Cultural Benchmarking of LLMs in MSA and Arabic Dialectal Dialogue

Anonymous ACL submission

Abstract

There is a significant gap in evaluating cultural reasoning in LLMs using conversational datasets that capture culturally rich and dialectal contexts. In Arabic NLP, most prior work focuses on Modern Standard Arabic (MSA) and short text snippets, overlooking the cultural nuances that naturally arise in dialogue. To address this gap, we introduce a culturally grounded conversational dataset covering 13 Arabic-speaking countries, including MSA and corresponding dialects, spanning 12 daily-life domains and 54 fine-grained subtopics. We define three tasks: (i) multiple-choice cultural reasoning, (ii) machine translation between MSA and dialects, and (iii) dialect-steering generation. Experiments with open-weight LLMs reveal substantial challenges: models struggle with dialectal data and perform significantly worse on all three tasks compared to MSA, highlighting the need for culturally aware dialogue systems. The data and code will be made available upon acceptance.

1 Introduction

Arabic is spoken by over 400 million people worldwide, making it one of the most widely used languages in the world (UNESCO, 2025). While Modern Standard Arabic (MSA) serves as the formal written standard, most everyday communication occurs in diverse regional dialects that vary widely across and within countries (Habash, 2010). These dialects differ from MSA not only phonologically, but also lexically, grammatically, and pragmatically, encoding culturally grounded norms and practices. For most speakers, dialect is the primary medium for expressing and transmitting cultural knowledge in daily conversation (Kwaik et al., 2018).

On the other hand, recent years have seen substantial progress in Arabic NLP, with the emergence of Arabic-centric LLMs such as Jais (Sengupta et al., 2023; Anwar et al., 2025), SILMA (silma-ai, 2024), and ALLaM (Bari et al., 2025), alongside multilingual models that increasingly support Arabic. Evaluation benchmarks also have expanded accordingly. Among these, cultural commonsense reasoning has emerged as a particularly important dimension, as it probes whether models can reason about the shared knowledge, customs, and social expectations that underlie human communication. ArabCulture (Sadallah et al.,

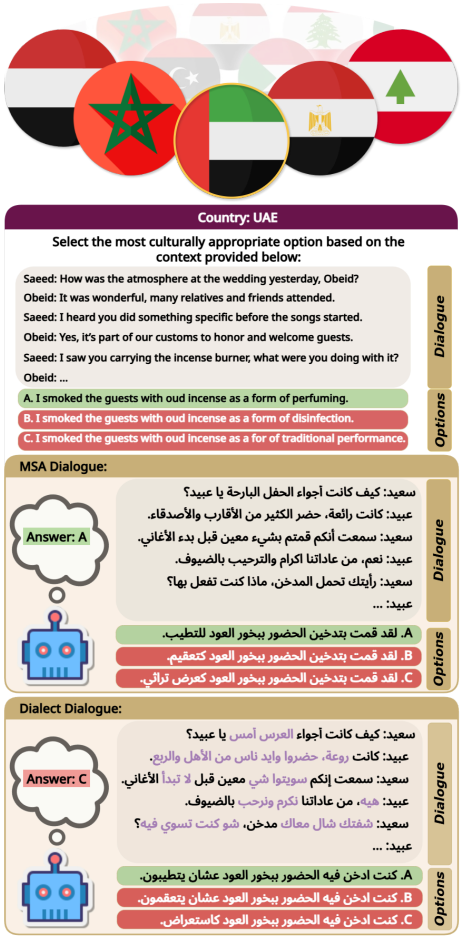

Figure 1: Example from **ArabCulture-Dialogue** showing a UAE wedding scenario in both MSA and UAE (Emirati) dialect.

2025) is a notable example, providing a native-speaker–curated benchmark of 3,482 questions across 13 countries and 54 cultural topics.

However, existing cultural reasoning benchmarks, including ArabCulture, rely exclusively on isolated, single-turn multiple-choice questions presented in MSA. This evaluation paradigm, while useful for controlled assessment, diverges fundamentally from how cultural knowledge is actually exchanged and applied. In natural settings, cultural reasoning unfolds across conversational turns, where speakers must interpret implicit norms, respond appropriately to culturally situated utterances, and maintain pragmatic coherence throughout an interaction. Moreover, such exchanges are expected to be in dialects, suggesting that current benchmarks may systematically overestimate model capabilities by evaluating in a register that is both simpler and less culturally laden than authentic usage. This raises a critical question: *can models that perform adequately on MSA-based cultural questions actually apply this knowledge in natural and dialect-mediated dialogue?*

To address this gap, we introduce **ArabCulture-Dialogue**, a human-curated conversational dataset that extends ArabCulture into multi-turn dialogues in both MSA and country-specific dialects. As illustrated in Figure 1, each instance consists of a culturally grounded conversation followed by three candidate responses, only one of which is culturally appropriate in both MSA and the local dialect. To our knowledge, this is the first dataset to benchmark Arabic cultural commonsense reasoning in a dialogue-based setting across MSA and regional dialects. We also define three evaluation tasks on ArabCulture-Dialogue: (i) dialogue-based multiple-choice cultural reasoning, which requires selecting the culturally appropriate response from three answer options; (ii) dialect translation between MSA and country-specific varieties; and (iii) dialect steering, which tests controlled generation in a specified dialect. Together, these tasks evaluate cultural reasoning in context, cross-register linguistic competence, and dialect-aware generation.

We evaluate a range of Arabic-centric, multilingual, and proprietary LLMs. Results show consistent degradation in performance on dialectal dialogues compared to MSA, with smaller open-source models performing especially poorly. Cultural reasoning in MSA often fails to transfer to dialectal settings, and fine-grained country-level knowledge remains difficult. These findings highlight substantial limitations in current LLMs for culturally grounded, dialect-rich Arabic dialogue.

Our contributions are threefold:

1. We introduce **ArabCulture-Dialogue**, the first parallel MSA–dialect cultural dialogue dataset covering 13 Arab countries, created through rigorous human curation by 26 native speakers;
2. We define three evaluation tasks: cultural MCQ, dialect translation, and dialect steering, to comprehensively assess culturally grounded dialogue capabilities; and
3. We conduct extensive experiments showing that dialectal cultural reasoning remains challenging for current open models, highlighting the need for culturally and dialectally aware systems.

2 Related Work

Dialect and Cultural Reasoning in NLP Dialectal variation often encodes culturally grounded meaning beyond surface-level linguistic differences. Studies on English, Hindi, and Chinese dialects show that dialect choice signals social identity, politeness, norms, power relations, and pragmatic conventions (Hovy, 2015; Blodgett et al., 2016; Jurgens et al., 2017; Hershcovich et al., 2022). Despite this, many NLP approaches historically treat dialects as noise to be normalized toward a standard variety, causing large language models to degrade in performance and exhibit bias on dialectal inputs (Hofmann et al., 2024; Cao et al., 2023). These findings highlight the need for culturally grounded evaluation. Yet, existing benchmarks rarely capture dialectal cultural reasoning in interactive settings. Our work addresses this gap by evaluating cultural reasoning in dialogue, where dialect-mediated norms emerge across turns rather than isolated prompts.

Arabic and Dialectal NLP Arabic presents an informative case due to its entrenched diglossic structure: Modern Standard Arabic (MSA) dominates formal writing, education, and most NLP benchmarks, while everyday communication across the Arab world occurs primarily in regional dialects (Dialectal Arabic, DA). These dialects carry culturally situated meaning, encoding region-specific idioms, politeness strategies, humor, and social norms (Holes, 2006) that are often limited or absent in MSA, making dialect choice closely tied

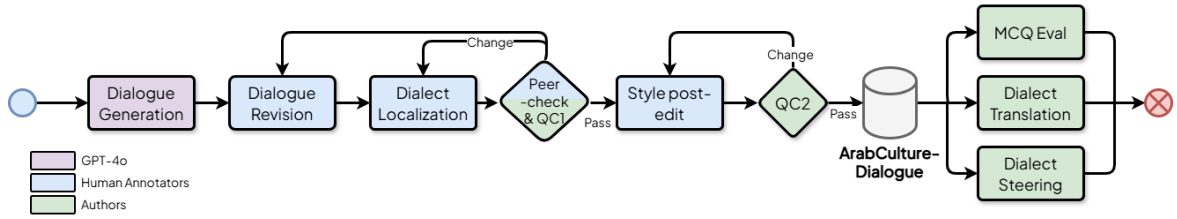

Figure 2: Dataset construction pipeline of **ArabCulture-Dialogue**. After the initial dialogue generation by GPT-5, all subsequent stages, including revision, dialect localization, style post-editing, and quality control, are performed through human annotation, resulting in a fully human-curated dataset.

to cultural identity and pragmatic intent (Abdul-Mageed et al., 2021; Bouamor et al., 2018).

Despite this centrality in daily communication, most Arabic NLP resources have historically prioritized MSA due to its standardized orthography and data availability, treating dialects mainly as a technical challenge through identification, normalization, or conversion to MSA (Abdul-Mageed et al., 2021; Abdelali et al., 2021; Zaidan and Callison-Burch, 2014). Shared tasks such as NADI (Abdul-Mageed et al., 2024) exemplify this focus by emphasizing dialect labeling rather than cultural interpretation. In contrast, our work evaluates cultural reasoning without collapsing dialectal input into MSA, allowing models to be assessed on their ability to interpret culturally meaningful dialectal cues in context.

Task-Specific Cultural Evaluation in Arabic Recent Arabic-specific benchmarks expose the limitations of MSA-centric and single-turn evaluation. `ArabCulture` (Sadallah et al., 2025) and `PALM` (Alwajih et al., 2025a) introduce culturally grounded QA and instruction datasets spanning 13 and 22 Arab countries respectively, with prompts in MSA and local dialects, and reveal substantial regional performance disparities even for strong models. While these datasets establish the importance of culturally grounded evaluation in Arabic, they focus on single-turn questions or instructions, whereas our work extends this line of research to *multi-turn conversational interactions* where cultural knowledge must be applied and maintained.

Conversational and Multimodal Cultural Resources Recent studies show that Arabic cultural reasoning becomes more challenging under realistic evaluation conditions. The PALM-X shared task (Alwajih et al., 2025b) reports limited gains from task-specific fine-tuning, while Beyond MCQ (Bhatti and Alam, 2025) shows that open-ended and dialectal formats lead to notable performance degradation; SaudiCulture (Ayash et al., 2025) further demonstrates difficulties with fine-grained regional customs, even within a single country. Together, these findings indicate that dialectal variation, open-ended generation, and cultural specificity expose limitations that are less visible in simplified evaluations and motivate conversational and multimodal resources for cultural reasoning. JAWAHER (Magdy et al., 2025) focuses on culturally grounded proverbs, NileCHAT (El Mekki et al., 2025) provides dialect-heavy conversational data, and multimodal benchmarks such as Peacock (Alwajih et al., 2024) and JEEM (Kadaoui et al., 2025) show that cultural understanding often requires grounding across linguistic and visual modalities. While these efforts broaden cultural evaluation, they do not model how cultural norms are negotiated across conversational turns, a gap our work directly addresses through multi-turn Arabic dialogue.

In summary, while prior work shows that dialects are central to Arabic cultural expression and that models struggle with dialectal inputs, existing benchmarks remain fragmented and largely single-turn; we address this gap with a multi-country conversational benchmark for evaluating cultural competence in realistic, multi-turn discourse.

3 Dataset Construction

We construct a human-curated, culturally grounded dialogue dataset by transforming the ArabCulture benchmark (Sadallah et al., 2025) into multi-turn conversations in both MSA and 13 Arabic dialects using the pipeline depicted in Figure 2. ArabCulture provides culturally relevant scenarios paired with one correct and two incorrect continuations. We preserve the exact country distribution and subtopic coverage from ArabCulture, and use each instance as the starting point for creating richer conversational data.

