Star at PalmX 2025 Shared Task: Arabic Cultural Understanding via Targeted Pretraining and Lightweight Fine-tuning

Eman Elrefai[1]
[1]Alexandria University
eman.lotfy.elrefai@gmail.com

Esraa Khaled[2]
[2]Cairo University
esraa.k.fouad@gmail.com

Alhassan Ehab[3]
[3]Minia University
alhassanehab186@gmail.com

Abstract

We present a two-stage framework for enhancing Arabic cultural understanding in small language models, specifically designed for PalmX 2025(Alwajih et al., 2025) Shared Task 1: General Culture Evaluation. Our approach combines continuous pretraining on a culturally-enriched Arabic corpus spanning 10 Arab countries and different cultural domains, followed by supervised fine-tuning on cultural question-answering data. Using Parameter-Efficient Fine-Tuning (PEFT) (Zhang et al., 2025) with LoRA on the Qwen3-4B base model, we achieve 74% accuracy on the development set and 64% on the blind test set, ranking our team ninth in the competition. Our system demonstrates the effectiveness of targeted cultural pretraining for improving Arabic language models' cultural competency while maintaining computational efficiency.

1 Introduction

Arabic cultural understanding represents a critical challenge in natural language processing, as existing large language models often lack the nuanced cultural knowledge necessary to serve Arabic-speaking communities effectively. The PalmX 2025 Shared Task 1 focuses on evaluating models' ability to understand and reason about Arabic cultural concepts, traditions, and knowledge across diverse Arab regions.

Our main system strategy employs a two-stage training paradigm: (1) continuous pretraining on culturally-diverse Arabic content to build foundational cultural knowledge. (2) supervised fine-tuning on structured cultural question-answering data to enhance reasoning capabilities. This approach addresses the fundamental challenge of cultural representation in language models while maintaining computational efficiency through parameter-efficient techniques.

Key findings from our work include achieving competitive performance (74% development accuracy, 64% test accuracy) while using only 4B parameters, demonstrating that targeted cultural pretraining significantly improves performance over baseline models, and identifying that multi-domain cultural coverage is essential for robust cultural understanding. The main challenge discovered was balancing broad cultural coverage with deep domain-specific knowledge within computational constraints.

2 Literature Review

The PalmX 2025 Subtask 1 presents a multiple-choice question-answering challenge focused on Arabic cultural knowledge. The input consists of cultural questions in Modern Standard Arabic (MSA) with four possible answers (A, B, C, D), and the output is the correct answer choice.

Example:

Question:
ما هي العاصمة التاريخية للدولة الأموية؟
Choices:
A.بغداد B.دمشق C.القاهرة D.مكة
Answer: B

2.1 Dataset Details

Our pretraining corpus was constructed from Arabic Wikipedia articles covering 10 Arab countries (Bahrain, Egypt, UAE, Iraq, Kuwait, Jordan, Lebanon, Palestine, Syria, Saudi Arabia) and cultural domains like (Media, Sport, Transport, Healthcare, Education, Religion, Economy, History, Festivals, Tourism).

The coverage was limited to these 10 countries due to data availability and quality constraints: some Arab countries had very limited or incomplete Wikipedia content across the chosen domains, which would have introduced imbalance and sparsity into the corpus. Focusing on countries with richer and more representative cultural data ensured

Proceedings of The Third Arabic Natural Language Processing Conference, pages 825–829
November 8-9, 2025 ©2025 Association for Computational Linguistics

both consistency and reliability of the pretraining resource.

The dataset for instruction fine tuning stage comprises cultural questions covering various aspects of Arab heritage, including history, literature, traditions, geography, and social customs. The training set contains 2,000 examples, with a development set of 500 examples for validation. Questions are formulated in MSA and span knowledge from multiple Arab countries and cultural domains.

2.2 Related Work

Previous work in Arabic NLP has focused primarily on general language understanding tasks like (El Mekki et al., 2025; Bari et al., 2024; Sengupta et al., 2023). Cultural understanding in language models has been explored for various languages (Pawar et al., 2025; Nayak et al., 2024), but limited work exists specifically for Arabic cultural knowledge. Our work bridges this gap by combining cultural corpus pretraining with parameter-efficient fine-tuning (LoRA/PEFT) to enhance cultural awareness in small LMs. To the best of our knowledge, this is the first contribution focusing specifically on Arabic cultural evaluation within the PalmX framework.

3 System Overview

3.1 Architecture

Our system builds upon the Qwen3-4B (Yang et al., 2025) base model, selected for its strong multilingual capabilities and computational efficiency. We employ Low-Rank Adaptation (LoRA)(Singhapoo et al., 2025) for parameter-efficient fine-tuning, enabling effective adaptation while minimizing computational overhead.

The LoRA adaptation is applied to multiple attention and feed-forward layers:

$$h = W_0 x + \Delta W x = W_0 x + BAx \qquad (1)$$

where W_0 represents the frozen pre-trained weights, $\Delta W = BA$ is the low-rank adaptation with matrices $B \in R^{d\times r}$ and $A \in R^{r\times d}$, and $r \ll d$ is the rank.

