ADAPT–MTU HAI at PalmX 2025: Leveraging Full and Parameter-Efficient LLM Fine-Tuning for Arabic Cultural QA

Shehenaz Hossain[1] **& Haithem Afli**[1]

[1]ADAPT Centre, Computer Science Department,
Munster Technological University, Cork, Ireland

Correspondence: shehenaz.hossain@mymtu.ie, haithem.afli@mtu.ie

Abstract

We present ADAPT–MTU HAI's submission to PalmX 2025, targeting Arabic cultural question answering through large language model (LLM) adaptation. We apply full fine-tuning on NileChat-3B for general cultural comprehension, and parameter-efficient LoRA-based tuning on ALLaM-7B for Islamic knowledge reasoning. Our models achieved first place in the General Culture subtask and third place in the Islamic Culture subtask. This paper outlines our methodology and results, demonstrating the effectiveness of aligning LLM fine-tuning strategies with cultural knowledge domains.

1 Introduction

Language is not merely a tool for communication—it embodies the cultural, historical, and religious identities of its speakers. In Arabic, this interplay is particularly intricate: expressions are shaped by centuries of regional diversity, theological tradition, and social customs (Habash, 2010; Zitouni, 2011; Farghaly and Shaalan, 2009; Darwish et al., 2021). As large language models (LLMs) become increasingly central to NLP applications (Antoun et al., 2020; Touvron et al., 2023; Huang et al., 2024b), a pressing question arises—can these models truly reason over culturally embedded content, especially in linguistically rich and context-dependent settings such as Arabic?

The PaLMX 2025 shared task (Alwajih et al., 2025) [1] directly addresses this challenge through two subtasks. **Subtask 1** focuses on Arabic cultural comprehension, evaluating LLMs on multiple-choice questions (MCQs) covering general cultural knowledge like geography, customs, historical figures, dialectal expressions, and more.

Subtask 2 targets Islamic knowledge reasoning, assessing understanding of Quranic principles, Hadith, and theology. Both subtasks require models to go beyond surface-level fluency and demonstrate genuine cultural and contextual alignment.

Our team submitted systems to both subtasks, building tailored solutions to address their unique requirements. For Subtask 1, we fine-tuned **NileChat-3B** (Mekki et al., 2025), a culturally grounded decoder-only model adapted for North African Arabic under the Language–Heritage–Values (LHV) framework. For Subtask 2, we employed **ALLaM-7B-Instruct** (Bari et al., 2024), an Arabic instruction-tuned model, and applied **parameter-efficient fine-tuning** using LoRA (Brown et al., 2020) with **8-bit quantization** (Dettmers et al., 2023) to reduce memory usage without sacrificing accuracy.

On the official leaderboard, our systems ranked **first** in Subtask 1 with prompt-aligned full fine-tuning for cultural QA, and **third** in Subtask 2, where efficient adaptation highlighted the strength of lightweight tuning in resource-constrained settings.

This paper presents our unified approach to both subtasks. Section 2 summarizes related work, Section 3 details our methodology and training setups, Section 4 discusses results and analysis, and Section 5 concludes with reflections on cultural modeling in Arabic LLMs.

2 Related work

Research on embedding Islamic cultural knowledge into NLP systems is still emerging, though select initiatives have begun to address this need (Saadaoui et al., 2024). The Qur'an QA Shared Task (Malhas et al., 2022, 2023)[2][3] introduced the Qur'anic Reading Comprehension Dataset (QRCD), composed of approximately 1,093 question-passage pairs derived from the Holy Qur'an in Modern Standard Arabic. Participating systems, including AraBERT-based

[1]https://palmx.dlnlp.ai/

[2]https://sites.google.com/view/quran-qa-2022

[3]https://sites.google.com/view/quran-qa-2023

Proceedings of The Third Arabic Natural Language Processing Conference, pages 802–808

models, achieved modest Exact Match (EM) scores below 35%, highlighting the challenge of reasoning over sacred religious text (Mostafa and Mohamed, 2022).