3.1 From Cultural Premises to MSA Dialogues

For each ArabCulture sample, we first generate a short MSA dialogue based on the original premise and answer descriptions. GPT-4o is used to produce an initial draft of the dialogue, which is then manually revised by two native Arabic annotators from the corresponding country. The samples are split equally between the annotators within each country. Annotators are required to be native to the country, familiar with its cultural norms, and fluent in both MSA and the local dialect. The use of large language models by annotators is strictly prohibited throughout the data construction pipeline.

During revision, annotators verify linguistic correctness, naturalness, and cultural appropriateness. They also ensure internal consistency while addressing two common issues identified during early inspection: (1) **information leakage**, where the dialogue reveals the correct answer too explicitly, and (2) **stylistic cues**, where the correct answer differs noticeably in tone or structure from the incorrect options. These refinements ensure that selecting the correct option requires genuine cultural reasoning rather than reliance on superficial patterns.

3.2 Dialect Localization and First Quality Check

Each revised MSA dialogue is translated into the dialect of the corresponding country by the annotators who revised the dialogue. Annotators are instructed to avoid literal translation and instead produce natural, utterance-level conversational speech. Once dialect translation is completed by the two annotators, a different annotator performs an independent cross-review quality check (QC) of the translated dialogue, checking for dialect consistency, cultural correctness (including eliminating offensive content, if any), and adherence to the original MSA version. This multi-annotator workflow: MSA revision, dialect translation, and dialect cross-review, follows the formal guideline that is created for the annotators to help maintain consistent quality across all countries.

After the cross-annotator quality check is completed, we conduct an additional individual QC step. In this step, we randomly sample 50 instances per country and assess whether the dialogues meet the predefined quality criteria described above. If any instance fails to meet these standards, annotators are instructed to revise the dialogue or answer options accordingly. This QC process is carried out independently for each country, allowing the reviews to proceed in parallel and thereby improving efficiency.

3.3 Post-Editing for Style Consistency and Second Quality Check

During the first quality checks, we observe that some answer options were not stylistically aligned. For instance, the correct option might begin with a common discourse marker or be noticeably longer than the incorrect ones. These stylistic discrepancies can introduce unintended cues that make the correct answer easier to identify. Consequently, we introduce a post-editing stage, where annotators adjust all three answer options, in both MSA and dialect, to achieve comparable length, tone, and stylistic structure, while ensuring that only one option remains culturally correct. This step reduces unintentional stylistic cues and ensures that successful prediction relies on cultural reasoning rather than surface-level patterns.

After the answer-option refinement stage, we conduct a second round of quality check independently, without involving the original annotators, to ensure that the final dataset aligns with our intended goals. In this QC stage, we manually inspect 60 dialogues (30 in MSA and 30 in the respective dialect) per country and verify the following criteria: (1) Minimal stylistic differences exist among the three answer options, (2) The key information conveyed in the correct answer (as preserved from the original ArabCulture data) is retained, (3) Each answer option constitutes a natural and contextually appropriate response to the preceding dialogue (e.g., responses appropriately address preceding questions), and (4) The edited MSA answer options and their dialect counterparts are parallel in content and intent. Based on our evaluation, almost all samples from each country pass these criteria, with only one or two samples exhibiting minor, non-critical issues.

Finally, we merge and finalize the validated instances to construct the first parallel MSA-dialect cultural dialogue dataset, which we refer to as '**ArabCulture-Dialogue**'. The entire data construction pipeline involves contributions from 26 annotators in total, resulting in a fully human-curated dataset designed to preserve both conversational quality and cultural authenticity.[1]

[1]All workers were compensated fairly, and the dataset

stats	MSA	Dialect	Merge
#dialogues	3,471	3,471	6,942
#country-specific dialogue	1,390	1,390	2,780
avg. words per dialogue	50.57	48.48	49.53
avg. utterance per dialogue	6.06	6.06	6.06
avg. words per utterance	7.36	7.01	7.18
#words	175,515	168,289	343,804
#unique words	22,116	30,894	41,109
#unique words (MSA ∧ Dialect)	11,901 (38.52% of Dialect unique words)		

Table 1: Dataset statistics. 'Merge' denotes the full **ArabCulture-Dialogue** dataset, while 'MSA' and 'Dialect' represent its two partitions.

3.4 ArabCulture-Dialogue

Each final instance in our dataset consists of an MSA dialogue, its localized dialect version, and three answer options written in both MSA and the corresponding dialect. Table 1 presents an overview of the **ArabCulture-Dialogue** dataset, while Appendix A provides detailed statistics broken down by country, region, and topic.

As shown in Table 1, the MSA portion of **ArabCulture-Dialogue** contains more total tokens than the dialect portion, whereas the dialect data exhibits a larger vocabulary size. This trend is expected, as Arabic dialects tend to employ more diverse lexical forms and expressive variations. To the best of our knowledge, **ArabCulture-Dialogue** is the first dataset to benchmark Arabic cultural commonsense grounding across both MSA and 13 Arabic dialects within a dialogue-based setting, where cultural interactions are naturally expressed.

Through a carefully designed data construction pipeline, i.e., comprising generation, human revision, dialect translation, post-editing, and several quality checks, we produce a parallel MSA–dialect dialogue dataset that preserves the cultural grounding of ArabCulture while introducing a richer conversational context. This dataset provides a strong foundation for evaluating cultural reasoning, translation, and dialect-aware language generation in large language models.

4 Experimental Setup

Using **ArabCulture-Dialogue**, we evaluate dialogue-based cultural commonsense reasoning in Arabic across (1) Arabic-centric large language models, (2) multilingual large language models, and (3) proprietary large language models. All model inferences are conducted using a single run.

The Arabic-centric models include Jais-Adapted-7B-Chat (Sengupta et al., 2023), Jais-2-8B-Chat (Anwar et al., 2025), ALLaM-7B-Instruct (Bari et al., 2025), SILMA-9B-Instruct (silma-ai, 2024), c4ai-command-r7b-arabic (Alnumay et al., 2025), Fanar-1-9B (Team et al., 2025), and Hala-9B (Hammoud et al., 2025). The multilingual category includes Gemma-2-9B-Instruct (Team, 2024), Qwen3-8B (Team, 2025), and LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024). For proprietary models, we evaluate GPT-5 (with the reasoning level set to normal) and Gemini-2.5-Pro. To ensure a fair comparison, all Arabic-centric and multilingual models are constrained to a similar parameter scale, ranging from 7 billion to 9 billion parameters. In contrast, proprietary models are included to reflect the current state of the art.

Building on the manually curated parallel dialogue dataset described in the previous section, we evaluate these models across three complementary tasks: (1) dialogue-based cultural commonsense reasoning in multiple-choice question (MCQ) evaluation, (2) dialect translation, and (3) dialect steering. All prompts are written in English and provided in the Appendix B.

MCQ Evaluation For the MCQ evaluation task, models are presented with a dialogue and three answer options, only one of which is correct. We use the dataset in its original format, as it is already structured for this evaluation. In addition to this standard setting, we assess evaluation robustness by optionally providing explicit geographic context (region, or both region and country). Since cultural knowledge encoded in LLMs can vary across locations, this additional information may help models better reason about culturally grounded dialogues (Sadallah et al., 2025; Koto et al., 2024). We report simple accuracy as the evaluation metric.

Dialect Translation This task evaluates a model's ability to translate dialogue utterances between Modern Standard Arabic (MSA) and country-specific Arabic dialects across 13 countries. Since the dataset contains parallel MSA–dialect dialogues for each country, each utterance naturally forms a translation pair. We assess translation quality in both directions using BLEU (Papineni et al., 2002), BERTScore with mBERT as the scoring model (Devlin et al., 2019), and an LLM-as-a-judge framework based

creation cost was approximately USD 10K.

Model	Context: None						Context: Region + Country					
	MSA			Dialect			MSA			Dialect		
	Acc	CS	~CS	Acc	CS	~CS	Acc	CS	~CS	Acc	CS	~CS
Random	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333
Proprietary Models												
Gemini-2.5-pro	0.942	**0.920**	0.954	0.939	0.917	0.956	0.950	**0.933**	0.961	0.939	**0.924**	0.950
GPT-5	**0.943**	0.908	**0.966**	**0.948**	**0.924**	**0.964**	**0.953**	0.927	**0.971**	**0.948**	0.923	**0.965**
Arabic Centric Models												
Jais-7B-chat	0.554	0.504	0.589	0.498	0.460	0.524	0.531	0.492	0.558	0.469	0.440	0.489
ALLaM-7B-Instruct	0.418	0.387	0.437	0.398	0.374	0.413	0.506	0.465	0.533	0.471	0.440	0.491
Cohere-Arabic-7B	0.355	0.366	0.348	0.342	0.351	0.337	0.375	0.385	0.368	0.357	0.361	0.355
Jais-2 8B-Chat	0.740	0.689	0.774	0.710	0.650	0.751	0.766	0.704	0.808	0.731	0.667	0.773
Fanar-1-9B	0.391	0.373	0.403	0.350	0.354	0.348	0.618	0.553	0.662	0.534	0.486	0.566
SILMA-9B-Instruct	0.783	0.715	0.829	0.716	0.663	0.751	0.784	0.724	0.825	0.714	0.665	0.747
Hala-9B	0.779	0.711	0.826	0.733	0.660	0.782	0.820	0.751	0.866	0.763	0.692	0.810
Multilingual Models												
Llama-3.1-8B-it	0.456	0.424	0.478	0.412	0.406	0.416	0.473	0.441	0.495	0.430	0.422	0.435
Qwen-3-8B-it	0.354	0.350	0.357	0.354	0.346	0.359	0.380	0.368	0.387	0.368	0.355	0.377
Gemma-2-9B-it	0.671	0.619	0.707	0.609	0.558	0.644	0.710	0.661	0.744	0.643	0.582	0.684

Table 2: MCQ evaluation under two geographic settings: None (no context) and Country + Region. Results are averaged and reported separately for country-specific (CS) and non-country-specific (∼CS) dialogues. The best overall model is shown in **bold**, and the best within each category is underlined.

on GPT-5 with the reasoning level set to low. We evaluate models in both zero-shot and supervised fine-tuning settings; however, due to resource constraints, only multilingual models are fine-tuned.

Dialect Steering The dialect steering task evaluates a model's ability to control the dialectal variety of its generated responses. Given a dialogue context and an utterance in MSA, the model is instructed to produce a single response either in MSA or in a specified target dialect. This setting tests whether the model can both recognize and generate the intended dialect, which varies across countries. The model receives the dialogue context and completes it with one utterance in the target variety. We evaluate performance under both zero-shot and supervised fine-tuning settings, using an LLM-as-a-judge framework based on GPT-5. We also apply the GlotLID language identification model (Kargaran et al., 2023) to verify whether the generated output matches the target dialect automatically.

5 Results and Analysis

In this section, we report the results for the three tasks. Two observations apply to all of them:

1. Arabic-centric models outperform multilingual models of similar sizes.
2. Much larger proprietary models perform well in the dialog completion setup, with potential gaps in producing dialectal outputs.

5.1 MCQ Evaluation

For the Arabic-centric models, Hala-9B and SILMA-9B seem to be better than the other models, with Jais2-8B being another competitive model, as shown in Table 2. In contrast, the three Arabic-centric 7B models lag, even for relatively recent models such as ALLaM-7B and Cohere-Arabic-7B. This might hint that models of this size are incapable of performing dialogue-based cultural commonsense reasoning; however, other confounding factors might be causing this.

As expected, most models have higher accuracy picking the right answer when fed with MSA dialogues than with their respective DA ones. However, the gap is not drastic. For the dialogue's topics, all models almost categorically perform better on non-country-specific (~CS) dialogues than on country-specific (CS) ones. This is consistent with the results of the original ArabCulture benchmark from which our dialogue dataset was created (Sadallah et al., 2025). Lastly, providing information about the region and the country to which the dialogue is relevant in general increases the models' ability to pick the right answer.[2]

5.2 Dialect Translation

For both MSA-to-DA and DA-to-MSA translation directions, each dialogue's relevant country and

[2]Only providing the region as a context is also better than not, as shown in Tables C7 and C8 of Appendix C.

