CultranAI at PalmX 2025: Data Augmentation for Cultural Knowledge Representation

Hunzalah Hassan Bhatti[1*], **Youssef Ahmed**[1*], **Md Arid Hasan**[2], **Firoj Alam**[3]
[1]Qatar University, [2]University of Toronto, Canada
[3]Qatar Computing Research Institute
hunzalahhassan@gmail.com, fialam@hbku.edu.qa

Abstract

In this paper, we report our participation to the PalmX cultural evaluation shared task. Our system, ***CultranAI***, focused on data augmentation and LoRA fine-tuning of large language models (LLMs) for Arabic cultural knowledge representation. We benchmarked several LLMs to identify the best-performing model for the task. In addition to utilizing the PalmX dataset, we augmented it by incorporating the Palm dataset and curated a new dataset of over 22K culturally grounded multiple-choice questions (MCQs). Our experiments showed that the Fanar-1-9B-Instruct model achieved the highest performance. We fine-tuned this model on the combined augmented dataset of 22K+ MCQs. On the blind test set, our submitted system ranked 5th with an accuracy of 70.50%, while on the PalmX development set, it achieved an accuracy of 84.1%. We made experimental scripts publicly available for the community.[1]

1 Introduction

Cultural information plays a pivotal role in shaping human identity, behavior, and social interactions. It encompasses the shared beliefs, values, customs, languages, traditions, and collective knowledge of a community or society. In today's interconnected information, communication, and interaction ecosystem, hundreds of millions of users engage with LLMs for everyday queries - many of which involve aspects of local culture, traditions, cuisine, and more (Pawar et al., 2025; Hasan et al., 2025). A central challenge lies in evaluating how effectively LLMs comprehend and generate responses to such culturally embedded queries, particularly in multilingual settings characterized by significant dialectal variation. Other challenges include how to develop culturally aligned LLMs (Wang et al., 2023) and make them available in low-compute environments (Hu et al., 2022). Recent initiatives have introduced evaluation resources - such as culturally relevant datasets, task-specific benchmarks, and performance metrics - to assess LLM capabilities in this domain (Myung et al., 2024; Li et al., 2024b; Mousi et al., 2025).

Yet these efforts remain limited, especially in achieving deeper, dialect-specific advancements. Addressing this gap requires sustained, targeted, rigorous initiatives. The PalmX Shared Task at ArabicNLP 2025 (Alwajih et al., 2025b) is a step in this direction, offering a dedicated benchmark for culturally specific evaluation with a special emphasis on Arabic - thereby advancing the development of LLMs that are both linguistically and culturally aligned. Other recent relevant efforts for Arabic include the development of Arabic-centric LLMs (Team et al., 2025; Sengupta et al., 2023; Bari et al., 2025), leaderboards (Al-Matham et al., 2025), and culturally specific datasets (Alwajih et al., 2025a; Ayash et al., 2025).

To advance the state of the art in Arabic cultural knowledge representation within LLMs, in this paper, we report our participation in the shared task. We specifically focus on the cultural evaluation subtask. To address the challenges of training and deploying LLMs in low-compute resource settings, we conducted a comparative analysis of quantized vs. full-precision models. In parallel, we employed LLM-driven data augmentation strategies to improve the model accuracy. To summarise, the contributions of our study are as follows.

- We provide a performance comparison of different LLMs (Arabic-centric and multilingual) in a zero-shot setup.
- We demonstrate that the performance gap between quantized models and their full-precision counterparts is minimal.
- We show that data augmentation contributes to improving model performance.

[*] The contribution was made while the author was interning at the Qatar Computing Research Institute.
[1] https://github.com/hunzed/CultranAI

Proceedings of The Third Arabic Natural Language Processing Conference, pages 809–817

2 Related Work

General Capabilities of LLMs. LLMs have demonstrated remarkable capabilities across a wide range of natural language processing (NLP) tasks, including text classification, question answering, summarization, and dialogue generation (Bubeck et al., 2023; Abdelali et al., 2024). Their ability to leverage vast amounts of pretraining data and adapt to downstream tasks with minimal supervision has enabled strong performance in both zero-shot and few-shot settings (Abdelali et al., 2024). These advances have accelerated the integration of LLMs into diverse real-world applications spanning education, healthcare, finance, and customer support.

