Data-Efficient Fine-Grained Cross-Cultural Transfer of Commonsense Reasoning in LLMs

Anonymous EMNLP submission

Abstract

Large language models (LLMs) pretrained primarily on English data often reflect Western-centric biases, limiting their effectiveness in diverse cultural contexts. While some work has explored cultural alignment, the potential for cross-cultural transfer, using alignment in one culture to improve performance in others, remains underexplored. This paper investigates cross-cultural transfer in the Arab world, where linguistic and historical similarities coexist with local cultural differences. Using a culturally grounded commonsense reasoning dataset covering 13 Arab countries, we evaluate lightweight alignment methods such as in-context learning (ICL) and demonstration-based reinforcement (DITTO), alongside baselines like instruction fine-tuning (IFT) and Direct Preference Optimization (DPO). Our results show that just 12 culture-specific examples from one country can improve performance in others by 15–20% on average. These findings demonstrate that efficient cross-cultural alignment is possible and offer a promising approach to reducing Western bias in LLMs while advancing culturally fair NLP in low-resource settings.

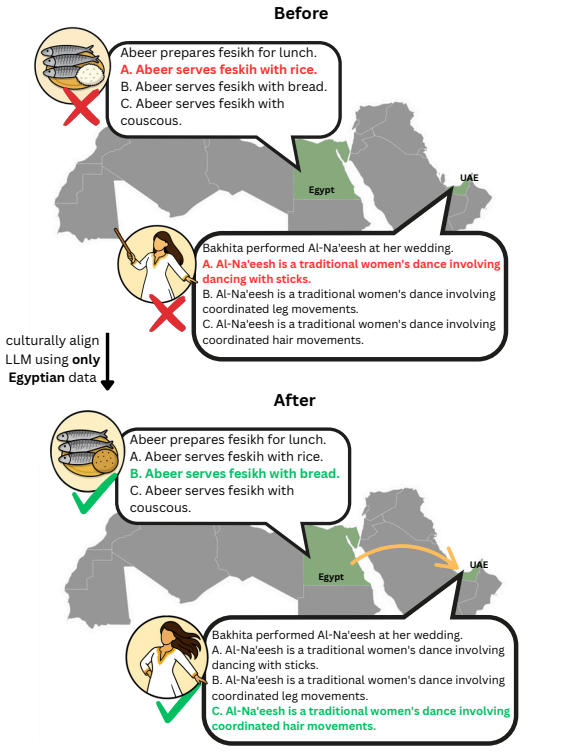

Figure 1: An example to demonstrate the concept of cross-cultural transfer. If culturally aligning an LLM on only Egyptian data improves the performance of the LLM on UAE culture, then cultural knowledge is transferred.

1 Introduction

Large Language Models (LLMs) are increasingly deployed across diverse cultural contexts, yet they often reflect a Western-centric worldview, misaligning with local customs, values, and norms (Naous et al., 2024; Sadallah et al., 2025; Wang et al., 2024). Prior studies have explored broad East-West cultural misalignments in LLMs, but little is known about how these models handle intra-regional cultural variation, such as that found across the 22 Arab countries. For example, despite sharing linguistic ties, Emirati culture differs significantly from Egyptian or Syrian traditions in food, festivals, and gender roles. However, most Arabic LLMs are trained on translated English data or regionally-aggregated corpora (Sengupta et al., 2023; Sadallah et al., 2025), potentially flattening these cultural distinctions.

A central challenge in aligning LLMs to country-specific cultural knowledge is data scarcity. High-resource countries like Egypt have vastly more online content than low-resource ones like the UAE (114M vs. 1.3M population) (Insight, 2025; United Nations Population Fund, 2025), leading to under-representation. This raises a key question: *Can knowledge from one culture be transferred to benefit another with limited data?*

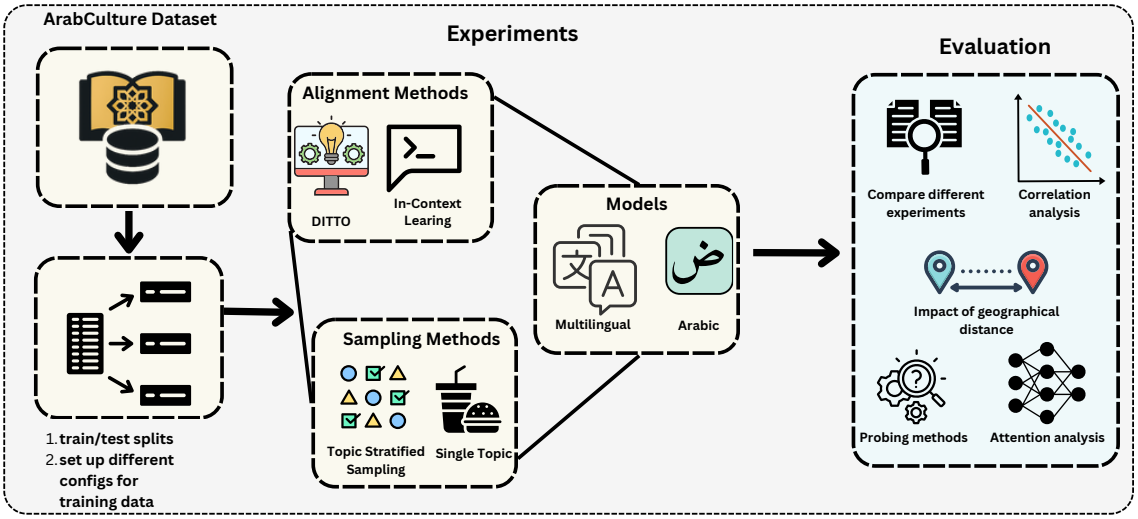

Figure 2: This figure illustrates an overview of our alignment and evaluation pipeline. The ArabCulture dataset is split into train/test subsets, aligned via either In-Context Learning or DITTO on different models with different sampling methods, then evaluated and probed (stimulus, attention, correlation) to quantify cross-cultural transfer.

In this paper, we investigate the feasibility of cross-cultural commonsense transfer within the Arab world. Specifically, we ask: *Can aligning an LLM to the culture of one Arab country improve performance on others?* We explore this question through two lightweight alignment strategies: In-Context Learning (ICL) and Demonstration-based Iterative Task Tuning Optimization (DITTO) (Shaikh et al., 2024). Although ICL is a strong few-shot baseline, DITTO offers a reinforcement learning alternative that requires only a handful of high-quality demonstrations, making it particularly suitable for low-resource cultural domains.

We construct experiments over a 13-country, 3.2k-example ArabCulture dataset spanning diverse domains such as food, rituals, relationships, and social norms. Using only 12 cultural demonstrations per source country, we test transfer to unseen target cultures across four LLMs (Qwen2.5, Gemma-2, ALLaM, and SILMA) (Team, 2024a; Team et al., 2024; Bari et al., 2024; Team, 2024b). We further probe whether cross-cultural improvement is predictable from geographic proximity, and whether alignment reshapes latent cultural representations in the model's internal space.

Our contributions are:

- We pioneer the use of DITTO for cultural alignment, achieving up to 20% accuracy gains in Arab commonsense reasoning with only 12 demonstrations per country.
- We show that cross-cultural transfer is feasible: cultural knowledge from high-resource countries improves LLM performance on culturally distinct, low-resource ones.
- We perform probing and correlation analyses to show that improvements are driven more by cultural salience than geographic proximity, and that targeted alignment enhances the linear separability of specific cultures in the model's latent space.

Our findings offer a compelling path toward culturally adaptive NLP systems using minimal, targeted supervision, a crucial step for equitable and globally relevant AI.

2 Related Work

2.1 Cultural Reasoning

Disparities in cultural knowledge persist, often favoring dominant cultures (Shen et al., 2024; Wang et al., 2024; Naous et al., 2024). Recent work shows that including geographical context in prompts significantly boosts model performance on low-resource cultural reasoning tasks (Koto et al., 2024). Likewise, culturally aware data collection, targeted model adaptation, and robust evaluation frameworks are crucial in addressing linguistic diversity and cultural biases (Hershcovich et al., 2022). Several Arabic cultural datasets and benchmarks have been developed, including the ACVA Arabic Culture benchmark, which includes general

true/false statements about Arab culture as a whole (Huang et al., 2024), and the AraDiCE-Culture benchmark, which includes cultural questions from only six Arab countries (Mousi et al., 2025).

The assumption that underlies most of the work related to Arabic cultural alignment of LLMs is that Arabs share the same culture, either entirely or regionally, raising the question of whether Arab culture is a homogeneous culture or a diverse set of cultures (Keleg, 2025). We further investigate this assumption by experimenting with country-level cultural alignment using the Arabic Culture dataset (Sadallah et al., 2025), which consists of cultural data from 13 Arab countries, and conducting a fine-grained comprehensive analysis.

2.2 Cultural Alignment Approaches

There has been growing effort by researchers to improve the cultural awareness of models by incorporating cultural data with methods for cultural alignment such as fine-tuning (Li et al., 2024), and in-context learning (such as few-shot prompting) (Wang et al., 2024; AlKhamissi et al., 2024).

Although fine-tuning can be an effective approach to cultural alignment, it can cause models to forget previous knowledge (Choenni et al., 2024; AlKhamissi et al., 2024) and larger models need a sizeable amount of data, whereas reinforcement learning iteratively uses feedback from a reward model to optimize its responses requiring a small set of human demonstrations to align effectively. Recent advances in preference alignment involve reinforcement learning by using iterative feedback to guide LLMs toward desired norms, exemplified by Direct Preference Optimization (DPO) (Rafailov et al., 2023). Additionally, methods like DITTO (Shaikh et al., 2024) extend DPO to align effectively without requiring large-scale data. While DITTO was initially developed for stylistic adaptation tasks, it is adopted in this work for the novel application of cultural alignment.

We build on these existing approaches to align LLMs with regional cultural nuances while preserving broader commonsense capabilities. Additionally, we investigate how LLMs generalize cultural knowledge in the Arab world, focusing on whether training on specific regions enhances performance elsewhere and how geographical or cultural distance influences model outcomes.

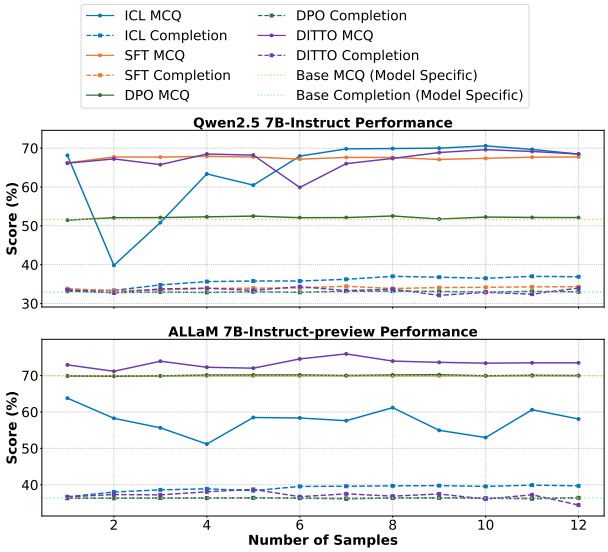

Figure 3: Demonstrates sample efficiency of different alignment methods for cultural alignment task, on both multi-choice questions (MCQ) and completion aligned using cultural demonstrations of UAE's culture.

3 Methodology

An overview of our methodology is outlined in Figure 2. The dataset is split into train/test sets, aligned using ICL or DITTO across models and sampling strategies, then evaluated and analyzed to measure cross-cultural transfer.

3.1 Arabic Culture Dataset

We use the Arabic culture dataset (Sadallah et al., 2025), which consists of approximately 3,200 handcrafted cultural statements and the corresponding multiple choice options (one correct, two incorrect). The dataset spans 12 topics and 40+ subtopics from 13 countries grouped into 4 regions of the Arab world (North Africa, Gulf region, Nile Valley, Levant). Each country subset consists of roughly 250 pairs of statements and choices. For each country, we split these 10% for training/alignment examples and 90% held-out for evaluation. This ensures that our evaluation always assesses the model on unseen cultural statements.