Model	BLEU				BERTScore			LLM-as-Judge (0–5)				
	B1	B2	B3	B4	P	R	F1	Adeq.	Flu.	Reg.	Term.	Overall
Proprietary Models (0-shot)												
Gemini-2.5-pro	0.582	0.451	0.354	0.273	0.877	0.877	0.877	4.518	4.440	4.513	4.623	4.188
GPT-5	**0.588**	**0.456**	**0.359**	**0.276**	**0.876**	**0.881**	**0.879**	**4.915**	**4.704**	**4.651**	**4.925**	**4.530**
Arabic-Centric Models (0-shot)												
Jais-7B-chat	0.395	0.269	0.188	0.129	0.808	0.808	0.808	4.021	3.528	1.482	4.184	2.439
ALLaM-7B-Instruct	0.499	0.362	0.269	0.196	0.848	0.846	0.847	4.199	3.731	3.019	4.241	3.408
Cohere-Arabic-7B	0.448	0.312	0.220	0.152	0.834	0.836	0.835	3.850	3.339	2.422	3.953	2.995
Fanar-1-9B	0.460	0.322	0.227	0.156	0.840	0.840	0.840	4.007	3.214	2.206	3.986	2.943
SILMA-9B-Instruct	0.159	0.092	0.056	0.035	0.750	0.709	0.728	1.443	1.492	1.278	1.576	1.302
Hala-9B	0.381	0.260	0.180	0.122	0.781	0.800	0.790	2.870	2.829	1.060	3.403	1.764
Multilingual Models (0-shot)												
Llama-3.1-8B-it	0.306	0.169	0.099	0.058	0.784	0.790	0.787	1.656	1.305	1.156	1.651	1.342
Qwen-3-8B-it	0.407	0.266	0.177	0.115	0.819	0.817	0.818	3.418	2.752	1.512	3.341	2.354
Gemma-2-9B-it	0.333	0.195	0.118	0.071	0.790	0.800	0.795	1.884	1.408	1.306	1.708	1.495
Multilingual Models (SFT)												
Llama-3.1-8B-it	0.142	0.098	0.066	0.046	0.776	0.721	0.747	1.500	2.076	1.832	2.268	1.747
Qwen-3-8B-it	0.380	0.261	0.180	0.122	0.805	0.786	0.795	2.054	2.269	1.562	2.998	2.040
Gemma-2-9B-it	0.388	0.274	0.194	0.135	0.818	0.795	0.806	2.071	2.515	2.185	2.940	2.210

Table 3: **MSA-to-Dialect** translation quality under **Context: Country + Region**. We report BLEU and BERTScore, along with LLM-as-Judge scores (0–5) for **Adeq.** (semantic adequacy), **Flu.** (fluency and grammaticality), **Reg.** (dialectal and regional correctness), **Term.** (terminology and lexical choice), and **Overall** (holistic quality). The best overall model is shown in **bold**, and the best model within each category is underlined.

region are provided in the prompt. We focus on analyzing the MSA-to-DA translation results in Table 3, with the Dialect-to-MSA results reported in Table F13 of Appendix F.1.

Table F13 shows that GPT-5 and Gemini-2.5-Pro achieve similar scores on BLEU and BERTScore, indicating comparable lexical and semantic accuracy. However, their performance diverges under LLM-as-a-Judge evaluation using a fixed prompt (see Appendix B5), where GPT-5 consistently ranks higher on adequacy and overall quality. Additionally, LLM-as-a-Judge also demonstrates strong alignment with our sampled human evaluations for Moroccan Arabic (see Appendix G21).

Among Arabic-centric models, ALLaM-7B-Instruct is the strongest performer, surpassing other Arabic-focused systems on both automatic metrics and LLM-as-Judge evaluations—particularly in adequacy and terminology. However, its scores remain well below proprietary models, with the largest gap observed in dialectal and regional correctness (Reg). This suggests that while Arabic-centric pretraining improves semantic preservation, achieving fine-grained control over regional dialectal realizations remains a significant challenge.

Supervised fine-tuning improves multilingual models in some dimensions but does not close the gap with zero-shot Arabic-centric models. Fine-tuned Gemma-2-9B-it shows moderate gains in BLEU and LLM-as-Judge fluency and terminology, but register scores remain low. This suggests that limited supervised data is insufficient to robustly encode dialectal distinctions, especially across multiple countries.

5.3 Dialect Steering

We evaluate *dialect steering* as a controlled generation task with two targets: Modern Standard Arabic (MSA) and country-dialect Arabic. For each prompt, we ask the model to continue a short dialogue either in MSA or in the target dialect, and score outputs with (i) an LLM-as-a-judge quality metric (1–5, reported as $(s-1)/4 \in [0,1]$) and (ii) dialect identity via GlotLID (Kargaran et al., 2023). GlotLID is reported in two ways: (1) *strict ISO-code accuracy* where exact ISO 639-3 match against the country target code, and (2) *macro-region accuracy* where a coarser mapping that collapses close dialects into Gulf/Levant/Nile River/North Africa, following Bhatti and Alam (2025).

Dialect steering overview. Table 4 summarizes the performance of the different models. GPT-5 achieves the best judged quality for both MSA and dialect continuations, while Gemini-2.5-pro is slightly weaker on judged quality but noticeably

Model	Target: MSA		Target: Dialect	
	Judge	$\text{Acc}_{\text{Dialect}}$	Judge	$\text{Acc}_{\text{Dialect}}$
Proprietary Models				
Gemini-2.5-pro	0.8950	0.7570	0.9240	**0.5051**
GPT-5	**0.9305**	0.7181	**0.9555**	0.4539
Arabic-centric Models				
Jais-7B-chat	0.7819	0.8270	0.5994	0.0221
ALLaM-7B	0.8942	0.8270	0.8168	0.3625
Cohere-Arabic-7B	0.8433	0.7943	0.7162	0.1855
Jais-2.8B-chat	0.8603	0.7890	0.7304	0.2077
Fanar-1.9B	0.7915	0.7755	0.6214	0.0383
SILMA-9B-it	0.7676	0.6662	0.5908	0.0782
Hala-9B	0.8291	**0.8742**	0.6122	0.0162
Multilingual Models				
LLaMA-3.1-8B-it	0.7393	0.7851	0.5156	0.0640
Qwen3-8B-it	0.5487	0.8273	0.4295	0.0406
Gemma-2-9B-it	0.7925	0.7801	0.6042	0.1638
Multilingual Models (SFT)				
LLaMA-3.1-8B-it	0.7498	0.7002	0.5144	0.1268
Qwen3-8B-it	0.7999	0.7435	0.5360	0.0766
Gemma-2-9B-it	0.8245	0.7686	0.6035	0.1700

Table 4: Dialect steering results averaged across all prompts. Zero-shot performance is reported for all models, and supervised fine-tuning (SFT) results are reported for multilingual ones. Judge scores $\in [0,1]$. $\text{Acc}_{\text{Dialect}}$ reports GlotLID's strict ISO-code accuracy. Underline indicates the best score for each model category, and **bold** indicates the overall best.

stronger on strict-code GlotLID. Within Arabic-centric models, ALLaM-7B is the most reliable overall, whereas several Arabic-specialized baselines produce fluent continuations that nevertheless collapse toward wider regional varieties under strict ISO coding. The results hint that most models can respond in MSA. Moreover, some can adequately respond in DA. However, the GlotLID results indicate that they are not always using the correct dialect. This is further shown in Table G18, where the models' responses do not always follow the intended country-level dialect, as indicated by the varying GlotLID accuracy scores. Refer to Appendix G for further discussion.

Supervised steering shifts quality and dialect identity. Fine-tuning improves the quality of the multilingual models' responses as indicated by the judge's scores, especially for MSA. However, the models' ability to reply in MSA decreases, as indicated by the lower dialect accuracy scores. When the target is generating outputs in a specific dialect, SFT improves the models' ability to choose the right dialect in aggregate, yet the gap is still significant. However, the gains vary from one country-level dialect to another (See Table G19). For instance, Moroccan responses seem to benefit more from SFT than other dialects. The following excerpts suggest that SFT could be pushing the model to generate distinctive cues of each dialect (bolded) rather than pan-Arabic colloquialism.

Gemma-2-9B-it (UAE, zero-shot):
كل عام وانت بالف خير يا هزيم، وربي يبارك فيك ويحفظك.
(Wishing you well every year, Hazim. May God bless you and protect you.)

Gemma-2-9B-it (UAE, SFT):
لدي، نقول كل عام وأنت بخير **عشان نشوف** الفرحة في عيون الناس.
(Alright, let's say 'happy every year' and see what happens.)

Gemma-2-9B-it (Morocco, zero-shot):
والليلة **هادفنا** نعمل السُمْن
(And tonight, our goal is to make ghee.)

Gemma-2-9B-it (Morocco, SFT):
بصح، باش نستاهل الشهر الفضيل، نعمل السلو والسحلب **باش نحتفظو** به حتى رمضان.
(Right, so that we're ready for the blessed month, we make sellou and sahlab and keep them until Ramadan.)

6 Conclusion and Future Work

We introduce **ArabCulture-Dialogue**, the first culturally grounded conversational dataset covering 13 Arabic-speaking countries, spanning both MSA and corresponding dialects across 12 everyday domains and 54 fine-grained subtopics, with a total of 343,804 words, and use it to evaluate three tasks: (i) multiple-choice cultural reasoning, (ii) translation between MSA and dialects, and (iii) dialect-steered generation.

Our results show that while proprietary models perform strongly on cultural reasoning MCQs, open-source models, particularly at the 7B scale, often struggle, in some cases approaching random guessing; similar weaknesses appear in dialect translation and dialect steering, where all model types exhibit limited dialectal competence. These findings expose substantial gaps in current open-source LLMs' ability to model culturally grounded, dialect-rich Arabic, especially in conversational settings. These limitations point to promising directions for future work, including dialect-aware pretraining and instruction tuning, expanding coverage to additional Arab countries and dialects, and developing models that better integrate cultural knowledge in conversation.

Limitations

While our dialogue data provides translations into 13 different country-level dialects covering the different regions of the Arab world, it still does not cover all Arab countries. Additionally, we acknowledge the interspeaker dialectal variation that exists within each Arabic-speaking country.

Despite the efforts to ensure a high-quality translation of the MSA dialogues, the translators were inevitably impacted by the MSA dialogues' style (e.g., syntax). Hence, signs of translationese can still be noticed in some translations.

Ethics and Broader Impact

This benchmark is designed to evaluate LLMs' cultural reasoning abilities across Arabic-speaking countries in both MSA and regional dialects. Beyond evaluation, the dataset can also be used for training models to improve their understanding of culturally grounded Arabic language use. However, several considerations must be acknowledged. Cultural practices often overlap across countries, and not all instances in the dataset represent strictly country-specific culture; such distinctions are explicitly annotated. Additionally, the benchmark does not aim to capture the full cultural diversity of the Arab world, as it covers 13 of the 22 Arab countries and therefore represents only a subset of Arab cultural practices. These limitations should be taken into account when interpreting results or deploying models trained or evaluated using this dataset.

Additionally, annotators provided agreement to participate in this initiative and were informed that the data would be used for benchmarking purposes. Since their work involved refining existing content rather than creating data from scratch, no personally identifiable information is included in the dataset.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Yazeed Alnumay, Alexandre Barbet, Anna Bialas, William Darling, Shaan Desai, Joan Devassy, Kyle Duffy, Stephanie Howe, Olivia Lasche, Justin Lee, Anirudh Shrinivason, and Jennifer Tracey. 2025. Command r7b arabic: A small, enterprise focused, multilingual, and culturally aware arabic llm. *Preprint*, arXiv:2503.14603.

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: Shared Tasks*, pages 774–789, Suzhou, China. Association for Computational Linguistics.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of Arabic multimodal large language models and benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12753–12776, Bangkok, Thailand. Association for Computational Linguistics.

