Phoenix at Palmx: Exploring Data Augmentation for Arabic Cultural Question Answering

Houdaifa Atou[λ*], **Issam Ait Yahia**[λ*], **Ismail Berrada**[λ],
[λ]College of Computing
Mohammed VI Polytechnic University, Ben Guerir, Morocco
{houdaifa.atou, issam.aityahia, ismail.berrada}@um6p.ma

Abstract

Large Language Models (LLMs) have become central to natural language processing, but their performance in low-resource cultural domains remains limited, mainly due to the dominance of English data in training. This limitation is especially evident in open small models. Evaluating and improving LLMs' performance in Arabic culture is therefore necessary. This paper presents ***Phoenix*** and ***PhoenixIs***, two models fine-tuned for the ***Palmx 2025*** general culture and Islamic culture subtasks. Phoenix uses the *Palmx-GC* and *Palmx-IC* datasets as seed data and applies diverse data augmentation strategies to construct an enriched fine-tuning dataset. *Phoenix* achieves an accuracy of 71.35% on the general culture subtask, while *PhoenixIs* reaches 83.82% on the Islamic culture subtask.

1 Introduction

Culture refers to the shared knowledge, beliefs, values, practices, and traditions that shape how a community understands and interacts with the world. Although Large Language Models (LLMs) have achieved strong performance across a wide range of natural language processing tasks, they have been shown to exhibit cultural bias toward Western culture (Cecilia Liu et al., 2024; Navigli et al., 2023; Cao et al., 2023). Such bias arises from the dominance of English data in their pre-training and post-training corpora. This limitation may affect their ability to adapt and generate culturally appropriate responses for diverse communities. To tackle this bias, efforts have been made to align LLMs with different cultures (Joshi et al., 2025; Mekki et al., 2025; Li et al., 2024) and to assess their cultural knowledge on specific domains (AlKhamissi et al., 2024; Alwajih et al., 2025a). Beyond cultural adaptation, progress in Arabic NLP has been supported by benchmarks and resources across a variety of tasks, including machine translation (Akallouch and Fardousse, 2025), named entity recognition (Yahia et al., 2024; Jarrar et al., 2024), and question answering (Mozannar et al., 2019).

In this context, the ***Palmx 2025*** shared task was introduced to evaluate the ability of LLMs to capture Arabic cultural knowledge and to promote the development of culturally aware systems for the Arab world (Alwajih et al., 2025b). It includes two subtasks, General Culture and Islamic Culture, each based on datasets of multiple-choice questions in Modern Standard Arabic (MSA), namely *Palmx-GC* and *Palmx-IC*.

In this paper, we present our participating systems for the General Culture subtask (*Phoenix*) and the Islamic Culture subtask (*PhoenixIs*) of Palmx 2025. Starting from Palmx-GC as seed data, we applied three data augmentation strategies: question paraphrasing, which generates semantically equivalent variants of existing questions, sample-based augmentation, which produces new multiple-choice questions by conditioning on individual question–answer pairs, and dataset-based augmentation, which creates thematically related questions by leveraging the full dataset (Section 4.1). For the Islamic Culture subtask, we only explored question paraphrasing. The augmented data was then used to fine-tune dedicated models for each subtask. Our experiments demonstrate that the proposed augmentation strategies improve performance on both subtasks. Phoenix obtains an accuracy of 71.35% on the General Culture subtask, and PhoenixIs attains 83.82% on the Islamic Culture subtask.

Our contributions in this work are: **(1)** we present Phoenix and PhoenixIs, two systems developed for the General Culture and Islamic Culture subtasks, respectively. **(2)** We design and evaluate three data augmentation strategies that enrich the available training data and improve model perfor-

[*]Equal Contribution.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 843–850

mance. **(3)** We provide an extensive analysis of these strategies, showing that they enhance accuracy on both subtasks.