3.2 Alignment Methods

We adopt two main alignment approaches for LLMs to align on country specific cultural examples: in-context learning (ICL) and DITTO, a recently proposed lightweight method that extends Direct Preference Optimization (DPO) (Rafailov et al., 2023) and iteratively aligns model outputs to a small set of user-provided demonstra-

Country	Qwen2.5 7B-Inst				Gemma-2 9B-it				ALLaM 7B-Inst				SILMA 9B-Inst				Avg.
	DITTO		ICL		DITTO		ICL		DITTO		ICL		DITTO		ICL		
	Comp.	MCQ	Comp.	MCQ	Comp.	MCQ	Comp.	MCQ	Comp.	MCQ	Comp.	MCQ	Comp.	MCQ	Comp.	MCQ	
Algeria	0.19	16.74	2.80	18.50	-0.22	25.26	**3.86**	4.52	-0.31	1.57	3.99	-10.49	-0.88	2.14	1.82	0.79	4.39
Egypt	-0.09	17.56	2.39	**19.67**	-1.61	28.34	1.72	0.66	-2.77	-1.10	3.61	-25.32	-0.97	-1.22	0.03	-0.66	2.52
Jordan	-0.31	18.91	2.77	17.09	1.47	**33.93**	**4.80**	11.21	-4.84	-12.82	3.86	-4.84	1.13	1.57	**2.07**	3.11	**4.94**
KSA	-0.06	17.84	3.24	**19.92**	3.30	27.46	3.42	6.00	0.78	-1.91	3.14	-16.27	0.22	0.66	1.26	1.39	4.40
Lebanon	0.91	18.38	**3.52**	18.66	1.28	7.19	3.52	-0.03	0.34	**3.71**	3.99	-14.92	0.88	-3.01	0.63	-1.54	2.72
Libya	-0.12	15.11	3.27	16.71	0.37	**33.55**	2.38	-0.06	-2.07	-0.28	2.64	-13.22	0.82	-0.34	1.35	-6.97	3.32
Morocco	1.64	17.25	3.33	18.79	-0.41	13.70	3.77	6.91	-0.09	3.14	3.93	-11.72	1.13	2.71	1.89	**3.49**	4.34
Palestine	-0.03	17.97	1.45	18.03	0.31	24.94	3.42	0.47	-3.33	0.82	2.80	-19.76	0.38	2.05	1.54	0.00	3.19
Sudan	1.07	18.98	3.21	16.15	1.44	15.11	3.26	14.32	-1.67	1.73	3.20	-22.87	1.70	2.74	1.85	1.10	3.83
Syria	0.98	17.00	3.30	19.04	0.15	31.82	2.60	0.44	-1.45	2.55	**4.21**	-5.90	-0.28	1.95	0.57	**3.40**	**5.02**
Tunisia	-0.81	17.18	2.01	18.13	1.35	20.58	2.29	0.60	0.22	1.35	**4.33**	-18.60	0.28	2.33	0.38	0.35	3.25
UAE	1.07	16.84	**3.99**	16.81	2.38	28.15	3.55	1.57	-2.07	**3.58**	3.36	-11.84	1.70	2.20	**2.04**	2.27	4.73
Yemen	-0.91	18.57	2.14	12.00	-0.35	5.72	2.67	0.22	0.53	-0.50	2.86	-17.65	-0.12	0.44	0.63	-0.75	1.59
Avg.	0.27	17.56	2.88	**17.65**	0.73	**22.75**	**3.17**	3.60	-1.29	0.14	**3.53**	-14.88	0.46	1.09	1.24	0.46	

Table 1: Overall accuracy improvements for Arab cultural commonsense reasoning when training on country-specific knowledge across different models with topic-based sampling. Results show performance on Completion and MCQ tasks using DITTO and ICL methods. Base accuracies (MCQ%/Completion%): Qwen2.5 (51.65/32.89), Gemma-2 (34.56/32.52), ALLaM (69.9/36.35), SILMA (70.81/32.39). **Bold** and highlighted cells top two MCQ and Completion values for each model, as well as the top two country-based improvements.

tions (Shaikh et al., 2024). DITTO treats high-quality user-provided demonstrations as strictly preferred over intermediate model outputs, guiding the model toward better alignment through iterative preference-based updates. DITTO offers a data-efficient alternative to large-scale supervised fine-tuning or full-scale reinforcement learning from human feedback (RLHF) (Bai et al., 2022), enabling precise cultural alignment from a small number of carefully selected examples as highlighted in Figure 3, resulting in significant improvement in overall performance in Arab cultures.

We use 2 multilingual models (Qwen2.5B 7B-Instruct (Team, 2024a) and gemma-2 9b-it (Team et al., 2024)), and 2 Arabic-centric models (ALLaM 7B-Instruct-preview (Bari et al., 2024) and SILMA 9B-Instruct (Team, 2024b)) as the base language models.

3.3 Demonstration Sampling

We employ two complementary in-context sampling strategies for both ICL and DITTO: topic-based sampling and food-based sampling. For topic-based sampling, we select 12 demonstration examples from the training subset of a specific country, ensuring one example per main topic to capture a broad thematic spectrum. In contrast, food-based sampling draws all 12 demonstrations exclusively from the "food" topic, covering a range of subtopics within this domain. In both setups, the model is prompted with these demonstrations and tasked with selecting the most culturally appropriate completion for an unseen statement–choice pair. Demonstration examples are curated to represent diverse topics, promoting comprehensive coverage of region-specific cultural knowledge and reasoning patterns.

3.4 Evaluation

We evaluate the effects of cultural alignment by calculating the country-level accuracy improvements of the culturally aligned models over the baseline models for both multiple-choice (MCQ) and completion settings. We use the `lm-eval` framework (Gao et al., 2024), which computes accuracy based on log-likelihood. For completion, the log-likelihood assigned to the gold continuation is used to calculate accuracy. To analyze the impact of geographical distance on cultural alignment, we calculate the Pearson correlation between distances and accuracy improvements. In addition to country-level analysis, we also analyze performance by topic and report our findings.

4 Results

Our experiments reveal key findings across the four language models, as evident from Table 1 and Table 2, highlighting accuracy improvements when training on data from one country and evaluating across others. Each cell shows percentage-point

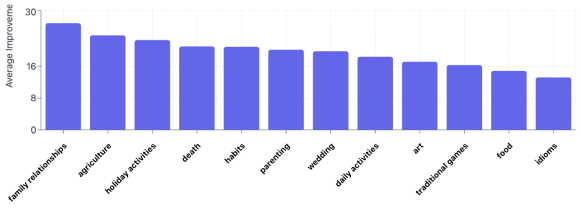

Figure 4: Topics Ranked by Average Improvement.

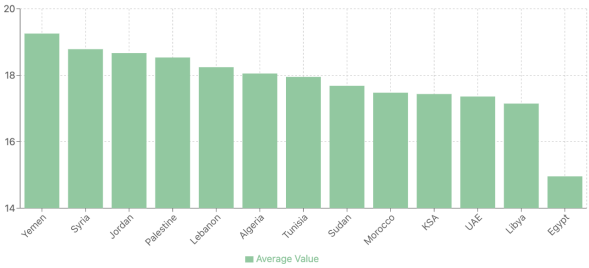

Figure 5: Countries as "Cultural Teachers" exhibiting the highest cross-cultural effect.

gains relative to the respective baseline models.

Country	Gemma-2 9B-it				SILMA 9B-Inst				Avg.
	DITTO		ICL		DITTO		ICL		
	Comp.	MCQ	Comp.	MCQ	Comp.	MCQ	Comp.	MCQ	
Algeria	0.44	28.71	**3.92**	0.00	-0.03	1.86	**1.54**	-2.73	4.21
Egypt	-0.66	37.51	2.35	5.50	-0.03	2.27	0.00	2.49	6.18
Jordan	1.10	22.78	3.58	11.37	0.63	1.98	1.19	2.20	5.60
KSA	-1.26	30.88	1.54	4.46	-0.19	0.98	0.69	0.26	4.67
Lebanon	0.97	**32.01**	**3.52**	3.27	1.13	2.61	**1.32**	3.81	6.08
Libya	-0.85	20.77	2.23	2.45	0.10	2.36	1.19	1.79	3.76
Morocco	1.19	20.45	1.76	3.45	-1.10	**3.46**	0.69	1.89	3.97
Palestine	0.84	28.53	2.92	0.25	-0.34	2.55	1.13	1.64	4.69
Sudan	0.75	31.45	3.14	8.36	-0.28	1.73	1.19	3.08	6.18
Syria	0.00	**37.38**	2.82	8.20	0.72	0.69	0.85	2.61	**6.66**
Tunisia	-1.42	21.58	2.86	1.16	0.22	1.54	0.66	2.61	3.65
UAE	0.50	29.06	2.42	17.97	-0.34	**3.24**	0.10	2.45	**6.93**
Yemen	-0.91	18.66	2.26	4.59	-1.10	-0.18	0.28	1.86	3.18
Avg.	0.05	**27.67**	**2.72**	**5.46**	-0.05	1.93	**0.83**	1.84	

Table 2: Overall accuracy improvements for Arab cultural commonsense reasoning when training on country-specific knowledge across different models with food-based sampling. Results show performance on Completion and MCQ tasks using DITTO and ICL methods. **Bold** and highlighted cells top two MCQ and Completion values for each model, as well as the top two country-based improvements.

Strong Cross-Cultural Transfer. Training on small demonstration sets from a single Arab country consistently improves model performance on other Arab cultures, that is, cross-cultural, averaging 2-5% gains in MCQ and completion tasks across models and methods. Interestingly, Syria as a source country ('teacher') results in the highest average improvement (5.02%) across all models and methods, followed by Jordan (4.94%) and UAE (4.73%). Furthermore, Jordan-trained Gemma-2 exhibits strong cross-cultural improvements, yielding a 4.8% completion gain with ICL and a 33.9% MCQ gain with DITTO. This cultural transfer occurs despite the geographical and cultural differences between countries, suggesting that cultural knowledge effectively transfers across the Arab region regardless of model architecture. Consistent cross-cultural improvements suggest that these models develop broader Arab cultural understanding rather than just memorizing country-specific features.

Multilingual VS. Arabic-centric. Comparing multilingual models (Qwen2.5 7B-Inst (Team, 2024a), Gemma-2 9B-it (Team et al., 2024)) with Arabic-centric models (ALLaM 7B-Inst (Bari et al., 2024), SILMA 9B-Inst (Team, 2024b)) reveals distinct performance patterns in Table 1. Correlated with lower baseline for multilingual models, Gemma-2 shows the largest relative gains in MCQ tasks (34.56_{base} +22.75% with DITTO), outperforming both Qwen2.5 ($51.65_{base} + 17.56\%$) and Arabic-centric models. In contrast, ALLaM shows the strongest improvement in the completion task (+3. 53% with ICL) despite its relatively high baseline (36.35%).

This pattern suggests that multilingual models excel at adapting cultural multiple-choice capabilities through cultural demonstrations, while Arabic-centric models more effectively enhance their generative understanding. Notably, Jordan's data produces exceptional gains with Gemma-2 (+33.93% MCQ), while Syria shows the highest cross-model improvement (5.02% average). The performance dichotomy persists in food-based sampling (Table 2), though with narrower gaps in completion tasks, indicating domain-specific demonstrations may partially neutralize architectural advantages. SILMA shows more balanced cross-task improvements with food-based sampling, suggesting Arabic-centric models benefit distinctly from fine-grained cultural knowledge.

Performance Comparison of DITTO and ICL. While ICL provides modest but highly consistent improvements with minimal negative transfers, DITTO demonstrates dramatically higher performance ceilings particularly on MCQ tasks with the Gemma-2 model, showing improvements up to 33.93% gain but with greater variability and occa-

sional negative transfers. For completion tasks, ICL consistently outperforms DITTO across all models, with ALLaM showing the largest gap (3.53% for ICL vs. -1.29% for DITTO). However, for MCQ tasks, DITTO excels with multilingual models, particularly Gemma-2 (22.75% with DITTO vs. 3.60% with ICL). This asymmetry suggests that iterative optimization benefits recognition tasks in multilingual models, while in-context demonstration better enhances generative capabilities, especially in Arabic-centric models. The gap between methods narrows in food-based sampling demonstrated in Table 2, indicating that domain-specific examples may reduce method-dependent variance. This suggests that optimal method selection depends on specific goals. DITTO excels when maximum potential improvement is the priority, while ICL offers better reliability for balanced performance across both knowledge and generation tasks with lower implementation complexity.

