AYA at PalmX 2025: Modeling Cultural and Islamic Knowledge in LLMs

Jannatul Tajrin[1*], **Bir Ballav Roy**[2], **Firoj Alam**[1]
[1]Qatar Computing Research Institute, Qatar
[2]BRAC University, Bangladesh
jannatultajrin33@gmail.com, bir.ballav.roy@g.bracu.ac.bd, fialam@hbku.edu.qa

Abstract

Culture fundamentally shapes human perception and reasoning, while religion—often embedded within cultural contexts—provides cohesive moral frameworks and a sense of community. The *PalmX 2025* shared task introduced two subtasks aimed at evaluating the capability of large language models (LLMs) to capture and represent culturally and Islamically grounded knowledge. In this paper, we present our participation in this shared task, leveraging parameter-efficient fine-tuning (PEFT) techniques in conjunction with targeted data augmentation strategies. We further conducted extensive zero-shot evaluations across a range of Arabic-centric and multilingual models to establish strong baselines and guide model selection. Our submitted system achieved competitive performance on the blind test sets, ranking 3^{rd} in Subtask 1 with an accuracy of 71.45% and 1^{st} in Subtask 2 with an accuracy of 84.22%.

1 Introduction

Culture is the shared system of meanings—values, norms, language, and rituals—that organizes how people perceive, decide, and relate (Hofstede, 2011). Religion, often a core strand of culture, provides moral frameworks, practices, and communities that guide conduct and purpose (Geertz, 2013). Attending to cultural and religious aspects improves communication, trust, and legitimacy, while reducing unintended harm and inequity across groups (Betancourt et al., 2003). Without culturally and religiously grounded priors, LLMs misinterpret idioms and taboos, amplify toxicity, and encode systematic biases, including documented anti-Muslim stereotypes (Gehman et al., 2020; Blodgett et al., 2020; Abid et al., 2021). Recent studies also show Western-leaning value biases and uneven cultural performance, demonstrating the need to encode diverse cultural and Islamic values in training data, safety policies, and evaluation suites (Li et al., 2024; Hasan et al., 2025).

To encode cultural, religious, and everyday knowledge, recent work has developed resources, methods, language-centric models, and benchmarks (Pawar et al., 2024). For Arabic, several LLMs have been pre-trained, including Jais (Sengupta et al., 2023), AceGPT (Huang et al., 2024), ALLaM (Bari et al., 2025), and Fanar (Team et al., 2025). While these models exhibit strong generative capabilities, many instruction-tuned variants rely heavily on synthetic or machine-translated data (e.g., Jais, AceGPT), which limits cultural knowledge and coverage. Moreover, most evaluations remain confined to general NLP and capability-oriented benchmarks (Abdelali et al., 2024; Mousi et al., 2025), with comparatively little attention to cultural and religious dimensions (Alwajih et al., 2025a). To advance the encoding of cultural and religious knowledge in Arabic-centric LLMs, the *PalmX 2025* shared task (Alwajih et al., 2025b) introduced a benchmark targeting Arabic cultural and islamic knowledge at both general and domain-specific levels, thereby enabling more inclusive and representative evaluations for the Arabic language and its diverse heritage. The shared task offered two subtasks. The annotated datasets for each subtask consists of human-validated multiple-choice question–answer pairs in MSA, ensuring both linguistic precision and cultural authenticity.

In this work, we benchmark multiple instruction-tuned LLMs across four configurations for both subtasks: (i) base, (ii) domain-specific fine-tuning, (iii) combined fine-tuning across subtasks, and (iv) data augmentation. Fine-tuning consistently improves performance on both subtasks: *Fanar-1-9B-Instruct* attains the higher accuracy on the cultural subtask (Subtask 1) under combined fine-tuning (80.8%), while *ALLaM-7B-Instruct* achieves the

[*] The contribution was made while the author was interning at the Qatar Computing Research Institute.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 830–836
November 8-9, 2025 ©2025 Association for Computational Linguistics

best accuracy on the Islamic subtask (Subtask 2) with augmented data (77.33%). Accordingly, we select the LoRA-based, combined fine-tuned Fanar model for Subtask1 and the ALLaM model with augmentation for Subtask 2 as our final systems. To summarize, our main contributions are:

- We present extensive baseline results for multiple LLMs under a zero-shot learning setup.
- Our proposed models achieved 3^{rd} place in Subtask 1 and 1^{st} place in Subtask 2.
- We show that paraphrase-based data augmentation yields notable performance gains for the islamic culture subtask.