Following this, Hajj-FQA (Aleid and Azmi, 2025) was released in 2025 as the first Arabic dataset targeting pilgrimage-related fatwa questions, offering realistic legal and religious Q&A reflective of common Hajj scenarios (Alyemny et al., 2023). The Hadith-QA corpus expands Islamic QA further by focusing on Prophetic narrations, while IslamicPCQA provides a rich Persian multi-hop benchmark (12,282 QA pairs) over Islamic encyclopedic content, illustrating cross-lingual interest in knowledge reasoning even beyond Arabic contexts (Ghafouri et al., 2023). Recent work has also introduced large-scale QA resources for deep religious understanding, (Qamar et al., 2024) presented a 73,000-question dataset spanning Quranic Tafsir and Ahadith, enriched with contextual explanations and interpretations to support nuanced QA system development.

Additionally, the CAMeL cultural bias benchmark evaluates Arabic LLMs' performance on culturally sensitive prompts, confirming consistent issues with Western-centric bias and cultural misalignment in language models (Naous et al., 2024). In recent years, several Arabic and Arabic-English LLMs have been introduced — including FANAR (Team et al., 2025), JAIS (Sengupta et al., 2023), AceGPT (Huang et al., 2024a), and ALLaM (Bari et al., 2024).In parallel, Arabic cultural and dialectal (Hossain et al., 2025; de Francony et al., 2019) evaluation benchmarks such as CamelEval (Qian et al., 2024) and AraDiCE (Mousi et al., 2024) have foregrounded the importance of cultural alignment, dialect robustness, and domain sensitivity in LLM evaluation—factors directly relevant to legal-religious reasoning While these models demonstrate impressive general reasoning and instruction-following ability, independent evaluations reveal that they still inherit cultural biases and struggle with nuanced religious and historical content. For example, (Mohammed et al., 2025) show that even GPT-4 can produce factually incorrect or inconsistent responses to Islamic content due to misinterpreting context, lacking grounding in authoritative sources, and being sensitive to minor wording changes. Similarly, (Alnefaie et al., 2023) report that GPT-4 struggles with Quranic questions, largely because of challenges in classical Arabic, semantic ambiguity, and contextual interpretation.

Despite the advances, structured MCQ-style benchmarks focused specifically on Islamic cultural literacy in Arabic remain rare. PalmX2025 addresses this gap directly, framing cultural understanding explicitly as a multiple-choice reasoning format — making it one of the first shared tasks to assess not just fluency but deep cultural and theological accuracy.

3 Dataset Composition

3.1 Subtask 1: Arabic Cultural Comprehension

This dataset contains culturally grounded MCQs in Modern Standard Arabic on customs, history, geography, arts, cuisine, and dialects, each with four options (A–D) and one correct answer. It includes 2,000 training, 500 development, and 2,000 blind test questions.

3.2 Subtask 2: Islamic Knowledge Reasoning

This dataset contains MCQs on Islamic practices, theology, Quranic knowledge, jurisprudence, and historical context, following the same format as Subtask 1. For training, we combined 600 Subtask 2 MCQs with 2,000 from Subtask 1 to leverage shared linguistic patterns and reasoning structures. It includes 300 development and 1,000 blind test questions.

4 Methodology

4.1 Subtask 1: Full Fine-Tuning of NileChat-3B

For Subtask 1, which focuses on Arabic cultural comprehension, we employ **NileChat-3B** (Mekki et al., 2025)[4], a 3-billion-parameter decoder-only language model built upon Qwen-2.5. NileChat-3B has been instruction-tuned on Egyptian and Moroccan Arabic under the Language–Heritage–Values (LHV) framework, enabling it to capture culturally nuanced responses across Arabic dialects. The model natively supports both Arabic script and Arabizi, making it well-suited for culturally grounded language tasks.