2 Related Work

Although large language models have achieved strong performance across a variety of languages, adapting them to specific cultural contexts, particularly those that are low-resource, remains a significant challenge, as they often display a bias toward Western culture (Cecilia Liu et al., 2024; Navigli et al., 2023; Cao et al., 2023; Naous et al., 2024). To mitigate this issue, several adaptation strategies have been explored, including continuous pretraining (Mekki et al., 2025), prompt tuning (Masoud et al., 2024), prompt engineering (Shen et al., 2024; Tao et al., 2024; AlKhamissi et al., 2024), and supervised fine-tuning (Li et al., 2024). All of these approaches rely, to varying degrees, on the availability of well-constructed cultural datasets, which remain scarce. In response, a growing body of work has focused on building resources that capture cultural knowledge across different languages and communities (Alwajih et al., 2025a; Myung et al., 2024).

Since the manual annotation of cultural data is resource-intensive and difficult to scale, researchers have increasingly turned to data augmentation to expand training sets (Liu et al., 2025; Li et al., 2024; Joshi et al., 2024). Nonetheless, this approach requires careful design to ensure that the generated data maintains quality and reliability (Liu et al., 2024).

The limited representation of Arabic in pretraining corpora has motivated a growing effort to develop LLMs specifically designed for the Arab world. One approach has been to rely on translation, where large volumes of English data are translated into Arabic to supplement training resources (Sengupta et al., 2023). Other work has emphasized the inclusion of native Arabic data without translation in order to better capture the linguistic and cultural features of the language (Huang et al., 2024). To address the bias toward English and the resulting cultural misalignment, some approaches have relied on continual pretraining with carefully curated cultural data (Mekki et al., 2025), while others have explored training models entirely from scratch (Bari et al., 2024; Team et al., 2025). In this work, we build on these efforts by finetuning such models for the Palmx shared task subtasks.

3 Palmx

The Palmx shared task was established to evaluate the ability of LLMs to capture Arabic cultural knowledge and to encourage the creation of systems that are culturally aware within the Arab world. It is divided into two subtasks: General Culture and Islamic Culture. The General Culture subtask examines the ability of LLMs to reason about different aspects of Arabic culture. Its questions span a wide range of domains such as customs, etiquette, and arts from across Arab countries, including Palestine, Morocco, Egypt, and others. The Islamic Culture subtask is designed to evaluate models' understanding of central elements of Islamic culture. The questions address topics including religious practices, Quranic knowledge, and Hadith literature.

3.1 Datasets

The Palmx shared task provides two datasets, Palmx-GC for the General Culture subtask and Palmx-IC for the Islamic Culture subtask. Both datasets consist of multiple-choice questions in MSA and are split into training, development, and blind test sets. Table 1 presents the detailed statistics for each split.

Split	Palmx-GC	Palmx-IC
Training	2000	600
Development	500	300
Blind Test	1000	1000

Table 1: Distribution of samples in Palmx-GC and Palmx-IC.

4 Phoenix

We propose two systems for the Palmx shared task: *Phoenix* for the General Culture subtask and *PhoenixIs* for the Islamic Culture subtask. Both systems build on the official provided datasets, and each incorporates data augmentation to expand the training data before finetuning task-specific models. An overview of the augmentation strategies is presented in Figure 1.

4.1 Data Augmentation

4.1.1 Question Paraphrasing

In the paraphrasing setup, the LLM was provided with a single question and instructed to generate n_1 semantically equivalent variants. This strategy

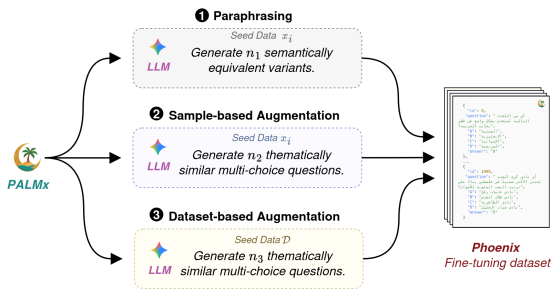

Figure 1: Overview of Phoenix data augmentation strategies

increases data diversity while preserving the original meaning, thereby helping the model generalize to different phrasings of the same cultural concept. We employed Gemini 2.5 Pro for this augmentation.

