MarsadLab at PalmX Shared Task: An LLM Benchmark for Arabic Culture and Islamic Civilization

Md. Rafiul Biswas[1], Shimaa Ibrahim[2], Kais Attia[3], Firoj Alam[4], Wajdi Zaghouani[2]
[1]Hamad Bin Khalifa University, Qatar, [2]Northwestern University in Qatar, Qatar
[3]Independent Researcher, Tunisia, [4]Qatar Computing Research Institute, Qatar
{mbiswas,fialm}@hbku.edu.qa, wajdi.zaghouani@northwestern.edu

Abstract

This paper presents our submission to the PalmX 2025 Shared Task on Arabic cultural and religious knowledge comprehension. We focus on training large language models capable of representing domain-specific cultural and religious knowledge in Arabic. Our approach leverages parameter-efficient fine-tuning of the instruction-tuned Qwen2.5-7B model using Low-Rank Adaptation (LoRA). To address the challenges of limited training data, we apply quantization-aware fine-tuning with 4-bit precision, enabling efficient adaptation under constrained resources. The model is further aligned with the multiple-choice evaluation format to enhance task-specific reasoning. Without relying on external data augmentation, our system achieves competitive performance across both the *Arabic General Culture* and *Islamic Culture* subtasks, demonstrating the effectiveness of targeted fine-tuning for enriching cultural and religious knowledge representation in LLMs. On the blind test sets, our systems ranked 7^{th} and 4^{th} in the cultural and Islamic subtasks, respectively. To ensure reproducibility, we make our full codebase and experimental configurations available at https://github.com/rafiulbiswas/PalmX.

1 Introduction

Culturally aware language technologies are essential for high-stakes applications—education, public services, healthcare, and content moderation—where responses must be accurate, respectful, and contextually appropriate. In Arabic settings, a lack of cultural and religious grounding can lead to biased or inappropriate outputs, partly due to the predominance of Western-centric training data in large language models (LLMs) (Ayash et al., 2025; Alwajih et al., 2025b). Addressing this gap requires models that can represent and reason over Arabic cultural heritage and Islamic knowledge, as well as standardized evaluations that make such competence measurable (Sadallah et al., 2025a).

To advance cultural and islamic capabilities in Arabic-centric LLMs PalmX 2025 shared task offered two subtasks—*General Culture* and *Islamic Culture*—using multiple-choice (MCQ) datasets in Modern Standard Arabic (MSA) (Alwajih et al., 2025a). These subtasks probe models' ability to reason about customs, cuisine, history, and Islamic practices, providing a focused testbed for culturally grounded reasoning in Arabic.

Developing such capabilities is challenging. Beyond data imbalance, Arabic presents diglossia, rich morphology, and strong context dependence, all of which complicate knowledge representation and question answering (Hasan et al., 2025). Practical constraints—limited labeled data and domain-specific MCQ formats—further motivate resource-efficient adaptation strategies.

We adapt an instruction-following LLM to these subtasks using parameter-efficient fine-tuning. Concretely, we fine-tune the 7B-parameter Qwen2.5-Instruct (Team, 2025) with Low-Rank Adaptation (LoRA) (Hu et al., 2022) on the official PalmX training sets (Alwajih et al., 2025a), enabling effective domain adaptation under modest compute. At inference, we employ prompt-based strategies to inject expert priors and enforce output constraints (e.g., instructing the model to act as an "expert in Arabic culture and Islamic studies" and to output only the option letter). Empirically, careful prompt design yields consistent but modest gains in MCQ accuracy; closing the remaining gap will likely require richer cultural grounding and more structured supervision. To summarize, our contributions include:

- We adapt Qwen2.5-7B-Instruct to Arabic cultural and religious knowledge using Low-Rank Adaptation (LoRA) with 4-bit quantization-aware fine-tuning, achieving effective domain specialization under modest computational budgets.
- We introduce inference-time instruction templates and output-space constraints that align the

Proceedings of The Third Arabic Natural Language Processing Conference, pages 818–824
November 8-9, 2025 ©2025 Association for Computational Linguistics

model with the multiple-choice setting (expert prior + option-letter output), yielding consistent accuracy gains without additional supervision.

