

ISL-NLP at PalmX 2025: Retrieval-Augmented Fine-Tuning for Arabic Cultural Question Answering

Mohamed Gomaa, Nouredin Elmadany

Arab Academy for Science, Technology and Maritime Transport

Intelligent Systems Laboratory

Alexandria, Egypt

m.g.abdalla1@student.aast.edu, noureldmadany@aast.edu

Abstract

Cultural understanding is essential for large language models (LLMs), particularly in the Arabic context where many models struggle to capture nuanced cultural elements. To address this gap, we propose a novel approach for Arabic cultural multiple-choice question answering that integrates retrieval-based training data augmentation with parameter-efficient fine-tuning. Our system employs Gemini¹ to retrieve contextual evidence for each question, selects candidate pairs, and adapts NileChat-3B by fine-tuning only three projection layers, reducing trainable parameters by 68.2% while preserving general language proficiency. On the PalmX 2025 Subtask 1 benchmark², our system attains 67.60% accuracy on the blind test set, ranking 6th overall and outperforming the NileChat-3B baseline by 3% on the development set. The model weights are publicly available at [MohamedGomaa30/Ibn-Al-Nafs](#).

1 Introduction

The PalmX 2025 ([Alwajih et al., 2025](#)) provides a rigorous Arabic cultural benchmark³ for evaluating AI systems in Arabic, particularly their ability to reason within complex cultural, religious, and historical contexts. This task addresses a key gap in Arabic natural language processing (NLP) by focusing on multiple-choice questions that require cultural reasoning rather than surface-level fact recall.

We tackle this challenge with a two-stage architecture that combines contextual retrieval and parameter-efficient model adaptation, motivated by two observations: (1) Arabic LLMs often lack cultural knowledge available in existing data, and (2) full fine-tuning of large models is computationally

expensive and risks catastrophic forgetting of general linguistic abilities.

- **Contextual Retrieval** – We employ Gemini’s retrieval features with structured prompts to automatically attach concise (≤ 50 words) contextual evidence to each question-answer pair in the PalmX 2025 subtask1 dataset. The retrieved evidence captures cultural, geographical, and historical information.
- **Model Adaptation** – We adapt NileChat-3B ([Mekki et al., 2025](#)) by fine-tuning only three projection layers: q_proj for question representation, v_proj for value transformation, and $gate_proj$ for information routing. This yields a 68.2% reduction in trainable parameters compared to full fine-tuning.

Our system ranked 6th on the Palmx 2025 leaderboard with 67.70% accuracy on the blind test set, surpassing the NileChat-3B baseline by 3% on the development set. These results demonstrate that targeted architectural choices can improve cultural reasoning in LLMs while preserving computational efficiency and real-world deployability. The remainder of this paper is organized as follows. The background is presented in Section 2. In Section 3, we provided The details of our proposed system are described in Section 3. In Section 4, the experimental results and their analysis are given. Finally, we conclude this paper in Section 5.

2 Background

2.1 Task Setup

The task evaluates the large language model’s ability to understand Arabic culture, covering history, geography, arts and customs and traditions for Arabic countries. The input consists of text-based multiple-choice questions (MCQs) and context-aware text that is related to the question that will help the model distinguish the correct answer. This

¹We use the [Gemini 2.5 Pro](#) API for contextual evidence retrieval.

²<https://example.com/palmx2025>

³<https://palmx.dlnlp.ai>

is for the training phase only in modern standard Arabic, with the model selecting one correct answer (A, B, C, or D) from four options.

2.1.1 Input Example for Training Phase

ما الملاح الأدبية التي تميز إبداعات محمد الماغوط في السياق الثقافي السوري؟

Options:

- A. تطوير نمط القصة القصيرة الاجتماعية في الصحافة المحلية
- B. تأليف روايات تاريخية مستوحاة من الثورة السورية
- C. دمج القصيدة النثرية مع المسرح السياسي الساخر في أعماله
- D. إحياء الشعر العمودي التقليدي مع إضافة عناصر فلسفية

Context:

يُعرف الماغوط بتأسيسه لـ الشعر الحُر في سوريا، مبتعدًا عن عمودية الشعر التقليدي، كما برع في كتابة القصيدة النثرية. أعماله المسرحية، مثل أشقائنا النعمان وكاسك يا وطن، اتسمت بالجرأة والسخرية السياسية، وهي سمة بارزة في الأدب السوري المعاصر

Output: C

2.1.2 Dataset Preparation

The training dataset is enriched with evidence-based context retrieved through Gemini, which provides historical, geographical, and cultural facts for each multiple-choice pair. This contextual information guides the model in learning cultural cues and improves its ability to select the correct answer.