4.1.1 Input Formatting and Tokenization

To ensure strict compatibility with the shared task's evaluation pipeline, each training example is formatted using the official multiple-choice question (MCQ) template provided by the organizers. The

[4] https://huggingface.co/UBC-NLP/NileChat-3B

input consists of a question followed by four answer options prefixed with "A." through "D.", and concludes with the Arabic keyword used to prompt the model's autoregressive completion:

{question text}

A. {option A}
B. {option B}
C. {option C}
D. {option D}

الجواب:

This formatting aligns precisely with the evaluation script, which expects the model to autoregressively generate a single-letter label (e.g., "A") immediately following الجواب:.

Tokenization is performed using the model's associated AutoTokenizer, with inputs truncated or padded to a maximum length of 512 tokens. As the tokenizer does not define a dedicated padding token, we explicitly assign the end-of-sequence token (eos_token) as the pad_token to ensure consistency in attention masking and loss computation across batches.

4.1.2 Training Configuration

Fine-tuning is conducted on a single NVIDIA A100 (40GB) GPU using Hugging Face's Trainer with BF16 precision for 3 epochs, batch size 1, and gradient accumulation of 16 (effective batch size 16). Inputs are truncated or padded to 512 tokens, with full-sequence supervision achieved by copying input_ids into labels and masking padding tokens with -100. This implements standard causal language modeling (CLM), training the model to predict each token from preceding context, including question and answer. We use AdamW (LR 2e–5, no weight decay, without warm-up steps), evaluating and checkpointing at each epoch, and selecting the best model by validation loss. Preprocessing via datasets.map() removes irrelevant columns to reduce memory use and prevent data leakage.

4.2 Subtask 2: LoRA-Based Fine-Tuning of ALLaM-7B

For Subtask 2, which centers on Islamic cultural and legal knowledge reasoning, we adopt **ALLaM-7B-Instruct-preview**(Bari et al., 2024)[5], a 7-billion-parameter Arabic instruction-tuned language model developed to handle Modern Standard Arabic (MSA), Arabic dialects, and culturally grounded textual inputs. Due to its scale and resource requirements, we fine-tune ALLaM-7B using **Low-Rank Adaptation (LoRA)**(Hu et al., 2021), a parameter-efficient approach that significantly reduces memory consumption and training time while preserving task-specific adaptation capabilities.

4.2.1 Input Formatting and Tokenization

To encourage more structured reasoning during training while maintaining compatibility with the evaluation protocol, we introduced an augmented version of this prompt for fine-tuning:

{question text}

A. {option A}
B. {option B}
C. {option C}
D. {option D}

دعنا نفكر خطوة بخطوة:
أجب بالحرف فقط للإجابة الصحيحة (A أو B أو C أو D):
الجواب:

While the evaluation prompt does not contain these (e.g.,:دعنا نفكر خطوة بخطوة) reasoning cues, prior work in prompt engineering has shown that such instructions during fine-tuning can enhance a model's internal reasoning processes without impairing its ability to follow simpler formats at inference(Wei et al., 2022; Kojima et al., 2023). We applied full-sequence causal language modeling (CLM) supervision by duplicating input_ids into labels and used a custom collator for dynamic padding.

4.2.2 LoRA Configuration

To efficiently fine-tune ALLaM-7B, we employ Low-Rank Adaptation (LoRA) using Hugging Face's peft library. Only low-rank matrices injected into the attention projection layers are updated, while the base model remains frozen. Specifically, we target the q_proj and v_proj modules with a LoRA rank of 16, scaling factor (alpha) of 32, and dropout of 0.05.The task is set to CLM, updating under 1% of parameters for efficient adaptation on limited hardware.

4.2.3 Quantization and Memory Optimization

To further reduce GPU memory usage, ALLaM-7B is loaded in 8-bit precision via bitsandbytes and trained in FP16 mixed precision for efficiency. GPU cache clearing and checkpoint pruning control

[5] https://huggingface.co/ALLaM-AI/ALLaM-7B-Instruct-preview

memory usage, with all experiments run on a single NVIDIA RTX 4090 (24GB VRAM).