Mohamed Anwar, Abdelhakim Freihat, George Ibrahim, Mostafa Awad, Abdelrahman Atef Mohamed Ali Sadallah, Gurpreet Gosal, Gokul Ramakrishnan, Biswajit Mishra, Sarath Chandran, Ahmed Frikha, Rituraj Joshi, Etienne Goffinet, Abhishek Maiti, Ali El Filali, Sarah Al Barri, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Awantika Shukla, and 41 others. 2025. Jais 2: A family of Arabic-centric open large language models. Technical report, IFM.

Lama Ayash, Hassan Alhuzali, Ashwag Alasmari, and Sultan Aloufi. 2025. Saudiculture: A benchmark for evaluating large language models cultural competence within saudi arabia. *Preprint*, arXiv:2503.17485.

M. Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Hunzalah Hassan Bhatti and Firoj Alam. 2025. Beyond mcq: An open-ended arabic cultural qa benchmark with dialect variants. *Preprint*, arXiv:2510.24328.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abdellah El Mekki, Houdaifa Atou, Omer Nacar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. NileChat: Towards linguistically diverse and culturally aware LLMs for local communities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10978–11002, Suzhou, China. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Hasan Abed Al Kader Hammoud, Mohammad Zbeeb, and Bernard Ghanem. 2025. Hala technical report: Building arabic-centric instruction & translation models at scale.

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.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts ai decisions about people's character, employability, and criminality. *Preprint*, arXiv:2403.00742.

Clive Holes. 2006. The arabic dialects of arabia. In *Proceedings of the Seminar for Arabian Studies*, pages 25–34. JSTOR.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Karima Kadaoui, Hanin Atwany, Hamdan Al-Ali, Abdelrahman Mohamed, Ali Mekky, Sergei Tilga, Natalia Fedorova, Ekaterina Artemova, Hanan Aldarmaki, and Yova Kementchedjhieva. 2025. Jeem: Vision-language understanding in four arabic dialects. *Preprint*, arXiv:2503.21910.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. Glotlid: Language identification for low-resource languages. *arXiv preprint arXiv:2310.16248*.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). 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.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. A lexical distance study of arabic dialects. *Procedia Computer Science*, 142:2–13. Arabic Computational Linguistics.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. JAWAHER: A multidialectal dataset of Arabic proverbs for LLM benchmarking. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12320–12341, Albuquerque, New Mexico. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7695–7710, Vienna, Austria. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

silma-ai. 2024. Silma 9b instruct v1.0. https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An arabic-centric multimodal generative ai platform.

Gemma Team. 2024. Gemma.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

UNESCO. 2025. World arabic language day.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

A Data Statistics Details

Table A1 presents dataset statistics grouped by region. The North Africa region includes Algeria, Libya, Morocco, and Tunisia; the Nile River region includes Egypt and Sudan; the Levant includes Jordan, Lebanon, Palestine, and Syria; and the Gulf includes Saudi Arabia (KSA), the UAE, and Yemen. Detailed statistics for each region are reported in Tables A3, A4, A5, and A6, respectively.

Consistent with the overall dataset statistics, we observe across all regions that the MSA data contains more total tokens than the dialect data, while the dialect data exhibits a larger number of unique words. This pattern reflects the greater lexical diversity and expressive richness of Arabic dialects compared to MSA. Table A2 further presents dataset statistics broken down by topic. **ArabCulture-Dialogue** covers 12 topics that reflect culturally grounded aspects of everyday Arabic life.

B Prompts for All Downstream Tasks

To support in-context learning across all tasks, all prompts are written in English. This design choice follows the findings of Sadallah et al., 2025, which show that current Arabic-specific and multilingual models achieve better performance when prompted in English rather than in MSA. For the MCQ evaluation task, we use the prompt shown in Figure B1. Figure B2 presents the prompts used for dialect translation in both MSA-to-dialect and dialect-to-MSA directions, while Figures B3 and B4 show the prompts used for the dialect steering task. Figure B5 shows the prompt used for LLM-based evaluation, where a language model is instructed to act as a strict Arabic translation judge and score outputs along multiple linguistic and cultural dimensions.

Metric	North Africa	Nile River	Levant	Gulf
General Dialogue Statistics				
# Dialogues	1,036	521	1,097	817
# Country Specific Dialogues	448	341	192	409
Modern Standard Arabic (MSA) Data				
Avg. Words per Dialogue	51.08	50.74	50.34	50.10
Avg. Utterances per Dialogue	6.04	6.03	6.05	6.09
Avg. Words per Utterance	7.46	7.41	7.33	7.24
# Words	52,917	26,436	55,228	40,934
# Unique Words	10,444	6,607	10,883	9,204
Dialect Data				
Avg. Words per Dialogue	50.16	50.03	46.72	47.75
Avg. Utterances per Dialogue	6.04	6.03	6.05	6.09
Avg. Words per Utterance	7.30	7.30	6.71	6.85
# Words	51,968	26,066	51,247	39,008
# Unique Words	12,579	70,610	12,631	9,973

Table A1: Dataset Statistics across Regions

You are a professional Arabic that able to reasoning in the Arabic culture.

Rules:
- Output only the OPTIONS [A/B/C]
- Do not add explanations, comments, or quotation marks. Only the option label [A/B/C]

You are tasked with selecting the most culturally appropriate option based on the context provided below.

Location: {country}, {region}
Conversation:
{dialogue}

Consider the cultural nuances of the specified location and choose the most suitable next utterance! Give the option label only [A/B/C]

Options: {choices}

Figure B1: The prompt used for MCQ evaluation across all types of LLMs

C MCQ Evaluation Details

Tables C7 and C8 report model performance on MSA and dialect data, respectively, for the dialogue-based multiple-choice cultural commonsense reasoning task in Arabic. Overall, performance on MSA dialogues is substantially higher than on dialect dialogues, and this trend is consistent across all Arabic-centric and multilingual models. In contrast, proprietary models exhibit comparable performance across both MSA and dialect settings. This robustness holds across all context configurations, including no geographic context, region-only context, and full context with both region and country information.

D MCQ Evaluation Analysis per Country

We further analyze the performance of the strongest multilingual model (Gemma-2-9B-Instruct) and the strongest Arabic-centric model (Hala-9B) across countries, regions, and topics. As shown in Tables D9 and D10, both models achieve relatively strong performance on dialogues from Jordan and Palestine, suggesting that the cultural cues in these countries may be easier to infer compared to others. In contrast, dialogues from Yemen and the UAE are consistently the most challenging.

At the regional level, North Africa emerges as the most difficult region, with performance dropping to 0.663 on country-specific dialogues in the dialect setting. This result highlights the greater complexity and diversity of dialectal and cultural expressions in North Africa, which often differ substantially from the cultural norms and linguistic patterns typically represented in MSA.

E MCQ Evaluation Analysis per Topic

Tables E11 and E12 present the performance of Gemma-2-9B and Hala-9B on country-specific and

Metric	Food	Daily	Holiday	Habits	Wedd.	Death	Art	Games	Idiom	Parent.	Fam.	Agri.
General Dialogue Statistics (By Topic)												
#dialogues	724	520	461	342	279	239	234	223	118	116	108	107
#country Specific	357	158	175	85	145	42	166	85	82	23	21	51
Modern Standard Arabic (MSA) Data												
avg. words per dial.	49.56	50.09	50.84	51.65	49.39	51.19	50.56	50.50	53.91	50.90	50.71	52.64
avg. utt per dial.	6.13	6.03	6.03	6.04	6.04	5.99	5.98	6.01	6.13	6.03	6.00	6.31
avg. words per utt.	7.11	7.32	7.44	7.56	7.18	7.54	7.45	7.41	7.79	7.45	7.45	7.40
#words	35,883	26,047	23,437	17,665	13,780	12,234	11,832	11,262	6,361	5,904	5,477	5,633
#unique words	6,138	5,899	4,845	5,149	3,377	2,888	3,168	3,061	2,238	2,374	2,244	1,933
Dialect Data												
avg. words per dial.	47.22	47.97	48.51	49.75	47.83	48.97	48.84	48.84	52.84	48.59	48.09	49.98
avg. utt. per dial.	6.13	6.03	6.03	6.04	6.04	5.99	5.98	6.01	6.13	6.03	6.00	6.31
avg. words per utt.	6.70	6.96	7.05	7.24	6.92	7.17	7.16	7.13	7.60	7.08	7.02	6.99
#words	34,186	24,945	22,361	17,014	13,345	11,703	11,429	10,892	6,235	5,637	5,194	5,348
#unique words	8,703	7,632	6,509	6,309	4,286	3,792	3,954	3,966	2,616	2,789	2,564	2,316

Table A2: Detailed Dataset Statistics by Topic. The 'Daily', 'Holiday', 'Wedd.', 'Games', 'Parent.', 'Fam.', 'Agri.' represents daily activities, holiday activities, weddings, traditional games, parenting, family relationships, and agriculture, respectively.

Metric	Algeria	Libya	Morocco	Tunisia
General Dialogue Statistics (North Africa)				
#dialogues	271	239	276	250
#country Specific	81	100	103	164
Modern Standard Arabic (MSA) Data				
avg. words per dial.	51.66	52.26	50.35	50.12
avg. utt per dial.	6.06	6.05	6.02	6.04
avg. words per utt.	7.53	7.64	7.36	7.31
#words	13,999	12,491	13,896	12,531
#unique words	4,309	4,133	4,181	3,929
Dialect Data				
avg. words per dial.	48.46	51.08	50.96	50.24
avg. utt. per dial.	6.06	6.05	6.02	6.04
avg. words per utt.	7.01	7.41	7.47	7.34
#words	13,134	12,208	14,065	12,561
#unique words	4,259	4,204	4,355	3,846

Table A3: Detailed Dataset Statistics: North Africa

Metric	Egypt	Sudan
General Dialogue Statistics (Nile River)		
#dialogues	265	256
#country Specific	197	144
Modern Standard Arabic (MSA) Data		
avg. words per dial.	50.46	51.03
avg. utt per dial.	6.04	6.03
avg. words per utt.	7.36	7.47
#words	13,372	13,064
#unique words	4,246	3,983
Dialect Data		
avg. words per dial.	49.29	50.79
avg. utt. per dial.	6.04	6.03
avg. words per utt.	7.17	7.43
#words	13,063	13,003
#unique words	4,100	4,071

Table A4: Detailed Dataset Statistics: Nile River

non-country-specific dialogues, grouped by topic. As observed, performance on country-specific dialogues is consistently lower than on non-country-specific dialogues, in both MSA and dialect settings. In addition, the easiest topics for the models are 'agriculture' and 'family relationships', while the most challenging topics are 'death' and 'food'.

F Dialect Translation

This appendix provides a detailed analysis of Dialect-to-MSA translation, along with country-wise and region-wise breakdowns for both translation directions.

F.1 Dialect-to-MSA Translation

Table F13 shows that Dialect-to-MSA translation is consistently easier than MSA-to-Dialect. All models achieve higher BLEU, BERTScore, and LLM-as-Judge scores in this direction. GPT-5 again achieves the best overall performance, with near-ceiling scores on adequacy, fluency, and register. Gemini-2.5-pro follows closely, indicating that normalization into MSA benefits from strong general language modeling even without explicit dialect specialization.