2.2 Dataset Details

The PalmX 2025 Subtask 1 dataset targets Arabic cultural knowledge, covering customs, traditions, and general background across different Arab countries. The task is evaluated through multiple-choice questions (MCQs), organized as follows:

- **Training Set:** 2,000 MCQ pairs.
- **Development Set:** 500 MCQ pairs for intermediate evaluation.
- **Blind Test Set:** 2,000 unseen MCQ pairs, balanced across countries and domains.

2.3 Related Work

Cultural Alignment in LLMs: Cultural alignment for Large Language Models (LLMs) has received growing attention due to concerns over the dominance of Western perspectives and the marginalization of non-Western cultures (AlKhamissi et al., 2024; Wang et al., 2024). Prior studies show that

existing models often fail to capture nuanced cultural variables, leading to irrelevant or biased outputs (Mihalcea et al., 2024; Ryan et al., 2024). This challenge is particularly pronounced for underrepresented linguistic communities such as Arabic speakers, whose cultural diversity is frequently oversimplified (Keleg, 2025).

Arabic Cultural Nuances and LLMs: Several Arabic-centric LLMs have recently been introduced to address these gaps. NileChat-3B is the first Arabic model adapted for Egyptian and Moroccan communities, designed to incorporate dialects, customs, and traditions. Jais (Sengupta et al., 2023) is a bilingual Arabic–English model trained on hundreds of billions of tokens, demonstrating improved reasoning and knowledge in Arabic. AceGPT (Huang et al., 2023) is tailored for Arabic-speaking communities by aligning cultural and linguistic features. Fanar (Abbas et al., 2025) is trained on one trillion Arabic and English tokens and explicitly aligned with Islamic values and Arab cultures. ALLaM (Bari et al., 2024) achieves state-of-the-art performance across several Arabic benchmarks, including Arabic MMLU (Hendrycks et al., 2020), ACVA, and Arabic Exams.

Cultural QA Benchmarks and Technical Adaptation: New benchmarks have advanced cultural evaluation in Arabic NLP, including:

- **ArabicMMLU** (Koto et al., 2024), focusing on educational and academic subjects.
- **ArabDCE-Culture** (Mousi et al., 2024), targeting cultural fact-based QA across diverse Arab countries.
- **BLEnD** (Myung et al., 2024), evaluating everyday Algerian contexts.

From the perspective of model adaptation, improvements in cultural QA have been supported by Parameter-Efficient Fine-Tuning (PEFT) techniques (Xu et al., 2023). Rather than updating all parameters—which is computationally expensive and risks catastrophic forgetting—PEFT updates only a small subset of weights. This reduces memory and compute requirements while enabling targeted adaptation to culturally specific datasets.

3 Proposed System

Our system follows a two-stage pipeline that combines contextual retrieval with parameter-efficient

fine-tuning to address the challenges of Arabic cultural multiple-choice question answering (MCQ). In the first stage, we leverage Gemini’s retrieval capabilities to enrich the dataset with culturally relevant evidence. In the second stage, we adapt NileChat-3B through partial fine-tuning of selected layers, reducing computational cost while preserving performance.

3.1 Key Algorithms and Design Decisions

We adopt NileChat-3B as the base model due to its strong performance on Arabic language understanding tasks, particularly in Egyptian and Moroccan contexts. Instead of full fine-tuning—which is computationally expensive and risks catastrophic forgetting—we selectively update only three projection layers: **q_proj** for question representations, **v_proj** for value transformations in attention layers, and **gate_proj** for information routing in feed-forward layers. This strategy reduces trainable parameters by **68.2%** compared to full fine-tuning, improving training efficiency while preserving general linguistic capabilities.