Transferability with Fine-Grained Sampling. When alignment data is restricted to a single domain (food), cross-cultural effects remain strong across methods as demonstrated in Table 2. Training on country-specific food-related examples can yield significant accuracy improvements, with Syria and UAE showing the highest overall average gains (6.66% and 6.93% respectively). The results demonstrate asymmetry in knowledge transfer effectiveness. Lebanon consistently performs well as a source of transfer learning, appearing in the top performers for both Gemma-2 and SILMA completion tasks. Notably, MCQ tasks show higher variability, with Gemma-2's DITTO method achieving remarkably strong improvements (averaging 27.67% across countries), particularly when trained on Syrian and Lebanese examples (37.38% and 32.01%, respectively). For completion tasks, Gemma-2 with ICL yields the strongest average improvement (2.72%), while SILMA benefits more modestly but consistently across methods. These findings indicate that the selection of fine-grained demonstrations fosters robust cross-cultural adaptation, but the degree of reciprocity in knowledge transfer varies substantially by country, model architecture, and assessment method.

5 Discussion

5.1 Topic Learnability

Our analysis reveals significant patterns in the cross-cultural transfer of commonsense reasoning within Arab cultural contexts. As illustrated in Figure 4, there is notable variation in both topic learnability and country transferability. Family relationships emerged as the most effectively aligned topic (26.83% improvement), followed by agriculture (23.79%) and holiday activities (22.59%), suggesting that explicitly structured social domains with clear cultural rules are most amenable to alignment techniques. Conversely, idioms (13.20%) and food (14.84%) showed the lowest improvements, indicating that linguistically embedded and context-dependent cultural elements present greater challenges. This hierarchy of topic learnability provides valuable insights for prioritizing cultural alignment efforts across different domains.

5.2 Country-Based Transfer Effectiveness

Our analysis across both alignment methods reveals a clear hierarchy in cultural knowledge transferability among Arab countries. Yemen emerges as the strongest cultural knowledge source (approximately 19.2% average improvement), with Syria, Jordan, and Palestine following closely behind (all with improvements between 18.5-18.8%). These four countries consistently demonstrate superior performance in their ability to transfer cultural knowledge across diverse Arab contexts. The middle tier includes Lebanon, Algeria, Tunisia, Sudan, Morocco, and KSA (ranging from approximately 17.5-18.2%). UAE and Libya show slightly lower transferability (approximately 17.0-17.3%), while Egypt consistently shows the weakest transferability (approximately 15%) among all countries examined, with a notable gap compared to all other countries.

5.3 Impact of Geographical Distance on Cross-Cultural Transfer

To measure the effect of geographical distance on cross-cultural transfer, we used the geographical distances between the capitals of each country shown in Table 5 in Appendix A and the accuracy improvements per country over the baseline to calculate the Pearson correlation between distance and accuracy improvement for each country. The average correlations coefficient across all countries and training methods are shown in Table 3. The Pearson correlation results for the Qwen2.5-7B model are shown in Figure 6, and a more detailed breakdown of the results for all the models is shown in the figures in Appendix B.

The results reveal significant variation in how the

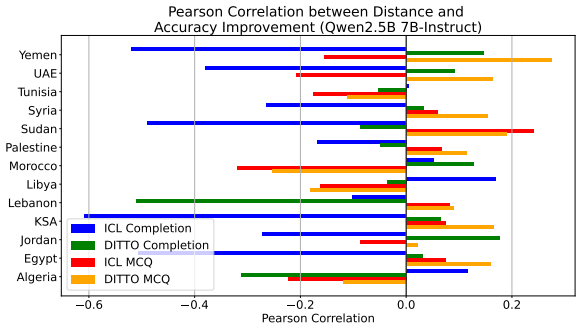

Figure 6: Pearson Correlation Coefficient between Distance from Training Country and Evaluation Accuracy Improvement for four different train/eval methods (Qwen2.5B 7B-Instruct base model).

Model	DITTO		ICL	
	Completion	MCQ	Completion	MCQ
ALLaM 7B-Inst	-0.0936	0.1277	-0.2506	-0.0024
Qwen2.5 7B-Inst	-0.0288	0.0515	-0.2280	-0.0564
SILMA 9B-Inst	-0.0672	0.0320	-0.0637	-0.0687
Gemma-2 9B-it	-0.0654	0.0557	-0.1661	-0.0543

Table 3: Mean Pearson correlations across countries between distance and accuracy improvement across models for Completion and MCQ tasks.

four models perform across the 13 countries using different evaluation methods. The data shows that performance varies not only by country but also by testing approach, with ICL Completion generally producing the most varied results and DITTO MCQ typically showing more positive correlations, as shown in Table 3. Notable patterns include the UAE consistently showing negative correlations across most models, while Morocco tends toward positive correlations, particularly with Gemma-2. The Gemma-2 model exhibits the most extreme correlation values, with correlation coefficients ranging from -0.8 to 0.65. These disparities likely reflect differences in cultural contexts, and potentially imbalanced training data representation from these regions, highlighting the challenges in developing language models that perform consistently across diverse Arabic-speaking populations.

5.4 Cross-Cultural Transfer Beyond Arab Culture

We additionally explore the effect of cross-cultural transfer when training on Indonesian cultural data and evaluating on Arabic cultural data. The results are shown in Table 4.

Table 4: Performance Comparison of Qwen-2.5 7B and ALLaM 7B Models

Context	MCQ Scores (%)			Completion Scores (%)		
	ICL	DiTTO	Base	ICL	DiTTO	Base
Qwen-2.5 7B-Instruct						
Arab LB	63.65	66.76		34.34	31.98	
Papua (ID)	71.13	67.17		34.49	32.14	
Arab AVG	69.30	69.21	51.65	35.77	33.16	32.89
Aceh (ID)	69.09	67.17		34.15	32.14	
Arab UB	71.57	70.63		36.88	34.53	
ALLaM 7B-Instruct-preview						
Arab LB	44.58	57.08		38.99	31.51	
Papua (ID)	71.22	71.88		37.76	36.44	
Arab AVG	55.02	70.04	69.90	39.88	35.06	36.35
Aceh (ID)	65.63	72.60		38.14	37.29	
Arab UB	65.06	73.61		40.68	37.13	

Note: Base scores are constant across contexts. ICL = In-Context Learning. ID contexts (Aceh, Papua) are out-of-culture, while Arab contexts represent in-culture testing.

5.5 Cultural Representation in Model Latent Space

To understand how different Arab cultures are internally represented within the model, we conducted a probing analysis across all layers of the Qwen model, using both one-vs-all and multiclass linear classifiers to assess the linear separability of cultural knowledge. The results are illustrated in Figure 7.

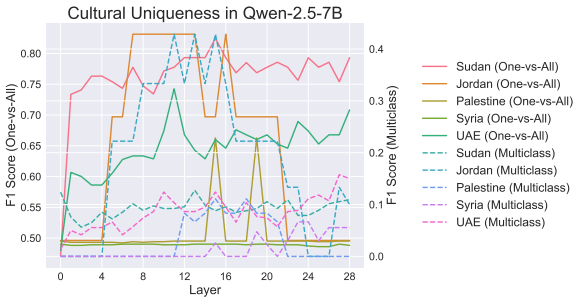

Figure 7: F1 scores across model layers for different countries using one-vs-all and multiclass classifiers.

Our probing analysis shown in Figures 7 and 8 reveals that Qwen encodes Arab cultures with varying distinctness, showing high linear separability for Sudan and Jordan but much lower for Palestine and Syria. Multiclass probing confirms the difficulty of jointly distinguishing multiple cultures, though Sudan and Jordan remain relatively more separable. After UAE-specific alignment, only the UAE showed improved cultural encoding, while other countries remained largely unchanged, yet reasoning performance improved across all countries. This suggests that targeted cultural alignment

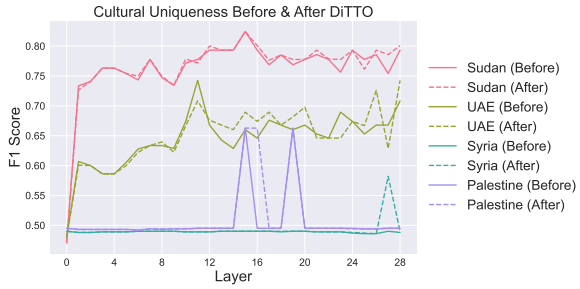

Figure 8: F1 scores across model layers for Sudan, UAE, Syria, and Palestine before and after UAE-specific alignment.

can enhance specific representations while indirectly benefiting generalization, offering a viable path toward culturally adaptive NLP systems.

6 Conclusion

Our study demonstrates that large language models can effectively achieve cross-cultural adaptation using lightweight alignment methods like ICL and DITTO, yielding consistent gains across Arab countries—even with limited, culturally specific data. Our experiments show strong cross-cultural generalizability, training on one country's dataset can significantly improve accuracy in other countries, with gains frequently exceeding 15–20%. Some country pairs show modest gains even at large distances, while others see minimal improvement despite close proximity, suggesting that cultural proximity is not strictly tied to geographic location. Probing analyses further show that targeted alignment enhances cultural encoding (e.g., for the UAE) without harming overall performance, highlighting the feasibility and benefits of culturally adaptive NLP in multilingual settings. In general, our results highlight that lightweight alignment methods can effectively align on the cultural commonsense reasoning task by incorporating region-specific cultural demonstrations. Whether through ICL or DITTO, LLMs can learn robust cultural representations that transfer to new countries. This work therefore provides evidence that cross-cultural adaptation is both feasible and beneficial in multilingual NLP settings, particularly in the Arab world.

7 Limitations

While our findings illuminate promising insights into pathways of cross-cultural transfer, several critical limitations constrain the scope of our conclusions.

Task Diversity. Our primary focus was on the evaluation of cultural multiple choice questions and a completion task. The realm of open-ended tasks (e.g., dialogue, narrative generation) introduces additional layers of complexity for cross-cultural alignment, underscoring the necessity for a deeper investigation into how cultural knowledge extrapolates across open-ended text generation.

Country Coverage While there are 22 countries that are members of the Arab League, the data set we use only represents 13 of them, which although more representative than other datasets, still does not completely represent the Arab world. This further proves the point we bring up in the introduction about the discrepancies in data availability by country, and emphasizes the importance of investigating cross-cultural transfer in low-resource settings.

Fine-Grained Cultural Nuances. Our analysis highlights performance variations even within topic categories, such as family relationships and idioms. In practice, cultural norms can be more nuanced and context-dependent than captured by any small demonstration set. A larger set of demonstrations and supervised fine-tuning may be required to mastering the intricacy of cultural knowledge that required memorization.

Despite these constraints, our work demonstrates that meticulously chosen examples, irrespective of being derived from broad topics or targeted domains, can significantly improve performance in varied cultural settings. These findings pave the way for future work that refines cross-cultural alignment strategies and investigates the interplay between linguistic diversity and cultural distance in multilingual NLP.

References

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.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless

assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

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, and 6 others. 2024. Allam: Large language models for arabic and english. *arXiv preprint arXiv: 2407.15390*.

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15042–15058, Bangkok, Thailand. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

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.

Global Media Insight. 2025. United arab emirates (uae) population statistics 2025. Accessed: 2025-05-20.

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, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Preprint*, arXiv:2402.10946.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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

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*.

Omar Shaikh, Michelle Lam, Joey Hejna, Yijia Shao, Michael Bernstein, and Diyi Yang. 2024. Show, don't tell: Aligning language models with demonstrated feedback. *Preprint*, arXiv:2406.00888.

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.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv: 2408.00118*.

Qwen Team. 2024a. Qwen2.5: A party of foundation models.

Silma Team. 2024b. Silma.

United Nations Population Fund. 2025. Egypt - world population dashboard. Accessed: 2025-05-20.

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.

A Distance Between Countries

The approximate distances[1] between the capitals of each country are shown in Table 5.

B Correlations between Geographical Distance and Accuracy Improvement

Table 6 shows the overall median correlation scores across models and settings to supplement the means in Table 3. Figures 20 through 12 show the correlation scores for the remaining three models (ALLaM 7B-Instruct, SILMA 9B-Instruct, and Gemma-2 9B-It), which are also displayed in heatmaps in Figure 9. Additionally, Figure 13 shows the correlation scores averaged across the Arabic models ALLaM 7B-Instruct and SILMA 9B-Instruct, while Figure 14 shows the correlation scores averaged across the multilingual models Qwen2.5 7B-Instruct and Gemma-2 9B-It. To demonstrate what the correlations look like, Figure 15 shows the accuracy improvement vs distance graph for the strongest correlation, while Figure 16 is for the weakest correlation.