2 Related Work

Recent advances in LLMs have demonstrated remarkable capabilities across a wide spectrum of natural language processing (NLP) tasks (Bubeck et al., 2023; Touvron et al., 2023; Abdelali et al., 2024; Dalvi et al., 2024). Beyond sheer model size, instruction tuning and preference optimization enhance both generalization and alignment, enabling models to follow user intent while delivering strong zero- and few-shot performance.

2.1 Cultural Knowledge

Recent work has begun to move beyond general Arabic capability benchmarks toward explicit evaluation of cultural competence. AraDiCE introduces a fine-grained dialect–culture suite spanning Gulf, Egypt, and Levantine, enabling targeted assessment of cultural awareness alongside dialectal understanding (Mousi et al., 2025). Country-specific evaluation is advancing as well: *SaudiCulture* focuses on regionally grounded cultural knowledge within Saudi Arabia (Ayash et al., 2025). Another recent effort proposed a framework, which highlights the significance of benchmarking LLMs with culturally embraced data, underlining the performance disparity between high and low resource language (Alam et al., 2025; Hasan et al., 2025). These efforts complement broader Arabic benchmarks such as LAraBench, which established multi-task capability evaluations but did not directly target cultural facets (Abdelali et al., 2024).

2.2 Islamic Knowledge

In contrast to the breadth of general cultural evaluation, Islamic/religious benchmarking remains limited in scale and linguistic coverage. *QUQA* evaluates GPT-4 on Classical-Arabic Qur'anic QA and reports modest exact-match and F1 scores, revealing limits even for state-of-the-art models in scripture-centric settings (Alnefaie et al., 2023). A cross-lingual Qur'anic QA effort expands a small Arabic set to 1,895 Arabic–English pairs and assesses pre-trained LMs/LLMs mainly with retrieval metrics (MAP@10, MRR, Recall@10), offering a first but narrow bilingual baseline (Oshallah et al., 2025). Retrieval-augmented studies over Qur'anic summaries examine faithfulness and citation via human ratings for open-source LLMs, but remain English-only and task-specific (Khalila et al., 2025). Multimodal cultural VQA benchmarks include religious practices and iconography across many languages; however, they target vision–language models and do not provide text-only, source-grounded Islamic QA suitable for doctrinal assessment (Vayani et al., 2025).

On the resource side, *Hajj-FQA* offers a human-annotated QA set over Hajj fatwas (Aleid and Azmi, 2025); *Fatwaset* compiles a large Arabic fatwa corpus with rich metadata for downstream NLP (Alyemny et al., 2023); and Qur'anic QA resources—such as the Qur'anic Reading Comprehension Dataset (QRCD) and subsequent retrieval/QA studies—provide task-specific testbeds while exposing issues of hallucination and domain brittleness (Basem et al., 2025). Collectively, these works lay important groundwork for Islamic-knowledge evaluation in Arabic; nevertheless, coverage remains narrow (few languages beyond Arabic/English), datasets are modest, and critical scholarly dimensions—madhhab/fiqh context, *hadith* authenticity, *tafsīr* grounding, awareness of abrogation (*naskh*), and dialectal/diacritic variation—are largely unencoded, leaving a clear gap relative to the methodological rigor now common in broader multicultural evaluation.

3 Tasks and Dataset

3.1 Tasks

The PalmX 2025 Shared Task evaluates Arabic General culture and Islamic knowledge through multiple-choice question answering in Modern Standard Arabic. It consists of two subtasks:

- **Subtask 1 – General Culture Evaluation:** This subtask evaluates the ability of LLMs to comprehend and reason about diverse aspects of Arabic general culture, including traditional customs, social etiquette, cuisine, historical events, notable figures, geography, arts, and dialectal expressions across different Arab countries. The focus is on assessing models' capacity to apply broad cultural knowledge that is relevant across the Arab world.

- **Subtask 2 – General Islamic Evaluation:** This subtask measures models' understanding of core elements of Islamic culture, which forms a foundational component of many Arabic societies. It covers topics such as Islamic rituals and practices (e.g., prayer, fasting), Quranic knowledge, Hadith literature, major historical developments in Islam, and religious holidays. Models are evaluated on multiple-choice questions designed to test both religious literacy and contextual sensitivity, ensuring they can handle culturally and theologically significant content with accuracy and respect.