3.2 Resources Beyond Provided Training Data

While PalmX 2025 subtask1 is the primary training dataset, we augmented it with retrieval-augmented context. Using Gemini, we generated concise evidence from trusted cultural, historical, and geographical sources for each MCQ pair. This additional context strengthens the model’s ability to make culturally informed decisions beyond surface-level associations.

3.3 Rationale for Training-Time Context Augmentation

The positive effect of training-time context augmentation comes from latent concept alignment rather than memorization. The model is taught to link superficial cues in questions and answers with their underlying cultural principles through the supervisory signal provided by the (question, context, answer) triplets. The internal representations of the model are improved during training in order to encode these patterns of cultural reasoning. As a result, the model exhibits enhanced generalization without explicit context when it is tested, recognizing pertinent cultural cues in unaugmented questions and deducing the right response from its learned conceptual understanding.

3.4 Addressing Task Challenges

The task presents two main challenges. First, Arabic cultural questions require nuanced contextual knowledge beyond factual recall. To address this, we utilized Gemini to retrieve concise, culturally grounded evidence for each MCQ pair, enabling the model to reason with supporting information rather than relying solely on memorization. Second, limited computational resources constrained model training. To mitigate this, we employed parameter-efficient fine-tuning, updating only the **q_proj**, **v_proj**, and **gate_proj** layers of NileChat-3B. This approach reduces computational overhead and mitigates catastrophic forgetting while maintaining strong performance.

3.5 Implementation Details

We implemented our system using **PEFT**⁴, the **SFTrainer** from the **TRL** (0.8.2) library⁵, and the **Transformers** library (v≥4.41.0)⁶. The dataset was formatted into instruction-response pairs, with a structured Arabic prompt guiding the model to analyze each question, consider candidate answers, and output a single-letter choice (**A, B, C, D**). Training was conducted for three epochs with eight gradient accumulation steps and a per-device batch size of two. To optimize memory and efficiency, we used the **AdamW** optimizer⁷, **FP16** mixed precision, a learning rate of 2×10^{-4} , and non-reentrant gradient checkpointing.

4 Experimental Results

4.1 Data Splits

The official palmx_2025_subtask1_culture dataset is divided into a training set of 2,000 multiple-choice question–context pairs, a development set of 500 pairs for validation, and a blind test set of 2,000 unseen pairs balanced across cultural domains.

4.2 Data Preprocessing

Each question was augmented with culturally relevant evidence retrieved using Gemini. For every question–answer pair, we constructed a structured prompt in Modern Standard Arabic. The prompt

⁴<https://github.com/huggingface/peft>

⁵We use the supervised fine-tuning component (SFTrainer) from <https://github.com/huggingface/trl>

⁶<https://github.com/huggingface/transformers>

⁷<https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html>

instructed Gemini to retrieve concise historical, geographical, or cultural evidence (≤ 50 words) that distinguishes the correct answer from distractors, without explicitly revealing the answer.

4.3 Experimental Settings

We fine-tuned NileChat-3B using the PEFT approach, updating only three projection layers (**q_proj**, **v_proj**, and **gate_proj**). Training was conducted on two NVIDIA T4 GPUs (15 GB each) for three epochs with a per-device batch size of 2, gradient accumulation steps of 8, a learning rate of 2×10^{-4} , and FP16 mixed precision. Optimization used AdamW.

Our implementation relied on the **Transformers** ($v \geq 4.41.0$), **TRL** (0.8.2), and **PEFT** libraries from Hugging Face, with dataset handling via **Datasets** and retrieval through Gemini’s API. All preprocessing and training scripts will be released publicly for reproducibility.

4.4 Results

We compare our approach on the development split against the model base NileChat-3B to measure the improvement from our method. , and against general-purpose state-of-the-art Arabic models (**Qwen2.5-1.5B** and **Qwen1.5-1.8B**). Using the official metrics of precision, recall, F1-score, and accuracy at Table 1. Our system achieves the best performance across all metrics, with notable improvements over both baselines. The proposed system outperforms Qwen2.5-1.5B by approximately 10% in precision, recall, F1-score, and accuracy, demonstrating its effectiveness.