[1] https://www.distance.to/

Model	DITTO		ICL	
	Completion	MCQ	Completion	MCQ
ALLaM 7B-Inst	-0.0986	0.1492	-0.2422	0.0033
Qwen2.5 7B-Inst	0.0311	0.1149	-0.2636	-0.0859
SILMA 9B-Inst	-0.0421	0.0477	-0.0553	-0.1741
Gemma-2 9B-it	-0.2122	0.0868	-0.3097	-0.1176

Table 6: Median Pearson correlations between distance and accuracy improvement across models for Completion and MCQ tasks.

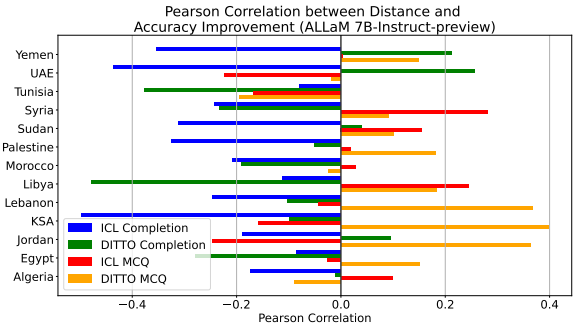

Figure 10: Pearson Correlation Coefficient between Distance from Training Country and Evaluation Accuracy Improvement for four different train/eval methods (ALLaM 7B-Instruct-preview base model).

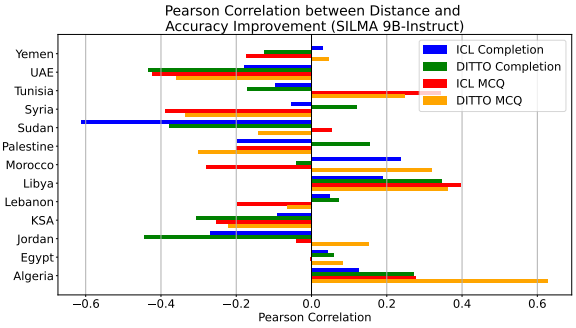

Figure 11: Pearson Correlation Coefficient between Distance from Training Country and Evaluation Accuracy Improvement for four different train/eval methods (SILMA 9B-Instruct base model).

Table 5: Distance Matrix Between Countries (in kilometers)

From/To	Morocco	Algeria	Tunisia	Libya	Egypt	Sudan	Palestine	Jordan	Syria	Lebanon	KSA	UAE	Yemen
Morocco	0	948	1,569	1,859	3,596	4,435	3,913	3,968	3,958	3,876	5,234	5,946	5,493
Algeria	948	0	630	1,016	2,706	3,755	2,996	3,048	3,027	2,944	4,340	5,032	4,695
Tunisia	1,569	630	0	518	2,090	3,245	2,368	2,419	2,397	2,314	3,717	4,403	4,117
Libya	1,859	1,016	518	0	1,739	2,753	2,077	2,135	2,148	2,071	3,377	4,098	3,680
Egypt	3,596	2,706	2,090	1,739	0	1,596	432	494	613	485	1,639	2,363	2,104
Sudan	4,435	3,755	3,245	2,753	1,596	0	1,794	1,821	1,997	2,027	1,738	2,426	1,201
Palestine	3,913	2,996	2,368	2,077	432	1,794	0	63	213	234	1,369	2,036	2,039
Jordan	3,968	3,048	2,419	2,135	494	1,821	63	0	177	219	1,328	1,984	2,027
Syria	3,958	3,027	2,397	2,148	613	1,997	213	177	0	86	1,408	2,019	2,170
Lebanon	3,876	2,944	2,314	2,071	485	2,027	234	219	86	0	1,494	2,107	2,240
KSA	5,234	4,340	3,717	3,377	1,638	1,738	1,369	1,328	1,408	1,494	0	773	1,070
UAE	5,946	5,032	4,403	4,098	2,363	2,426	2,036	1,984	2,019	2,107	773	0	1,467
Yemen	5,493	4,695	4,117	3,680	2,104	1,201	2,039	2,027	2,170	2,240	1,070	1,467	0

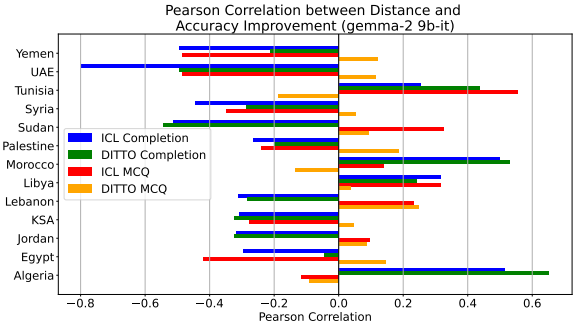

Figure 12: Pearson Correlation Coefficient between Distance from Training Country and Evaluation Accuracy Improvement for four different train/eval methods (gemma-2 9b-it base model).

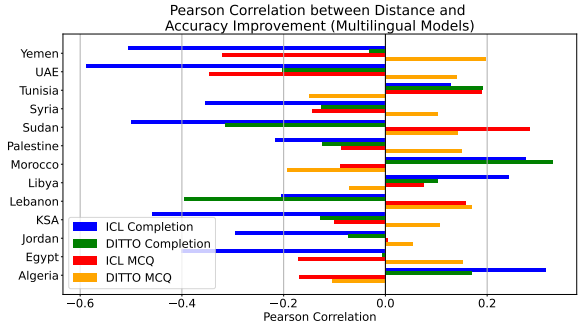

Figure 14: Pearson Correlation Coefficient between Distance from Training Country and Evaluation Accuracy Improvement for four different train/eval methods (Multilingual Models Averaged).

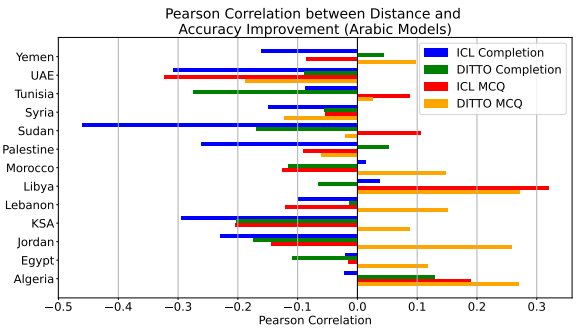

Figure 13: Pearson Correlation Coefficient between Distance from Training Country and Evaluation Accuracy Improvement for four different train/eval methods (Arabic Models Averaged).

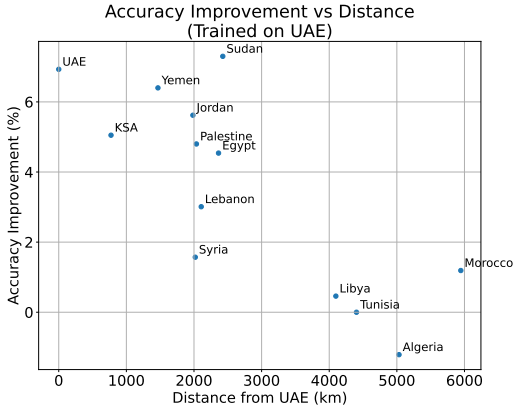

Figure 15: Evaluation Accuracy Improvement vs. Distance for ICL Topic-based Training on Samples from the UAE with completion evaluation (gemma-2 9b-it). Pearson Correlation = -0.796699.

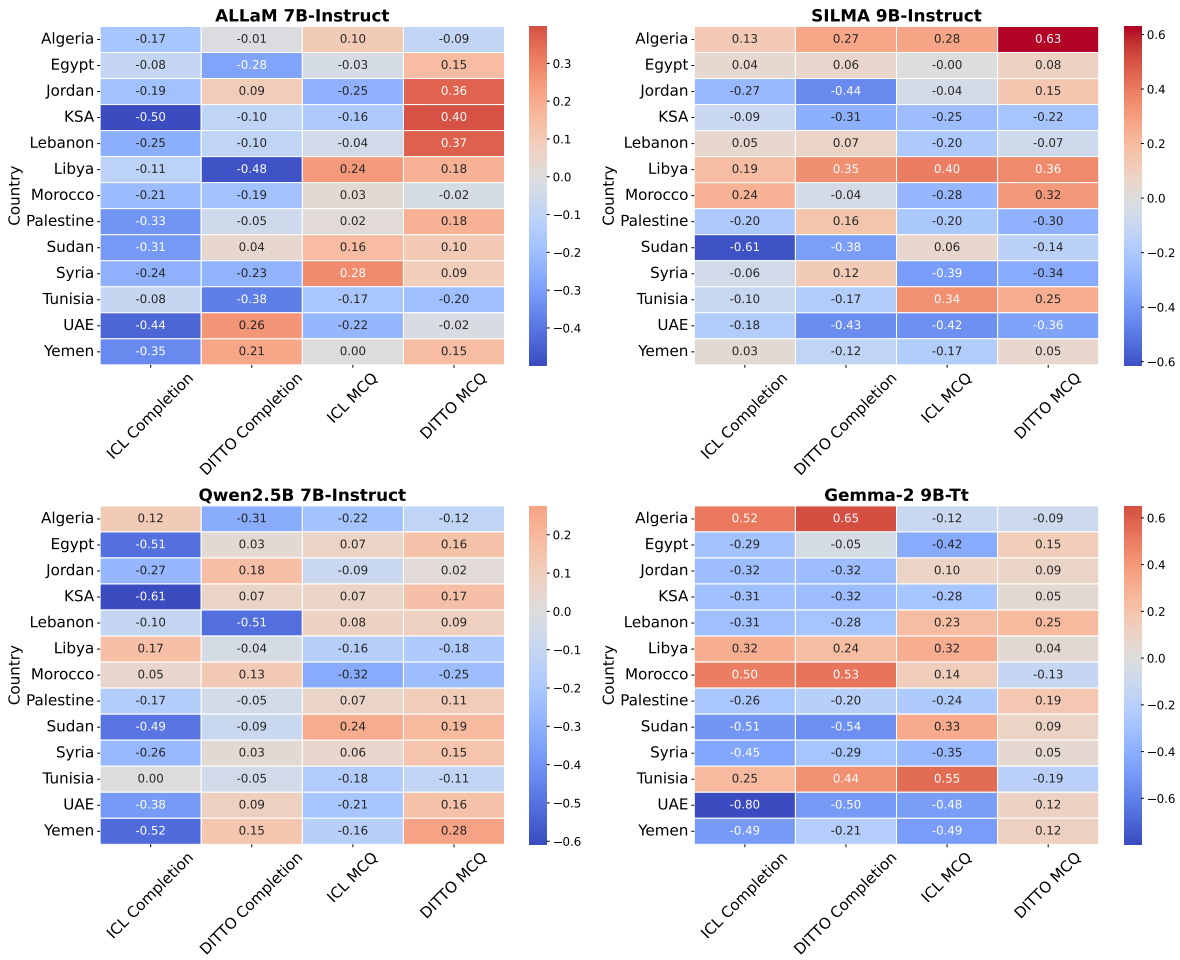

Figure 9: Pearson Correlation Coefficient between Distance from Training Country and Evaluation Accuracy Improvement for four different train/eval methods with topic-based sampling.

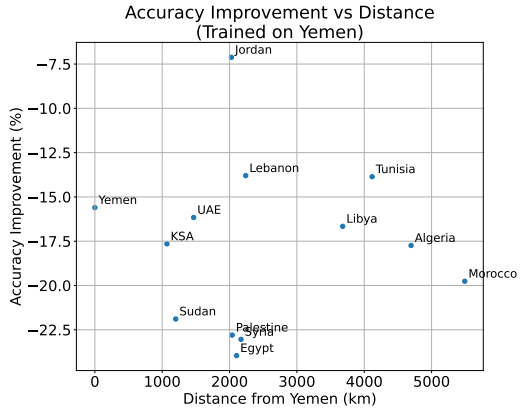

Figure 16: Evaluation Accuracy Improvement vs. Distance for ICL Topic-based Training on Samples from Yemen with MCQ Evaluation (ALLaM 7B-Instruct-preview). Pearson Correlation = 0.003255.