3.2 Dataset

We used the dataset released as a part of PalmX 2025 Shared Task. For both subtasks the datasets has been formulated as MCQ format.

Data Augmentation. We employed paraphrase-based data augmentation to increase the diversity and robustness of the questions in the training data. In this approach, original questions were reworded into semantically equivalent variants while strictly preserving their intended meaning and correct answers. The resulting augmented dataset introduced controlled variations in phrasing, complexity, and syntactic structure, thereby encouraging better generalization. We used the GPT-4.1 model to paraphrase the questions. Listing 1 shows the exact prompt we used.

In Table 1, we report the detail distribution of the dataset. As reported in the Table, for both subtasks we applied data augmentation to increase training set size. As for the development and test set we have used same dataset released as a part of the shared task. Test set in the table refers to the blind test set. Note that in our initial set of experiments we have used dev set as a test set to evaluate models' performance.

Listing 1: Prompt for paraphrase based data augmentation.

```
system_prompt = (
    "You are a high-quality data augmentation
    assistant for Arabic multiple-choice
    question answering. "
    "Your job is to create adversarial variants
    of questions: rephrase or make the question
    more challenging "
    "or tricky, but do not alter its meaning or
    change which answer is correct. "
    "The answer options and the correct answer
    must remain valid for the new question."
    )
user_prompt = (
        "Below is an Arabic multiple-choice
        question with options and the correct
        answer indicated. "
        "Rewrite the question to make it
        slightly more challenging or confusing
        for test-takers "
        "(e.g., use more complex language, add
        subtle ambiguity, or require deeper
        understanding), "
        "but do not change its intended meaning
        or the correct answer. "
        "Return your answer as a JSON object
        with the new question, all original
        options, and answer letter
        preserved.\n\n"
        f"Question: {data['question']}\n"
        f"A: {data['A']}\n"
        f"B: {data['B']}\n"
        f"C: {data['C']}\n"
        f"D: {data['D']}\n"
        f"Answer: {data['answer']}"
    )
```

Subtasks	Train	Dev	Test
Culture	2,000	500	2,000
Culture + Islamic	2,600	500	2,000
Culture + Aug	4,000	500	2,000
Islam	600	300	1,000
Islam + Aug	1,200	300	1,000

Table 1: Dataset statistics for PalmX 2025 subtasks. Aug refers to data augmentation.

4 Experiments

4.1 Models

We have selected several instruction-tuned LLMs for the zero-shot evaluation and fine-tuning models on two subtasks – *General Cultural* and *Islamic Cultural* – under four configurations: base, fine-tuned, combined fine-tuning (culture + Islamic), and augmented data. The models considered included `Qwen2.5-7B-Instruct` (Wang et al., 2024), `Jais-13B-Chat` (Sengupta et al., 2023), `Miraj Mini`,[1] `Llama-3.1-8B-Instruct` (Touvron et al., 2023), `NileChat-3B` (Mekki et al., 2025), `ALLaM-7B-Instruct` (Bari et al., 2025), Gemma-7b-it[2] and `Fanar-7B-Instruct` (Team et al., 2025).

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

[2] https://huggingface.co/google/gemma-7b-it

4.2 Training

We fine-tuned the models using the LoRA approach. The LoRA configuration used a rank (r) of 16, an alpha value of 32, a base learning rate of 2e-4 and a dropout rate of 0.05, a maximum sequence length of 512 tokens, trained for three epochs, targeting the query and value projection layers of the transformer architecture. LoRA adapters were loaded from a prior checkpoint and the implementation of *attention* was set to *'eager'* for compatibility. The tokenizer for the base model was used, with the padding token aligned with the end-of-sequence token. The evaluation was performed with a batch size of 4 to accommodate the memory requirements of the 9B parameter model. The prompts were formatted for multiple choice answer with predefined choice prefixes (A, B, C, D).

Evaluation We evaluated models using accuracy, calculated as the percentage of correctly answered questions. For model training and internal evaluation, we were limited to the development dataset. Final evaluation and ranking were carried out by the organizers on the blind test set.

5 Results

The evaluation results across the PalmX subtasks demonstrate notable improvements through fine-tuning and data combination.

Cultural Subtask. In Table 2, we report the performance of cultural evaluation on the development set. For the PalmX General Cultural subtask, the ALLaM-7B-Instruct model improved from a base accuracy of 63.8 to 75.8 after fine-tuning, maintaining the same performance when combined with the Islamic dataset. Fanar-1-9B-Instruct outperformed ALLaM on this subtask, achieving 72.4 at base and improving to 80.2 after fine-tuning, with a slight increase to 80.8 using the combined data.