4.1.2 Sample-based Augmentation

In sample-based augmentation, the LLM was given an original question together with its multiple-choice answers and asked to generate new thematically and structurally similar multiple-choice questions. This approach expands the dataset by producing additional questions that maintain the original format while introducing controlled variation. We used Gemini 2.5 Flash for sample-based augmentation.

4.1.3 Dataset-based Augmentation

For dataset-based augmentation, the LLM was provided with the full Palmx-GC dataset of 2,000 samples and prompted to generate n_3 new multiple-choice questions that are thematically related. Unlike the previous strategies, this method leverages the dataset as a whole, encouraging the model to create questions that capture broader patterns across domains. Gemini 2.5 Pro was used for this setup. To ensure quality and cultural fidelity, a random subset of the LLM-generated questions from each augmentation strategy was manually inspected by human annotators. This verification step helped confirm semantic correctness and adherence to cultural context before including the data in training (see Appendix B for a detailed error analysis).

4.2 Finetuning Data

For finetuning, we constructed task-specific datasets by combining the original seed data with the augmented questions. In the General Culture subtask, *Phoenix* was finetuned on a total of 18,742 questions, consisting of 2,000 from Palmx-GC, 6,000 from question paraphrasing, 6,411 from sample-based augmentation, and 4,331 from dataset-based augmentation. For the Islamic Culture subtask, *PhoenixIs* was finetuned on 4,400 questions, including 600 from Palmx-IC, 1,800 from question paraphrasing, and 2,000 from Palmx-GC.

4.3 Model Pre-selection

To identify suitable base models for finetuning, we first evaluated several state-of-the-art Arabic-focused LLMs in a zero-shot setting on the Palmx-GC and Palmx-IC validation sets. The results are summarized in Table 2. Based on this evaluation, we selected *Fanar-1-9B-Instruct* (Team et al., 2025) for *Phoenix* and *ALLaM-7B-Instruct* (Bari et al., 2024) for *PhoenixIs*.

Model	Accuracy (%)
Category: General Culture (GC)	
Fanar-1-9B-Instruct	**72.40**
ALLaM-7B-Instruct	70.60
NileChat-3B	70.00
AceGPT-v2-8B-Chat	65.00
Falcon-H1-7B-Instruct	39.20
Category: Islamic Culture (IC)	
ALLaM-7B-Instruct	**73.00**
Fanar-1-9B-Instruct	67.00
NileChat-3B	64.00
AceGPT-v2-8B-Chat	63.67
Falcon-H1-7B-Instruct	34.00

Table 2: Zero-shot accuracy of different LLMs on the Palmx-GC (GC) and Palmx-IC (IC) validation sets. The top-performing model in each category is highlighted.

5 Experiment and Results

5.1 Experimental Setup

Experiments were conducted on the Palmx shared-task datasets for General Culture (GC) and Islamic Culture (IC). Fanar-1-9B-Instruct was selected as the base model for GC, and ALLaM-7B-Instruct for IC, following the pre-selection analysis in Subsection 4.3. Model performance was evaluated using accuracy on the validation and blind test splits. For GC, the fine-tuning corpus combined the original Palmx data with progressively applied augmentation strategies: paraphrasing (PA), sample-based

(SA), and dataset-based (DA). For IC, augmentation was deliberately restricted to paraphrasing in order to safeguard the theological fidelity of religious material. All results are reported on the validation set, with the exception of the official leaderboard scores, which are based on the blind test set. All experiments were repeated three times with different random seeds, and we report the average accuracy. We use Lora (Hu et al., 2022) to finetune both models with a learning rate of 0.0002, an effective batch size of 128, and LoRA hyperparameters $R = 64$, $\alpha = 16$, and dropout 0.1. All experiments were conducted on one NVIDIA A100 GPU.