Cultural and Everyday Knowledge. Despite successes in several downstream NLP tasks, LLMs often underperform on tasks requiring culturally grounded knowledge, particularly in low-resource languages and dialects (Pawar et al., 2025; Hasan et al., 2025; Alam et al., 2025). A culturally aligned model should accurately interpret and generate content that reflects local linguistic forms, social norms, and lived experiences across domains such as healthcare, education, and cuisine (Li et al., 2024b,a; Shi et al., 2024). However, current models frequently fail to capture region-specific expressions and indigenous knowledge, limiting their effectiveness in culturally nuanced contexts (Myung et al., 2024; Chiu et al., 2025). To address these limitations, recent research has focused on developing benchmarks and datasets that evaluate and enhance LLMs' performance for both cultural and everyday information-seeking queries. These resources span mono- and multilingual settings and are sourced from diverse origins, including Wikipedia (Yang et al., 2018; Kwiatkowski et al., 2019), Google Search QA (Khashabi et al., 2021), Reddit forums (Fan et al., 2019), and native speaker-authored question–answer pairs (Clark et al., 2020). Other approaches combine native and machine-translated content or employ LLMs to generate culturally relevant QA datasets (Putri et al., 2024; Li et al., 2024b).

Although English and multilingual resources have advanced the state of the art in culturally aligned LLMs, the richness and diversity of the Arabic language and its dialects require dedicated efforts in both resource creation and culturally aligned model development. Recent initiatives have begun addressing this gap through the development of datasets for benchmarking and fine-tuning Arabic-centric models (Mousi et al., 2025; Alwajih et al., 2025a). The PalmX Shared Task at ArabicNLP 2025 is a targeted initiative to advance culturally aligned LLM development through a benchmark for culturally grounded evaluation in Arabic.

3 Task and Dataset

3.1 Task Overview

The PalmX 2025 shared task offered two subtaks, one of which is *General Culture Evaluation* (Subtask 1). The goal of the task is to benchmark Arabic language models on their ability to answer culturally grounded multiple-choice questions in Modern Standard Arabic (MSA). The questions span various domains such as history, customs, geography, literature, and food, and are designed to reflect general cultural literacy in Arab countries.

Participants are provided with a training and development set of MCQs, each with four answer options. The final evaluation is performed on a held-out test set of 2,000 questions, with accuracy as the primary metric. The task encouraged the use of external data for model enhancement, provided that models remain under 13 billion parameters and final checkpoints are submitted for evaluation.

3.2 Dataset

The *PalmX 2025 Cultural Evaluation* dataset consists of 2,000 training examples and 500 development examples, each formulated as a MCQ with four answer options and a single correct answer. We used the training split for fine-tuning and reserved the development set for evaluation, except for our final iterations, where we use both training and evaluation splits for fine-tuning.

3.3 Data Augmentation

Palm. To complement PalmX dataset, we incorporated the Palm dataset (Alwajih et al., 2025a), a broader community-curated resource created by contributors from the 22 Arab countries. Unlike PalmX, which is entirely in MSA, Palm spans both MSA and various dialects, offering instruction-style QA pairs on 20 culturally relevant topics such as heritage, cuisine, history, and proverbs. All examples are manually written by native speakers with cultural familiarity, ensuring authenticity and regional diversity. Although Palm includes training and test splits, only the test portion, comprising 1,926 QA pairs, is publicly available.

We split the available Palm test set into two halves: one for fine-tuning, and the other for evaluation. The splits were created using stratified sampling based on country, ensuring balanced representation across regions in both halves. To bring its free-form QA format in line with PalmX, we converted each example into MCQ format using GPT-4.1. Specifically, we generated three plausible distractors per question, preserving semantic coherence and cultural plausibility. In Appendix 3, we provided the prompt that we used for MCQ version of the palm dataset.

Extending PalmX Dataset. To further diversify and expand our training data, we leveraged the NativQA framework (Alam et al., 2025) in combination with GPT-4.1. The NativQA framework can seamlessly curate large-scale QA pairs based on user queries, ensuring cultural and regional alignment in native languages. GPT-4.1 was selected for its optimal trade-off between cost and performance at the time of experimentation. In all cases, we employed zero-shot prompting with GPT-4.1.