Islamic Subtask. In Table 3, we report the performance of Islamic evaluation on the development set. In the PalmX Islamic Culture subtask, ALLaM-7B showed a base accuracy of 72.7, which improved to 76.33 after fine-tuning and further to 77.33 with additional Islamic data augmentation. On the final hidden test set, Fanar-1-9B achieved an accuracy of 71.45 on the General Culture evaluation, while ALLaM-7B attained 84.22 on the General Islamic evaluation, indicating strong performance in their respective domains.

Error analysis. Figure 1, in Appendix, presents the confusion matrix for Subtask 1 on the hidden test set. The raw confusion matrix shows the proportion of correct and incorrect predictions per true label. Off-diagonal entries reveal misclassification patterns, with slightly higher confusion between classes **A** and **B** as well as **C** and **D**. Overall, the model demonstrates balanced accuracy across categories, with no single class dominating the error distribution.

Model	Base	Train Set	Comb.	Aug.
Gemma-7B-it	49.6	67.6	39.6	49.6
ALLaM-7B	63.8	75.8	75.8	70.6
Llama3.1-8B	66.6	66.6	66.6	66.6
Qwen2.5-7B	69.2	69.2	69.2	69.2
NileChat-3B	70.0	76.0	72.8	70.0
Fanar-1-9B	72.4	80.2	**80.8**	79.2

Table 2: Results on development set with a comparison of different models on PalmX *General Cultural subtask*. FT: Fine-tuned only on the PalmX training set, Comb.: Combined (Culture + Islamic), Aug.: Cultural + Augmented.

Model	Base	Train Set	Comb.	Aug.
Gemma-7B-it	40.0	40.0	25.3	25.3
Qwen2.5-7B	57.3	57.3	57.3	57.3
NileChat-3B	64.0	64.0	64.0	63.7
Fanar-1-9B	67.0	66.0	70.3	67.0
ALLaM-7B	72.7	76.3	76.3	**77.3**

Table 3: Performance comparison of different models on PalmX *Islamic Cultural subtask*. Comb.: Culture + Islamic, Aug.: Islamic + Augmented.

Subtasks	Model	Acc
Culture	Fanar-1-9B	71.5
Islamic	ALLaM-7B	84.2

Table 4: Performance on final hidden test set.

Figure 2, in Appendix, illustrates the confusion matrix for Subtask 2 on the hidden test set. The model correctly predicted **127**, **483**, **179**, and **54** instances for classes **A**, **B**, **C**, and **D**, respectively. The largest proportion of correct predictions occurred for class B, with relatively low misclassification rates. Notable confusion patterns include A → B (32 instances) and D → B (10 instances). While diagonal dominance is evident, indicating that the model captures the underlying class distinctions well, performance on class D is comparatively lower, suggesting a need for more targeted learning on that category. Overall, the results reflect strong performance, with class B exhibiting the highest classification accuracy in the Islamic

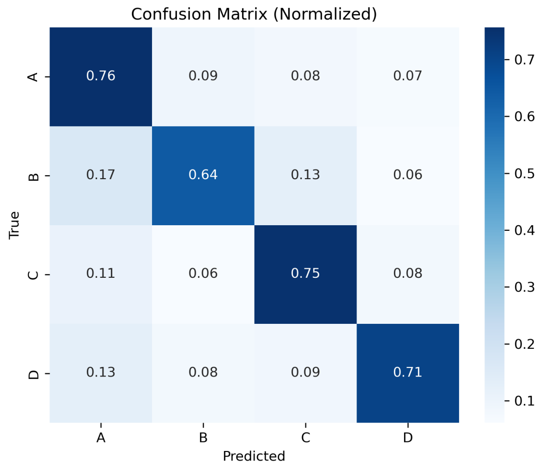

Figure 1: Confusion matrices for Subtask 1 (General Cultural Evaluation) on the hidden test set.