5.2 Official Leaderboard Results

Table 3 presents the official Palmx 2025 leaderboard. Phoenix achieved fourth place in GC with an accuracy of 71.35%, performing within one percentage point of the leading system. PhoenixIs achieved 83.82% in the Islamic Culture subtask, ranking second among all submitted systems. These results indicate that our augmentation strategies enabled competitive performance across both subtasks.

Rank	Team	Accuracy (%)
Category: General Culture (GC)		
1	HAI	72.15
2	Pulkit Chatwal	71.65
3	AYA_Team	71.45
4	**Phoenix (ours)**	**71.35**
5	CultranAI	70.50
6	ISL-NLP	67.60
7	Rafiul Biswas	67.55
8	Hamyaria	65.90
9	Star	64.05
Category: Islamic Culture (IC)		
1	AYA Team	84.22
2	**PhoenixIs (ours)**	**83.82**
3	HAI	82.52
4	Rafiul Biswas	74.13
5	Hamyaria	70.83
6	TarnishedLab	62.84

Table 3: Official Palmx 2025 results. Our team's entries are highlighted.

Fine-tuning data	Acc.
General Culture (GC)	
Palmx	77.73 ± 1.21
Palmx + PA	80.07 ± 1.21
Palmx + PA + SA	80.60 ± 1.06
Palmx + PA + SA + DA	$\mathbf{80.93 \pm 0.76}$
Islamic Culture (IC)	
Palmx Islamic	73.73 ± 2.92
Palmx Islamic + PA	75.11 ± 1.91
Palmx Islamic + PA + Palmx-GC	$\mathbf{78.56 \pm 0.78}$

Table 4: Ablation on validation sets (mean $\pm$ std over three runs). GC uses Fanar-1-9B-Instruct; IC uses ALLaM-7B-Instruct.

5.3 Ablation Study: Impact of Augmentation

To assess the contribution of each augmentation strategy, controlled ablations were performed on the validation sets (Table 4). In GC, performance improved consistently with each additional augmentation step, reaching 80.93% with the full combination of PA, SA, and DA. This trend illustrates that diversity introduced at both the question and dataset level substantially enhances generalization. In IC, we explored the effect of paraphrasing and the inclusion of Palmx-GC in the finetuning mixture. The base model achieved 73.73% accuracy. Incorporating paraphrasing increased performance to 75.11%, and adding Palmx-GC further raised accuracy to 78.56%. Overall, our study shows that each component of the proposed augmentation strategy contributed to the final performance.

6 Conclusion

In this work, we presented Phoenix and PhoenixIs, two systems developed for the Palmx 2025 shared task on Arabic cultural understanding. By leveraging the Palmx-GC and Palmx-IC datasets and applying a range of data augmentation strategies, we constructed enriched fine-tuning sets. Our experiments showed that our proposed data augmentation strategies enabled consistent improvements across both General Culture and Islamic Culture subtasks.

Limitations.

Our approach relies on synthetic augmentation for the General Culture task, which, while effective, may introduce distributional biases or artifacts. Human verification was applied to sampled augmented data, but the majority of generated content

remained unreviewed. For the Islamic Culture task, augmentation was deliberately restricted to paraphrasing to preserve theological fidelity, which limited the exploration of richer augmentation strategies.

References

Oussama Akallouch and Khalid Fardousse. 2025. In-context learning for low-resource machine translation: A study on tarifit with large language models. *Algorithms*, 18(8):489.

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, 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.

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, and 1 others. 2024. Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. 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 2016–2039, Mexico City, Mexico. 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.

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.