Arabic-centric models show a substantial improvement relative to their MSA-to-Dialect performance. ALLaM-7B-Instruct approaches proprietary models on several country and region settings, particularly on register and terminology. This sug-

Metric	Jordan	Lebanon	Palestine	Syria
General Dialogue Statistics (Levant)				
#dialogues	290	255	273	279
#country Specific	17	99	29	47
Modern Standard Arabic (MSA) Data				
avg. words per dial.	50.79	51.39	49.90	49.36
avg. utt per dial.	6.10	6.05	6.24	5.81
avg. words per utt.	7.34	7.50	7.04	7.46
#words	14,728	13,104	13,624	13,772
#unique words	4,125	4,415	4,259	4,518
Dialect Data				
avg. words per dial.	50.21	47.53	45.61	43.42
avg. utt. per dial.	6.10	6.05	6.24	5.81
avg. words per utt.	7.24	6.88	6.29	6.44
#words	14,560	12,121	12,451	12,115
#unique words	4,249	4,631	4,674	4,314

Table A5: Detailed Dataset Statistics: Levant

Metric	KSA	UAE	Yemen
General Dialogue Statistics (Gulf)			
#dialogues	261	283	273
#country Specific	98	105	206
Modern Standard Arabic (MSA) Data			
avg. words per dial.	50.26	48.95	51.15
avg. utt per dial.	6.10	6.08	6.09
avg. words per utt.	7.26	7.06	7.42
#words	13,118	13,853	13,963
#unique words	4,112	4,565	4,251
Dialect Data			
avg. words per dial.	47.02	45.84	50.42
avg. utt. per dial.	6.10	6.08	6.09
avg. words per utt.	6.72	6.53	7.30
#words	12,271	12,972	13,765
#unique words	4,111	4,428	4,325

Table A6: Detailed Dataset Statistics: Gulf

You are a professional Arabic translation system specialized in Modern Standard Arabic (MSA) and regional Arabic dialects.
Rules:

- Output only the translated text in Arabic.
- Do not add explanations, comments, labels, or quotation marks.
- Preserve meaning, tone, and level of formality.
- Do not change names, numbers, or entities.
- If the input already matches the target variety, return it unchanged.

(a) System prompt shared across all models and directions.

Translate the following text from Modern Standard Arabic (MSA) into the {region} dialect of {country}.
Text: {record}

(b) User prompt for MSA-to-Dialect translation.

Translate the following text from the {region} dialect of {country} into Modern Standard Arabic (MSA).
Text: {record}

(c) User prompt for Dialect-to-MSA translation.

Figure B2: Prompting strategy used in all zero-shot translation experiments. A fixed system prompt defines the translation role and constraints, while direction-specific user prompts specify the translation task and input text.

gests that mapping dialectal input to a standardized target reduces the burden of dialect-specific generation.

Multilingual zero-shot models also benefit from this direction. Gemma-2-9B-it becomes the strongest multilingual model, outperforming Qwen-3-8B-it and Llama-3.1-8B-it across all metrics. Nevertheless, a clear gap remains between multilingual and Arabic-centric models, especially on adequacy and terminology.

Supervised fine-tuning improves register and fluency for multilingual models but often reduces adequacy scores. This trade-off is visible across countries and regions and suggests mild overfitting to stylistic normalization rather than semantic fidelity.

F.2 Country-wise and Region-wise Trends

Country-wise results in Tables F14 and F15 show consistent performance differences across countries for both MSA-to-Dialect and Dialect-to-MSA translation, respectively. In the MSA-to-Dialect direction, several countries exhibit lower BLEU and LLM-as-Judge register scores for non-proprietary models, with particularly large drops for Morocco and Tunisia. This pattern indicates greater difficulty in generating accurate country-specific dialect forms, especially for models without explicit dialect specialization. Proprietary and Arabic-centric models reduce these gaps but do not fully eliminate them.

For Dialect-to-MSA translation, country-wise differences are smaller across all metrics. BLEU, BERTScore, and LLM-as-Judge scores remain high and relatively stable across countries, suggesting that mapping dialectal input into standardized MSA is less sensitive to country-level variation than dialect generation.

Region-wise results in Tables F16 and F17 fur-

Model	Context: None			Context: Reg.			Context: Reg. + Cou.		
	Acc	CS	~CS	Acc	CS	~CS	Acc	CS	~CS
Random	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333
Arabic Centric Models									
Jais-7B-chat	0.554	0.504	0.589	0.532	0.488	0.563	0.531	0.492	0.558
Jais-2 8B-Chat	0.740	0.689	0.774	0.754	0.701	0.789	0.766	0.704	0.808
ALLaM-7B-Instruct	0.418	0.387	0.437	0.474	0.437	0.499	0.506	0.465	0.533
SILMA-9B-Instruct	**0.783**	**0.715**	**0.829**	0.780	0.714	0.824	0.784	0.724	0.825
c4ai-command-r7b (Ar)	0.355	0.366	0.348	0.340	0.351	0.334	0.375	0.385	0.368
Fanar-1-9B	0.391	0.373	0.403	0.444	0.410	0.467	0.618	0.553	0.662
Hala-9B	0.779	0.711	0.826	**0.810**	**0.739**	**0.858**	**0.820**	**0.751**	**0.866**
Multilingual Models									
Gemma-2-9B-it	0.671	0.619	0.707	0.702	0.648	0.739	0.710	0.661	0.744
Qwen-3-8B-it	0.354	0.350	0.357	0.370	0.361	0.375	0.380	0.368	0.387
Llama-3.1-8B-it	0.456	0.424	0.478	0.463	0.430	0.486	0.473	0.441	0.495
Proprietary Models									
GPT-5	0.943	0.908	0.966	0.946	0.912	0.968	0.953	0.927	0.971
Gemini-2.5-pro	0.942	0.920	0.954	0.943	0.922	0.958	0.950	0.933	0.961

Table C7: MCQ evaluation for MSA dataset. Results are averaged and reported separately for country-specific (CS) and non-country-specific (∼CS) dialogues. The best overall model is shown in **bold**, and the best within each category is underlined.

You are a helpful assistant who writes only in Modern Standard Arabic (MSA). Continue the dialogue with a single natural utterance and avoid extra explanations.

Continue the following dialogue in Modern Standard Arabic (MSA): {record}

Figure B3: The prompt used for the dialogue completion across all types of LLMs

You are a helpful assistant who writes in the {dialect_name} dialect (code {code}). Continue the dialogue with one natural utterance in that dialect without translation or commentary.

Continue the following dialogue in {dialect_name} while keeping the conversational tone: {record}

Figure B4: The prompt used for dialect steering across all types of LLMs

ther clarify these trends. In the MSA-to-Dialect direction, North Africa consistently yields lower scores across all model categories, while Gulf, Levant, and Nile River regions achieve higher and more stable performance. In contrast, Dialect-to-MSA results show much smaller regional differences, with all regions reaching similar levels of translation quality.

Across both country-wise and region-wise settings, LLM-as-Judge register scores display the largest variation between models. This indicates that dialectal correctness remains the primary challenge in Arabic dialect translation, even when semantic adequacy and fluency scores are relatively high.

G Dialect Steering

As mentioned in Section 5.3, the models' ability to use an intended dialect in a dialogue setup is variable. However, this is partially caused by the misalignment between GlotLID's labels and the set of country-level dialects we have. For instance, a strict ISO-code GlotLID of "0" is realized for Gulf prompts, as shown in Table G18, even when the continuation is clearly colloquial and regionally plausible. This is most visible for the UAE split, for which the responses are predicted to be in the broad Gulf Arabic code (`afb`) rather than a UAE-exclusive dialect label, and for Saudi Arabia,

Model	Context: None			Context: Reg.			Context: Reg. + Cou.		
	Acc	CS	~CS	Acc	CS	~CS	Acc	CS	~CS
Random	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333
Arabic Centric Models									
Jais-7B-chat	0.498	0.460	0.524	0.471	0.440	0.491	0.469	0.440	0.489
Jais-2-8B-Chat	0.710	0.650	0.751	0.715	0.647	0.761	0.731	0.667	0.773
ALLaM-7B-Instruct	0.398	0.374	0.413	0.435	0.411	0.450	0.471	0.440	0.491
SILMA-9B-Instruct	0.716	**0.663**	0.751	0.715	0.664	0.749	0.714	0.665	0.747
c4ai-command-r7b (Ar)	0.342	0.351	0.337	0.338	0.345	0.334	0.357	0.361	0.355
Fanar-1-9B	0.350	0.354	0.348	0.379	0.369	0.386	0.534	0.486	0.566
Hala-9B	**0.733**	0.660	**0.782**	**0.750**	**0.676**	**0.800**	**0.763**	**0.692**	**0.810**
Multilingual Models									
Gemma-2-9B-it	0.609	0.558	0.644	0.641	0.585	0.679	0.643	0.582	0.684
Qwen-3-8B-it	0.354	0.346	0.359	0.366	0.351	0.376	0.368	0.355	0.377
Llama-3.1-8B-it	0.412	0.406	0.416	0.427	0.415	0.436	0.430	0.422	0.435
Proprietary Models									
GPT-5	0.948	0.924	0.964	0.943	0.915	0.962	0.948	0.923	0.965
Gemini-2.5-pro	0.939	0.917	0.956	0.939	0.912	0.960	0.939	0.924	0.950

Table C8: MCQ evaluation for Dialect dataset. Results are averaged and reported separately for country-specific (CS) and non-country-specific (∼CS) dialogues. The best overall model is shown in **bold**, and the best within each category is underlined.

where the same country label spans multiple major varieties (e.g., Najdi, Hijazi, and Gulf-adjacent Eastern speech). In both cases, models often predict a neighboring code (commonly Najdi `ars`), which strict exact-match scoring penalizes. We nevertheless report strict-code GlotLID because many downstream pipelines treat dialect as a discrete label, but we interpret it jointly with judged quality and the macro-region rows. We also acknowledge that GlotLID is not a perfect dialect identification system, and that dialect identification is not a single-label classification task (Keleg and Magdy, 2023).

Case studies across Gulf, Darija, and Levant. Table G20 grounds the aggregate trends in concrete generations for UAE (Gulf), Morocco (Darija), and Syria (Levant). Two consistent phenomena stand out. First, code stability is hardest in the Gulf: even when models produce unmistakably colloquial Gulf continuations, GlotLID often assigns a neighboring label (frequently `ars`, Najdi) rather than the Gulf ISO label (`afb`) used for the UAE split, which explains the persistent strict-code zeros for UAE in Table G18. This is not surprising given that Saudi Arabic is not a single uniform target in practice—Najdi and Hijazi are both prominent, and Eastern (Gulf-adjacent) speech shares many cues with UAE-style Gulf—so the UAE/KSA boundary is an especially fragile place to demand exact-code agreement. Second, the "dialect" target is not a single knob: Morocco behaves like a distinctive lexical style that supervision can amplify (matching the large Morocco gains above), whereas Syria is easier to keep fluent but easier to drift toward pan-Levantine or even MSA-like realizations, especially when the continuation content is generic. A useful way to read the supervised results, therefore, is not as "SFT always improves dialect" but as "SFT improves controllability and fluency, and it improves dialect identity when the target dialect has separable cues that the training signal reinforces."

G.1 Agreement Between Human Evaluation and LLM-as-Judge

To assess the reliability of the LLM-as-Judge evaluation, we analyze its agreement with human judgments on a subset of the data. Due to the cost of human annotation, this analysis is conducted on a subset of the Moroccan dialect samples, for which human evaluation scores are available.

Table G21 reports the agreement between human judgments and LLM-as-a-Judge scores for the adequacy and fluency rubric. We compute

Country	Gemma-2-9B-it		Hala-9B	
	CS	~CS	CS	~CS
Algeria	0.654	0.653	0.716	0.826
Egypt	0.751	0.706	0.802	0.721
Jordan	0.882	0.777	0.882	0.923
KSA	0.704	0.772	0.847	0.877
Lebanon	0.636	0.635	0.758	0.821
Libya	0.730	0.727	0.790	0.878
Morocco	0.728	0.809	0.854	0.948
Palestine	0.690	0.783	0.897	0.869
Sudan	0.611	0.750	0.743	0.929
Syria	0.660	0.759	0.723	0.836
Tunisia	0.628	0.756	0.689	0.861
UAE	0.552	0.753	0.657	0.865
Yemen	0.597	0.723	0.675	0.723
Gulf	0.611	0.756	0.712	0.847
Levant	0.672	0.749	0.781	0.869
Nile River	0.692	0.733	0.777	0.850
N. Africa	0.679	0.731	0.755	0.879

Table D9: Cultural reasoning MCQ evaluation in MSA dataset, grouped by country and region.