C Cultural Representation in Models

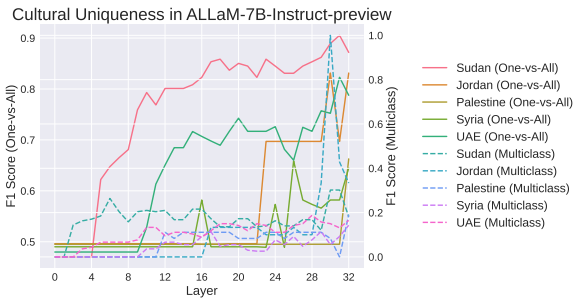

Figure 17: F1 scores across model layers for different countries using one-vs-all and multiclass classifiers.

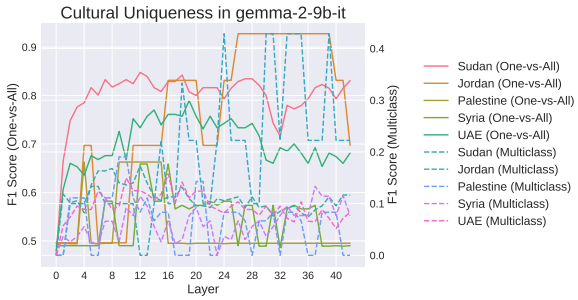

Figure 18: F1 scores across model layers for different countries using one-vs-all and multiclass classifiers.

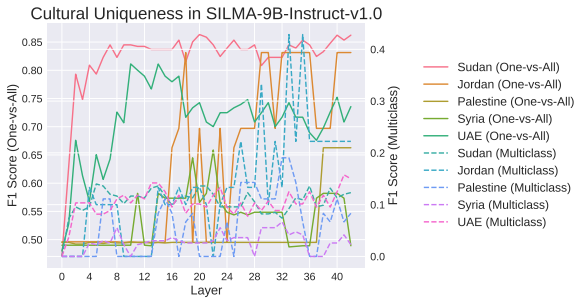

Figure 19: F1 scores across model layers for different countries using one-vs-all and multiclass classifiers.

D Cross-Cultural Transfer between Indonesian and Arab cultures

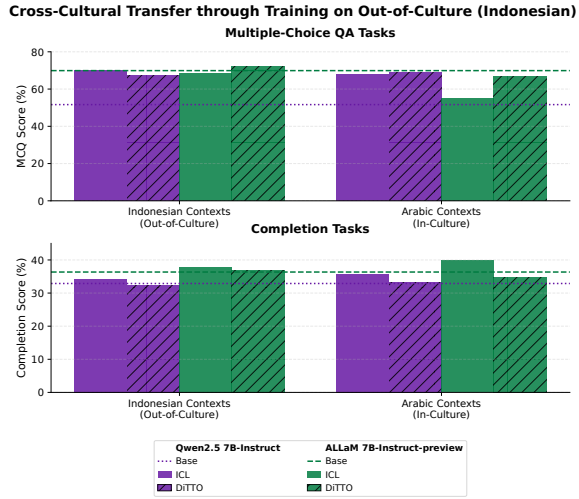

Figure 20: .

E Detailed Accuracy Improvement for All Models

Table 7: Cross-country evaluation results for Multilingual language models. Models are evaluated on different countries (columns) after being trained on specific countries (rows). Values represent score difference from the base model.

Model	Method	Trained On	Δ MCQ vs. Base													Δ Completion vs. Base											
			Alg	Egy	Jor	KSA	Leb	Lib	Mor	Pal	Sud	Syr	Tun	UAE	Yem	Alg	Egy	Jor	KSA	Leb	Lib	Mor	Pal	Sud	Syr	Tun	UAE	Yem
Qwen2.5 7B-Inst	ICL	Algeria	18.1	20.7	15.7	19.3	17.2	28.2	21.7	24.8	15.9	16.4	13.9	25.0	4.4	0.4	6.2	7.9	2.1	-0.4	1.9	3.2	3.2	8.6	0.8	1.3	1.9	-0.8
		Egypt	22.2	22.3	16.5	22.3	15.1	29.6	24.5	27.2	15.0	17.2	13.4	25.4	5.6	-0.4	8.7	6.4	2.5	-1.7	-1.4	0.8	4.0	7.3	2.4	1.3	0.8	0.0
		Jordan	15.3	18.2	17.2	15.1	13.4	23.6	17.8	22.8	15.0	16.8	13.0	24.6	9.2	-2.8	3.7	13.9	0.0	0.0	1.9	3.5	0.4	12.9	3.5	0.4	0.8	-2.8
		KSA	18.6	22.3	18.0	19.3	15.1	27.8	22.9	27.2	17.2	19.5	13.9	27.3	10.0	-1.6	6.2	6.4	3.4	1.7	0.5	1.2	3.6	7.7	3.1	2.1	4.2	3.2
		Lebanon	21.4	18.6	16.1	19.8	15.5	29.2	21.0	23.2	12.9	18.4	12.2	26.2	8.8	-1.2	7.0	12.4	-0.8	1.7	1.9	4.7	1.6	13.7	-0.8	0.8	2.7	1.6
		Libya	15.3	19.8	15.7	16.8	11.6	26.9	15.8	22.4	12.9	16.4	11.3	24.6	8.0	-1.2	6.6	9.0	0.4	-0.4	2.3	1.2	5.6	8.2	2.0	2.1	5.0	1.2
		Morocco	19.8	16.9	13.9	20.2	15.5	28.7	24.5	24.0	15.9	18.0	15.5	26.2	6.0	-0.8	6.2	10.9	0.0	1.7	1.4	3.2	4.4	10.3	2.7	2.9	-0.4	0.4
		Palestine	19.3	17.4	16.9	19.8	16.8	30.1	20.2	21.6	13.7	16.0	11.8	25.8	6.0	-2.8	2.5	4.5	1.3	-0.4	0.0	4.0	3.6	6.4	-0.4	0.0	0.0	0.0
		Sudan	16.9	15.3	14.2	19.3	16.4	25.0	16.2	19.6	10.7	16.4	8.4	21.5	10.4	0.0	5.4	13.5	2.1	2.2	0.5	0.8	2.8	11.2	1.2	2.1	1.5	-2.0
		Syria	18.9	20.2	16.5	16.0	17.2	32.9	22.1	22.8	16.3	18.8	11.8	27.7	7.2	-1.6	6.2	9.7	4.2	1.3	1.9	3.2	3.2	10.3	1.2	2.9	2.3	-2.0
		Tunisia	20.6	17.8	16.1	19.8	13.8	30.1	21.4	22.0	14.2	18.8	10.1	23.8	8.0	0.8	6.2	1.1	0.8	1.7	0.5	0.0	4.0	6.0	1.2	2.9	2.7	-1.6
		UAE	13.7	14.0	15.7	19.8	13.4	25.5	16.6	20.4	13.7	16.8	13.4	24.6	11.2	0.8	7.9	7.9	4.2	1.3	1.4	2.8	3.6	11.6	2.7	0.4	6.2	0.8
		Yemen	6.4	12.4	12.7	16.8	8.6	18.0	9.5	16.4	15.4	11.7	8.4	17.3	2.8	-1.2	3.7	5.6	0.4	1.7	2.8	1.2	1.2	4.7	1.2	0.4	1.9	4.0
	Ditto	Algeria	18.6	16.5	14.6	14.7	16.4	21.3	21.7	22.8	15.0	13.3	9.7	23.5	9.6	0.8	2.9	3.0	0.0	1.3	1.9	-2.8	-0.4	2.6	-0.8	1.3	-5.0	-1.6
		Egypt	23.0	18.6	16.9	16.0	15.1	24.5	22.9	21.6	15.4	15.6	7.6	23.1	8.0	-1.2	1.2	1.5	0.0	-0.4	-0.9	1.2	1.6	2.6	-5.5	-0.4	1.2	-2.0
		Jordan	19.4	20.7	15.4	18.9	19.0	25.5	24.5	24.0	17.2	19.5	8.8	23.5	10.0	-1.6	6.2	-0.7	-0.4	-2.6	1.4	2.4	-1.6	2.2	-3.5	-1.3	-2.3	-1.6
		KSA	18.9	19.4	16.1	16.8	18.1	23.6	23.3	24.8	15.4	16.4	8.4	21.2	9.6	1.2	3.3	2.6	0.8	-1.3	0.0	0.0	-3.6	4.3	-4.7	-1.3	-2.7	0.8
		Lebanon	21.0	17.4	16.5	16.4	18.5	23.2	22.5	24.4	16.3	16.4	12.6	24.2	9.6	1.6	3.7	3.0	2.1	0.9	-0.5	-3.6	2.0	2.6	-0.8	1.7	-0.8	0.0
		Libya	17.3	13.2	15.0	15.6	13.8	19.9	19.0	18.8	14.6	16.4	9.7	20.8	2.4	-1.2	3.3	0.0	2.1	-0.9	0.0	-5.5	1.2	0.9	-2.7	2.5	-1.5	0.8
		Morocco	20.6	16.1	15.7	17.6	13.8	24.5	23.7	21.2	15.4	14.5	8.4	23.8	8.8	0.8	3.3	6.0	1.3	1.3	-0.9	2.0	-0.4	9.0	-2.3	0.0	-1.2	2.4
		Palestine	18.6	21.9	16.1	17.6	15.1	25.9	22.9	22.8	18.0	13.7	10.1	21.9	9.6	-0.4	6.2	0.0	2.5	-1.3	0.5	-1.6	-1.6	2.6	-3.9	2.1	-1.5	-3.2
		Sudan	19.4	22.7	16.5	19.3	16.0	24.5	24.1	22.8	18.4	18.4	9.7	23.1	12.0	2.4	2.5	6.0	1.3	-0.4	0.0	3.6	-2.0	6.9	-2.0	1.7	-3.8	-2.0
		Syria	16.9	17.8	15.4	16.4	16.4	22.7	23.3	21.2	16.3	14.1	8.8	22.3	9.6	2.8	4.1	3.8	0.0	-1.3	1.9	-1.2	1.2	5.2	-3.1	2.1	-1.2	-1.2
		Tunisia	16.9	17.4	15.7	15.1	17.7	23.6	21.0	22.0	17.2	13.7	11.8	21.9	10.0	2.0	2.1	-5.2	1.7	-0.4	0.0	-2.0	-1.2	1.3	-5.1	-0.8	-1.5	-0.4
		UAE	18.1	15.7	16.9	17.6	16.0	24.5	22.1	20.8	18.4	12.1	8.8	21.2	7.2	1.6	4.1	3.4	0.8	-0.4	0.5	-1.2	2.0	6.9	-1.2	1.3	-3.5	0.0
		Yemen	18.6	19.8	15.7	17.6	16.8	23.6	23.3	24.8	19.3	16.0	11.8	23.5	10.8	0.8	2.9	-3.0	0.4	-1.7	-2.3	0.4	-0.4	1.7	-4.7	0.0	-4.6	-0.8
Gemma-2 9b-it	ICL	Algeria	6.5	4.5	0.7	3.8	3.5	2.3	5.9	8.4	1.7	10.6	2.1	5.8	0.4	-0.4	2.9	9.4	8.0	3.5	-0.5	2.8	5.2	8.6	2.0	2.9	3.1	2.4
		Egypt	-0.4	0.8	0.7	1.3	0.0	0.0	0.4	1.2	0.9	1.6	0.8	0.8	0.0	-2.8	3.7	3.8	3.8	1.7	-1.4	2.4	3.6	5.6	-1.6	3.4	-0.4	0.8
		Jordan	14.5	6.6	1.1	12.6	6.9	13.0	12.6	16.4	4.7	20.7	10.9	15.0	0.4	-2.4	3.3	17.6	6.3	3.0	-0.9	4.7	2.8	10.7	2.7	3.4	3.9	6.0
		KSA	7.3	5.8	0.4	8.0	4.8	3.7	2.4	10.4	4.3	16.8	3.8	6.5	2.8	-2.4	1.2	8.6	4.2	3.5	0.9	5.9	4.8	6.9	0.0	1.7	3.1	5.6
		Lebanon	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.4	0.4	-0.4	2.0	-3.6	4.5	8.2	9.7	4.7	0.0	4.0	0.8	6.4	2.7	2.5	3.1	2.4
		Libya	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.4	0.0	-0.4	2.8	-4.0	2.1	7.5	4.2	0.0	-0.9	3.6	5.2	3.9	0.4	4.2	1.9	2.4
		Morocco	6.9	6.6	0.7	8.4	4.8	4.6	5.5	10.4	2.1	17.6	6.3	10.0	4.4	-3.2	5.0	10.1	8.0	1.7	0.0	3.6	3.6	8.6	1.2	2.5	3.5	4.0
		Palestine	0.0	0.8	0.4	1.3	0.0	0.0	0.4	0.0	0.9	1.6	0.4	0.4	-1.2	-2.0	1.2	9.7	7.2	-0.9	0.5	0.4	4.4	10.3	1.6	3.8	3.9	4.0
		Sudan	16.1	12.0	3.0	16.0	6.0	12.5	14.6	28.4	3.0	24.6	10.9	21.5	0.0	-2.8	0.8	10.9	4.2	4.3	-2.3	4.7	0.8	11.6	1.2	2.5	1.9	4.0
		Syria	0.0	0.8	0.0	0.4	0.9	0.0	0.4	0.0	0.9	1.6	0.0	0.4	0.8	-6.5	2.1	6.0	4.6	3.9	-1.9	4.4	5.6	7.3	3.9	0.0	1.2	2.8
		Tunisia	0.0	0.8	0.0	0.4	0.4	0.0	0.4	1.2	1.3	1.6	0.4	0.8	1.6	-4.0	3.7	5.6	4.6	1.7	-1.4	2.4	5.2	3.4	0.4	5.0	1.9	0.8
		UAE	0.8	0.4	0.4	2.5	3.0	0.5	0.8	2.8	0.9	4.7	1.3	2.3	2.0	-1.2	4.5	5.6	5.0	3.0	0.5	1.2	4.8	7.3	1.6	0.0	6.9	6.4
		Yemen	-0.4	0.8	0.0	0.8	0.0	0.0	0.4	0.0	0.4	0.8	0.0	-0.4	1.2	-0.8	2.9	4.5	5.5	0.4	0.9	1.6	6.4	3.9	-1.2	3.8	1.5	5.2
	Ditto	Algeria	25.4	16.1	33.0	26.9	10.4	32.4	32.8	28.8	25.8	29.3	18.5	26.2	21.6	-4.0	-1.2	-1.9	1.7	2.2	-1.9	-4.4	1.6	4.7	-0.8	-0.4	-1.2	3.2
		Egypt	24.6	18.6	43.8	26.5	15.5	35.2	41.5	31.2	39.5	30.1	15.1	30.0	15.2	-4.8	1.2	-6.0	0.0	3.5	-3.7	-3.2	-5.2	1.7	-4.7	1.7	-1.5	1.2
		Jordan	31.9	23.1	49.4	32.8	20.3	40.7	44.7	38.4	41.2	35.2	21.8	33.5	26.4	-5.2	2.1	5.6	2.1	1.7	-0.9	0.0	0.0	7.7	0.0	2.9	1.5	1.6
		KSA	18.2	20.2	44.6	23.9	13.4	32.4	37.6	25.6	40.8	30.9	18.9	31.9	16.8	-1.6	7.4	6.4	2.1	6.5	1.9	0.8	0.0	9.5	1.6	3.8	0.4	4.8
		Lebanon	7.3	1.2	2.2	3.8	1.7	15.3	15.0	11.2	9.0	15.2	0.0	6.9	4.8	-5.6	3.3	4.5	3.0	4.3	0.9	-1.2	-4.4	7.7	0.8	2.1	0.4	1.6
		Libya	32.3	21.1	49.8	32.4	16.0	40.7	41.1	40.0	43.3	36.3	20.2	30.4	30.8	-2.4	4.1	0.8	5.5	3.0	-1.9	-4.7	-0.8	6.9	-3.5	0.8	-1.9	0.0
		Morocco	8.5	9.9	30.3	8.4	4.3	14.8	23.7	14.4	26.2	12.1	6.3	11.5	6.0	-6.5	-2.1	4.1	3.0	1.3	-1.9	-2.0	-5.2	4.3	-3.1	0.0	0.4	2.4
		Palestine	22.6	18.6	30.3	23.5	13.4	34.7	33.2	26.8	29.2	30.1	17.2	23.5	20.8	-2.8	2.9	0.8	0.0	0.4	2.3	-4.0	-2.4	1.7	-0.8	3.4	-0.8	4.0
		Sudan	14.1	10.7	21.3	13.0	5.2	17.1	23.3	18.0	20.6	17.6	10.1	14.2	10.0	-4.4	0.4	3.0	2.5	2.2	-0.9	0.4	3.2	3.0	-1.2	2.9	2.7	4.8
		Syria	29.0	24.8	42.7	32.4	18.1	40.3	40.7	37.2	36.9	35.9	19.3	32.7	22.4	-5.6	3.7	4.1	0.9	3.9	0.9	-2.0	-4.8	6.9	-2.3	-0.4	-1.9	-0.4
		Tunisia	21.0	14.9	23.6	18.9	12.1	28.7	27.7	22.4	21.9	25.8	14.7	20.0	15.6	-1.2	2.5	5.6	3.8	2.6	-1.4	-2.0	-1.2	7.3	-3.1	0.4	-1.5	6.0
		UAE	24.2	17.8	43.8	24.8	17.2	37.0	37.9	30.4	34.3	28.9	17.2	30.8	20.0	-2.4	5.0	6.4	3.8	2.6	1.9	0.0	1.6	4.3	0.8	1.3	0.4	5.6

