Hamyaria at PalmX2025: Leveraging Large Language Models to Improve Arabic Multiple-Choice Questions in Cultural and Islamic Domains

Walid Al-Dhabyani
Hadhramout University / Hadhramout, Yemen
Cairo University / Cairo, Egypt
w.aldhabyani@grad.fci-cu.edu.eg

Hamzah A. Alsayadi
Ibb University / Ibb, Yemen
hamzah.sayadi@gmail.com

Abstract

Large language models (LLMs) have been widely used recently. Adapting these models to multiple languages would enhance the accuracy and precision of the other languages. Applying LLMs with Arabic language could improve the prediction of Arabic language. This work applies LLMs with MCQs of Arabic in both the cultural and Islamic domain. The dataset used is PalmX, which is an **MCQ** benchmark dataset. In this work, traditional and AI generation data augmentations are used. For the cultural domain, we applied data augmentation techniques, including paraphrasing using *Fanar-1-9B-Instruct* model and answer shuffling. For the Islamic domain, we used the original dataset without augmentation to maintain content integrity. We then fine-tuned the *Qwen2.5-3B-Instruct* model on both datasets and evaluate its performance, achieving 65.90% accuracy on the cultural set and 70.83% on the Islamic set. Experiment and evaluation are discussed and the best accuracy achieved in this work is explained in both domains.

1 Introduction

Due to their exceptional performance in a wide range of applications, LLMs are becoming more and more well-liked in both academia and industry. Since LLMs are still essential for research and everyday applications, it is becoming more and more important to evaluate them at the task level as well as the societal level in order to better comprehend the hazards they may pose (Chang et al., 2024). Adapting LLMs in Arabic language is still challenging (Mashaabi et al., 2024). Due to grammatical complexity, semantic diversity, and domain specialization, answering multiple choice questions in Arabic is a challenging NLP task, especially in cultural and Islamic contexts. Building strong language-understanding systems requires improving the quality of MCQ datasets in these areas. In order to increase model performance, our work uses LLMs to enhance and optimize Arabic MCQ data. This work was conducted as part of the ArabicNLP 2025 competition named **"PalmX 2025: The First Shared Task on Benchmarking LLMs on Arabic Culture"** (Alwajih et al., 2025b)[1].

We use both conventional augmentation methods (answer shuffling) and AI-based methods (paraphrasing using QCRI/Fanar-1-9B-Instruct [2]) (Team et al., 2025) on the general culture dataset. To maintain the authenticity of religion, the Islamic dataset is left unchanged. With significant accuracy gains, Qwen2.5-3B-Instruct [3] (Team, 2024) is refined using both datasets.

The rest of the paper, an introduction is explained in Section 1. Section 2 contains the background. Section 3 illustrated the system overview. Section 4 has the experimental setup. The results are explained in section 5. Finally, conclusion is in section 6.

2 Background

Task setup, dataset details, and related work are explained in this part of the paper.

2.1 Task Setup

Enhancing the quality and precision of Arabic MCQ in two different areas—general culture and Islamic knowledge—is the challenge at hand. Arabic questions with several possible answers make up the input, and choosing the right response from the list of options is the output. Examples about PalmX dataset are in Appendix A.

2.2 Dataset Details

We make use of the PalmX dataset (Alwajih et al., 2025b) which is an MCQ benchmark Arabic dataset. Palmx dataset is created from the Palm

[1] https://palmx.dlnlp.ai/index.html
[2] https://huggingface.co/QCRI/Fanar-1-9B-Instruct
[3] https://huggingface.co/Qwen/Qwen2.5-3B-Instruct

Proceedings of The Third Arabic Natural Language Processing Conference, pages 790–796

Dataset	#Train	#Dev	#Test
Culture	2000	500	2000
Islamic	1000	300	1001

Table 1: PalmX dataset characteristics.

dataset [4] (Alwajih et al., 2025c) (Alwajih et al., 2025a), which is an instruction dataset. The Palm dataset is a comprehensive benchmark created to assess Large Language Models on tasks involving Arabic in a range of dialects and contexts.

The PalmX dataset is especially well-suited to our study goals because it offers broad coverage of both cultural and Islamic knowledge categories. In order to ensure thorough evaluation coverage, the PalmX dataset contains questions covering a range of topics and difficulty levels within each domain. The data set characteristics are illustrated in the table 1. The PalmX dataset has 3000 questions of the training data, 800 questions of the development data and 3001 questions for the test data. The PalmX dataset contains both Cultural and Islamic domains. The separation of both domains are illustrated in table 1.