- Our system attains resonable performances on PalmX 2025 *General Culture* and *Islamic Culture*, ranking 7^{th} and 4^{th} on the blind test sets, respectively, without recourse to external data augmentation.
- We provide a concise pipeline demonstrating that low-compute PEFT can reliably enrich cultural/religious knowledge in Arabic LLMs.

2 Related Works

Benchmarking language models for Arabic has progressed along two complementary lines: inclusion within multilingual suites and dedicated evaluations of large language models (LLMs) for Arabic. Early efforts commonly incorporated Arabic into broad benchmarks such as XGLUE, XTREME, XTREME-R, GEM, and Dolphin, covering a spectrum of tasks that emphasized classification (e.g., natural language inference), sequence labeling (part-of-speech tagging, named entity recognition), and generation (summarization) (Liang et al., 2020; Hu et al., 2020; Ruder et al., 2021; Gehrmann et al., 2021; Nagoudi et al., 2023). More recent work has turned to Arabic-focused LLM assessment, evaluating standard and Arabic-centric models on task suites and datasets (Sengupta et al., 2023; Khondaker et al., 2023; Abdelali et al., 2024; Dalvi et al., 2024), probing the effects of prompting in native (Arabic) versus non-native (English) languages (Kmainasi et al., 2025), and extending analyses to multimodal settings (Alwajih et al., 2024; Das et al., 2024).

Within cultural evaluation, prior studies quantify representational bias in entity mentions toward Western versus Arab contexts (Naous et al., 2024), assess cultural alignment using constructs from the World Values Survey (AlKhamissi et al., 2024), and introduce culture-aware diagnostic and QA resources (Arora et al., 2024; Myung et al., 2024; Alam et al., 2025). Complementing these efforts, Arabic-focused benchmarks have begun to appear: ARADICE targets dialect comprehension and cultural QA (Mousi et al., 2024), while other resources probe cultural values and regional knowledge via translated survey instruments and Wikipedia-derived questions (Al-Matham et al., 2025). Despite these advances, converging evidence indicates that general-purpose LLMs still underperform on culturally grounded reasoning and Arabic commonsense, underscoring the need for benchmarks, resources, and model-development methods explicitly tailored to Arabic cultural and dialectal contexts (Sadallah et al., 2025b; Yakhni and Chehab, 2025; Qian et al., 2024).

PalmX 2025 (Alwajih et al., 2025a) advances this research area with a curated, competition-driven evaluation of Arabic cultural capabilities in Modern Standard Arabic, spanning *General Culture* and *Islamic Culture*. QAs are designed to cover all Arab countries and key Islamic concepts, providing a focused MCQ testbed and strong baselines (e.g., *NileChat-3B*) (Mekki et al., 2025; Alwajih et al., 2025b). Our work aligns with this direction by adapting an instruction-tuned LLM to PalmX via parameter-efficient fine-tuning and expert-persona prompting, and by analyzing remaining performance gaps relative to culturally trained baselines.

3 Dataset

PalmX 2025 provides two Modern Standard Arabic (MSA) multiple-choice (four-option) datasets that target complementary facets of cultural knowledge.

Task 1: General Culture comprises 4,500 questions spanning Arab culture across 22 countries, with official splits of 2,000 training, 500 development, and 2,000 test items.

Task 2: Islamic Culture contains 1,900 questions focused on Islamic cultural knowledge, split into 600 training, 300 development, and 1,000 test items. All experiments in this work use the organizers' official splits without external data augmentation.

Both subtasks follow a consistent data distribution structure, previously unseen questions for blind testing, with accuracy serving as the primary evaluation metric (see in figure 1).