Method	Trained On	Δ MCQ vs. Base													Δ Completion vs. Base											
		Alg	Egy	Jor	KSA	Leb	Lib	Mor	Pal	Sud	Syr	Tun	UAE	Yem	Alg	Egy	Jor	KSA	Leb	Lib	Mor	Pal	Sud	Syr	Tun	UAE	Yem
ICL	Algeria	18.1	20.7	15.7	19.3	17.2	28.2	21.7	24.8	15.9	16.4	13.9	25.0	4.4	0.4	6.2	7.9	2.1	-0.4	1.9	3.2	3.2	8.6	0.8	1.3	1.9	-0.8
	Egypt	22.2	22.3	16.5	22.3	15.1	29.6	24.5	27.2	15.0	17.2	13.4	25.4	5.6	-0.4	8.7	6.4	2.5	-1.7	-1.4	0.8	4.0	7.3	2.4	1.3	0.8	0.0
	Jordan	15.3	18.2	17.2	15.1	13.4	23.6	17.8	22.8	15.0	16.8	13.0	24.6	9.2	-2.8	3.7	13.9	0.0	0.0	1.9	3.5	0.4	12.9	3.5	0.4	0.8	-2.8
	KSA	18.6	22.3	18.0	19.3	15.1	27.8	22.9	27.2	17.2	19.5	13.9	27.3	10.0	-1.6	6.2	6.4	3.4	1.7	0.5	1.2	3.6	7.7	3.1	2.1	4.2	3.2
	Lebanon	21.4	18.6	16.1	19.8	15.5	29.2	21.0	23.2	12.9	18.4	12.2	26.2	8.8	-1.2	7.0	12.4	-0.8	1.7	1.9	4.7	1.6	13.7	-0.8	0.8	2.7	1.6
	Libya	15.3	19.8	15.7	16.8	11.6	26.9	15.8	22.4	12.9	16.4	11.3	24.6	8.0	-1.2	6.6	9.0	0.4	-0.4	2.3	1.2	5.6	8.2	2.0	2.1	5.0	1.2
	Morocco	19.8	16.9	13.9	20.2	15.5	28.7	24.5	24.0	15.9	18.0	15.5	26.2	6.0	-0.8	6.2	10.9	0.0	1.7	1.4	3.2	4.4	10.3	2.7	2.9	-0.4	0.4
	Palestine	19.3	17.4	16.9	19.8	16.8	30.1	20.2	21.6	13.7	16.0	11.8	25.8	6.0	-2.8	2.5	4.5	1.3	-0.4	0.0	4.0	3.6	6.4	-0.4	0.0	0.0	0.0
	Sudan	16.9	15.3	14.2	19.3	16.4	25.0	16.2	19.6	10.7	16.4	8.4	21.5	10.4	0.0	5.4	13.5	2.1	2.2	0.5	0.8	2.8	11.2	1.2	2.1	1.5	-2.0
	Syria	18.9	20.2	16.5	16.0	17.2	32.9	22.1	22.8	16.3	18.8	11.8	27.7	7.2	-1.6	6.2	9.7	4.2	1.3	1.9	3.2	3.2	10.3	1.2	2.9	2.3	-2.0
	Tunisia	20.6	17.8	16.1	19.8	13.8	30.1	21.4	22.0	14.2	18.8	10.1	23.8	8.0	0.8	6.2	1.1	0.8	1.7	0.5	0.0	4.0	6.0	1.2	2.9	2.7	-1.6
	UAE	13.7	14.0	15.7	19.8	13.4	25.5	16.6	20.4	13.7	16.8	13.4	24.6	11.2	0.8	7.9	7.9	4.2	1.3	1.4	2.8	3.6	11.6	2.7	0.4	6.2	0.8
	Yemen	6.4	12.4	12.7	16.8	8.6	18.0	9.5	16.4	15.4	11.7	8.4	17.3	2.8	-1.2	3.7	5.6	0.4	1.7	2.8	1.2	1.2	4.7	1.2	0.4	1.9	4.0
Ditto	Algeria	18.6	16.5	14.6	14.7	16.4	21.3	21.7	22.8	15.0	13.3	9.7	23.5	9.6	0.8	2.9	3.0	0.0	1.3	1.9	-2.8	-0.4	2.6	-0.8	1.3	-5.0	-1.6
	Egypt	23.0	18.6	16.9	16.0	15.1	24.5	22.9	21.6	15.4	15.6	7.6	23.1	8.0	-1.2	1.2	1.5	0.0	-0.4	-0.9	1.2	1.6	2.6	-5.5	-0.4	1.2	-2.0
	Jordan	19.4	20.7	15.4	18.9	19.0	25.5	24.5	24.0	17.2	19.5	8.8	23.5	10.0	-1.6	6.2	-0.7	-0.4	-2.6	1.4	2.4	-1.6	2.2	-3.5	-1.3	-2.3	-1.6
	KSA	18.9	19.4	16.1	16.8	18.1	23.6	23.3	24.8	15.4	16.4	8.4	21.2	9.6	1.2	3.3	2.6	0.8	-1.3	0.0	0.0	-3.6	4.3	-4.7	-1.3	-2.7	0.8
	Lebanon	21.0	17.4	16.5	16.4	18.5	23.2	22.5	24.4	16.3	16.4	12.6	24.2	9.6	1.6	3.7	3.0	2.1	0.9	-0.5	-3.6	2.0	2.6	-0.8	1.7	-0.8	0.0
	Libya	17.3	13.2	15.0	15.6	13.8	19.9	19.0	18.8	14.6	16.4	9.7	20.8	2.4	-1.2	3.3	0.0	2.1	-0.9	0.0	-5.5	1.2	0.9	-2.7	2.5	-1.5	0.8
	Morocco	20.6	16.1	15.7	17.6	13.8	24.5	23.7	21.2	15.4	14.5	8.4	23.8	8.8	0.8	3.3	6.0	1.3	1.3	-0.9	2.0	-0.4	9.0	-2.3	0.0	-1.2	2.4
	Palestine	18.6	21.9	16.1	17.6	15.1	25.9	22.9	22.8	18.0	13.7	10.1	21.9	9.6	-0.4	6.2	0.0	2.5	-1.3	0.5	-1.6	-1.6	2.6	-3.9	2.1	-1.5	-3.2
	Sudan	19.4	22.7	16.5	19.3	16.0	24.5	24.1	22.8	18.4	18.4	9.7	23.1	12.0	2.4	2.5	6.0	1.3	-0.4	0.0	3.6	-2.0	6.9	-2.0	1.7	-3.8	-2.0
	Syria	16.9	17.8	15.4	16.4	16.4	22.7	23.3	21.2	16.3	14.1	8.8	22.3	9.6	2.8	4.1	3.8	0.0	-1.3	1.9	-1.2	1.2	5.2	-3.1	2.1	-1.2	-1.2
	Tunisia	16.9	17.4	15.7	15.1	17.7	23.6	21.0	22.0	17.2	13.7	11.8	21.9	10.0	2.0	2.1	-5.2	1.7	-0.4	0.0	-2.0	-1.2	1.3	-5.1	-0.8	-1.5	-0.4
	UAE	18.1	15.7	16.9	17.6	16.0	24.5	22.1	20.8	18.4	12.1	8.8	21.2	7.2	1.6	4.1	3.4	0.8	-0.4	0.5	-1.2	2.0	6.9	-1.2	1.3	-3.5	0.0
	Yemen	18.6	19.8	15.7	17.6	16.8	23.6	23.3	24.8	19.3	16.0	11.8	23.5	10.8	0.8	2.9	-3.0	0.4	-1.7	-2.3	0.4	-0.4	1.7	-4.7	0.0	-4.6	-0.8

Table 8: Cross-country evaluation results for Qwen2.5 7B-Instruct. Models are evaluated on different countries (columns) after being trained on specific countries (rows). Values represent score difference from the base model.