As illustrated in Figure 1, our process for extending the PalmX dataset began by identifying the country associated with each question using GPT-4.1. The prompt used for this task is provided in Listing 3. This country information was then combined with the NativQA framework to curate location-specific QA pairs.

The NativQA framework's retrieval process was carried out in two iterations to maximize topical diversity. To maintain factual quality, all answers were filtered using NativQA's Domain Reliability Check (DRC), which retains only those sourced from NativQA-verified web domains. Furthermore, GPT-4.1 was employed to filter and refine the answers described in Listing 1. The idea is to remove culturally irrelevant or factually incorrect QA pairs and refine answers for conciseness and the overall quality of the dataset. Similar to the Palm test set, these new entries were converted into MCQ format to match PalmX, using the same prompt applied to the original Palm data. This process augmented the original dataset with culturally rich examples while preserving structural and contextual consistency with the PalmX questions. We also manually reviewed 50 samples, which received an average score of about 7.4 on a scale of 10 for accuracy and clarity. We refer to this dataset as the PalmX-ext set. In Table 1, we report the distribution of the dataset that we used for training and evaluation.

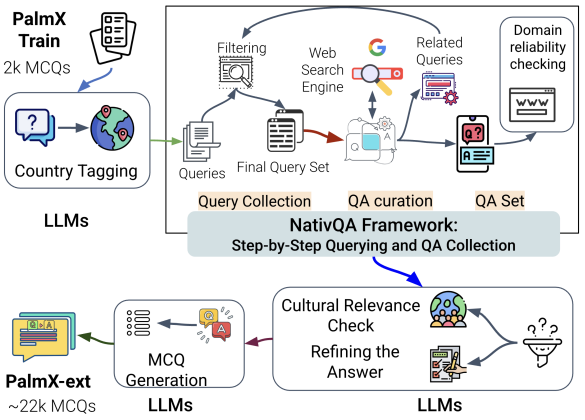

Figure 1: Pipeline for extending the PalmX dataset using the NativQA framework and GPT-4.1.

Data	Train	Dev	Test
PalmX	2,000	500	2,000
Palm	950	950	-
PalmX-ext	22,000	-	-

Table 1: Distribution of the datasets used for training, development and test.

4 Experiments

Models. We began by evaluating a set of open-sourced LLMs in a zero-shot setup on both evaluation datasets. This initial comparison helped us identify the most promising model for fine-tuning, and further demonstrated the utility of the Palm test set as an effective evaluation benchmark.

To identify the most suitable model for the task, we evaluated a set of models based on their performance on the PalmX development set. The models for experiments include `tiny-random-LlamaForCausalLM`,[2] `Qwen2.5-7B-Instruct` (Wang et al., 2024), `Jais-13B-Chat` (Sengupta et al., 2023), `Miraj Mini`,[3] `Llama-3.1-8B-Instruct` (Touvron et al., 2023), `NileChat-3B` (Mekki et al., 2025), `ALLaM-7B-Instruct` (Bari et al., 2025), and `Fanar-7B-Instruct` (Team et al., 2025). We selected both Arabic-centric and multilingual models to compare the effectiveness of models tailored to Arabic with those trained on broader multilingual corpora. The `tiny-random-LlamaForCausalLM` model was included for baseline results.

Training Setup. We experimented with two fine-tuning approaches: LoRA and QLoRA. LoRA trained only a set of low-rank adapter layers while keeping the rest of the model frozen, whereas

[2] https://huggingface.co/HuggingFaceH4/tiny-random-LlamaForCausalLM

[3] https://huggingface.co/arcee-ai/Meraj-Mini

QLoRA combined 4-bit quantization with LoRA adapters to reduce memory usage without a substantial drop in performance.

Both methods were trained on the same mix of datasets: PalmX Train, Palm, and PalmX-ext. We used Fanar's native tokenizer with its default tokenization strategy, a batch size of 4, and gradient accumulation steps of 4. Training was conducted for 3 epochs with a learning rate of 2×10^{-4}, saving the best-performing checkpoint at the end of each run. Fine-tuning followed a specific prompt - each question was prefixed by a system prompt, followed by the question and answer choices, as shown in Figure 2.