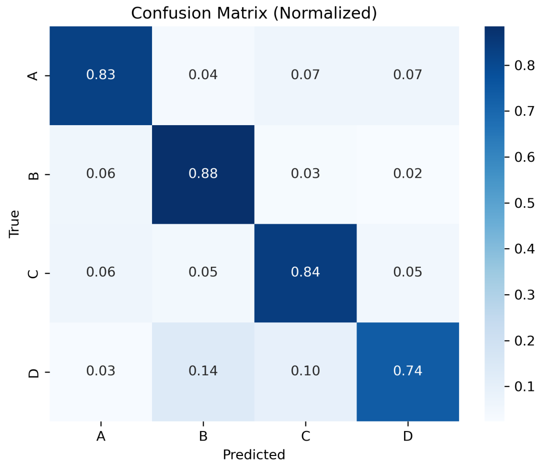

Figure 2: Confusion matrices for Subtask 2 (Islamic Cultural Evaluation) on the hidden test set.

cultural knowledge domain.

6 Conclusions and Future Work

In this paper, we presented our system developed for the Palmx 2025 Shared Task on MSA multiple-choice question answering in the Cultural and Islamic Evaluation. Our approach focused on fine-tuning state-of-the-art large language models, specifically ALLaM-7B-Instruct and Fanar-1-9B-Instruct, with data augmentation on domain-specific datasets, leveraging combined cultural and Islamic data to enhance performance. The Fanar-1-9B-Instruct model achieved the highest accuracy on the General Cultural subtask with 80.8 after fine-tuning and data combination, while ALLaM-7B-Instruct showed strong results in the Islamic subtask, reaching 77.33 accuracy with augmented data. On the final hidden test set, Fanar-1-9B-Instruct scored 71.45 on the General Culture evaluation, and ALLaM-7B-Instruct achieved 84.22 accuracy on the General Islamic evaluation. These results demonstrate the effectiveness of fine-tuning and data augmentation strategies in improving performance across different subtasks.

7 Limitations

While our experiments provide valuable insights into cultural and Islamic evaluation tasks, several limitations remain. Despite dataset variations (combined Islamic + General Cultural data and augmentation), some models such as *google/gemma-7b-it* and *Qwen2.5-7B-Instruct* showed nearly identical accuracy across settings, indicating limited sensitivity to data scale or diversity. We also observed accuracy decreases when training on combined datasets. However, performance overall across training setups is better than the base (unfine-tuned) model. These findings highlight the need for deeper error analysis, improved fine-tuning methods, and more robust data integration to better adapt language models for nuanced cultural understanding.

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.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

F. Alam, Md Asif Hasan, S. R. Laskar, M. Kutlu, Kareem Darwish, and S. A. Chowdhury. 2025. NativQA framework: Enabling llms with native, local, and everyday knowledge. *arXiv preprint arXiv:2504.05995*.

Hayfa A. Aleid and Aqil M. Azmi. 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:Article 135.

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. 2023. Is GPT-4 a good islamic expert for answering Quran questions? In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 124–133, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

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.

Ohoud Alyemny, Hend Al-Khalifa, and Abdulrahman Mirza. 2023. A data-driven exploration of a new islamic fatwas dataset for arabic nlp tasks. *Data*, 8(10):155.

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*.

Mohamed Basem, Islam Oshallah, Ali Hamdi, and Ammar Mohammed. 2025. Few-shot prompting for extractive quranic qa with instruction-tuned llms. *arXiv preprint arXiv:2508.06103*.

Joseph R Betancourt, Alexander R Green, J Emilio Carrillo, and Owusu Ananeh-Firempong 2nd. 2003. Defining cultural competence: a practical framework for addressing racial/ethnic disparities in health and health care. *Public health reports*, 118(4):293.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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.

Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shamur Absar Chowdhury, Hamdy Mubarak, Ahmed Ali, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2024. LLMeBench: A flexible framework for accelerating llms benchmarking. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Malta. Association for Computational Linguistics.

Clifford Geertz. 2013. Religion as a cultural system. In *Anthropological approaches to the study of religion*, pages 1–46. Routledge.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. 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.

Geert Hofstede. 2011. Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture*, 2(1):8.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 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: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Zahra Khalila, Arbi Haza Nasution, Winda Monika, Aytug Onan, Yohei Murakami, Yasir Bin Ismail Radi, and Noor Mohammad Osmani. 2025. Investigating retrieval-augmented generation in quranic studies: A study of 13 open-source large language models. *arXiv preprint arXiv:2503.16581*.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. 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.

Islam Oshallah, Mohamed Basem, Ali Hamdi, and Ammar Mohammed. 2025. Cross-language approach for quranic qa. *arXiv preprint arXiv:2501.17449*.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

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*.

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*.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, and 1 others. 2025. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19565–19575.

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*.