Method	Trained On	Δ MCQ vs. Base													Δ Completion vs. Base											
		Alg	Egy	Jor	KSA	Leb	Lib	Mor	Pal	Sud	Syr	Tun	UAE	Yem	Alg	Egy	Jor	KSA	Leb	Lib	Mor	Pal	Sud	Syr	Tun	UAE	Yem
ICL	Algeria	6.5	4.5	0.7	3.8	3.5	2.3	5.9	8.4	1.7	10.6	2.1	5.8	0.4	-0.4	2.9	9.4	8.0	3.5	-0.5	2.8	5.2	8.6	2.0	2.9	3.1	2.4
	Egypt	-0.4	0.8	0.7	1.3	0.0	0.0	0.4	1.2	0.9	1.6	0.8	0.8	0.0	-2.8	3.7	3.8	3.8	1.7	-1.4	2.4	3.6	5.6	-1.6	3.4	-0.4	0.8
	Jordan	14.5	6.6	1.1	12.6	6.9	13.0	12.6	16.4	4.7	20.7	10.9	15.0	0.4	-2.4	3.3	17.6	6.3	3.0	-0.9	4.7	2.8	10.7	2.7	3.4	3.9	6.0
	KSA	7.3	5.8	0.4	8.0	4.8	3.7	2.4	10.4	4.3	16.8	3.8	6.5	2.8	-2.4	1.2	8.6	4.2	3.5	0.9	5.9	4.8	6.9	0.0	1.7	3.1	5.6
	Lebanon	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.4	0.4	-0.4	2.0	-3.6	4.5	8.2	9.7	4.7	0.0	4.0	0.8	6.4	2.7	2.5	3.1	2.4
	Libya	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.4	0.0	-0.4	2.8	-4.0	2.1	7.5	4.2	0.0	-0.9	3.6	5.2	3.9	0.4	4.2	1.9	2.4
	Morocco	6.9	6.6	0.7	8.4	4.8	4.6	5.5	10.4	2.1	17.6	6.3	10.0	4.4	-3.2	5.0	10.1	8.0	1.7	0.0	3.6	3.6	8.6	1.2	2.5	3.5	4.0
	Palestine	0.0	0.8	0.4	1.3	0.0	0.0	0.4	0.0	0.9	1.6	0.4	0.4	-1.2	-2.0	1.2	9.7	7.2	-0.9	0.5	0.4	4.4	10.3	1.6	3.8	3.9	4.0
	Sudan	16.1	12.0	3.0	16.0	6.0	12.5	14.6	28.4	3.0	24.6	10.9	21.5	0.0	-2.8	0.8	10.9	4.2	4.3	-2.3	4.7	0.8	11.6	1.2	2.5	1.9	4.0
	Syria	0.0	0.8	0.0	0.4	0.9	0.0	0.4	0.0	0.9	1.6	0.0	0.4	0.8	-6.5	2.1	6.0	4.6	3.9	-1.9	4.4	5.6	7.3	3.9	0.0	1.2	2.8
	Tunisia	0.0	0.8	0.0	0.4	0.4	0.0	0.4	1.2	1.3	1.6	0.4	0.8	1.6	-4.0	3.7	5.6	4.6	1.7	-1.4	2.4	5.2	3.4	0.4	5.0	1.9	0.8
	UAE	0.8	0.4	0.4	2.5	3.0	0.5	0.8	2.8	0.9	4.7	1.3	2.3	2.0	-1.2	4.5	5.6	5.0	3.0	0.5	1.2	4.8	7.3	1.6	0.0	6.9	6.4
	Yemen	-0.4	0.8	0.0	0.8	0.0	0.0	0.4	0.0	0.4	0.8	0.0	-0.4	1.2	-0.8	2.9	4.5	5.5	0.4	0.9	1.6	6.4	3.9	-1.2	3.8	1.5	5.2
Ditto	Algeria	25.4	16.1	33.0	26.9	10.4	32.4	32.8	28.8	25.8	29.3	18.5	26.2	21.6	-4.0	-1.2	-1.9	1.7	2.2	-1.9	-4.4	1.6	4.7	-0.8	-0.4	-1.2	3.2
	Egypt	24.6	18.6	43.8	26.5	15.5	35.2	41.5	31.2	39.5	30.1	15.1	30.0	15.2	-4.8	1.2	-6.0	0.0	3.5	-3.7	-3.2	-5.2	1.7	-4.7	1.7	-1.5	1.2
	Jordan	31.9	23.1	49.4	32.8	20.3	40.7	44.7	38.4	41.2	35.2	21.8	33.5	26.4	-5.2	2.1	5.6	2.1	1.7	-0.9	0.0	0.0	7.7	0.0	2.9	1.5	1.6
	KSA	18.2	20.2	44.6	23.9	13.4	32.4	37.6	25.6	40.8	30.9	18.9	31.9	16.8	-1.6	7.4	6.4	2.1	6.5	1.9	0.8	0.0	9.5	1.6	3.8	0.4	4.8
	Lebanon	7.3	1.2	2.2	3.8	1.7	15.3	15.0	11.2	9.0	15.2	0.0	6.9	4.8	-5.6	3.3	4.5	3.0	4.3	0.9	-1.2	-4.4	7.7	0.8	2.1	0.4	1.6
	Libya	32.3	21.1	49.8	32.4	16.0	40.7	41.1	40.0	43.3	36.3	20.2	30.4	30.8	-2.4	4.1	0.8	5.5	3.0	-1.9	-4.7	-0.8	6.9	-3.5	0.8	-1.9	0.0
	Morocco	8.5	9.9	30.3	8.4	4.3	14.8	23.7	14.4	26.2	12.1	6.3	11.5	6.0	-6.5	-2.1	4.1	3.0	1.3	-1.9	-2.0	-5.2	4.3	-3.1	0.0	0.4	2.4
	Palestine	22.6	18.6	30.3	23.5	13.4	34.7	33.2	26.8	29.2	30.1	17.2	23.5	20.8	-2.8	2.9	0.8	0.0	0.4	2.3	-4.0	-2.4	1.7	-0.8	3.4	-0.8	4.0
	Sudan	14.1	10.7	21.3	13.0	5.2	17.1	23.3	18.0	20.6	17.6	10.1	14.2	10.0	-4.4	0.4	3.0	2.5	2.2	-0.9	0.4	3.2	3.0	-1.2	2.9	2.7	4.8
	Syria	29.0	24.8	42.7	32.4	18.1	40.3	40.7	37.2	36.9	35.9	19.3	32.7	22.4	-5.6	3.7	4.1	0.9	3.9	0.9	-2.0	-4.8	6.9	-2.3	-0.4	-1.9	-0.4
	Tunisia	21.0	14.9	23.6	18.9	12.1	28.7	27.7	22.4	21.9	25.8	14.7	20.0	15.6	-1.2	2.5	5.6	3.8	2.6	-1.4	-2.0	-1.2	7.3	-3.1	0.4	-1.5	6.0
	UAE	24.2	17.8	43.8	24.8	17.2	37.0	37.9	30.4	34.3	28.9	17.2	30.8	20.0	-2.4	5.0	6.4	3.8	2.6	1.9	0.0	1.6	4.3	0.8	1.3	0.4	5.6
	Yemen	2.8	2.9	12.7	2.5	0.4	6.9	12.2	4.0	18.0	4.3	2.5	3.5	1.2	1.2	2.5	3.4	0.0	2.2	-2.3	-2.4	-5.2	3.9	-7.4	0.0	-1.9	2.0

Table 9: Cross-country evaluation results for Gemma-2 9B-it. Models are evaluated on different countries (columns) after being trained on specific countries (rows). Values represent score difference from the base model.

Method	Trained On	Δ MCQ vs. Base													Δ Completion vs. Base											
		Alg	Egy	Jor	KSA	Leb	Lib	Mor	Pal	Sud	Syr	Tun	UAE	Yem	Alg	Egy	Jor	KSA	Leb	Lib	Mor	Pal	Sud	Syr	Tun	UAE	Yem
ICL	Algeria	-9.7	-15.7	0.0	-11.8	-11.2	-14.4	-11.9	-14.4	-9.0	-10.2	-9.2	-9.6	-10.8	-0.4	2.1	11.2	4.2	3.0	3.7	5.1	3.6	3.4	4.7	8.8	0.8	1.2
	Egypt	-25.4	-26.0	-17.2	-26.9	-20.7	-23.6	-28.9	-34.8	-22.3	-31.3	-17.6	-29.2	-24.4	-1.6	0.4	10.5	5.0	1.3	2.3	3.6	4.0	4.7	2.0	7.6	2.7	4.0
	Jordan	-3.2	-8.3	-0.4	-5.5	-5.2	-6.9	-7.9	-4.0	-3.9	-5.9	-6.7	-2.7	-3.2	0.0	0.8	13.5	3.4	2.6	1.4	5.1	3.2	9.0	2.3	5.0	1.2	2.0
	KSA	-16.9	-19.4	-6.7	-18.1	-16.4	-18.1	-19.4	-23.2	-16.7	-17.6	-13.4	-13.5	-13.2	-4.0	2.1	9.4	5.5	2.2	1.4	2.4	3.6	4.7	5.1	3.8	1.5	2.8
	Lebanon	-14.5	-17.4	-2.3	-16.4	-15.5	-14.8	-15.8	-21.6	-12.0	-17.2	-15.5	-16.5	-15.2	-0.8	1.2	10.9	5.0	1.3	5.6	2.8	4.0	4.7	4.3	5.9	3.5	3.2
	Libya	-14.5	-16.5	-6.0	-13.5	-13.4	-13.4	-16.6	-15.6	-13.3	-13.7	-13.0	-11.5	-11.6	-2.8	0.8	9.0	3.8	2.6	4.2	2.0	2.8	3.4	3.9	3.8	-0.8	1.6
	Morocco	-9.3	-13.6	-6.4	-13.9	-13.4	-13.0	-15.0	-14.0	-13.3	-12.5	-8.0	-8.9	-12.0	0.0	1.7	11.2	3.4	2.2	3.7	8.7	3.2	5.2	2.3	4.6	2.3	2.0
	Palestine	-19.8	-21.9	-12.7	-18.1	-19.0	-22.7	-20.6	-24.8	-22.3	-21.5	-14.3	-21.2	-18.8	0.0	0.8	9.7	4.2	0.9	4.2	1.6	4.0	0.4	2.3	5.0	-0.4	3.2
	Sudan	-21.8	-25.6	-21.7	-23.5	-18.5	-21.3	-27.7	-27.2	-26.6	-23.4	-17.2	-21.5	-20.8	-2.4	0.4	9.4	2.1	1.3	1.9	4.8	3.6	8.6	2.0	5.9	1.9	2.0
	Syria	-2.4	-13.2	-1.1	-8.4	-6.9	-5.1	-5.5	-8.8	-6.4	-5.1	-5.5	-4.2	-4.8	-1.2	0.8	10.1	5.0	0.9	4.2	5.1	6.0	5.6	6.6	7.1	1.9	2.0
	Tunisia	-19.4	-22.7	-7.1	-23.1	-15.1	-17.1	-24.5	-26.4	-15.9	-22.7	-11.3	-20.0	-16.4	-0.4	0.0	8.6	2.9	3.0	5.6	5.9	5.2	4.7	6.3	6.7	3.5	4.0
	UAE	-12.1	-16.5	-6.0	-11.8	-10.8	-11.6	-14.2	-18.0	-10.7	-11.7	-10.9	-9.6	-10.4	-2.0	0.0	8.2	3.8	1.3	4.2	-0.8	5.2	6.9	5.5	5.0	3.5	2.8
	Yemen	-17.7	-24.0	-7.1	-17.7	-13.8	-16.7	-19.8	-22.8	-21.9	-23.0	-13.9	-16.2	-15.6	0.0	-0.8	6.0	2.5	0.9	5.6	0.4	2.8	3.9	4.3	4.2	1.5	6.0
Ditto	Algeria	-2.8	-5.0	3.0	4.6	-1.7	9.3	3.2	0.0	0.4	4.3	3.8	3.5	-1.6	-5.2	0.8	-5.2	-2.5	-0.9	0.0	0.4	1.6	2.2	3.1	3.8	-2.7	1.2
	Egypt	-4.8	-6.2	0.0	0.8	-2.2	-0.9	-1.6	2.0	0.9	-4.3	4.2	-0.4	-1.6	-6.1	1.7	-9.0	-6.3	0.4	0.0	-4.4	-2.4	-4.7	0.4	2.1	-5.4	-1.2
	Jordan	-10.5	-13.2	-20.6	-13.9	-9.9	-8.3	-13.8	-18.0	-16.3	-10.2	-7.1	-13.9	-9.6	-6.5	-2.1	-12.4	-7.6	-3.0	-0.9	-4.4	-6.8	-6.4	1.2	1.7	-6.9	-7.6
	KSA	-1.2	-4.5	-0.8	-0.4	-7.3	0.0	-0.4	-4.0	-3.0	-0.8	2.1	-1.9	-2.8	-4.0	4.1	-0.4	1.3	-0.4	3.2	0.4	1.6	0.0	1.6	3.4	-1.9	2.0
	Lebanon	2.8	0.0	3.0	6.7	1.7	8.8	5.1	3.6	1.3	3.5	5.1	3.8	3.2	-0.4	2.9	0.0	-8.0	0.4	2.8	-1.2	2.8	0.9	0.4	5.5	-3.5	2.4
	Libya	-0.4	-6.6	2.2	-0.4	-0.4	-0.5	-1.2	-0.4	0.0	0.8	0.8	0.0	2.0	-4.4	1.2	-7.5	-6.7	-3.0	0.9	-1.2	-0.8	-4.7	1.6	2.9	-3.5	-1.2
	Morocco	2.0	-2.5	5.6	2.9	-0.9	7.0	2.4	2.8	-0.4	6.6	5.5	4.2	5.2	-3.6	3.3	4.9	-2.5	2.2	-0.9	0.0	-0.8	-2.1	1.6	2.9	-5.4	-0.8
	Palestine	2.8	-5.0	1.9	0.4	0.9	5.1	-2.4	0.8	-2.6	-0.4	4.2	1.5	3.6	-6.5	-0.8	-10.5	-7.6	-0.4	0.0	-5.9	-4.8	-6.0	2.0	2.9	-6.9	2.4
	Sudan	-2.4	-2.9	2.6	2.5	-2.2	5.1	3.2	0.0	2.1	2.7	6.3	1.9	3.6	-3.2	2.1	-4.9	-3.4	-0.9	-2.8	-2.0	1.2	-2.6	0.8	2.5	-5.0	-3.2
	Syria	2.0	-2.5	0.8	2.1	0.9	7.9	-0.4	4.8	0.4	3.9	8.0	1.5	4.4	-4.0	2.1	-5.6	-5.0	-0.4	0.0	-2.0	-0.4	-3.0	4.3	2.9	-6.5	-0.4
	Tunisia	-0.4	-0.8	-0.8	1.7	1.3	7.4	3.2	0.0	-1.3	0.4	2.1	1.5	4.0	-2.4	3.3	2.6	-0.8	2.2	0.0	0.8	0.0	-3.0	0.0	2.9	-3.1	0.4
	UAE	0.0	-2.9	5.6	5.9	-1.7	8.3	4.0	4.8	1.7	5.9	5.5	2.7	6.8	-5.7	-0.4	-8.6	-3.4	-1.7	-0.9	-1.2	-1.6	-3.0	0.0	5.5	-3.5	-1.6
	Yemen	-2.4	-3.3	0.4	-1.7	-5.6	6.0	-1.2	0.0	-2.2	-0.4	3.4	-0.8	1.6	-3.2	4.6	0.0	-4.2	1.7	1.4	0.8	2.4	-3.9	2.7	5.9	-2.3	1.2