Model	Pre.	Recall	F1-S	Acc.
Qwen2.5-1.5B	64.73	63.89	63.59	63.60
Qwen1.5-1.8B	63.24	60.88	59.15	59.80
NileChat-3B	71.74	70.00	69.92	70.00
Our system	73.81	73.88	73.54	73.60

Table 1: Performance on the development set of PalmX 2025 subtask1, Values are percentages.

5 Conclusion

This paper introduced a parameter-efficient, retrieval-augmented approach for Arabic cultural multiple-choice question answering. Our method combines Gemini-based contextual evidence retrieval with selective fine-tuning of NileChat-3B’s projection layers. The approach achieves a **3.0%** improvement over the base model on the development set and ranks **6th** on the official Palmx 2025

leaderboard, showing that targeted architectural adjustments can enhance cultural reasoning while remaining computationally feasible.

However, two limitations remain. First, the cultural knowledge base depends on the coverage and quality of retrieved evidence, which may miss region-specific details. Second, the selective fine-tuning strategy, while efficient, may restrict improvements in tasks requiring broad cross-cultural reasoning or temporal understanding.

Future work will extend the retrieval corpus to cover richer regional variations, integrate temporal reasoning modules for handling historical timelines, and explore hybrid adaptation strategies that combine parameter-efficient fine-tuning with lightweight full-layer updates. These directions aim to further strengthen cultural comprehension in Arabic NLP systems.

5.1 Acknowledgments

We thank the organizers of the Palmx 2025 shared task for providing the dataset and evaluation platform, as well as the anonymous reviewers for their valuable feedback. We also acknowledge the support of the Kaggle platform for providing GPU resources used in this work.

References

- Fanar Team Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur A. Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed G. Elfeky, Ahmed K. Elmagarmid, Mohamed Y. Eltabakh, Masmoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus’ab Husaini, Soon-Gyo Jung, Ji Lucas, Walid Magdy, Safa Mes-saoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Ahmad Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. [Fanar: An arabic-centric multimodal generative ai platform](#). *ArXiv*, abs/2501.13944.
- 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, Abdellah El Mekki, Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed, and

- Muhammad Abdul-Mageed. 2025. [PalmX 2025: The First Shared Task on Benchmarking LLMs on Arabic and Islamic Culture](#). In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.
- 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, Majed Alrubaian, Ali Alammari, Zaki Alawami, AbdulMohsen O. Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairish, Areeb Alowisheq, and Haidar Khan. 2024. [Allam: Large language models for arabic and english](#). *ArXiv*, abs/2407.15390.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. 2020. [Measuring massive multitask language understanding](#). *ArXiv*, abs/2009.03300.
- Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2023. [Acegpt, localizing large language models in arabic](#). In *North American Chapter of the Association for Computational Linguistics*.
- Amr Keleg. 2025. [LLM alignment for the Arabs: A homogenous culture or diverse ones](#). In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 1–9, Albuquerque, New Mexico. Association for Computational Linguistics.
- Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Al-mubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. [ArabicMMLU: Assessing massive multitask language understanding in Arabic](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. 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*, abs/2505.18383.
- Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Tamar Solorio. 2024. [Why AI is WEIRD and should not be this way: Towards AI for everyone, with everyone, by everyone](#). *arXiv preprint*, arXiv:2410.16315.
- Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arif Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur A. Chowdhury, and Firoj Alam. 2024. [Aradice: Benchmarks for dialectal and cultural capabilities in llms](#). *ArXiv*, abs/2409.11404.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Pérez-Almendros, Abinew Ali Ayele, V'ictor Guti'erez-Basulto, Yazm'in Ib'anez-Garc'ia, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Djouhra Ousidhoum, José Camacho-Collados, and Alice Oh. 2024. [Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages](#). *ArXiv*, abs/2406.09948.
- Michael J Ryan, William Held, and Diyi Yang. 2024. [Unintended impacts of LLM alignment on global representation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140, Bangkok, Thailand. Association for Computational Linguistics.
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Arun Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, A. Jackson, Preslav Nakov, Timothy Baldwin, and Eric P. Xing. 2023. [Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models](#). *ArXiv*, abs/2308.16149.
- Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. [Not all countries celebrate thanksgiving: On the cultural dominance in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.
- Lingling Xu, Haoran Xie, S. Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. [Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment](#). *ArXiv*, abs/2312.12148.