Mustafa Jarrar, Nagham Hamad, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Wojoodner 2024: The second arabic named entity recognition shared task. *arXiv preprint arXiv:2407.09936*.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2024. Adapting multilingual llms to low-resource languages using continued pre-training and synthetic corpus. *arXiv preprint arXiv:2410.14815*.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2025. Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus: A case study for Hindi LLMs. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 50–57, Abu Dhabi. Association for Computational Linguistics.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134, Vienna, Austria. Association for Computational Linguistics.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi

Yang, Denny Zhou, and 1 others. 2024. Best practices and lessons learned on synthetic data. *arXiv preprint arXiv:2404.07503*.

Reem I Masoud, Martin Ferianc, Philip Colin Treleaven, and Miguel RD Rodrigues. 2024. Llm alignment using soft prompt tuning: The case of cultural alignment. In *Workshop on Socially Responsible Language Modelling Research*.

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*.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

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.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, and 1 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. 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 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, and 1 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

Issam Yahia, Houdaifa Atou, and Ismail Berrada. 2024. Addax at WojoodNER 2024: Attention-based dual-channel neural network for Arabic named entity recognition. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 867–873, Bangkok, Thailand. Association for Computational Linguistics.

A Effectiveness Analysis

We investigated the effect of increasing the size of augmented data on model performance, as generating too many samples can hurt the performance. For Phoenix, we fine-tuned the model with 2,000, 8,000, 15,000, and 18,742 samples, with results shown in Figure 2. The best performance was achieved with 18,742 samples. Similarly, for PhoenixIs, we fine-tuned with 2,600, 3,200, 3,800, and 4,400 samples, as shown in Figure 3, where 4,400 samples yielded the strongest results. The composition of each set is detailed in Tables 5 and 6.

	S1	S2	S3	S4
Palmx-GC	2,000	2,000	2,000	2,000
PA	0	2,000	4,000	6,000
SA	0	2,000	4,000	6,411
DA	0	2,000	3,000	4,331
Total	2,000	8,000	13,000	18,742

Table 5: Composition of the dataset for each experiment on Phoenix.

	S1	S2	S3	S4
Palmx-IC	600	600	600	600
Palmx-GC	2000	2000	2000	2000
PA	0	600	1200	1800
Total	2600	3200	3800	4400

Table 6: Composition of the dataset for each experiment on PhoenixIs.

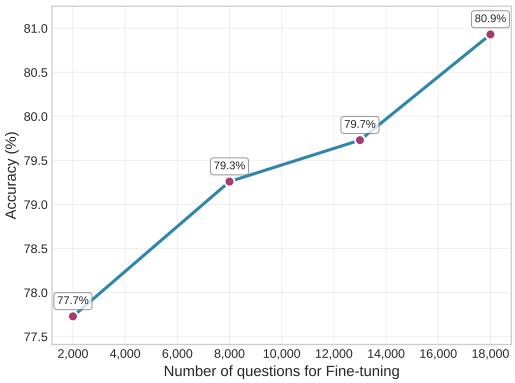

Figure 2: Influence of the number of fine-tuning samples on Phoenix.

B Human Verification and Error Analysis

To ensure the reliability of the augmented data, we conducted a manual verification of randomly sampled questions from each augmentation strategy:

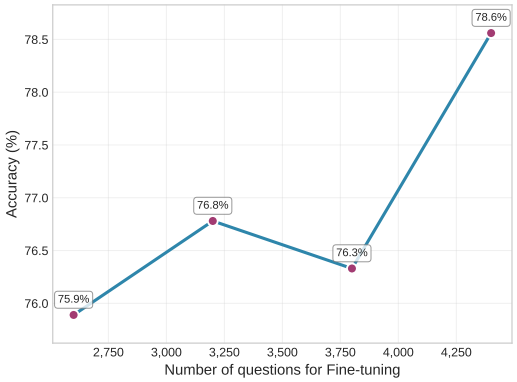

Figure 3: Influence of the number of fine-tuning samples on PhoenixIs.