2.3 Related Work

Previous research in Arabic NLP has highlighted the unique challenges posed by the language's morphological complexity and the need for culturally appropriate content generation. While significant progress has been made in general Arabic NLP tasks, specialized domains such as cultural and Islamic knowledge require targeted approaches that respect content integrity and cultural sensitivities. The remainder of this section is in Appendix B.

The novelty of our contribution lies in the domain-specific approach to Arabic MCQ enhancement, particularly the differentiated treatment of cultural versus Islamic content, recognizing that religious content requires special consideration to maintain authenticity and accuracy.

3 System Overview

In this section, our proposed solutions for both tasks and the used resources are explained in detail. Furthermore, the challenging that we faced through the work with the dataset and other models.

3.1 Key Algorithms and Design Decisions

Our system architecture employs a multi-stage approach involving data preprocessing, augmentation, and model fine-tuning. The core design decision revolves around treating cultural and Islamic domains differently based on their respective requirements for content modification. Figure 1 shows the approaches used in fine-tuning and evaluating the LLMs and Trained Model. In the next two section, we explained the steps used in fine-tunning the modles with Palm dataset.

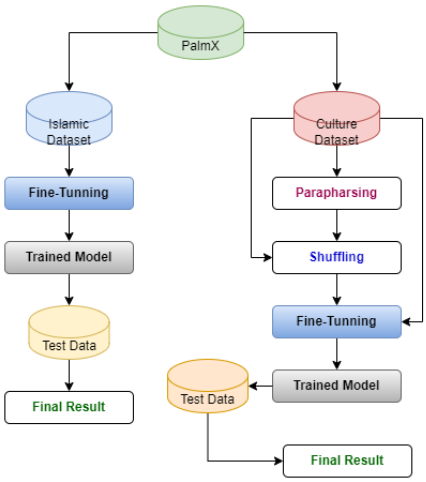

Figure 1: The used model for fine-tunning the LLM for the both dataset (Cultural and Islamic)

3.1.1 Cultural Domain Processing

1. Load original cultural MCQ dataset
2. Generate paraphrased questions using *QCRI/Fanar-1-9B-Instruct*. Paraphrasing techniques follow these approaches:
   - Two questions are generated for each question.
   - Concatenate the original questions with the generated questions separated by the phases " ".
3. Combine original and paraphrased datasets in one dataset.
4. Apply traditional augmentation (answer shuffling). For every question in the new dataset, there are three shuffling in the answers for each question.
5. Apply final answer shuffling, combined the new dataset that contains shuffling answers with the new dataset(orginal dataset + paraphrased dataset)
6. Load the pretrained model *Qwen2.5-3B-Instruct* from Hugging face hub.
7. Fine-tune *Qwen2.5-3B-Instruct* on augmented data.
8. Evaluate accuracy and performance of the

[4] https://github.com/UBC-NLP/palm

Name	Explanation
QCRI/Fanar-1-9B-Instruct	Large language model (LLM) used for paraphrasing the questions.
Qwen2.5-3B-Instruct	Utilized for fine-tuning the model specifically for the Arabic language.
PalmX Dataset	The primary dataset used for model training and evaluation.
Colab A100 GPU (40GB)	Provided the computational resources for experiments and training.
Transformers (Vaswani et al., 2017)	Open-source library used for model loading, fine-tuning, and inference.
PyTorch	Underlying deep learning framework enabling implementation and optimization.

Table 2: Execution environment and resources used in experiments with the PalmX dataset.

trained model in the test dataset.

3.1.2 Islamic Domain Processing

1. Load original Islamic MCQ dataset.
2. Preserve original structure without augmentation.
3. Load the pretrained model *Qwen2.5-3B-Instruct* from Hugging face hub.
4. Fine-tune *Qwen2.5-3B-Instruct* on original Islamic MCQ dataset.
5. Evaluate accuracy and performance of the trained model in the test dataset.

3.2 Resources Used, External Tools and Libraries

The resources used, External Tools and Libraries are explained in table 2. Model access, versioning, and deployment are managed through the *Hugging Face Hub*.