Data Augmentation. We then studied the effect of our data augmentation strategy by comparing LoRA training on **PalmX Train** alone *vs.* LoRA training on **PalmX Train** combined with our augmented **Palm** and **PalmX-ext** datasets. This experiment used the same configuration as the earlier LoRA *vs.* QLoRA comparison. After identifying the best-performing approach, we performed hyperparameter tuning to optimise its performance. In Appendix B and C, we report complete experimental setup and results, respectively.

After identifying the best-performing model, the most effective fine-tuning strategy, and the optimal hyperparameters, the final submission was trained for 3 epochs with a learning rate of 2×10^{-4}, LoRA rank 64, dropout 0.1, and scaling factor $\alpha = 16$, using PalmX Train and Dev, Palm, and PalmX-ext.

5 Results

Zero-shot Performance. Table 2 reports the zero-shot performance of several multilingual and Arabic-centric instruction models. **Fanar-7B** achieved the highest accuracy on PalmX Dev, making it our choice for fine-tuning.

Model	PalmX Dev	Palm
tiny-random-Llama	23.40	26.51
Qwen2.5-7B-Inst.	69.20	74.32
Jais-13B-chat	61.00	55.72
Miraj Mini	70.20	**75.99**
Llama3.1-8B-Inst.	66.60	74.06
Nilechat-3B	70.00	66.89
ALLaM-7B-Inst.	70.60	74.32
Fanar-7B	**72.40**	73.34

Table 2: Zero-shot performance of base models.

Comparison on PEFT methods. Table 3 compares LoRA with its quantized variant (QLoRA) under identical settings. LoRA achieved a slight improvement over QLoRA on PalmX Dev, suggesting that full-precision adapters were marginally more effective.

Method	PalmX Dev (%)
QLoRA (4-bit)	80.00
LoRA	**80.60**

Table 3: Results using PEFT methods.

Effect of Data Augmentation. Table 4 evaluates the impact of adding augmented Palm and PalmX-ext data to PalmX Train. The augmented dataset led to substantial gains on PalmX Dev, indicating improved generalization. A more detailed error analysis is provided in Appendix D.

Training Data	PalmX Dev (%)
PalmX	76.6
PalmX + PalmX-ext + Palm	**80.6**

Table 4: Results with and without augmented data.

The final submitted model achieved an accuracy of 84.1% on the Palm test set. As the PalmX development set was included in the training data, it was excluded from evaluation on the submitted model. On the blind test set, the model obtained an accuracy of 70.5%. A more detailed analysis of the discrepancy between Dev and Test performance is provided in Appendix E.

6 Conclusions and Future Work

In this paper, we present our system, *CultranAI*, designed to enhance cultural knowledge representation in LLMs for Arabic. We conduct an extensive comparative evaluation in a zero-shot setting using various multilingual and Arabic-centric models, which led us to identify *Fanar* as the most suitable model for further experimentation. To assess performance in low-compute scenarios, we explored different PEFT methods. We also investigated data augmentation techniques aimed at improving model accuracy. Our proposed system achieved an accuracy of 84.1% on the Palm set, and ranked 5^{th} on the blind test set with an accuracy of 70.5%. Future work will focus on refining data augmentation pipelines and further exploring model generalizability.

7 Limitations

While augmentation brought clear improvements, we believe the performance could have been higher with more careful dataset preparation. In the Palm dataset, instructional QAs were directly converted

to MCQ, but some QAs exceeded the 512-token PalmX limit. PalmX-ext avoided this by reformatting MCQs in the first post-processing step. Another problem was distractor quality: in both PalmX-ext and Palm, distractors were often shorter than the correct answer. These issues can be addressed by refining the prompts for distractor generation and adding a processing step to truncate long Palm QAs.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 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 (Volume 1: Long Papers)*, pages 487–520, St. Julian's, Malta. Association for Computational Linguistics.