Table 10: Cross-country evaluation results for ALLaM 7B-Instruct-preview. Models are evaluated on different countries (columns) after being trained on specific countries (rows). Values represent score difference from the base model.

Method	Trained On	Δ MCQ vs. Base													Δ Completion vs. Base											
		Alg	Egy	Jor	KSA	Leb	Lib	Mor	Pal	Sud	Syr	Tun	UAE	Yem	Alg	Egy	Jor	KSA	Leb	Lib	Mor	Pal	Sud	Syr	Tun	UAE	Yem
ICL	Algeria	-1.2	-2.5	-1.5	1.7	3.5	-1.9	3.6	2.8	-0.9	2.0	-0.4	7.3	-2.8	0.4	0.0	6.0	1.7	-1.7	1.9	2.4	1.2	7.3	2.0	0.8	-0.4	2.0
	Egypt	-6.5	-4.1	-3.4	0.0	6.0	-3.7	2.4	2.4	0.4	-2.0	0.4	2.7	-3.2	-2.4	-1.2	1.9	1.7	-2.6	1.9	0.8	0.0	1.7	-0.4	0.4	-3.8	2.8
	Jordan	2.0	1.7	1.1	4.6	3.0	1.4	5.9	6.4	3.0	4.7	0.0	5.0	1.2	0.4	1.2	8.2	5.0	-0.4	2.8	2.4	3.2	3.9	1.2	-1.3	-3.1	3.2
	KSA	1.2	2.1	-2.6	4.6	0.9	-0.5	2.0	4.0	-2.1	0.8	0.8	4.6	2.0	0.8	0.0	5.6	4.6	-1.7	2.8	2.0	0.8	2.1	-1.6	-0.8	-2.3	4.0
	Lebanon	-6.1	-4.1	-2.6	-1.3	3.0	-8.3	0.4	-0.8	0.9	0.8	0.4	2.7	-5.6	-3.2	-1.2	1.9	3.8	-2.6	1.4	2.0	0.8	2.1	2.0	-0.4	-0.8	2.4
	Libya	-13.7	-12.0	-3.7	-5.5	-1.3	-14.8	-7.9	-4.4	-2.1	-9.8	-3.0	-0.8	-12.4	0.0	1.2	2.6	3.4	-0.9	1.9	3.2	2.8	2.1	-0.8	-0.4	-0.8	3.2
	Morocco	4.0	2.5	0.0	2.9	5.2	0.5	6.7	6.0	1.3	7.0	4.2	4.2	0.4	-1.6	2.9	6.0	4.6	0.0	2.8	2.8	1.6	3.9	0.8	-0.8	-0.4	2.0
	Palestine	-3.6	-3.7	-1.1	2.5	2.6	-4.2	1.2	2.4	1.3	2.0	-0.4	4.2	-3.6	-2.4	0.8	4.9	5.0	-0.9	2.3	2.0	1.6	3.9	1.2	0.8	-1.2	2.0
	Sudan	-1.6	0.0	-1.1	4.6	2.2	-0.9	3.2	1.6	1.7	0.0	2.1	3.8	-1.2	-1.6	1.2	6.4	4.6	0.4	2.3	2.0	0.8	9.0	0.4	-1.3	-2.3	2.4
	Syria	2.8	1.7	0.8	4.2	6.5	2.3	3.6	6.4	3.0	6.6	0.4	5.0	0.8	-1.6	-1.2	3.0	1.7	-1.7	1.9	0.8	0.8	4.3	1.2	-1.7	-3.1	3.2
	Tunisia	-0.8	1.2	-3.4	1.7	5.2	-5.1	1.6	2.8	-1.3	2.4	-1.3	2.7	-1.6	-0.8	2.1	0.8	1.3	-1.7	1.4	0.0	0.4	0.9	-0.4	1.3	-2.3	2.4
	UAE	-1.2	1.2	-1.5	5.9	3.5	-1.4	4.0	4.4	1.3	4.3	2.1	5.4	1.2	-0.4	2.1	3.4	2.9	0.0	2.8	2.4	2.8	5.2	2.3	-0.4	-0.4	4.0
	Yemen	-4.8	-5.4	-2.6	0.4	3.9	-5.1	0.8	0.0	-0.4	-0.4	2.1	1.9	-0.4	-0.4	2.1	2.6	2.1	0.9	1.4	2.0	-2.4	0.9	-0.8	1.7	-4.6	3.2
Ditto	Algeria	0.0	-1.7	0.0	4.2	5.6	-1.9	1.2	2.0	1.7	5.5	0.8	6.5	3.2	-2.8	-5.4	-3.0	-2.5	3.0	0.0	-0.8	-2.0	1.7	3.5	-3.8	-2.7	3.6
	Egypt	-3.2	-4.1	-3.0	0.4	1.3	-0.5	-2.4	-0.4	0.0	-2.7	0.8	0.8	-2.4	-2.4	0.0	-4.1	-3.0	0.9	-2.8	-0.4	-3.2	-0.4	1.2	0.0	-1.9	3.6
	Jordan	0.4	-2.1	-0.4	3.4	2.2	1.4	2.8	0.4	0.9	5.1	1.7	3.1	1.6	-3.2	-0.4	4.9	4.6	0.9	0.0	0.0	0.8	3.9	3.1	0.4	-3.1	2.8
	KSA	0.0	-1.2	0.4	0.8	2.6	2.3	-0.8	0.8	0.9	2.4	0.4	-0.4	0.8	-2.0	0.4	1.1	-1.7	1.3	-0.5	-0.8	0.0	1.7	1.6	0.4	-1.2	2.4
	Lebanon	-6.9	-9.1	-4.1	-3.8	0.0	-3.2	-3.2	-3.6	-1.3	0.0	-5.5	2.3	-1.2	-3.6	-2.9	0.4	3.4	0.4	1.4	4.4	1.6	2.1	3.1	-1.3	-0.4	2.8
	Libya	0.0	-6.2	-0.8	-0.8	3.5	-3.2	-1.2	-0.8	1.3	4.3	-2.1	-0.8	2.0	-2.0	-2.9	2.2	2.5	2.2	1.9	-2.8	1.6	4.3	4.3	-2.9	-1.5	4.0
	Morocco	1.2	-2.1	0.8	3.8	6.5	0.0	3.2	3.2	0.4	5.9	2.5	5.0	4.4	-0.4	-1.2	3.4	2.5	0.0	0.9	4.0	0.0	2.6	2.7	-0.8	-3.5	4.4
	Palestine	-1.2	-2.5	0.4	5.0	6.5	0.5	1.6	2.4	0.0	7.0	1.7	1.2	4.0	-2.0	-4.6	-2.6	-1.7	0.9	0.9	0.4	1.6	2.6	2.7	2.5	0.8	3.6
	Sudan	2.4	-1.7	1.1	5.5	6.0	1.9	0.4	4.8	0.9	6.3	1.7	2.3	4.0	-0.4	-3.7	6.0	3.8	-1.3	1.4	0.0	4.0	4.7	2.0	2.9	-0.8	3.2
	Syria	-1.6	-1.2	-0.8	2.9	4.7	0.5	0.8	2.4	1.7	6.6	-0.4	5.0	4.4	-3.2	-6.6	-3.0	-0.4	2.6	1.4	1.2	-1.6	2.1	2.3	-2.1	-1.9	6.0
	Tunisia	2.8	-2.5	0.8	2.1	2.6	2.3	0.0	3.6	-0.9	7.0	1.7	3.5	6.8	2.0	-1.7	1.9	0.8	0.9	0.9	-0.8	-0.8	1.3	0.8	0.0	-4.2	2.8
	UAE	1.2	-2.1	1.1	3.4	5.6	1.9	0.4	1.2	0.9	7.4	0.0	1.5	6.0	-1.6	-1.2	3.0	5.0	0.9	0.0	0.4	3.6	1.7	4.7	-0.4	-1.5	7.2
	Yemen	-1.2	-2.9	-1.5	-0.8	6.9	3.2	-3.2	-0.8	-2.6	2.4	4.6	3.5	-1.2	-2.0	-1.2	-3.0	0.4	2.6	-1.4	0.8	-1.6	1.3	0.0	2.9	-4.2	4.4

Table 11: Cross-country evaluation results for SILMA 9B-Instruct. Models are evaluated on different countries (columns) after being trained on specific countries (rows). Values represent score difference from the base model.