4.2.4 Training Configuration

Fine-tuning is performed using Hugging Face's Trainer with gradient checkpointing for memory efficiency. Training runs for 5 epochs with a per-device batch size of 8 and gradient accumulation over 4 steps (effective batch size 32), using a maximum sequence length of 256 tokens. Optimization employs AdamW (default β), a learning rate of 3e–5 with cosine decay, 100 warmup steps, and weight decay 0.01. A custom data collator applies dynamic padding and masks padded tokens with -100 to ensure loss is computed only on valid token positions.

4.2.5 Adapter Merging and Deployment

Following fine-tuning, LoRA adapters are merged into the base model resulting in a self-contained checkpoint. The merged model is uploaded to Hugging Face for submission.

4.3 Evaluation Protocol

All final test results were computed by the organizers using the official evaluation script [6] on a held-out blind test set. We submitted our fine-tuned models via Hugging Face, and accuracy was reported based on the organizers' execution of the shared evaluation pipeline.

5 Results

We report results for both subtasks on development and blind test sets (Table 1). Development scores were computed locally with the official evaluation script, while blind test scores were obtained through centralized evaluation by the organizers on a held-out set.

Table 1: Model Accuracy (%) on Development and Test Sets for Both Subtasks

Task	Dev Set (%)	Test Set(Blind) (%)
Subtask 1	78.60	72.15
Subtask 2	75.60	82.52

In Subtask 1, which targets general Arabic cultural awareness, our model achieved 78.60% accuracy on the development set, with a slight drop to 72.15% on the blind test set, likely due to domain shift or question-style variation. For instance, in the development set, it misclassified a question on the main environmental factor affecting the distribution of the Kuhl's free-tailed bat in southwest Saudi Arabia (correct:الحاجة إلى شرب الماء بانتظام) despite predicting heat adaptation, while correctly answering a question on the precise academic trajectory of Dr. Nidal Shamoun in Syria.

Conversely, the Subtask 2 model, which targets domain-specific reasoning in Islamic knowledge, demonstrated strong generalization capacity. Despite a slightly lower dev set performance (75.60%), it achieved a significant improvement on the test set, reaching 82.52%. For example, in one development set question on why a man's testimony equals that of two women, the correct answer was "(B + C) صحيحتان" ("both B and C are correct"); our system selected option B ("النسيان لدى المرأة أكبر من الرجل" – "forgetfulness is greater in women than in men"), which is partially correct but incomplete. In contrast, it correctly answered a question on what is opened for a believer who engages in *tasbīḥ* (تسبيح – "glorification of God"), selecting "أبواب الجنة" ("the gates of Paradise").

These results underscore the methodological rigor of our approach in capturing culturally grounded linguistic patterns under minimal supervision. The coherence between development and test set performance attests to the generalizability and stability of our fine-tuning strategy across evaluation regimes.

6 Conclusion and Future Work

We introduced culturally aligned LLM adaptation strategies that achieved top rankings at PalmX 2025. The combination of full fine-tuning and lightweight LoRA techniques enabled scalable and effective performance across subtasks. In future work, we aim to incorporate retrieval-augmented generation and test robustness on dialectal and low-resource Arabic varieties. Despite these promising results, our approach has limitations. Full fine-tuning is computationally expensive and may not generalise well across domains. Additionally, both datasets are limited in scope, which may affect transferability to unseen topics. Lastly, performance remains sensitive to prompt formatting and initialisation choices, which can impact reproducibility.

Acknowledgments

This research was supported by the ADAPT Research Centre at Munster Technological University.

[6] https://palmx.dlnlp.ai/

ADAPT is funded by Taighde Éireann – Research Ireland through the Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) via Grant 13/RC/2106_P2. We also thank the PalmX 2025 shared task organisers for their efforts in preparing the datasets and evaluation platform, and the anonymous reviewers for their valuable feedback and constructive suggestions, which have helped to improve the quality and clarity of this work.

References

Hani 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(6):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, 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.

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).

Wissam Antoun, Fady Baly, and Hazem M. Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *CoRR*, abs/2003.00104.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2024. Allam: Large language models for arabic and english.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab world. *Commun. ACM*, 64(4):72–81.