3.3 Addressing Task Challenges

Assuring proper handling of culturally sensitive content, managing the complexity of the Arabic language with its morphological and dialectical variations, striking a balance between the advantages of data augmentation and the preservation of content integrity in domain-specific contexts, and optimizing performance within the limitations of computational resources are the main challenges this work attempts to address. Furthermore, in the paraphrasing stage, there was some words translated to English language that we have addressed and solved.

4 Experimental Setup

In this section of the work, we discussed in detail the experiments steps for both datasets such as Data Split usage, Preprocessing, and Hyperparameter Details, and Evaluation Metrics.

4.1 Data Split Usage

The experiments utilized the standard train/development/test split provided by the PalmX dataset. With evaluation performed directly on the designated test sets for both cultural and Islamic domains.

4.2 Preprocessing and Hyperparameter Details

Cultural Domain Configuration: We fine-tune *Qwen2.5-3B-Instruct* for 1 epoch (optionally extending to `NUM_EPOCHS = 5`) with a learning rate of `1e-5` and a batch size of `1`; for evaluation we use `BATCH_SIZE = 100`. Training uses `8` gradient-accumulation steps, `50` warm-up steps, and a maximum sequence length of `512`, optimized with AdamW (`adamw_torch`) and a cosine learning-rate scheduler. Gradient checkpointing is enabled. Checkpoints are saved every `1,000` steps, evaluation runs every `1,000` steps, and logging occurs every `200` steps.

Islamic Domain Configuration: We likewise use *Qwen2.5-3B-Instruct* for 1 epoch (with extended runs up to `NUM_EPOCHS = 10`) at a learning rate of `1e-5` and a batch size of `1`. The setup includes `8` gradient-accumulation steps, `50` warm-up steps, a maximum sequence length of `512`, the AdamW (`adamw_torch`) optimizer, and a cosine scheduler, with BF16 precision enabled. We save every `300` steps, evaluate every `300` steps, and log every `200` steps.

4.3 Evaluation Metrics

The primary evaluation metric used is accuracy, calculated as the percentage of correctly answered questions in the respective test sets. This metric provides a straightforward measure of model performance in the MCQ answering task. The accuracy, confusion matrix, and heatmap are discussed in detail in this section 5.

5 Results

5.1 Quantitative Findings

Our experiments yielded the following performance results:

Cultural Domain: The model achieved **65.90%** test accuracy using an augmented dataset that combined traditional and AI-based techniques with the

Class	Prec.	Rec.	F1	Sup.
A	0.58	0.77	0.66	497
B	0.67	0.63	0.65	491
C	0.66	0.67	0.66	500
D	0.79	0.57	0.67	512
Acc.			0.66	2000
Macro	0.67	0.66	0.66	2000
W. Avg	0.68	0.66	0.66	2000

Table 3: Classification (confusion matrix) report for the Culture dataset.

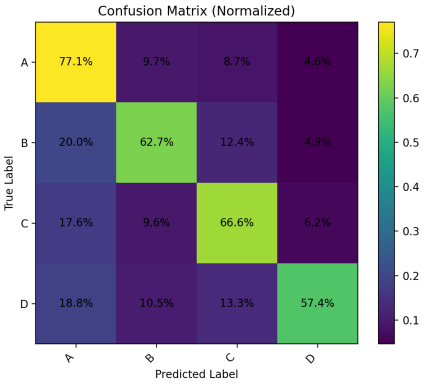

Figure 2: Normalized confusion matrix (heatmap) for cultural dataset

Class	Prec.	Rec.	F1	Sup.
A	0.48	0.73	0.58	153
B	0.89	0.75	0.81	546
C	0.71	0.69	0.70	213
D	0.41	0.56	0.47	73
OTHER	0.00	0.00	0.00	16
Acc.			0.71	1001
Macro	0.50	0.55	0.51	1001
W. Avg	0.74	0.71	0.72	1001

Table 4: Classification report (confusion matrix) for the Islamic dataset.

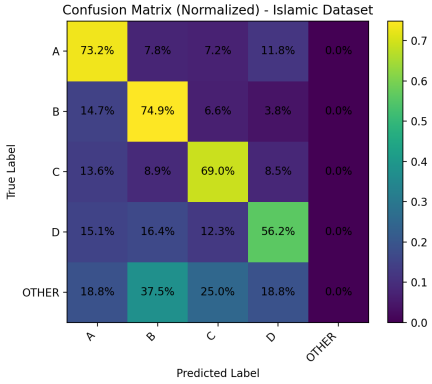

Figure 3: Normalized confusion matrix (heatmap) for Islamic dataset

original data. Training only on the original dataset yielded **65.40%**, while other models performed worse.