4 System Overview

We experimented with different open sources LLM such as NileChat-3B (Mekki et al., 2025), LLaMA3.1 8B (Touvron et al., 2023), Fanar-1-9B-Instruct (Team et al.) and Qwen2.5-7B-Instruct (Team, 2025). Qwen2.5-7B-Instruct outperformed over other LLM and so we adapt Qwen2.5-7B-Instruct to Arabic cultural understanding via parameter-efficient fine-tuning with Low-Rank Adaptation (LoRA) (Hu et al., 2022).

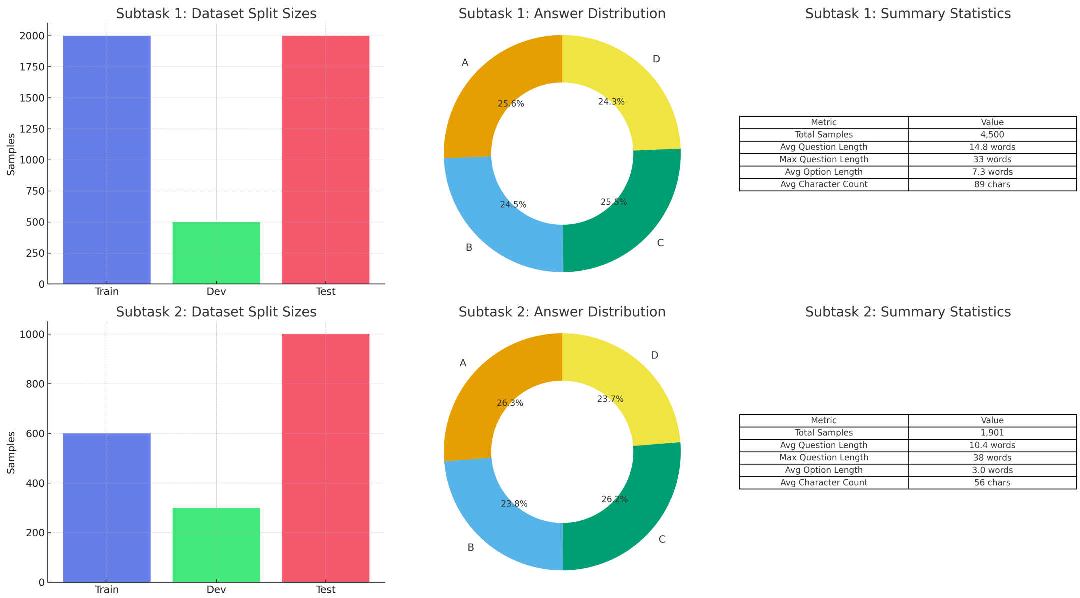

Figure 1: Dataset statistics in two subtasks.

4.1 Training Methodology

Prompting and supervision: We formatted training examples using a structured instruction-following template for Arabic cultural question-answering. Each instance comprises a system message, the user turn containing the question and the four labeled options, and an assistant turn with *only* the correct option letter. We implement this using the model's native chat template markers (`<|im_start|>` / `<|im_end|>`) to delimit turns. Explicitly constraining the target to the option letter suppresses verbosity, improves label consistency, and simplifies answer extraction at evaluation time; supervision is via standard next-token cross-entropy over the assistant turn.

Optimization setup: We train for three epochs (selected via development-set performance) with a learning rate of 2×10^{-4} and a linear warmup of 100 steps. We use an effective batch size of 16 via per-device batch size $= 4$ and gradient accumulation $\times 4$. Mixed precision uses `bfloat16` where supported (falling back to `fp16`), and the maximum sequence length is 512 tokens, which comfortably covers all MCQ contexts in our data.

Memory efficiency: To enable fine-tuning on commodity GPUs, we combine 4-bit NF4 quantization of the base weights with gradient checkpointing, trading additional compute for a reduced activation footprint. In practice, this configuration supports single-GPU training with $\sim$8 GB of memory. During preprocessing, we tokenize in mini-batches (size 100) to avoid holding the entire tokenized corpus in memory, and we periodically release cached CUDA memory to mitigate fragmentation during longer runs.