3.2 Two-Stage Training Framework

3.2.1 Stage 1: Cultural Pretraining

We perform continuous pretraining (Tack et al., 2025)on a curated Arabic cultural corpus to inject domain-specific knowledge into the model. The pretraining objective follows the standard causal language modeling loss:

$$\mathcal{L}_{pretrain} = -\sum_{i=1}^{T} \log P(x_i|x_{<i};\theta) \qquad (2)$$

3.2.2 Stage 2: Supervised Fine-tuning

Following cultural pretraining, we fine-tune the model on the PalmX cultural QA dataset using a chat-based instruction format. The fine-tuning process employs response-only training, where gradients are computed only on assistant responses:

$$\mathcal{L}_{finetune} = -\sum_{i\in\mathcal{R}} \log P(x_i|x_{<i}, context;\theta) \qquad (3)$$

where $\mathcal{R}$ denotes response tokens.

3.3 Cultural Corpus Construction

Our pretraining corpus spans 10 Arab countries (Bahrain, Egypt, UAE, Iraq, Kuwait, Jordan, Lebanon, Palestine, Syria, Saudi Arabia) and cultural domains like (Media, Sport, Transport, Healthcare, Education, Religion , Economy, History , Festivals, Tourism). Articles were systematically collected via Wikipedia API as shown in figure [1] and processed through a comprehensive cleaning pipeline including:

Algorithm 1 Cultural Corpus Processing Pipeline

```
Input: Raw Wikipedia articles $D = \{d_1, d_2, ..., d_n\}$
Initialize: Clean corpus $C = \emptyset$
for each article $d_i$ in $D$ do
    Remove HTML tags and formatting
    Filter by language (Arabic content only)
    Apply deduplication using content hashing
    Chunk into sequences $\leq$ 4096 tokens
    Attach article title as metadata (marker for cultural/contextual grounding)
    $C = C \cup \{processed\_chunks\}$
end for
Return: $C$
```

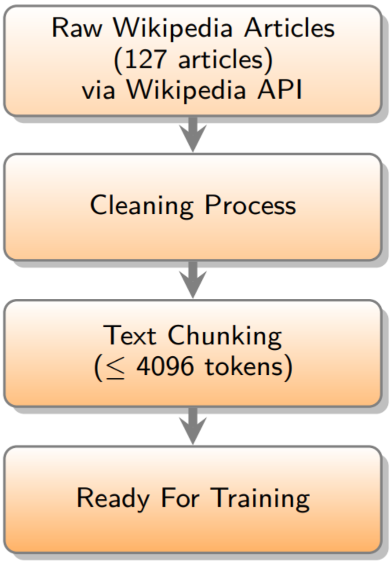

Figure 1: Data Collection Pipeline

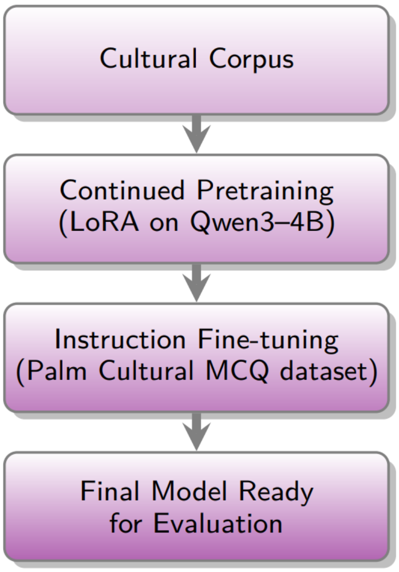

Figure 2: Full Training Pipeline

4 Experimental Setup

4.1 Data Splits

We utilized the official PalmX dataset splits: 2,000 training examples for supervised fine-tuning, 500 development examples for validation, and a blind test set for final evaluation. For cultural pretraining, we created a development split (3% of cultural corpus) to monitor pretraining progress.The training set contains 4480 examples, with a development set of 139 examples for validation.

4.2 Implementation Details

Our implementation leverages the Unsloth(Han et al., 2023) framework for efficient and scalable training. We summarize the Low-Rank Adaptation (LoRA) configuration and training hyperparameters in table [1].

Parameter	**Value**
Rank (r) and Alpha	128

Target modules	`q_proj,k_proj`
	`v_proj,o_proj`
	`gate_proj,up_proj`
	`down_proj,lm_head`
	`embed_tokens`

Dropout	0.05
Maximum sequence length	4096

Table 1: LoRA Configuration

The configuration adopts a rank and alpha of 128, applies LoRA to multiple attention and projection layers, and supports long-context training with sequences up to 4096 tokens.

The training pipeline consists of two phases: cultural pretraining and cultural QA fine-tuning. Each phase is optimized using the AdamW optimizer with a cosine learning rate schedule and a weight decay of 0.01, with learning rates, epochs, and batch sizes adjusted per phase to balance performance and convergence as shown in table [2].

4.3 Evaluation Metrics

The primary evaluation metric is the accuracy on the MMLU (Nacar et al., 2025). We employ exact match evaluation where the model's predicted letter (A, B, C, D) must exactly match the gold answer. We use MMLU since the competition itself is based on this benchmark because it is widely used to test broad knowledge ability, making it suitable for evaluating general-purpose language models.