Rawan Al-Matham, Kareem Darwish, Raghad Al-Rasheed, Waad Alshammari, Muneera Alhoshan, Amal Almazrua, Asma Al Wazrah, Mais Alheraki, Firoj Alam, Preslav Nakov, Norah Alzahrani, Eman alBilali, Nizar Habash, Abdelrahman El-Sheikh, Muhammad Elmallah, Haonan Li, Hamdy Mubarak, Mohamed Anwar, Zaid Alyafeai, and 24 others. 2025. BALSAM: A platform for benchmarking arabic large language models. *arXiv preprint arXiv:2507.22603*.

Firoj Alam, Md Arid Hasan, Sahinur Rahman Laskar, Mucahid Kutlu, and Shammur Absar Chowdhury. 2025. NativQA Framework: Enabling llms with native, local, and everyday knowledge. *arXiv preprint arXiv:2504.05995*.

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, and 25 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, 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.

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.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. Technical report, Microsoft Research.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. NativQA: Multilingual culturally-aligned natural query for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14886–14909, Vienna, Austria. Association for Computational Linguistics.

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

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. GooAQ: Open question answering with diverse answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 421–433, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. CultureLLM:: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. CulturePark: Boosting cross-cultural understanding in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 65183–65216.

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.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. BLEnD: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada.

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.

Rifki Putri, Faiz Haznitrama, Dea Adhista, and Alice Oh. 2024. Can llm generate culturally relevant commonsense qa data? case study in indonesian and sundanese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20571–20590.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, and 1 others. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An arabic-centric multimodal generative ai platform.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

A Prompts

<bos> You're a helpful Arabic assistant that answers multiple-choice questions accurately. Choose the best answer based only on the given question and options.
<start_of_turn>user
السؤال
A.
الخيار الأول
B.
الخيار الثاني
C.
الخيار الثالث
D.
الخيار الرابع
<end_of_turn>
<start_of_turn>model
Answer <end_of_turn>

Figure 2: Example of a formatted prompt used for Arabic MCQ fine-tuning.

```
system_prompt = """
You are an advanced NLP annotation assistant
specializing in evaluating Arabic questions and
answers. Your role is to classify questions,
assess answers, and refine them for conciseness
and accuracy.

Follow the structured guidelines for
classification:
- **Step 1: Evaluate and refine the answer**,
ensuring it is concise and factually correct.
- **Step 2: Determine if the question-answer
pair is relevant to the Arabic culture.

### **Annotation Task**
You are an expert Arabic NLP QA annotator. Your
task is to evaluate and refine a
question-answer pair based on the following
steps:

### **Step 1: Evaluate and Edit the Answer**
- **Answer Evaluation:**
  - **Correct:** Fully and accurately answers
the question.
  - **Incorrect:** Does not answer the question
or contains false information.
  - **Partially Correct:** Provides some
relevant information but is incomplete.
- **Answer Refinement:**
  - If correct or partially correct but **too
long, vague, or redundant**, rewrite it to be
**concise and precise**.

### **Step 2: Determine Arabic cultural
relevance**
- **Yes:** The question explicitly refers to
the Arabic culture.
- **No:** The question is about a different
culture than Arabic.
- **Unsure:** It is difficult to determine
whether the question refers to any specific
culture.
"""

user_prompt = f"""
### **Input Data:**

Question: {data['question']}
Answer: {data['answer']}

### **Your Response in JSON format:**
{
"answer_evaluation": "Correct" or "Incorrect"
or "Partially Correct",
"corrected_answer": "Provide a concise, precise
answer if needed, otherwise leave empty.",
"culture_relevance": "Yes" or "No" or "Unsure"
}
"""
```

Listing 1: Prompt for evaluating, refining, and filtering Arabic QA pairs.

```
system_prompt = """
You are an expert in educational content
creation specializing in Arabic language and
culture. Your task is to convert culturally
relevant question-answer pairs into
multiple-choice questions (MCQs) by generating
three plausible, culturally relevant, and
contextually appropriate incorrect answer
options (distractors) in Arabic for each
question.


Requirements:
- All options must be in Arabic.
- Distractors must be plausible and relevant to
the question.
- Avoid answers that are obviously incorrect,
unrelated, or closely paraphrase the correct
answer.
- Output only the 3 incorrect answers in the
following JSON format:

JSON Output format:
{{
"A.": "",
"B": "",
"C": ""
}}
"""

user_prompt = f"""
Given the following question and its correct
answer, generate 3 plausible but incorrect
answer options in Arabic.

Question: "{data['question']}"
Correct Answer: "{data['answer']}"
"""
```