- **Sample-based Augmentation (SA):** From a random set of 100 generated questions, we identified **10 problematic cases**. Of these, **3 were factually incorrect**, while the remaining **7 deviated from instructions** (e.g., not strictly following the required format or asking about tangential topics). Importantly, most of these still produced valid question–answer pairs despite the inconsistencies.
- **Dataset-based Augmentation (DA):** From a random set of 100 generated questions, we found **3 issues**, all of which were culturally valid but referenced **non-Arab countries**.

Overall, the error rate across both strategies was relatively low. The main sources of error were format deviation and domain drift rather than factual inaccuracies. This indicates that our augmentation pipeline is broadly reliable.

ما هي المدينة التي تعرف بأنها "مدينة الجسور المعلقة" في الجزائر؟
أ. وهران
ب. عنابة
ج. قسنطينة
د. تلمسان

ما هي أعلى قمة جبلية في الوطن العربي وتقع في المغرب؟
أ. جبل سانت كاترين
ب. جبل توبقال
ج. جبل شمس
د. جبل النبي

Figure 4: Cases from our generated data where the generation was correct. The proposed answer is highlighted in blue.

يُعرف الفنان السوداني الكبير محمد وردي بلقب "فنان أفريقيا الأول". ما هو اللقب الآخر الذي ارتبط به بشكل وثيق ويعكس مكانته في الغناء السوداني؟
أ. بلبل السودان
ب. عنقريب الفن
ج. فنان الشعب
د. فنان الوادي

ما هو أعلى جبل في اليابان؟
أ. جبل كيتا
ب. جبل هوتاكا
ج. جبل فوجي
د. جبل أينو

Figure 5: Cases from our generated data where the generation was incorrect (or deviated from instructions). The proposed answer is highlighted in blue.

C Error Analysis on Validation Set

We inspected a small sample of incorrect validation set predictions to illustrate typical failure cases. Figure 6 shows four representative errors, where the ground truth is marked in green and the model's predictions are in red.

ما هي الدولة التي تحتل المرتبة الثالثة في مساحة الأراضي في إفريقيا والعالم العربي؟
أ. السودان
ب. الجزائر
ج. نيجيريا
د. مصر

متى يُحتفل بيوم الرياضة الوطني في قطر؟
أ. في الثلاثاء من الأسبوع الثاني من فبراير
ب. في 1 مايو
ج. في 10 ذو الحجة
د. في 18 ديسمبر

أي من الشخصيات التالية يُعتبر روائياً ليبياً؟
أ. هشام مطر
ب. محمد الفيتوري
ج. جمال خشتة
د. عزالدين شكري الفيلالي

ما هو الدين الذي يعتنقه النوبيون والبجا بشكل تقليدي في السودان؟
أ. المسيحية
ب. اليهودية
ج. الهندوسية
د. الإسلام

Figure 6: Cases from the Palmx-GC validation set where the model's prediction was incorrect. Ground truth is marked in green, predictions in red.

Paraphrasing

من أي لونين تتألف الكوفية الفلسطينية؟
أ. أزرق وذهبي
ب. أحمر وأخضر
ج. أبيض وأسود
د. بني ورمادي

Sample-based Augmentation

من هو الكاتب الفلسطيني المعروف بقصصه القصيرة التي تناولت القضية الفلسطينية وعُرفت بأسلوبها الواقعي والرمزي؟
أ. محمود درويش
ب. سميح القاسم
ج. إبراهيم نصر الله
د. غسان كنفاني

Dataset-based Augmentation

أي من هذه الأنهار يشكل الحدود بين الأردن وفلسطين؟
أ. نهر اليرموك
ب. نهر الليطاني
ج. نهر الأردن
د. نهر العاصي

Figure 7: Examples of augmented question–answer pairs generated using the three strategies: paraphrasing, sample-based augmentation, and dataset-based augmentation. The proposed answer is highlighted in blue.