Multi-task adapters: For the two PalmX subtasks, we train separate LoRA adapters on the same quantized backbone to avoid negative transfer across cultural domains while retaining a unified deployment artifact. The *General Culture* adapter is fine-tuned on $\sim$2,000 instances, and the *Islamic Culture* adapter on $\sim$600 instances. This modular design permits task-specific specialization and lightweight "hot-swapping" at inference time without reloading the base model.

4.2 Ablation Study

To better understand the contribution of different components in our system, we conducted comprehensive ablation experiments examining the impact of LoRA hyperparameters, and prompt engineering choices.

LoRA Rank Analysis: We investigated the effect of LoRA rank on model performance and computational efficiency. Table 1 presents results for different rank configurations while keeping other hyperparameters constant ($\alpha = 32$, dropout=0.1).

The results reveal a clear performance improvement from rank 4 to 16, with diminishing returns beyond rank 16. Our chosen rank of 16 represents the optimal balance.

Target Module Selection: We evaluated different combinations of target modules for LoRA adaptation. Table 2 shows the impact of different mod-

Rank	Task 1 (%)	Task 2 (%)	Params (M)	Time (h)
4	63.2	69.8	5.24	2.1
8	65.8	72.1	10.49	2.4
16	**67.6**	**74.1**	**20.97**	**3.0**
32	67.9	74.3	41.94	3.8

Table 1: Effect of LoRA rank on performance on test dataset

ule combinations. Targeting all projection matrices yields the best performance, with attention modules alone outperforming Feed-Forward Neural (FFN) Network modules, suggesting that adapting attention patterns is more crucial for cultural understanding.

Target Modules	Task 1 (%)	Task 2 (%)
Attention (q, v)	64.3	70.2
Attention (q, k, v, o)	66.1	72.5
FFN only	63.7	69.4
All (Attn + FFN)	**67.6**	**74.1**

Table 2: Performance of different target module configurations

Prompt Engineering Variations: We tested several prompt variations to identify the most effective format for Arabic cultural questions.

Prompt Strategy	Task 1 (%)	Task 2 (%)
English system	64.7	71.2
Arabic system	66.3	72.8
Expert framing	**67.6**	**74.1**
Expert + few-shot	66.9	73.5

Table 3: Effect of prompt engineering strategies

The expert framing prompt that positions the model as "an expert in Arabic culture and Islamic studies" yields the best results. Adding chain-of-thought or few-shot examples slightly decreased performance.

5 Result

In Table 4, we report the performance of different models for both tasks before and after fine-tuning. Across both subtasks, performance varies markedly by model and adaptation strategy. The fine-tuned Qwen2.5-7B-Instruct yields the strongest overall results, attaining (67.55%) accuracy on *Task 1: General Culture* and (74.13%) on *Task 2: Islamic Culture*. Fine-tuning provides substantial improvements for all models except Fanar 7B on Task 2, with gains ranging from 7.8 to 12.2 percentage points on Task 1. Notably, Qwen2.5-7B demonstrates the most consistent improvement, gaining 7.75 points on Task 1 and 8.73 points on Task 2. Relative to the task with top ranked system (72.15%) and (84.22%), respectively, this places our best system within (4.60) percentage points on General Culture and (10.09) points on Islamic Culture, indicating substantial room for improvement, especially for the latter.

A cross-task comparison reveals a general trend of improved accuracy on the Islamic subtask after fine-tuning. For Qwen2.5-7B, the gain from Task 1 to Task 2 is (+6.58) percentage points. The NileChat-3B baseline is comparatively stable at (≈64%) on both tasks after fine-tuning, while Llama 3.1 8B-Instruct exhibits a modest uplift over this baseline on Islamic Culture (about (+4.9) points). An exception to the broader trend is Fanar 7B, which performs competitively on General Culture (66.0%) but declines on Islamic Culture (62.4%) compared to its baseline performance (49.6%), suggesting that while fine-tuning improves its general performance, domain- or data-mismatch effects persist that merit further analysis.