Gael de Francony, Victor Guichard, Praveen Joshi, Haithem Afli, and Abdessalam Bouchekif. 2019. Hierarchical deep learning for Arabic dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 249–253, Florence, Italy. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Ali Farghaly and Khaled Shaalan. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 8(4).

Arash Ghafouri, Hasan Naderi, Mohammad Aghajani asl, and Mahdi Firouzmandi. 2023. Islamicpcqa: A dataset for persian multi-hop complex question answering in islamic text resources.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

Shehenaz Hossain, Fouad Shammary, Bahaulddin Shammary, and Haithem Afli. 2025. Enhancing dialectal Arabic intent detection through cross-dialect multilingual input augmentation. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 44–49, Abu Dhabi, UAE. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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. 2024a. 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.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024b. Acegpt, localizing large language models in arabic.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Rana Malhas, Watheq Mansour, and Tamer Elsayed. 2022. Qur'an QA 2022: Overview of the first shared task on question answering over the holy qur'an. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 79–87, Marseille, France. European Language Resources Association.

Rana Malhas, Watheq Mansour, and Tamer Elsayed. 2023. Qur'an QA 2023 shared task: Overview of passage retrieval and reading comprehension tasks over the holy qur'an. In *Proceedings of ArabicNLP 2023*, pages 690–701, Singapore (Hybrid). 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.

Marryam Mohammed, Sama Ali, Salma Khaled, Ayad Majeed, and Ensaf Mohamed. 2025. Aftina: enhancing stability and preventing hallucination in ai-based islamic fatwa generation using llms and rag. *Neural Computing and Applications*, pages 1–26.

Ali Mostafa and Omar Mohamed. 2022. GOF at qur'an QA 2022: Towards an efficient question answering for the holy qu'ran in the Arabic language using deep learning-based approach. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 104–111, Marseille, France. European Language Resources Association.

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.

Tarek Naous, Michael J 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.

Faiza Qamar, Seemab Latif, and Rabia Latif. 2024. A benchmark dataset with larger context for non-factoid question answering over islamic text.

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

Zakia Saadaoui, Ghassen Tlig, and Fethi Jarray. 2024. Llms based approach for quranic question answering. pages 112–118.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models.

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, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan 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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Imed Zitouni. 2011. Introduction to arabic natural language processing n. y. habash. morgan & claypool (synthesis lectures on human language technologies, vol. 10), 2010, xvii+167pp; isbn 978-1-59829-795-9, ebook isbn 978-1-59829-796-6. *Comput. Linguist.*, 37(3):623–625.

A Additional Results

In this appendix, we provide detailed dev set results for both subtasks, comparing baseline (zero-shot) and fine-tuned variants of Fanar-9B-Instruct, NileChat-3B, and ALLaM-7B models. These results illustrate the consistent improvements achieved through fine-tuning, with larger models generally benefiting more from adaptation.

Table 2: Dev set accuracy of baseline and fine-tuned models for subtask 1.

Model	Fine-tune	Dev Acc.(%)
Fanar-1-9B-Instruct	(zero-shot)	69.80
Fanar-1-9B-Instruct	(fine-tuned)	75.40
NileChat-3B	(zero-shot)	70.00
NileChat-3B	(fine-tuned)	78.60

For Subtask 1 (cultural QA), Fanar-1-9B-Instruct improves from 69.80% in zero-shot to 75.40% after fine-tuning, while NileChat-3B achieves the highest dev accuracy of 78.60% after fine-tuning.

Table 3: Dev set accuracy of baseline and fine-tuned models for subtask 2.

Model	Fine-tuning	Dev Acc.(%)
NileChat-3B	(fine-tuned)	71.67
ALLaM-7B	(zero-shot)	68
ALLaM-7B	(PEFT)	75.60

For Subtask 2 (Islamic knowledge reasoning), NileChat-3B with fine-tuning reaches 71.67%, while ALLaM-7B shows stronger performance, improving from 68.00% zero-shot to 75.60% after PEFT-based adaptation.