Do LLMs Distinguish Between Halal and Haram? Benchmarking Islamic Cultural Alignment in General vs. Arabic-Centric SLMs

Md. Abdur Rahman [1] **Md. Tofael Ahmed Bhuiyan** [1] **Abdul Kadar Muhammad Masum** [2]

Abstract

Large Language Models (LLMs) frequently exhibit cultural misalignment in non-Western contexts, often failing to grasp theological and societal nuances inherent to the Arab world. This study introduces **ACE-Adapt**, a unified evaluation framework designed to assess the cultural fidelity of parameter-efficient Small Language Models (SLMs) under 10 billion parameters. Leveraging the PalmX-IC benchmark, eight distinct architectures stratified into general purpose and Arabic-centric categories are rigorously evaluated on tasks covering Islamic rituals, jurisprudence, and history. By transforming static multiple-choice queries into strict conversational constraints and applying Quantized Low-Rank Adaptation (QLoRA), a significant performance dichotomy is demonstrated. Empirical results reveal that Arabic-centric models consistently outperform their general-purpose counterparts, regardless of parameter scale. Notably, the Fanar-1 9B model achieves state-of-the-art accuracy of 79.60%, while the 3B-parameter NileChat surpasses the larger Llama-3.1 8B baseline. These findings offer nuance to scaling laws in cultural domains, suggesting that domain-aligned pre-training priors can rival the benefits of raw model size when resolving semantic ambiguities in Islamic culture.

1. Introduction

While Large Language Models (LLMs) demonstrate profound natural language capabilities, their efficacy is often circumscribed by "cultural misalignment." Driven by the hegemony of Anglocentric pre-training corpora, LLMs frequently exhibit latent biases toward Western perspectives (Liu et al., 2024; Cao et al., 2023). Consequently, when deployed in non-Western contexts, they generate outputs that are grammatically fluent yet culturally incongruent. This deficit is particularly acute in the Arab world, a region characterized by deep theological integration into daily life. Without culturally grounded priors, models fail to distinguish between permissible (*Halal*) and prohibited (*Haram*) actions—distinctly exemplified when a translation-based model erroneously recommended alcohol consumption immediately following prayer (Naous et al., 2024). Such incidents underscore the critical need for deep cultural alignment to prevent the propagation of systematic stereotypes (Gehman et al., 2020).

Recent advancements in Arabic NLP have yielded promising foundation models, such as Jais (Sengupta et al., 2023), ALLaM (Althnian et al., 2025), and Fanar (Team et al., 2025). However, the evaluation of these architectures has remained largely confined to traditional capability-oriented benchmarks, rather than assessing their semantic grasp of nuanced cultural and religious dimensions.

To address this gap, we introduce the **ACE-Adapt** (Arabic Cultural Evaluation and Adaptation) framework. We focus on the efficacy of parameter-efficient Small Language Models (SLMs) with fewer than 10 billion parameters, investigating a central hypothesis: *Can smaller, domain-aligned Arabic models outperform larger, general-purpose architectures in determining cultural nuance?* Leveraging the PalmX-IC benchmark, we evaluate eight distinct architectures to demonstrate that domain-aligned pre-training priors are vastly more critical than parameter scale for resolving semantic ambiguities in Islamic culture.

Our primary contributions are threefold: (1) we propose **ACE-Adapt**, a unified framework that formulates static multiple-choice queries as strict conversational constraints to rigorously evaluate cultural alignment; (2) we provide a comprehensive evaluation of 8 SLMs (<10B parameters), demonstrating that Arabic-centric models significantly outperform larger, general-purpose baselines such as Llama-3.1 and Phi-3.5; and (3) we report state-of-the-art results for

[1]Computational Intelligence Lab, Southeast University, Dhaka 1208, Bangladesh [2]Department of Computer Science and Engineering, Southeast University, Dhaka 1208, Bangladesh. Correspondence to: Abdul Kadar Muhammad Masum <akmmasum@seu.edu.bd>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

the sub-10B category using the Fanar-1 9B model which got 79.60% accuracy, establishing a new baseline for the PalmX-IC benchmark.

2. Related Works

2.1. Cultural Alignment and Adaptation Strategies

Multilingual LLMs frequently exhibit a "Western value drift," often failing to align with non-Anglocentric societal norms (Liu et al., 2024; Navigli et al., 2023). To mitigate this, researchers have explored prompt engineering (Shen et al., 2024), supervised fine-tuning (SFT) (El Mekki et al., 2025), and synthetic data augmentation (Li et al., 2024). However, ensuring the semantic reliability of synthetic cultural data remains challenging. Our work builds upon these adaptation foundations but pivots toward a parameter-efficient approach (QLoRA) rigorously evaluated on human-validated data.