Table 3 shows the classification performance. Precision, recall, and F1-scores are balanced across the four classes (A–D) with averages around **0.66**. Class A has high recall (**0.77**) but lower precision (**0.58**), while Class D shows the opposite (**recall 0.57**, **precision 0.79**). The results confirm consistent performance across classes.

The normalized confusion matrix in Figure 2 illustrates the model's classification performance across classes A–D. Correct predictions lie on the diagonal (e.g., **77.1%** of class A and **66.6%** of class C), while off-diagonals show misclassifications (**20%** of class B predicted as A, **18.8%** of class D as A). The model achieves higher recall for classes A and C but struggles with class D, often confused with A (**18.8%**) and C (**13.3%**), suggesting overlapping feature representations between A↔B and D↔A/C.

Islamic Domain: The model achieved **70.83%** test accuracy using the original Islamic dataset without augmentation.

Table 4 summarizes the classification results. Class B performed best (**F1=0.81**, **precision=0.89**, **recall=0.75**), showing robust and balanced performance. Class C also performed well (**F1=0.70**, precision=0.71, recall=0.69). Class A captured many relevant cases with higher recall (**0.73**) but lower precision (**0.48**, F1=**0.58**). Class D underperformed (**F1=0.47**), and the "OTHER" class had no correct predictions due to extremely low support (**16** samples). The overall weighted F1-score is **0.72**, but the lower macro-average F1-score (**0.51**) highlights poorer performance on underrepresented classes.

Figure 3 shows the normalized confusion matrix for the Islamic dataset. Class B achieved the highest recall (**74.9%**), followed by A (**73.2%**) and C (**69.0%**), while Class D had the lowest (**56.2%**). Correct predictions lie on the diagonal, while off-diagonals show key misclassifications: **37.5%** of OTHER samples were predicted as B, **14.7%** of B as A, and **16.4%** of D as B. The "OTHER" class, with very few samples, had no correct predictions, indicating difficulty in recognizing this minority class. Overall, the model performs well on majority classes but struggles with D and OTHER.

5.2 Analysis

The results highlight three key insights: **Domain-Specific Performance:** The Islamic domain achieved higher accuracy (**70.83%**) than the cultural domain (**65.90%**), suggesting that preserving original content structure benefits religious and cul-

turally sensitive queries.

Augmentation Impact: Despite using traditional and AI-based augmentation to expand the cultural dataset, the slightly lower accuracy implies that content preservation can be more critical than dataset size in certain domains. By employing AI-based data augmentation through concatenation of real questions with generated ones, our findings indicate that this approach is not particularly effective. Due to time constraints, we were unable to conduct additional experiments using the original questions combined with the generated ones in a merged dataset, which could potentially improve the accuracy of the trained model. Furthermore, the application of traditional augmentation techniques yielded only marginal benefits. In the context of Arabic MCQ datasets, it is crucial to apply traditional augmentation methods more selectively and precisely. Overall, both augmentation strategies led to an improvement of only 0.5%, which is considered negligible.

Model Configuration: Differences in hyperparameters—particularly more training epochs in the Islamic domain (up to **10** vs. **5**)—may also explain the performance gap.

Model Selection: We experimented with various models for AI-based data augmentation and model training. During the data augmentation phase, we encountered several issues. In particular, many models failed to correctly paraphrase the questions; for example, ALLAM (Bari et al., 2024) often transformed the original questions into different syntactic forms, resulting in outputs that were difficult to interpret. Additionally, some models inadvertently translated certain words into English, even though the questions were primarily in Arabic. For the training phase, we observed that most models produced lower accuracies compared to the QWEN2.5-3B-INSTRUCT model. For instance, LLAMA-3.2-3B-INSTRUCT[5] consistently underperformed relative to QWEN2.5-3B-INSTRUCT.

5.3 Error Analysis

The performance gap between domains suggests several potential factors:

1. **Content Integrity**: Islamic questions may benefit more from maintaining original phrasing and structure due to the precision required in religious knowledge.
2. **Augmentation Effects**: The paraphrasing process in cultural questions might introduce subtle semantic changes that affect answers accuracy.
3. **Training Dynamics**: The different training configurations (more epochs for Islamic domain) may have allowed for better convergence on the Islamic dataset.