Listing 2: Prompt for generating 3 plausible distractors.

```
system_prompt = "You are an AI assistant for
country identification."

user_prompt = """
You are an expert in Arab culture and geography.
Given a question in Arabic, your task is to
identify the most relevant Arab
country that the question is likely referring
to, either explicitly or implicitly.

Always return the name of a single Arab country
in English
(e.g., Qatar, Egypt, Saudi Arabia, UAE,
Morocco, etc.).

Even if the country is not directly named, use
cultural, linguistic,
environmental, or historical clues to infer the
closest matching Arab country.

Return your response in JSON format with a
single field "country"
containing only the country name.

QUESTION: "{question}"
"""
```

Listing 3: Prompt for identifying country.

Epochs	LR	r	Dropout	Alpha	PalmX Dev (%)
4	5e-5	64	0.15	16	79.6
5	1.2e-4	64	0.05	16	79.8
3	5e-5	64	0.05	16	79.8
4	5e-5	64	0.10	16	80.2
5	1e-4	32	0.05	32	80.4
3	2e-4	64	0.05	16	**80.5**

Table 5: PalmX Dev results from hyperparameter tuning.

B Hyperparameters

Hyperparameter tuning varied the number of epochs (3–5), the learning rates (5×10^{-5} to 2×10^{-4}), the dropout rates (0.05, 0.1, 0.15), and the LoRA-specific parameters such as the rank ($r = 32$ or 64) and the scaling factor ($\alpha = 16$ or 32). Starting from a baseline, we tested higher epochs, lower learning rates, and increased dropout for regularization effects, as well as a reduced-rank, higher-α variant ($r = 32$, $\alpha = 32$). Each configuration was trained and evaluated on the PalmX Dev set to ensure consistency in reporting.

C Results on the Hyperparameter Tuning

Fine-tuning experiments with Fanar-7B are summarized in Table 5. The top setup used 3 epochs, a 2×10^{-4} learning rate, LoRA rank 64, dropout 0.05, and $\alpha = 16$, yielding an average accuracy of 80.5 on PalmX Dev. We also observed a slight improvement when increasing the dropout to 0.1 in an earlier run with a similar configuration, and therefore incorporated this change into the top-performing setup to form our final configuration.

D Error Analysis: Effect of Augmentation

To better understand the impact of augmentation, we analyzed the subset of questions from the PalmX 2025 development set that the base model (Fanar-9B-Instruct) failed to answer correctly. Out of 500 questions, Fanar produced 138 errors.

Finetuning on PalmX alone corrected 38 of these errors. When augmented data was included, the model solved an additional 53 questions, while losing accuracy on only 3 of the 38 cases previously resolved. In total, the augmented model recovered 88 of the 138 initially incorrect items.

Representative examples of these improvements are shown in Figures 3 and 4. These illustrate how augmentation introduced broader topical coverage, especially on less-documented cultural and regional details. Without augmentation, the model remained limited to narrower knowledge encoded in PalmX.

Which of the following sequences accurately reflects the academic educational path of Dr. Nidal Shamoun in Syria?	أي من التسلسلات التالية يعكس بدقة مسار التحصيل العلمي الأكاديمي للدكتور نضال شمعون في سوريا؟
Which of the following sequences accurately reflects the official procedures followed to start an agricultural investment in the UAE?	أي من التسلسلات التالية يعكس بدقة الإجراءات الرسمية المتبعة لبدء استثمار زراعي في دولة الإمارات؟

Figure 3: Questions solved by both PalmX-only and Augmentation.

E Error Analysis: Dev vs. Test Performance

We also examined the discrepancy between the Dev and Test set performance. While our model showed strong results on Dev, its accuracy dropped considerably on Test. To better understand this, we compared representative samples of questions from Train, Dev, and Test.