2.2. Evolution of Arabic Language Models

The development of Arabic LLMs has shifted from translation-based adaptation, which often lost cultural nuance (Sengupta et al., 2023), to native-centric pre-training. Models like AceGPT (Huang et al., 2024) focus on localization via native instruction tuning, while foundation models like ALLaM (Bari et al., 2024) and Fanar (Team et al., 2025) are trained on massive Arabic corpora. While these models improve general NLP tasks, their alignment with the intricacies of Islamic jurisprudence and history particularly in smaller parameter regimes remains an area requiring empirical investigation.

2.3. Benchmarking Islamic and Regional Knowledge

The Arabic NLP evaluation landscape is shifting from generic capabilities to nuanced cultural assessments. Frameworks like LAraBench (Abdelali et al., 2024), AraDiCE (Mousi et al., 2025), and SaudiCulture (Ayash et al., 2025) address regional and dialectal variations. Conversely, strictly Islamic domain benchmarking has been fragmented across specific verticals like Qur'anic QA (Alnefaie et al., 2023) or Fatwa retrieval (Aleid & Azmi, 2025), often using English-centric metrics. The PalmX shared task (Alwajih et al., 2025) consolidates these efforts into a holistic testbed. We utilize this benchmark to rigorously compare the "cultural priors" of general versus specialized SLMs.

3. Methodology

We introduce **ACE-Adapt** (Arabic Cultural Evaluation and Adaptation), a unified framework designed to align and assess the semantic latent space of Small Language Models (SLMs) within the domain of Arabic Islamic culture. As illustrated in Figure 1, the methodology operates through a tripartite pipeline: (1) Context-Aware Data Formulation, (2) Manifold Alignment via Quantized Low-Rank Adaptation (QLoRA), and (3) Deterministic Inference.

3.1. Dataset Formulation and Preprocessing

The foundation of our evaluation is the PalmX-IC dataset (Alwajih et al., 2025), which assesses semantic understanding across cultural verticals like *Fiqh* and *Hadith*. As depicted in Figure 2, the raw data presents a structured multiple-choice format.

To bridge the gap between this structured format and the causal generation nature of decoder-only transformers, ACE-Adapt restructures the problem as an instruction-following task. Let the dataset be $\Omega = \{(q_i, \mathcal{C}_i, y_i)\}_{i=1}^{N}$, where q_i is the query, $\mathcal{C}_i$ represents the choices, and y_i is the label.

The preprocessing pipeline (Figure 3) linearizes the options and appends a strict Arabic instruction constraint $\mathcal{I}_{cons}$. The input prompt x_i is constructed as:

$$x_i = \text{BOS} \oplus \mathcal{H}(q_i \oplus \text{"\textbackslash n"} \oplus \textstyle\sum_{j \in \{A..D\}} (idx_j \cdot c_{ij}) \oplus \mathcal{I}_{cons}) \oplus \text{EOS} \quad (1)$$

where $\mathcal{H}$ represents the user-role header embedding and $\oplus$ denotes tensor concatenation. This forces internal reasoning before generating the character output y_i.

3.2. Architectural Adaptation and Optimization

To evaluate 8 sub-10B parameter models, we apply 4-bit Normal Float (NF4) quantization to compress pretrained weights $\Theta \in \mathbb{R}^{d_{in} \times d_{out}}$. Adaptation is performed via LoRA (Hu et al., 2022), targeting all linear projection layers within the transformer blocks. The forward pass of a refined layer is derived via:

$$\mathbf{h}_{out} = \text{Dequant}(\Theta_{NF4})\mathbf{h}_{in} + \eta \cdot (\mathbf{BA})\mathbf{h}_{in} \quad (2)$$

where $\mathbf{A} \in \mathbb{R}^{r \times d_{in}}$ and $\mathbf{B} \in \mathbb{R}^{d_{out} \times r}$ are trainable low-rank matrices. The scaling factor $\eta = \frac{\alpha}{r}$ balances updates. We set rank $r = 16$ and $\alpha = 16$ utilizing the Unsloth optimization library to reduce backpropagation overhead.

