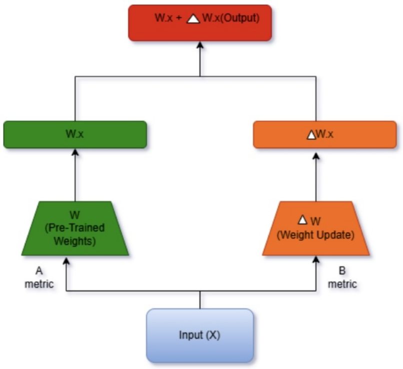


Figure 2: Schematic illustration of the Low-Rank Adaptation (LoRA) mechanism integrated into the fine-tuning process.

5.3 Prompt Formatting

To optimize the model's performance on the multiple-choice cultural reasoning task, each dataset instance is converted into a carefully designed prompt structure. This includes the question stem, four labeled options (A through D), and a clear instruction to generate only the letter of the correct choice. An illustrative prompt template is as follows:

```
Question: [Question text]
Options:
A) [Option A]
B) [Option B]
C) [Option C]
D) [Option D]
Output only the correct letter:
```

This structured format reduces output variability during inference, promotes focused discriminative reasoning by the model, and ensures tight alignment between the training objective and prevalent evaluation paradigms in cultural reasoning, such as zero-shot or few-shot settings.

5.4 Training Procedure

The fine-tuning process is executed via the `Transformers` library's `Trainer` API, incorporating mixed-precision training in `bfloat16` to enhance computational efficiency and reduce memory footprint. We utilize a per-device batch size of 2 on a single NVIDIA A100 GPU, augmented by gradient accumulation across 4 steps, resulting in an effective batch size of 8. A fixed learning rate of 2×10^{-4} is applied without scheduling, with training spanning three epochs to balance convergence and overfitting prevention. The `DataCollatorForLanguageModeling` is configured with `mlm=False` to uphold the causal autoregressive training objective, ensuring that the model learns to generate responses conditioned on the full prompt context. Throughout training, we monitor validation loss to confirm generalization to unseen cultural reasoning examples.

5.5 Adapter Merging and Deployment

Upon completion of fine-tuning, the LoRA adapters are integrated into the base model weights through the `merge_and_unload()` procedure, yielding a consolidated checkpoint devoid of external dependencies and maintaining the original model's inference speed. This merging step is crucial for production environments, as it eliminates the need for additional adapter loading during deployment. The resultant model, designated as `NileChat-3B-Arabic-QA-Merged-v2`, is primed for seamless inference and deployment in practical applications, such as interactive cultural education tools or multilingual question-answering systems.

6 Results

6.1 Evaluation

The official metric for the PalmX 2025 shared task was *accuracy*, measuring the proportion of correct predictions across all test questions. Given a test set of N questions, accuracy is calculated as:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N}\ (\hat{y}_i = y_i)}{N} \times 100\%, \qquad (1)$$

where $\hat{y}_i$ denotes the predicted answer for question i, y_i represents the gold standard label, and $(\cdot)$ is the truth indicator returning 1 if the argument is true and 0 otherwise. This metric equally weights all questions, ensuring that performance reflects general reasoning capabilities rather than domain-specific biases.

6.2 Leaderboard Performance

Our system obtained an overall accuracy of **71.65%**, securing the **second rank** among all participating teams. This performance demonstrates that our parameter-efficient LoRA fine-tuning method can effectively adapt a large Arabic LLM to culturally grounded multiple-choice reasoning with limited task-specific data.

Rank	Team	Score (%)
1	HAI research group	72.15
2	**Our Result**	**71.65**
3	AYA_Team	71.45
4	Phoenix	71.35
5	CultranAI	70.50
6	ISL-NLP	67.60
7	Rafiul Biswas	67.55
8	Hamyaria	65.90
9	Star	64.05

Table 2: Leaderboard results from the PalmX 2025 shared task.

6.3 Discussion

The narrow margin between the top three teams—less than one percentage point—indicates that small architectural or fine-tuning choices can substantially influence outcomes in culturally nuanced reasoning tasks. Our approach's ability to match and even surpass larger-scale fine-tuning efforts highlights the efficiency of targeted LoRA adaptation for Arabic cultural QA, while suggesting broader implications for resource-efficient multilingual NLP.

7 Future Work

Promising directions for extending this work include adapting the proposed framework to other low-resource languages, thereby assessing its efficacy in cross-lingual cultural reasoning tasks. Furthermore, integrating multimodal capabilities—such as fine-tuning on Visual Question Answering (VQA) datasets enriched with culturally pertinent images—could substantially improve model performance by synergistically combining visual and textual cues for more nuanced cultural understanding.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, et al. 2023. Larabench: Benchmarking arabic ai with large language models. *arXiv preprint arXiv:2305.14982*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, et al. 2023. Alghafa evaluation benchmark for arabic language models. In *Proceedings of ArabicNLP 2023*, pages 244–275.

Fakhraddin Alwajih, Abdellah El Mekki, Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed, and Muhammad Abdul-Mageed. 2025. PalmX 2025: The First Shared Task on Benchmarking LLMs on Arabic and Islamic Culture. 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.

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

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Prakash Aryan. 2024. Resource-aware arabic llm creation: Model adaptation, integration, and multidomain testing. In *International Conference on Advanced Network Technologies and Intelligent Computing*, pages 415–434. Springer.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Gary Marcus and Ernest Davis. 2019. *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

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.

Ahmad Mustapha, Hadi Al-Khansa, Hadi Al-Mubasher, Aya Mourad, Ranam Hamoud, Hasan El-Husseini, Marwah Al-Sakkaf, and Mariette Awad. 2024. Arastem: A native arabic multiple choice question benchmark for evaluating llms knowledge in stem subjects. *arXiv preprint arXiv:2501.00559*.

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.

Laura Cabello Piqueras and Anders Søgaard. 2022. Are pretrained multilingual models equally fair across languages? *arXiv preprint arXiv:2210.05457*.

Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Narsis A Kiani, David Gomez-Cabrero, and Jesper N Tegner. 2023. A parameter-efficient learning approach to arabic dialect identification with pre-trained general-purpose speech model. *arXiv preprint arXiv:2305.11244*.

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

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*.