These results demonstrate three key observations. First, parameter-efficient fine-tuning confers clear benefits over off-the-shelf models for culturally grounded question answering in Arabic, with consistent improvements observed across most model-task combinations. Second, the effectiveness of fine-tuning varies by model architecture and task domain, as evidenced by the differential improvements across models. Third, the persistent gap to the subtask best scores—particularly on Islamic Culture—highlights the difficulty of capturing nuanced, domain-specific knowledge and the need for richer supervision and/or targeted knowledge integration beyond instruction tuning alone.

Model	General Culture		Islamic Culture	
	Before fine tuning(%)	After fine tuning(%)	Before fine tuning(%)	After fine tuning(%)
NileChat-3B	52.30	64.50	51.80	64.00
LLaMA3.1 8B	58.40	65.90	61.70	69.20
Fanar 7B	54.20	66.00	49.60	62.40
Qwen2.5L-7B	**59.80**	**67.55**	**65.40**	**74.13**

Table 4: Performance comparison of language models on test dataset before and after parameter-efficient fine-tuning. All scores represent accuracy percentages.

Computational efficiency. Our approach is computationally lightweight: training *Task 1* (2,000 samples) completes in approximately three hours on a single NVIDIA RTX 3090; with 4-bit quantization, fine-tuning fits within 8 GB of GPU memory. At inference, throughput is about $\sim$2 s per question on GPU and $\sim$8 s on CPU. The resulting LoRA checkpoint occupies $\sim$1.2 GB, compared to $\sim$15 GB for the full model,

6 Error Analysis

To better understand the limitations and failure modes of our fine-tuned models, we conducted a comprehensive error analysis on a stratified sample of 200 incorrect predictions from our best-performing model (**QWEN2.5L-7B**). Our analysis reveals distinct error patterns across tasks: for General Culture, the primary failure modes include factual knowledge gaps (42%), cultural context misunderstanding (28%), and ambiguous question interpretation (18%). For Islamic Culture, errors predominantly stem from religious text interpretation challenges (35%), difficulty handling sectarian variations (24%), and historical timeline confusion (21%).

When comparing errors across tasks, we also observed common problems such as relying on surface-level patterns instead of deeper understanding, showing overconfidence in culturally ambiguous cases, and favoring Western or standardized views over regional cultural perspectives. These findings suggest that while parameter-efficient fine-tuning improves performance, the models still face challenges in handling complex cultural reasoning that requires deeper context and sensitivity to local variations.

7 Conclusion

This paper presented MarsadLab's approach to the PalmX 2025 shared task on Arabic Islamic and Cultural understanding. Through parameter-efficient fine-tuning using LoRA adaptation of Qwen2.5-7B-Instruct, we achieved competitive performance across both tasks. Our work demonstrates that parameter-efficient methods can effectively adapt LLMs for culturally-nuanced tasks without requiring extensive computational resources. By training only 0.27% of model parameters through LoRA while employing 4-bit quantization, we reduced memory requirements by approximately 75% compared to full fine-tuning, making our approach accessible to researchers with limited GPU resources. Future work includes investigating low-compute and minimal-data regimes for such tasks.

Acknowledgments

This study was supported by the grant NPRP14C-0916-210015, awarded by the Qatar Research, Development and Innovation Council (QRDI).

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, and 1 others. 2025. BALSAM: A platform for benchmarking arabic large language models. *arXiv preprint arXiv:2507.22603*.

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*.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Fakhraddin Alwajih, Gagan Bhatia, and Muhammad Abdul-Mageed. 2024. Dallah: A dialect-aware multimodal large language model for arabic. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 320–336.

Fakhraddin Alwajih, Abdellah El Mekki, Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed, and Muhammad Abdul-Mageed. 2025a. PalmX 2025: The first shared task on benchmarking LLMs on arabic culture. In *Proceedings of the Third Arabic Natural Language Processing Conference (ANLP 2025)*, Suzhou, China. Association for Computational Linguistics.