Models are optimized using Supervised Fine-Tuning (SFT) on the transformed train set (≈ 2000 samples). Let $\Phi = \{\mathbf{A}, \mathbf{B}\}$ represent the trainable adapter parameters. The cultural alignment loss $\mathcal{L}_{cult}$ minimizes the negative log-likelihood of the ground truth response:

$$\mathcal{L}_{cult}(\Phi) = -\frac{1}{M} \sum_{i=1}^{M} \log P(y_i | x_i; \Theta_{NF4}, \Phi) \quad (3)$$

3.3. Deterministic Inference and Extraction

Post-adaptation, models are evaluated on a held-out test set (500 samples). To ensure reproducibility, we employ greedy

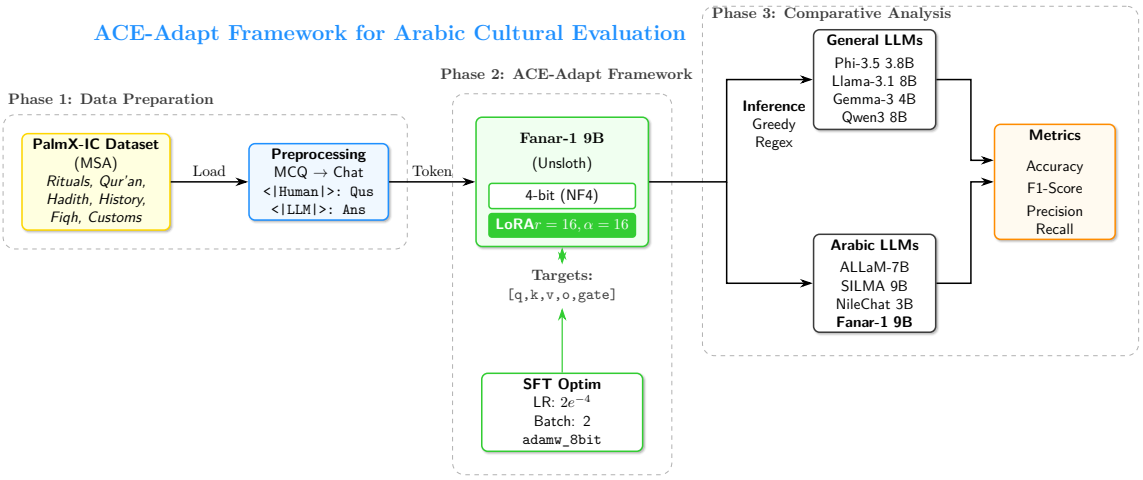

Figure 1. The ACE-Adapt Framework Architecture. The pipeline illustrates the transition from raw MSA data ingestion to 4-bit Quantized Adaptation and comparative analysis.

<Question>

A. <Choice 1>

B. <Choice 2>

C. <Choice 3>

D. <Choice 4>

Answer: _____

Figure 2. Foundational schema of the PalmX-IC dataset requiring specific cultural knowledge for resolution.

decoding (Temperature $T = 0$). The raw output string $\hat{s}_i$ is parsed using a strict Regular Expression filter to extract the valid class identifier:

$$\hat{y}_i = \text{Regex}(\hat{s}_i, \text{pattern} = \text{“}\hat{}[A - D]\text{”}) \quad (4)$$

This logic penalizes verbosity; failing to produce a clear identifier immediately following the instruction is marked as a failure.

4. Experimental Results and Discussion

We present a rigorous empirical evaluation of the ACE-Adapt framework to validate the *Cultural Alignment Hypothesis*: that models pre-trained on high-density Arabic corpora exhibit superior cultural reasoning compared to larger, polyglot architectures.

4.1. Experimental Setup

To ensure reproducibility and isolate architectural priors, all models were fine-tuned using a unified configuration on a single NVIDIA A100 GPU (40GB). We utilized Unsloth for optimized 4-bit backpropagation. Backbone parameters were frozen, and LoRA was applied to all linear projection layers. Hyperparameters, selected via validation set grid search, are detailed in Table 1.

Table 1. Hyperparameter Configuration for ACE-Adapt.

Hyperparameter	Value / Configuration
Model Optimization	
Quantization	4-bit Normal Float (NF4)
LoRA Rank (r)	16
LoRA Alpha (α)	16 (Scaling factor $\eta = 1.0$)
Target Modules	$[q, k, v, o, gate, up, down]$
Training Dynamics	
Optimizer	AdamW (8-bit)
Learning Rate	2×10^{-4} (Linear Decay)
Batch Size	8 (4 grad. accum.)
Max Sequence Length	2048 Tokens
Training Steps	1,250 ($\approx$ 5 epochs)

4.2. Quantitative Performance Analysis

Evaluation was conducted on the held-out PalmX-IC test split ($N = 500$) using deterministic greedy decoding ($T = 0$). Table 2 summarizes the performance across General and Arabic-centric SLMs.