4.4 System Pipeline

Our complete training pipeline consists of three sequential stages as shown in figure [2]:

1. **Cultural Pretraining**: Train on cultural corpus for 3 epochs

2. **Cultural QA Fine-tuning**: Train on PalmX cultural dataset for 2 epochs

Task	Hyper- parameters	Other Settings

Cultural Pretraining	`LR: 2e-5,Emb. LR: 5e-6,`
	`Epochs: 3, Batch: 16`
		`AdamW, cosine schedule,`
		`weight-decay:0.01`

Cultural QA Fine-tuning	`LR: 2e-5,`
	`Epochs: 2, Batch: 16`
		`Same as above`

Table 2: Training Hyperparameters

5 Results

5.1 Quantitative Results

Table [3] presents our official evaluation results on the PalmX 2025 Shared Task 1.

Dataset	Accuracy (%)
Development Set	74.0
Blind Test Set	64.0

Table 3: Official evaluation results on PalmX 2025 Subtask 1

5.2 Ablation Studies

We conducted ablation studies to assess the contribution of each training stage:

Configuration	Dev Accuracy (%)
Base Model Only	64.0
Star Model	**74.0**

Table 4: Ablation study showing contribution of each training stage

The results demonstrate that each training stage contributes significantly to final performance, with cultural pretraining providing the largest single improvement (10%) over the base model as shown in table [4].

5.3 Error Analysis

To better understand the behavior of the model, we performed a manual analysis of randomly sampled errors. We identified three major error types:

- **Ambiguous Knowledge:** The model struggled when multiple answers appeared plausible due to overlapping cultural concepts. For example, when asked about the founder of a specific Arab media outlet, the model confused the chief editor with the original founder.

- **Reasoning Gaps:** Some questions required multi step reasoning across history and religion, where the model failed to integrate knowledge.

- **Data Coverage Limitations:** Errors arise from missing representation of certain countries (e.g., Mauritania, Yemen) or underrepresented domains (e.g.,, folk traditions). This highlights the importance of broader cultural coverage in pre-training.

Overall, the errors suggest that while pretraining enriched the model with domain knowledge, deeper reasoning capabilities and broader cultural coverage remain key challenges.

5.4 Response Generation Quality

Our model successfully generates concise, accurate responses in the required format. Example model outputs demonstrate proper Arabic language usage and cultural sensitivity:

> **Input:**
> ما هو الطبق التقليدي الأشهر في المغرب العربي؟
> **Generated:** B
> **Gold:** B (الكسكس)

6 Conclusion

We presented a systematic approach to enhancing Arabic cultural understanding in language models through targeted pretraining and efficient fine-tuning. Our two-stage framework achieved competitive performance (74% development accuracy) while maintaining computational efficiency through LoRA adaptation.

Key contributions:

- A comprehensive cultural pretraining corpus spanning 10 Arab countries and more than 10 domains
- Demonstration that cultural pretraining significantly improves cultural QA performance
- An efficient training pipeline suitable for resource-constrained environments

Limitations:

- Limited coverage of dialectal variations across Arab regions
- Focus on factual knowledge may not capture implicit cultural understanding
- Performance gap between development and test sets suggests potential overfitting

Future Work: Future research directions include expanding corpus coverage to include dialectal content, investigating few-shot learning approaches for cultural adaptation, and developing more sophisticated evaluation metrics that capture cultural nuance beyond factual accuracy.

7 Acknowledgments

We thank the PalmX 2025 shared task organizers for providing this valuable evaluation framework. We also acknowledge the anonymous reviewers for their constructive feedback.

References

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.

M Saiful Bari, Yazeed Alnumay, Norah A Alzahrani, Nouf M Alotaibi, Hisham A Alyahya, Sultan Al-Rashed, 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*.

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*.

Daniel Han, Michael Han, and Unsloth team. 2023. Unsloth. https://github.com/unslothai/unsloth.

Omer Nacar, Serry Taiseer Sibaee, Samar Ahmed, Safa Ben Atitallah, Adel Ammar, Yasser Alhabashi, Abdulrahman S Al-Batati, Arwa Alsehibani, Nour Qandos, Omar Elshehy, and 1 others. 2025. Towards inclusive arabic llms: A culturally aligned benchmark in arabic large language model evaluation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 387–401.

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

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.

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*.

Kritsada Singhapoo, Akarachai Inthanil, and Attapon Pillai. 2025. Fine-tuning ai models with limited resources. In *2025 11th International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, pages 148–151. IEEE.

Jihoon Tack, Jack Lanchantin, Jane Yu, Andrew Cohen, Ilia Kulikov, Janice Lan, Shibo Hao, Yuandong Tian, Jason Weston, and Xian Li. 2025. Llm pretraining with continuous concepts. *arXiv preprint arXiv:2502.08524*.

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*.

Dan Zhang, Tao Feng, Lilong Xue, Yuandong Wang, Yuxiao Dong, and Jie Tang. 2025. Parameter-efficient fine-tuning for foundation models. *arXiv preprint arXiv:2501.13787*.