Mean Absolute Difference (MAD) to quantify the average absolute deviation between human and LLM-as-a-Judge scores, and Accuracy@1 to measure the proportion of instances in which the LLM score falls within one point of the averaged human score. Overall, MAD values are generally below 1, with average Accuracy@1 scores of 0.77 and 0.75 across rubrics, indicating strong alignment between LLM-as-a-Judge assessments and human evaluations.

Country	Gemma-2-9B-it		Hala-9B	
	CS	~CS	CS	~CS
Algeria	0.506	0.600	0.605	0.800
Egypt	0.650	0.706	0.782	0.750
Jordan	0.824	0.747	0.824	0.879
KSA	0.684	0.735	0.827	0.870
Lebanon	0.596	0.635	0.667	0.776
Libya	0.710	0.655	0.780	0.755
Morocco	0.485	0.665	0.738	0.867
Palestine	0.724	0.725	0.828	0.836
Sudan	0.528	0.643	0.660	0.804
Syria	0.574	0.659	0.702	0.759
Tunisia	0.518	0.651	0.573	0.756
UAE	0.533	0.713	0.667	0.826
Yemen	0.553	0.723	0.621	0.631
Gulf	0.579	0.723	0.682	0.812
Levant	0.630	0.699	0.714	0.819
Nile River	0.598	0.667	0.730	0.783
N. Africa	0.551	0.639	0.663	0.803

Table D10: Cultural reasoning MCQ evaluation in Dialect dataset, grouped by country and region.

Topic	Gemma-2-9B		Hala-9B	
	CS	~CS	CS	~CS
Agriculture	0.765	0.839	0.765	0.946
Art	0.669	0.779	0.789	0.882
Daily Act.	0.614	0.746	0.715	0.890
Death	0.595	0.645	0.762	0.878
Family Rel.	0.667	0.839	0.857	0.931
Food	0.644	0.701	0.703	0.775
Habits	0.682	0.821	0.788	0.883
Holiday Act.	0.646	0.717	0.737	0.871
Idioms	0.659	0.583	0.744	0.806
Parenting	0.826	0.849	0.783	0.936
Trd. Games	0.659	0.703	0.800	0.891
Wedding	0.710	0.806	0.807	0.843

Table E11: Cultural reasoning MCQ evaluation in MSA dataset, grouped by topic.

You are a very strict Arabic translation evaluator.
You will be told:

- the source language variety
- the target language variety
- the region and country of the dialect (if applicable)

Evaluate the translation using **four rubrics**, each scored on a scale from 1 to 5:

- **Adequacy**: meaning preservation from source to translation
- **Fluency**: grammatical correctness and naturalness
- **Register & Variety**: correctness of MSA or the specified dialect
- **Terminology & Named Entities**: accuracy and consistency

Scoring scale:

- 5 = Excellent / perfect
- 4 = Minor issues
- 3 = Noticeable problems
- 2 = Major problems
- 1 = Unacceptable

Return **only** a valid JSON object in the following format:
{ "adequacy": int, "fluency": int, "register": int, "terminology": int, "overall": int, "judge_rationale": "one short sentence explaining the rationale for the overall score" }
Do not include any text outside the JSON.

(a) System prompt used to instruct the LLM judge for translation evaluation.

Translation direction:
{source_name} → {target_name}
Region: {region}
Country: {country}
SOURCE:
{source}
TRANSLATION:
{translation}

(b) User prompt providing the source text (ground truth) and the model-generated translation to the LLM judge.

Figure B5: Prompting strategy used for LLM-based evaluation. A fixed system prompt defines the evaluation criteria and scoring format, while the user prompt supplies the translation direction, regional metadata, and the source–translation pair.

Topic	Gemma-2-9B		Hala-9B	
	CS	~CS	CS	~CS
Agriculture	0.647	0.839	0.686	0.893
Art	0.602	0.721	0.735	0.838
Daily Act.	0.563	0.635	0.715	0.834
Death	0.548	0.609	0.738	0.812
Family Rel.	0.762	0.736	0.762	0.851
Food	0.555	0.651	0.611	0.720
Habits	0.635	0.763	0.694	0.817
Holiday Act.	0.543	0.710	0.674	0.843
Idioms	0.573	0.556	0.732	0.833
Parenting	0.696	0.731	0.609	0.839
Trd. Games	0.506	0.667	0.718	0.790
Wedding	0.655	0.716	0.793	0.821

Table E12: Cultural reasoning MCQ evaluation in Dialect dataset, grouped by topic.

Model	BLEU				BERTScore			LLM-as-Judge (0–5)				
	B1	B2	B3	B4	P	R	F1	Adeq.	Flu.	Reg.	Term.	Overall
Proprietary Models (0-shot)												
GPT-5	**0.688**	**0.587**	**0.507**	**0.434**	**0.910**	**0.912**	**0.911**	**4.905**	**4.883**	**4.957**	**4.927**	**4.773**
Gemini-2.5-pro	0.687	0.584	0.504	0.430	0.906	0.911	0.909	4.810	4.836	4.936	4.875	4.654
Arabic-Centric Models (0-shot)												
Jais-7B-chat	0.463	0.333	0.245	0.178	0.835	0.837	0.836	4.447	2.422	1.563	3.870	2.299
ALLaM-7B-Instruct	0.654	0.552	0.475	0.405	0.891	0.891	0.891	4.253	4.509	4.666	4.489	4.079
SILMA-9B-Instruct	0.473	0.362	0.283	0.218	0.850	0.834	0.841	3.542	2.735	2.143	3.571	2.510
c4ai-command-r7b	0.577	0.463	0.380	0.310	0.873	0.872	0.873	4.145	3.738	3.494	4.177	3.342
Fanar-1-9B	0.612	0.512	0.436	0.367	0.875	0.887	0.881	3.782	4.386	4.516	4.179	3.779
Hala	0.561	0.459	0.383	0.315	0.827	0.852	0.839	3.561	3.518	4.383	3.969	3.422
Multilingual Models (0-shot)												
Gemma-2-9B-it	0.598	0.481	0.394	0.319	0.881	0.881	0.881	3.423	3.203	3.417	3.545	3.128
Qwen-3-8B-it	0.554	0.430	0.340	0.265	0.869	0.866	0.868	2.875	2.699	2.669	3.047	2.647
Llama-3.1-8B-it (Instruct)	0.532	0.404	0.316	0.242	0.859	0.859	0.859	2.660	2.461	2.576	2.805	2.468
Multilingual Models (SFT)												
Gemma-2-9B-it (SFT)	0.533	0.446	0.377	0.316	0.861	0.834	0.847	2.117	3.117	3.819	3.151	2.654
Qwen-3-8B-it (SFT)	0.524	0.426	0.351	0.285	0.856	0.835	0.845	2.101	2.617	3.257	2.834	2.397
Llama-3.1-8B-it (SFT)	0.209	0.167	0.135	0.108	0.813	0.747	0.779	1.666	2.483	3.088	2.502	2.040

Table F13: Dialect-to-MSA translation quality under *Context: Country + Region*. We report BLEU and BERTScore, along with LLM-as-Judge scores (0–5) for *Adeq.* (semantic adequacy), *Flu.* (fluency and grammaticality in MSA), *Reg.* (appropriateness of register and standardness), *Term.* (terminology and lexical choice), and *Overall* (holistic quality). The best overall model is shown in **bold**, and the best model within each category is underlined.

Country	BLEU				BERTScore			LLM-as-Judge (0–5)				
	B1	B2	B3	B4	P	R	F1	Adeq.	Flu.	Reg.	Term.	Overall
Best Proprietary Model: GPT-5												
Algeria	0.581	0.449	0.351	0.269	0.865	0.873	0.869	4.857	4.631	4.721	4.873	4.439
Egypt	0.650	0.531	0.443	0.366	0.891	0.893	0.892	4.954	4.866	4.870	4.954	4.741
Jordan	0.554	0.418	0.316	0.233	0.883	0.893	0.888	4.939	4.793	4.820	4.962	4.667
KSA	0.639	0.518	0.425	0.342	0.894	0.898	0.896	4.902	4.689	4.638	4.940	4.511
Lebanon	0.590	0.461	0.363	0.281	0.874	0.881	0.878	4.952	4.869	4.904	4.965	4.782
Libya	0.584	0.448	0.349	0.265	0.877	0.883	0.880	4.865	4.707	4.665	4.884	4.498
Morocco	0.590	0.459	0.357	0.269	0.882	0.882	0.882	4.871	4.577	4.726	4.851	4.436
Palestine	0.617	0.492	0.398	0.313	0.881	0.880	0.880	4.947	4.793	4.781	4.963	4.638
Sudan	0.602	0.467	0.367	0.283	0.889	0.889	0.889	4.913	4.596	4.587	4.939	4.430
Syria	0.558	0.422	0.323	0.239	0.857	0.872	0.864	4.952	4.877	4.940	4.964	4.805
Tunisia	0.490	0.339	0.235	0.158	0.837	0.851	0.844	4.884	4.716	4.764	4.924	4.502
UAE	0.654	0.536	0.443	0.361	0.893	0.900	0.897	4.918	4.671	4.545	4.929	4.482
Yemen	0.529	0.388	0.290	0.209	0.864	0.861	0.862	4.931	4.370	3.520	4.874	3.963
Best Arabic-Centric Model: ALLaM-7B												
Algeria	0.388	0.244	0.156	0.097	0.818	0.812	0.815	4.387	3.307	2.097	4.151	3.029
Egypt	0.541	0.404	0.314	0.241	0.856	0.855	0.856	3.912	3.958	4.167	4.243	3.678
Jordan	0.490	0.351	0.261	0.188	0.861	0.864	0.862	4.126	3.885	3.487	4.387	3.625
KSA	0.554	0.418	0.322	0.244	0.866	0.866	0.866	4.183	4.068	3.349	4.523	3.694
Lebanon	0.524	0.390	0.296	0.218	0.856	0.853	0.854	4.149	3.790	3.620	4.249	3.638
Libya	0.545	0.423	0.337	0.262	0.858	0.856	0.857	4.377	3.623	2.512	4.237	3.261
Morocco	0.380	0.237	0.145	0.082	0.799	0.780	0.789	4.202	2.968	2.057	3.887	2.835
Palestine	0.535	0.400	0.301	0.218	0.862	0.857	0.859	4.171	3.951	3.667	4.358	3.671
Sudan	0.538	0.396	0.301	0.221	0.862	0.861	0.861	4.244	3.778	2.439	4.026	3.148
Syria	0.512	0.376	0.277	0.197	0.852	0.859	0.856	4.132	3.896	3.749	4.295	3.685
Tunisia	0.412	0.255	0.163	0.101	0.818	0.815	0.817	4.293	3.373	2.387	4.102	3.129
UAE	0.596	0.470	0.377	0.296	0.877	0.878	0.877	4.094	3.890	3.188	4.255	3.596
Yemen	0.480	0.345	0.254	0.181	0.843	0.838	0.841	4.358	3.976	2.386	4.398	3.260
Best Multilingual Model (0-shot): Qwen-3-8B												
Algeria	0.334	0.195	0.111	0.064	0.800	0.794	0.797	3.353	2.635	1.402	3.143	2.197
Egypt	0.384	0.235	0.146	0.086	0.814	0.809	0.811	3.272	2.490	1.707	3.318	2.339
Jordan	0.400	0.253	0.160	0.100	0.833	0.833	0.833	3.548	2.835	1.556	3.513	2.444
KSA	0.464	0.320	0.226	0.154	0.837	0.843	0.840	3.672	3.289	1.455	3.643	2.498
Lebanon	0.387	0.249	0.159	0.099	0.815	0.809	0.812	3.424	2.638	1.594	3.415	2.389
Libya	0.506	0.378	0.288	0.215	0.851	0.850	0.850	3.423	2.721	1.623	3.191	2.419
Morocco	0.325	0.190	0.104	0.054	0.779	0.759	0.769	3.109	2.242	1.250	2.883	2.089
Palestine	0.395	0.253	0.166	0.105	0.819	0.812	0.816	3.557	2.837	1.524	3.541	2.439
Sudan	0.464	0.322	0.229	0.157	0.838	0.837	0.837	3.296	2.783	1.439	3.239	2.278
Syria	0.373	0.232	0.150	0.095	0.806	0.813	0.809	3.566	2.869	1.697	3.574	2.502
Tunisia	0.332	0.183	0.100	0.055	0.797	0.793	0.795	3.227	2.444	1.436	2.951	2.191
UAE	0.493	0.356	0.262	0.187	0.840	0.845	0.842	3.643	3.208	1.439	3.631	2.439
Yemen	0.440	0.302	0.207	0.135	0.828	0.826	0.827	3.317	2.744	1.549	3.329	2.370
Best Multilingual Model (SFT): Gemma-2-9B												
Algeria	0.339	0.224	0.142	0.088	0.799	0.776	0.787	2.062	2.307	1.971	2.656	2.094
Egypt	0.406	0.287	0.203	0.142	0.816	0.795	0.805	2.067	2.511	2.803	3.063	2.352
Jordan	0.375	0.261	0.179	0.122	0.830	0.808	0.819	2.084	2.632	2.376	3.103	2.303
KSA	0.450	0.333	0.247	0.180	0.836	0.815	0.825	2.089	2.728	2.396	3.098	2.311
Lebanon	0.386	0.272	0.188	0.127	0.815	0.788	0.801	2.092	2.546	2.424	3.022	2.319
Libya	0.457	0.357	0.282	0.217	0.833	0.809	0.820	2.051	2.535	1.944	2.884	2.154
Morocco	0.328	0.218	0.139	0.085	0.796	0.767	0.781	2.016	2.129	1.960	2.468	2.004
Palestine	0.388	0.269	0.190	0.130	0.819	0.793	0.806	2.069	2.557	2.455	3.029	2.260
Sudan	0.440	0.333	0.255	0.191	0.836	0.812	0.824	2.065	2.578	2.009	2.948	2.161
Syria	0.367	0.244	0.165	0.107	0.811	0.797	0.804	2.064	2.542	2.518	3.068	2.307
Tunisia	0.288	0.168	0.093	0.054	0.789	0.770	0.779	2.080	2.307	1.698	2.636	2.067
UAE	0.443	0.331	0.249	0.182	0.834	0.814	0.824	2.098	2.569	1.945	2.996	2.184
Yemen	0.383	0.273	0.193	0.134	0.818	0.793	0.805	2.085	2.752	1.854	3.224	2.207