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, and 1 others. 2025b. Palm: A culturally inclusive and linguistically diverse dataset for arabic llms. *arXiv preprint arXiv:2503.00151*.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2024. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. *arXiv preprint arXiv:2406.17761*.

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.

Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur 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*, pages 214–222, St. Julians, Malta. Association for Computational Linguistics.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791, Bangkok, Thailand. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, and 37 others. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, 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.

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.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 220–247, Singapore. Association for Computational Linguistics.

Mohamed Bayan Kmainasi, Rakif Khan, Ali Ezzat Shahroor, Boushra Bendou, Maram Hasanain, and Firoj Alam. 2025. Native vs non-native language prompting: A comparative analysis. In *Web Information Systems Engineering – WISE 2024*, pages 406–420, Singapore. Springer Nature Singapore.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, and 5 others. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

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. 2024. Aradice: Benchmarks for dialectal and cultural capabilities in llms. *arXiv preprint arXiv:2409.11404*.

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. *arXiv preprint arXiv:2406.09948*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. 2023. Dolphin: A challenging and diverse benchmark for Arabic NLG. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1404–1422, Singapore. Association for Computational Linguistics.

Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Zhaozhi Qian, Faroq Altam, Muhammad Alqurishi, and Riad Souissi. 2024. Cameleval: Advancing culturally aligned arabic language models and benchmarks. *arXiv preprint arXiv:2409.12623*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A. Sadallah, J. C. Tonga, K. Almubarak, S. Almheiri, F. Atif, C. Qwaider, K. Kadaoui, S. Shatnawi, Y. Alesh, and F. Koto. 2025a. Commonsense reasoning in arab culture. *Preprint*, arXiv:2502.12788.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025b. Commonsense reasoning in arab culture. *arXiv preprint arXiv:2502.12788*.

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. Fanar: An arabic-centric multimodal generative ai platform.

Qwen Team. 2025. Qwen2.5-vl.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Silvana Yakhni and Ali Chehab. 2025. Can llms translate cultural nuance in dialects? a case study on lebanese arabic. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 114–135.

Appendix

Model Architecture

We employ Qwen2.5-7B-Instruct, a 7.61B-parameter causal LLM comprising 28 transformer layers with Grouped Query Attention (GQA) and a context window of up to 131,072 tokens. For our setting, we use the instruction-tuned variant—optimized to follow complex prompts via supervised fine-tuning and RLHF. To reduce memory footprint, the base model is loaded with `4-bit` NF4 quantization (with double quantization) while retaining `bfloat16` compute through `bitsandbytes`; we enable `k-bit` training using `prepare_model_for_kbit_training`. Tokenization relies on the Qwen tokenizer with right padding; when a pad token is not defined, we map `<pad>` to `<eos>`.

LoRA Adaptation Strategy

To specialize both attention patterns and intermediate representations for culturally grounded reasoning, we attach LoRA adapters to the attention projections `q_proj`, `k_proj`, `v_proj`, and `o_proj`, as well as to the feed-forward projections `gate_proj`, `up_proj`, and `down_proj`. After development-set tuning, we adopt a rank $r = 16$, scaling $\alpha = 32$, dropout $= 0.1$, and no bias, a configuration that yields approximately **20.97M** trainable parameters ($\approx$**0.27%** of the base model). In practice, this supports single-GPU fine-tuning with $\sim$8 GB of memory while preserving sufficient capacity for the target tasks.

Hyperparameters

After empirical evaluation on the development set, we selected the following LoRA hyperparameters:

- **Rank (r)**: 16 - Balancing expressiveness with parameter efficiency
- **Scaling factor** (α): 32 - Controlling the magnitude of LoRA updates
- **Dropout**: 0.1 - Preventing overfitting on the limited training data
- **Bias**: None - Following standard LoRA practice

This configuration results in approximately **20.97M trainable parameters** (0.27% of total model parameters), enabling fine-tuning with only 8GB of GPU memory while maintaining model expressiveness for cultural reasoning tasks.