The results delineate a clear boundary between model families. The lowest-performing Arabic-centric model (ALLaM-7B, 76.2%) surpasses the highest-performing General model (Gemma-3 4B, 73.8%) by 2.4%. This confirms that cultural nuance cannot be fully compensated for by the massive, multilingual training corpora used in generalized models. Furthermore, the Fanar-1 9B model achieved the highest scores across all metrics, with an accuracy of **79.60%**. This represents a significant +7.0% improvement over the

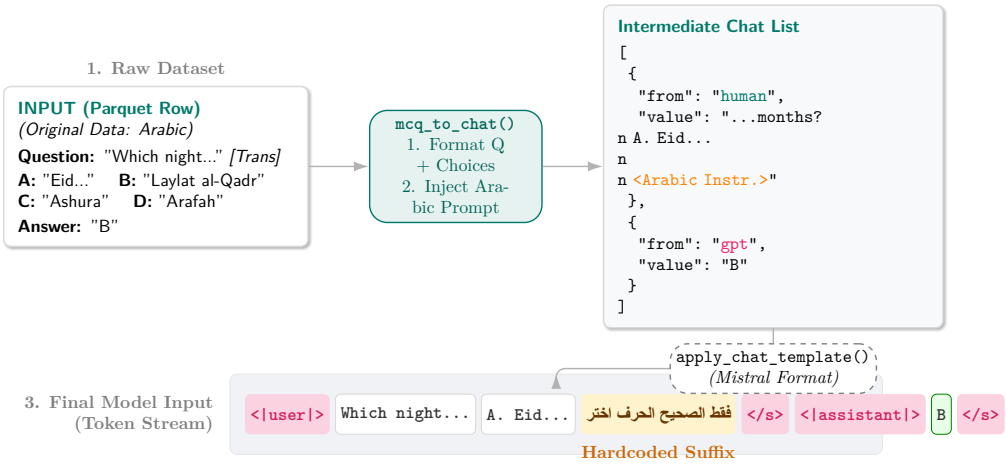

Figure 3. The Data Transformation Pipeline. The raw Parquet row is parsed, injected with Arabic instruction constraints, and serialized into a chat-compatible token stream.

Llama-3.1 8B baseline, despite having comparable parameter counts. Additionally, a critical observation is the efficiency of NileChat 3B. Despite having ≈60% fewer parameters than Llama-3.1 8B, it achieves a superior F1-score (0.7707 vs. 0.7268). This challenges traditional "scaling laws" in domain-specific contexts, proving that aligned cultural priors are more effective than raw parameter scale.

Table 2. Comparative Results on PalmX-IC. Bold denotes the best performance.

Category	Model	Acc.	Prec.	Rec.	F1
General	Phi-3.5 3.8B	0.7140	0.7217	0.7140	0.7151
	Llama-3.1 8B	0.7260	0.7314	0.7260	0.7268
	Gemma-3 4B	0.7380	0.7410	0.7380	0.7386
	Qwen3 8B	0.7380	0.7408	0.7380	0.7385
Arabic	ALLaM-7B	0.7620	0.7640	0.7620	0.7623
	SILMA 9B	0.7640	0.7687	0.7640	0.7646
	NileChat 3B	0.7700	0.7744	0.7700	0.7707
	Fanar-1 9B	**0.7960**	**0.7987**	**0.7960**	**0.7964**

4.3. Discussion: The Cultural Alignment Hypothesis

The empirical data strongly supports our central hypothesis. General LLMs predominantly encode knowledge in English, treating Arabic as a secondary linguistic mapping. When reasoning over Islamic jurisprudence (*Fiqh*) or Prophetic traditions (*Hadith*), they rely on cross-lingual transfer, which inevitably introduces semantic noise and strips theological specificity.

In contrast, Arabic-centric SLMs possess a "native" semantic topology for these concepts. The rapid training convergence (reaching stable loss in under 300 steps) indicates that ACE-Adapt did not teach these models cultural concepts *ab initio*; rather, the SFT process surfaced pre-existing native knowledge. For researchers deploying AI in the Arab world, data distribution quality drastically outperforms parameter scale. Fine-tuning a culturally pre-trained backbone yields competitive fidelity compared to adapting larger, Western-centric foundation models. However, because our evaluation follows SFT, these results reflect a combination of pre-training priors and fine-tuning efficiency. Future work will incorporate zero-shot and few-shot baselines across broader cultural datasets to strictly isolate the impact of pre-training priors from adaptation dynamics.