Table F14: Country-wise MSA-to-Dialect translation performance under *Context: Country + Region*, reporting results for the best-performing model in each category: proprietary (GPT-5), multilingual (Qwen-3-8B), Arabic-centric (ALLaM-7B), and multilingual SFT (Gemma-2-9B). We report BLEU, BERTScore, and LLM-as-Judge scores on a 0–5 scale.

Country	BLEU				BERTScore			LLM-as-Judge (0–5)				
	B1	B2	B3	B4	P	R	F1	Adeq.	Flu.	Reg.	Term.	Overall
Best Proprietary Model: GPT-5												
Algeria	0.643	0.528	0.439	0.360	0.902	0.898	0.900	4.893	4.865	4.959	4.918	4.742
Egypt	0.626	0.509	0.421	0.347	0.897	0.898	0.898	4.921	4.883	4.950	4.925	4.782
Jordan	0.766	0.683	0.612	0.545	0.935	0.936	0.935	4.985	4.927	4.962	4.973	4.885
KSA	0.686	0.581	0.496	0.420	0.912	0.913	0.912	4.936	4.860	4.957	4.970	4.792
Lebanon	0.698	0.597	0.519	0.448	0.913	0.917	0.915	4.935	4.878	4.961	4.943	4.782
Libya	0.720	0.628	0.554	0.486	0.922	0.924	0.923	4.870	4.893	4.991	4.912	4.758
Morocco	0.735	0.641	0.562	0.490	0.923	0.925	0.924	4.911	4.895	4.980	4.915	4.790
Palestine	0.704	0.609	0.536	0.468	0.907	0.914	0.910	4.923	4.923	4.951	4.931	4.813
Sudan	0.745	0.655	0.582	0.511	0.932	0.935	0.933	4.874	4.896	4.957	4.926	4.770
Syria	0.628	0.514	0.428	0.353	0.890	0.886	0.888	4.896	4.857	4.956	4.936	4.757
Tunisia	0.614	0.496	0.406	0.327	0.887	0.892	0.889	4.853	4.849	4.956	4.911	4.707
UAE	0.722	0.632	0.557	0.486	0.922	0.922	0.922	4.929	4.875	4.933	4.926	4.753
Yemen	0.657	0.552	0.469	0.394	0.895	0.898	0.896	4.829	4.878	4.935	4.858	4.707
Best Arabic-Centric Model: ALLaM-7B												
Algeria	0.623	0.507	0.420	0.343	0.892	0.884	0.888	4.091	4.482	4.745	4.465	3.988
Egypt	0.585	0.467	0.386	0.316	0.870	0.871	0.871	4.356	4.427	4.515	4.598	4.080
Jordan	0.740	0.657	0.589	0.525	0.914	0.914	0.914	4.513	4.636	4.740	4.755	4.291
KSA	0.649	0.548	0.470	0.399	0.889	0.889	0.889	4.451	4.562	4.647	4.702	4.200
Lebanon	0.678	0.579	0.506	0.438	0.900	0.898	0.899	4.179	4.524	4.707	4.555	4.070
Libya	0.684	0.591	0.519	0.452	0.899	0.898	0.899	4.079	4.488	4.642	4.381	3.954
Morocco	0.721	0.629	0.554	0.485	0.913	0.914	0.914	4.270	4.569	4.851	4.351	4.109
Palestine	0.661	0.564	0.491	0.422	0.887	0.893	0.890	4.455	4.561	4.683	4.724	4.203
Sudan	0.699	0.609	0.540	0.476	0.903	0.907	0.905	4.296	4.596	4.691	4.426	4.161
Syria	0.586	0.473	0.390	0.316	0.874	0.867	0.870	4.287	4.454	4.673	4.542	4.080
Tunisia	0.570	0.446	0.358	0.282	0.866	0.867	0.866	3.844	4.302	4.502	4.058	3.724
UAE	0.666	0.571	0.496	0.427	0.892	0.891	0.891	4.200	4.526	4.643	4.333	4.043
Yemen	0.629	0.525	0.445	0.372	0.883	0.885	0.884	4.203	4.463	4.602	4.415	4.069
Best Multilingual Model (0-shot): Gemma-2-9B												
Algeria	0.535	0.401	0.307	0.229	0.864	0.861	0.862	3.123	2.914	3.033	3.180	2.848
Egypt	0.560	0.429	0.335	0.260	0.877	0.874	0.875	3.653	3.276	3.565	3.770	3.280
Jordan	0.688	0.586	0.505	0.429	0.917	0.913	0.915	3.889	3.502	3.866	4.107	3.513
KSA	0.627	0.509	0.421	0.344	0.890	0.888	0.889	3.706	3.477	3.706	3.877	3.379
Lebanon	0.607	0.492	0.406	0.332	0.883	0.884	0.883	3.306	3.153	3.376	3.463	3.074
Libya	0.635	0.531	0.453	0.382	0.887	0.888	0.887	3.335	3.186	3.251	3.572	3.056
Morocco	0.531	0.419	0.336	0.266	0.832	0.857	0.843	2.782	2.710	3.008	2.891	2.609
Palestine	0.618	0.501	0.415	0.338	0.883	0.885	0.884	3.764	3.443	3.646	3.927	3.406
Sudan	0.668	0.557	0.473	0.397	0.910	0.909	0.909	3.570	3.296	3.500	3.661	3.209
Syria	0.548	0.422	0.333	0.258	0.868	0.861	0.864	3.434	3.171	3.502	3.614	3.139
Tunisia	0.497	0.361	0.265	0.190	0.853	0.854	0.854	2.844	2.702	2.693	2.813	2.653
UAE	0.651	0.545	0.459	0.382	0.903	0.898	0.900	3.471	3.396	3.608	3.553	3.200
Yemen	0.606	0.493	0.407	0.330	0.886	0.886	0.886	3.553	3.362	3.573	3.589	3.240
Best Multilingual Model (SFT): Gemma-2-9B												
Algeria	0.497	0.397	0.320	0.253	0.851	0.820	0.835	2.123	3.131	3.812	3.074	2.615
Egypt	0.486	0.386	0.312	0.248	0.850	0.824	0.837	2.117	3.067	3.778	3.297	2.665
Jordan	0.596	0.524	0.463	0.405	0.878	0.848	0.863	2.157	3.180	3.939	3.498	2.808
KSA	0.538	0.450	0.380	0.319	0.864	0.838	0.851	2.119	3.221	3.962	3.268	2.719
Lebanon	0.549	0.465	0.399	0.342	0.862	0.835	0.848	2.087	3.183	3.887	3.031	2.638
Libya	0.557	0.471	0.405	0.344	0.860	0.833	0.846	2.093	3.074	3.767	3.088	2.642
Morocco	0.554	0.468	0.401	0.339	0.865	0.837	0.851	2.101	3.133	3.794	3.081	2.633
Palestine	0.552	0.469	0.404	0.344	0.866	0.841	0.853	2.081	3.077	3.890	3.281	2.679
Sudan	0.590	0.514	0.451	0.392	0.872	0.845	0.858	2.152	3.048	3.783	3.200	2.674
Syria	0.465	0.373	0.303	0.243	0.852	0.822	0.837	2.124	3.100	3.749	3.040	2.590
Tunisia	0.467	0.363	0.285	0.221	0.842	0.818	0.830	2.084	3.196	3.742	2.831	2.600
UAE	0.534	0.454	0.388	0.329	0.867	0.839	0.853	2.118	3.031	3.741	3.059	2.584
Yemen	0.539	0.455	0.386	0.325	0.863	0.836	0.849	2.155	3.077	3.789	3.167	2.642

Table F15: Country-wise Dialect-to-MSA translation performance under *Context: Country + Region*, reporting results for the best-performing model in each category: proprietary (GPT-5), multilingual (Gemma-2-9B), Arabic-centric (ALLaM-7B), and multilingual SFT (Gemma-2-9B SFT). We report BLEU, BERTScore, and LLM-as-Judge scores on a 0–5 scale.

Region	BLEU				BERTScore			LLM-as-Judge (0–5)				
	B1	B2	B3	B4	P	R	F1	Adeq.	Flu.	Reg.	Term.	Overall
Best Proprietary Model: GPT-5												
Gulf	0.607	0.481	0.386	0.304	0.884	0.886	0.885	4.917	4.576	4.232	4.914	4.318
Levant	0.579	0.448	0.349	0.266	0.874	0.882	0.878	4.947	4.832	4.860	4.964	4.721
Nile River	0.627	0.500	0.406	0.325	0.890	0.891	0.891	4.934	4.734	4.731	4.947	4.589
North Africa	0.562	0.425	0.324	0.241	0.866	0.872	0.869	4.869	4.655	4.720	4.882	4.467
Best Arabic-Centric Model: ALLaM-7B-Instruct												
Gulf	0.544	0.412	0.318	0.241	0.862	0.861	0.861	4.211	3.976	2.972	4.389	3.515
Levant	0.515	0.379	0.283	0.205	0.858	0.858	0.858	4.144	3.883	3.629	4.324	3.655
Nile River	0.539	0.400	0.308	0.231	0.859	0.858	0.859	4.075	3.870	3.320	4.137	3.418
North Africa	0.428	0.286	0.197	0.132	0.822	0.814	0.818	4.312	3.306	2.253	4.089	3.055
Best Multilingual Model (0-shot): Qwen-3-8B-it												
Gulf	0.466	0.327	0.232	0.159	0.835	0.838	0.837	3.544	3.079	1.481	3.534	2.435
Levant	0.389	0.247	0.159	0.100	0.818	0.817	0.818	3.526	2.798	1.593	3.513	2.445
Nile River	0.423	0.277	0.186	0.121	0.826	0.822	0.824	3.284	2.633	1.576	3.279	2.309
North Africa	0.371	0.233	0.147	0.094	0.805	0.797	0.801	3.274	2.504	1.421	3.039	2.218
Best Multilingual Model (SFT): Gemma-2-9B-it												
Gulf	0.425	0.312	0.230	0.165	0.829	0.807	0.818	2.091	2.681	2.058	3.105	2.232
Levant	0.379	0.261	0.180	0.121	0.819	0.797	0.808	2.077	2.570	2.443	3.057	2.297
Nile River	0.422	0.310	0.229	0.166	0.826	0.803	0.814	2.066	2.544	2.414	3.006	2.258
North Africa	0.351	0.240	0.162	0.109	0.804	0.779	0.791	2.052	2.312	1.896	2.653	2.077

Table F16: Region-wise MSA-to-Dialect translation performance under *Context: Country + Region*, reporting results for the best-performing model in each category: proprietary (GPT-5), multilingual (Qwen-3-8B), Arabic-centric (ALLaM-7B), and multilingual SFT (Gemma-2-9B). We report BLEU, BERTScore, and LLM-as-Judge scores on a 0–5 scale.