What is the exact height of the central dome in the Emirates Palace from the ground?	ما الارتفاع الدقيق للقبة المركزية في قصر الإمارات من سطح الأرض؟
What is the most common recipe in Somalia that includes rice, pasta, and a mix of meats and vegetables?	ما هي الوصفة الأكثر شيوعًا التي تشمل الأرز والمعكرونة ومجموعة من اللحوم والخضروات في الصومال؟
How many religions are permitted to be practiced publicly in the Kingdom?	كم عدد الديانات المسموح بممارستها علنًا في المملكة؟
What is the most significant impact of the lack of cooperation between Tunisia and Europe in addressing the migration crisis, among the given options?	ما أهم تأثير لعدم تعاون تونس وأوروبا في معالجة أزمة الهجرة من بين الخيارات التالية؟

Figure 4: Questions solved only with Augmentation.

The train and dev sets are closely aligned, focusing on contemporary cultural, institutional, and social knowledge (see Figures 5 and 6). This alignment explains the stronger performance on dev: the model is effectively evaluated on material resembling what it was trained on.

By contrast, the test set introduces broader and less-represented domains, including ancient history, proverbs, zoology, and legal systems (Figure 7). These require background knowledge beyond the distribution covered in training, explaining the observed performance drop.

It should also be noted that model development and checkpoint selection relied on dev, while the test set remained hidden, reinforcing the discrepancy.

Which of the following factors primarily distinguishes the role of the Pope from that of the Patriarch in the Catholic ecclesiastical context?	أي العوامل التالية يُميّز بشكل أساسي دور البابا عن دور البطريرك في السياق الكنسي الكاثوليكي؟
Which of the following factors primarily distinguishes the Hebron clay pot (qudrah) from the regular clay pot in Palestinian cuisine?	أي العوامل التالية يُميّز بشكل أساسي القدرة الخليلية عن الفُخّارة في المطبخ الفلسطيني؟
Which of the following factors primarily distinguishes the Hebron clay pot (qudrah) from the regular clay pot in Palestinian cuisine?	كأي من الفعاليات التالية في الأردن يُركز بشكل رئيسي على دعم ريادة الأعمال التكنولوجية وتعزيز الابتكار؟
What is considered the most impactful achievement in the history of the Moroccan national football team?	ما الإنجاز الذي يُعتبر الأكثر تأثيرًا في تاريخ منتخب المغرب لكرة القدم؟
What was the main factor that enabled Yemeni football star Salem Said to gain widespread fame?	ما العامل الرئيسي الذي مكّن نجم كرة القدم اليمني سالم سعيد من اكتساب شهرة واسعة؟

Figure 5: Examples from PalmX Cultural Train Set.

Which of the following chess clubs affiliated with the Palestinian Chess Federation is directly linked to the city of Jerusalem?	أي من الأندية التالية المنتسبة للاتحاد الفلسطيني للشطرنج ترتبط مباشرةً بمدينة القدس؟
What are the main factors that contribute to the diversity of popular languages in the Republic of Djibouti?	ما العوامل الرئيسية التي تُساهم في تنوع اللغات الشعبية في جمهورية جيبوتي؟
Which of the following sectors are subject to restrictions on foreign investor ownership in the United Arab Emirates?	أي من القطاعات التالية تخضع لقيود على ملكية المستثمرين الأجانب في دولة الإمارات؟
Who is usually allowed to attend the bride's dance in Sudan?	من يُسمح له عادةً بحضور رقص العروس في السودان؟
Which of the following poets is the author of the book Dawa'ir al-Bouh in Jordan?	أي من الشُعراء التاليين هو مؤلف كتاب دوائر البوح في الأردن؟

Figure 6: Examples from PalmX Cultural Dev Set.

What was the ancient kingdom that the Syrian islands were part of?	ما هي المملكة القديمة التي كانت الجزر السورية جزءاً منها؟
What was the main factor that led to the end of the Damascus Spring period?	ما العامل الرئيسي الذي أدى إلى انتهاء فترة ربيع دمشق؟
What is the Levantine proverb commonly said about the month of July in Syria?	ما المثل الشامي الذي يُقال عن شهر تموز في سوريا؟
What is the maximum length that the hornless viper reaches in Saudi Arabia?	ما الطول الأقصى الذي يصل إليه الثعبان الأبتر في السعودية؟
What is the legal status of women's rights in Libya compared to men?	ما هو الوضع القانوني لحقوق النساء في ليبيا مقارنة بالرجال؟

Figure 7: Examples from PalmX Cultural Test Set.