5. Conclusion

We addressed the critical challenge of cultural alignment in LLMs within the context of Arabic Islamic heritage. Through the ACE-Adapt framework, we established a rigorous methodology for quantifying the semantic gap between general-purpose reasoning and culturally grounded understanding. To answer our titular question: Yes, LLMs can distinguish between Halal and Haram, but Arabic-centric models do so with greater semantic fidelity. While our evaluation on PalmX-IC yields suggestive evidence that architectural scale can be secondary to pre-training data distribution, we note the modest 2.4% performance gap between the lowest Arabic-centric and highest general-purpose models. Nonetheless, the efficiency of NileChat 3B and Fanar-1 9B demonstrates that native latent representations aid in resolving complex theological ambiguities. While limited to a single benchmark, these findings suggest that high-fidelity data curation provides a competitive alternative to pure parameter scaling in regional AI development.

References

Abdelali, A., Mubarak, H., Chowdhury, S., Hasanain, M., Mousi, B., Boughorbel, S., Abdaljalil, S., El Kheir, Y.,

Izham, D., et al. (2024). Larabench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 487–520).

Aleid, H. A., & Azmi, A. M. (2025). Hajj-fqa: A benchmark Arabic dataset for developing question-answering systems on Hajj fatwas. *Journal of King Saud University - Computer and Information Sciences*, *37*(6), 135.

Alnefaie, S., Atwell, E., & Alsalka, M. A. (2023). Is GPT-4 a good Islamic expert for answering Quran questions? In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing* (pp. 124–133).

Althnian, A., Alzahrani, N. A., Alsubaie, S. Z., Albilali, E., Abdelali, A., Alotaibi, N. M., Bari, M. S., Alnumay, Y., Alothaimen, A., Saif, M., et al. (2025). Araeval: An Arabic multi-task evaluation suite for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 33025–33049).

Alwajih, F., El Mekki, A., Mubarak, H., Hawasly, M., Mohamed, A., & Abdul-Mageed, M. (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* (pp. 774–789).

Ayash, L., Alhuzali, H., Alasmari, A., & Aloufi, S. (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.

Bari, M. S., Alnumay, Y., Alzahrani, N. A., Alotaibi, N. M., Alyahya, H. A., AlRashed, S., Mirza, F. A., Alsubaie, S. Z., Alahmed, H. A., et al. (2024). Allam: Large language models for Arabic and English. *arXiv preprint arXiv:2407.15390*.

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

El Mekki, A., Atou, H., Nacar, O., Shehata, S., & Abdul-Mageed, M. (2025). Nilechat: Towards linguistically diverse and culturally aware LLMs for local communities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 10978–11002).

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). LoRA: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.

Huang, H., Yu, F., Zhu, J., Sun, X., Cheng, H., Dingjie, S., Chen, Z., Alharthi, M., An, B., He, J., et al. (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* (pp. 8139–8163).

Li, C., Chen, M., Wang, J., Sitaram, S., & Xie, X. (2024). CultureLLM: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, *37*, 84799–84838.

Liu, C. C., Koto, F., Baldwin, T., & Gurevych, I. (2024). Are multilingual LLMs culturally-diverse reasoners? An investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 2016–2039).

Mousi, B., Durrani, N., Ahmad, F., Hasan, M. A., Hasanain, M., Kabbani, T., Dalvi, F., Chowdhury, S. A., & Alam, F. (2025). Aradice: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 4186–4218).

Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2024). Having beer after prayer? Measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 16366–16393).

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, *15*(2), 1–21.

Sengupta, N., Sahu, S. K., Jia, B., Katipomu, S., Li, H., Koto, F., Marshall, W., Gosal, G., Liu, C., Chen, Z., et al. (2023). Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Shen, S., Logeswaran, L., Lee, M., Lee, H., Poria, S., & Mihalcea, R. (2024). Understanding the capabilities and limitations of large language models for cultural commonsense. *arXiv preprint arXiv:2405.04655*.

Team, F., Abbas, U., Ahmad, M. S., Alam, F., Altinisik, E., Asgari, E., Boshmaf, Y., Boughorbel, S., Chawla, S., Chowdhury, S., et al. (2025). Fanar: An Arabic-centric multimodal generative AI platform. *arXiv preprint arXiv:2501.13944*.