Region	BLEU				BERTScore			LLM-as-Judge (0–5)				
	B1	B2	B3	B4	P	R	F1	Adeq.	Flu.	Reg.	Term.	Overall
Best Proprietary Model: GPT-5												
Gulf	0.689	0.589	0.508	0.434	0.910	0.911	0.910	4.898	4.871	4.942	4.917	4.750
Levant	0.700	0.602	0.525	0.454	0.911	0.913	0.912	4.935	4.897	4.957	4.946	4.811
Nile River	0.684	0.581	0.500	0.427	0.914	0.916	0.915	4.898	4.889	4.953	4.925	4.776
North Africa	0.678	0.573	0.490	0.416	0.908	0.910	0.909	4.883	4.876	4.971	4.914	4.750
Best Arabic-Centric Model: ALLaM-7B-Instruct												
Gulf	0.648	0.548	0.471	0.400	0.888	0.889	0.888	4.281	4.516	4.630	4.478	4.102
Levant	0.666	0.569	0.495	0.426	0.894	0.893	0.893	4.364	4.545	4.701	4.646	4.164
Nile River	0.641	0.537	0.462	0.394	0.887	0.889	0.888	4.326	4.510	4.601	4.514	4.119
North Africa	0.650	0.544	0.464	0.391	0.893	0.891	0.892	4.076	4.463	4.691	4.317	3.948
Best Multilingual Model (0-shot): Gemma-2-9B-it												
Gulf	0.628	0.516	0.429	0.353	0.893	0.891	0.892	3.573	3.410	3.628	3.669	3.270
Levant	0.616	0.501	0.416	0.340	0.888	0.886	0.887	3.607	3.322	3.605	3.787	3.290
Nile River	0.613	0.492	0.403	0.327	0.893	0.891	0.892	3.612	3.286	3.533	3.716	3.245
North Africa	0.548	0.426	0.338	0.265	0.858	0.864	0.861	3.014	2.871	2.995	3.105	2.785
Best Multilingual Model (SFT): Gemma-2-9B-it												
Gulf	0.537	0.453	0.385	0.324	0.865	0.838	0.851	2.130	3.107	3.827	3.162	2.647
Levant	0.541	0.458	0.393	0.334	0.865	0.837	0.850	2.114	3.135	3.866	3.219	2.681
Nile River	0.537	0.449	0.380	0.319	0.861	0.834	0.847	2.134	3.058	3.780	3.250	2.670
North Africa	0.519	0.425	0.353	0.289	0.855	0.827	0.840	2.101	3.134	3.780	3.020	2.622

Table F17: Region-wise Dialect-to-MSA translation performance under *Context: Country + Region*, reporting results for the best-performing model in each category: proprietary (GPT-5), multilingual (Gemma-2-9B), Arabic-centric (ALLaM-7B), and multilingual SFT (Gemma-2-9B SFT). We report BLEU, BERTScore, and LLM-as-Judge scores on a 0–5 scale.

Country/Region	Best Proprietary (GPT-5)		Best Multilingual (Gemma-2-9B-it)		Best Arabic-centric (ALLaM-7B)	
	Judge	GlotLID	Judge	GlotLID	Judge	GlotLID
Algeria	0.963	0.045	0.568	0.082	0.819	0.049
Egypt	0.962	0.908	0.695	0.682	0.865	0.816
Jordan	0.969	0.552	0.656	0.172	0.889	0.433
KSA	0.959	0.511	0.673	0.217	0.857	0.579
Lebanon	0.959	0.729	0.629	0.279	0.845	0.664
Libya	0.951	0.000	0.568	0.000	0.757	0.000
Morocco	0.968	0.988	0.551	0.182	0.799	0.660
Palestine	0.966	0.610	0.616	0.211	0.849	0.411
Sudan	0.947	0.000	0.570	0.000	0.799	0.000
Syria	0.952	0.733	0.625	0.224	0.815	0.545
Tunisia	0.958	0.893	0.522	0.084	0.762	0.604
UAE	0.947	0.000	0.592	0.000	0.803	0.000
Yemen	0.917	0.000	0.586	0.000	0.744	0.000
Gulf	0.941	0.609	0.616	0.273	0.801	0.533
Levant	0.963	0.646	0.633	0.220	0.853	0.506
Nile River	0.955	0.731	0.634	0.595	0.833	0.695
North Africa	0.960	0.929	0.553	0.284	0.786	0.693

Table G18: Dialect steering by country (zero-shot): best model per family. GlotLID is strict ISO-code accuracy; region rows use the macro-region mapping from Bhatti and Alam (2025).

Country/Region	SFT (Llama-3.1-8B-Instruct)		SFT (Qwen-3-8B)		SFT (Gemma-2-9B)	
	Judge	GlotLID	Judge	GlotLID	Judge	GlotLID
Algeria	0.466	0.016	0.529	0.008	0.560	0.012
Egypt	0.514	0.435	0.562	0.238	0.616	0.506
Jordan	0.566	0.115	0.557	0.054	0.674	0.153
KSA	0.556	0.145	0.582	0.213	0.645	0.187
Lebanon	0.512	0.210	0.533	0.048	0.584	0.175
Libya	0.499	0.000	0.549	0.000	0.605	0.000
Morocco	0.496	0.348	0.484	0.134	0.568	0.563
Palestine	0.545	0.126	0.537	0.077	0.632	0.199
Sudan	0.516	0.000	0.534	0.000	0.603	0.000
Syria	0.498	0.170	0.530	0.103	0.604	0.194
Tunisia	0.452	0.084	0.498	0.129	0.535	0.209
UAE	0.528	0.000	0.559	0.000	0.620	0.000
Yemen	0.525	0.000	0.515	0.000	0.589	0.000
Gulf	0.536	0.159	0.551	0.197	0.618	0.200
Levant	0.534	0.152	0.540	0.068	0.627	0.179
Nile River	0.515	0.407	0.548	0.222	0.610	0.478
North Africa	0.478	0.361	0.514	0.233	0.566	0.574

Table G19: Dialect steering by country: SFT models. GlotLID is strict ISO-code accuracy; region rows use the macro-region mapping from Bhatti and Alam (2025).

Country	Sample ID	Prompt excerpt	Model continuations (zero-shot) + predicted GlotLID
UAE	uae-15-traditions-during-religous-holidays-dialect	هزيم : كل عام وأنت بخير يا سالم ! سالم : وأنت بخير يا هزيم ، ما أجمل أجواء العيد بين الأهل والأصدقاء. هزيم : في كل عيد أحرص ابارك بنفس العبارة التي تعلمتها من والدي. سالم : نعم، هذه العبارة تملأ القلب بالفرح وتعتبر من عاداتنا. هزيم : من الجميل أن نحافظ على تقاليدنا وننقلها للأجيال القادمة. سالم : …	**GPT-5** (ars): صح لسانك يا هزيم ، ودامت أفراحكم ، شو رايك عقب صلاة العيد نمر على بيوت الربع نسلّم ونعايدهم، وبالليل تتفضلون عندنا للقهوة والحلو والعيديات؟ **ALLaM-7B** (ars): والله يا سالم ، العيد ما يحلا إلا بوجودك، وكل عام وأنت بخير وعيدك مبارك يا الغالي. **Gemma-2-9B-it** (ars): كل عام وانت بالف خير يا هزيم ، وربي يبارك فيك ويحفظك.
Morocco	morocco-2-traditions-before-religous-holidays-dialect	فاطمة: رمضان على الأبواب، هل بدأتِ الاستعدادات يا ليلى؟ ليلى: نعم، في هذا الوقت تبدأ الأمهات بتحضير ما يلزم للشهر الكريم . فاطمة: رائحة الشباكية والسلو تنتشر في كل بيت تقريبًا. ليلى: صحيح، كل أسرة تحرص على تجهيز كميات تكفي لأيام الشهر الفضيل. فاطمة: الجو كله يصبح مليئًا بالفرح والنشاط في المطبخ. ليلى: …	**GPT-5** (ary): إي بصح، كنفيقو بكري نديرو الشباكية والسلو، ونقسمو الوقت بين الطياب والعبادة، وما نساوش الحريرة والتمر والشاي، وكنشريو المقاضي قبل الزحام باش تهناو فالأيام اللّولة. **ALLaM-7B** (ary): والله حتى كاينة أجواء زوينة في الشهر الكريم ، الكل مجمعين على مائدة الإفطار وراجنين على تحضير الأكل. **Gemma-2-9B-it** (apc): والليلة هادفنا نعمل السُمْن
Syria	syria-10-gift-giving-practices-the-occasion-dialect	ليلى: سمعت أن جمال قدّم هدية كبيرة لأخيه مؤخراً. سامر: نعم، لقد اختار قطعة أثاث مميزة جداً. ليلى: هل كان هناك مناسبة خاصة لذلك؟ سامر: بالتأكيد، في مجتمعنا هناك تقاليد مرتبطة بمثل هذه الهدايا. ليلى: الهدايا عادةً ما تعكس المناسبة وأهميتها عندنا. سامر: …	**GPT-5** (apc): صح ، هيك الهدايا بتكون عادة بمناسبات متل الخطبة أو فتح بيت جديد، وعيب الواحد يروح فاضي الإيدين، فبيعطي شي إلو قيمة وبيضل ذكرى؛ وجمال اختار خزانة أنتيك بتجنّن لأن أخوه هالفترة عم يجهّز بيته. **ALLaM-7B** (apc): هيك شي، بهدايا كبيرة متل هيك بيضلوا فترة طويلة عند الناس وبيحكوا فيها، خاصة إذا كانت قطعة أثاث مميزة مثل ما قال سامر. **Gemma-2-9B-it** (arb): بالتأكيد، هدايا جمال كانت تتناسب مع أهمية المناسبة.

Table G20: Case studies for dialect steering (zero-shot), focusing on UAE (Gulf), Morocco (Darija), and Syria (Levant). All Arabic excerpts are rendered with arabtex; GlotLID codes are the strict-code predictions for each continuation.

Direction	Model	MAD ↓ [0-5]	Acc@1 ↑ [0-1]
Adequacy			
MSA → Dialect	GPT-5	0.76	0.92
MSA → Dialect	Qwen3-8B	2.20	0.28
Dialect → MSA	GPT-5	0.19	1.00
Dialect → MSA	Gemma2-9B	0.68	0.88
Fluency			
MSA → Dialect	GPT-5	0.97	0.68
MSA → Dialect	Qwen3-8B	1.20	0.60
Dialect → MSA	GPT-5	0.41	0.98
Dialect → MSA	Gemma2-9B	0.89	0.72

Table G21: Agreement between human judgments and LLM-as-a-judge scores for different models on the Moroccan dataset. Mean Absolute Difference (MAD) ranges from 0 to 5, with lower values indicating closer alignment between human and LLM-as-a-judge scores. Accuracy@1 (Acc@1) ranges from 0 to 1, with higher values indicating more frequent agreement within one point of the human average.