[5] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

5.4 Comparison with Baseline

The Cultural dataset had an accuracy of 65.90%, which was somewhat lower than its baseline of 70% as illustated in (Alwajih et al., 2025a), and the Islamic dataset had an accuracy of 70.83%, which was higher than its baseline of 65%, in comparison to the predetermined baselines. These findings show that the model performed well in the Islamic domain, outperforming the baseline by a significant margin, even while the Cultural dataset performed slightly worse than its baseline. Due to limited computational resources, we were unable to utilize large-scale LLMs such as **Qwen2.5-5B-Instruct** or **Qwen2.5-7B-Instruct**. Nevertheless, we acknowledge that employing such models could potentially yield higher accuracies than those achieved in our experiments.

6 Conclusion

This work presents a comprehensive approach to improving Arabic multiple-choice questions in cultural and Islamic domains using large language models. Our system demonstrates the importance of domain-aware processing, showing that different content domains benefit from tailored approaches to data handling and model training. The key findings indicate that Islamic domain questions achieve better performance when processed without augmentation (70.83% accuracy), while cultural domain questions, despite augmentation efforts, achieve 65.90% accuracy. This suggests that content integrity and cultural sensitivity are paramount considerations when working with specialized Arabic educational content.

Future research directions include the addition of more evaluation techniques. However, we need to investigate transfer learning between Islamic and cultural domains, creating more advanced augmentation techniques that maintain religious and cultural integrity. Extending the method to more effectively handle different Arabic dialects is required. Finally, we need to perform thorough comparisons with other Arabic Models and approaches.

References

Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. 2024. Arabicaqa: A comprehensive dataset for arabic question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2049–2059.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, Ahmed Oumar El-Shangiti, Aisha Alraeesi, Mohammed Anwar AL-Ghrawi, Abdulrahman S. Al-Batati, Elgizouli Mohamed, Noha Taha Elgindi, Muhammed Saeed, Houdaifa Atou, Issam Ait Yahia, Abdelhak Bouayad, Mohammed Machrouh, Amal Makouar, Dania Alkawi, Mukhtar Mohamed, Safaa Taher Abdelfadil, Amine Ziad Ounnoughene, Anfel Rouabhia, Rwaa Assi, Ahmed Sorkatti, Mohamedou Cheikh Tourad, Anis Koubaa, Ismail Berrada, Mustafa Jarrar, Shady Shehata, and Muhammad Abdul-Mageed. 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, Abdellah El Mekki, Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed, and Muhammad Abdul-Mageed. 2025b. 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.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, Abdelrahim A Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, et al. 2025c. Palm: A culturally inclusive and linguistically diverse dataset for arabic llms. *arXiv preprint arXiv:2503.00151*.

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

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

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Malak Mashaabi, Shahad Al-Khalifa, and Hend Al-Khalifa. 2024. A survey of large language models for arabic language and its dialects. *arXiv preprint arXiv:2410.20238*.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, et al. 2025. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

A Appendix A

PalmX Dataset Examples

Examples of PalmX dataset for Cultural and Islamic Domains.

Input (Cultural Domain):

- Question: هو ما هي عاصمة مصر؟ (What is the capital of Egypt?)
- Options: أ) القاهرة ب) الإسكندرية ج) الجيزة د) أسوان
- Expected Output: أ) القاهرة

Input (Islamic Domain):

- Question: كم عدد أركان الإسلام؟ (How many pillars of Islam are there?)
- Options: أ) ثلاثة ب) أربعة ج) خمسة د) ستة
- Expected Output: ج) خمسة

B Appendix B

Remaining of Related Work

Abdallah et al. (Abdallah et al., 2024) presented "ArabicaQA" which is the first extensive Arabic dataset for open-domain question answering (QA) and machine reading comprehension (MRC). It includes 3,701 difficult unanswerable questions and 89,095 answerable questions, together with open-domain annotations. In addition, the work

benchmarks several models, including as GPT-3.5,AraBERT (Antoun et al., 2020), PPLX, and Falcon, on Arabic QA tasks and presents AraDPR, the first dense passage retrieval model trained on Arabic Wikipedia. The results demonstrate that while dense retrieval techniques beat traditional approaches, fine-tuned Arabic-specific models perform better than traditional baselines, but LLMs still have difficulty successfully utilizing retrieved material. Their work advances Arabic natural language processing research by offering empirical insights and a useful resource.