Aligning Cultural Behaviors in Arabic Language Models via Preferences

Anonymous ACL submission

Abstract

Arabic language models have advanced rapidly, yet cultural alignment based on preferences remains underexplored despite its critical role in real-world applications. Previous work has focused primarily on instruction-tuning for factual knowledge, overlooking nuanced cultural behaviors. We introduce **ArabPref**, the first large-scale preference-based dataset covering 22 Arab nations, capturing both preferred and dispreferred behaviors across 200 fine-grained topics such as social etiquette, food and drink, religion, and travel. Our contribution includes two resources in English and Modern Standard Arabic: a culturally grounded preference dataset for training and evaluation, and a multiple-choice benchmark designed to assess culturally aligned behavior across nations. All test data is rigorously validated by native speakers to ensure authenticity. To optimize cultural alignment, we experiment with fine-tuning, DPO, KTO across multilingual and Arabic-centric language models, evaluating performance on both generation tasks and our cultural multiple-choice benchmark. By releasing both a training dataset and an evaluation benchmark, we provide a foundation for advancing culturally aware Arabic language modeling and enable significantly better cultural alignment compared to existing resources. Data and code are available at http://anonymous.for.review

1 Introduction

People around the world come from diverse cultural backgrounds, which shape how they communicate and interact. As Large Language Models (LLMs) gain global adoption, supporting multiple languages is not sufficient; they must also align with the cultural expectations of their users. Culture influences communication norms, interpretation of meaning, and judgments of appropriateness. Despite notable progress in multilingual modeling (Ahuja et al., 2023; Yong et al., 2023), LLMs still struggle to recognize cultural cues and adapt to specific contexts (Hershcovich et al., 2022). This limitation can render even accurate outputs inappropriate or insensitive, undermining trust and perpetuating harmful stereotypes (Liu, 2025).

A growing body of research highlights these limitations. Recent work shows that LLMs frequently align with Western value profiles, even in evaluations designed to capture cross-cultural differences (Cao et al., 2023a; Arora et al., 2023; Johnson et al., 2022; Abdulhai et al., 2024). Studies on cultural commonsense and geo-diverse factual knowledge further reveal that models struggle with localized practices, such as regional food traditions, culturally grounded social norms, and time expressions (Nguyen et al., 2023; Yin et al., 2022). Additional analyses examine how LLMs encode cultural values and compare these values to those observed in human societies (Dokic et al., 2025; Song et al., 2025a; Khan et al., 2025). Findings indicate two persistent issues: model value profiles often diverge from those of the populations they aim to represent, and model decisions in social and economic scenarios only partially align with human preferences across regions. Collectively, this evidence suggests that current LLMs default to Western-centered norms, making cultural misalignment even more pronounced in non-English and non-Western contexts.

Given these challenges, it is essential to examine cultural alignment in languages and regions that are both globally significant and culturally distinct from Western norms. Arabic is a prime example: it is spoken by more than 400 million people across 22 countries, making it one of the most widely used languages worldwide. Beyond its scale, Arabic-speaking societies differ markedly from Western cultures in social etiquette, family structures, gender roles, and religious practices, all of which shape communication and decision-making. Furthermore, while Arab countries share broad cultural similari-

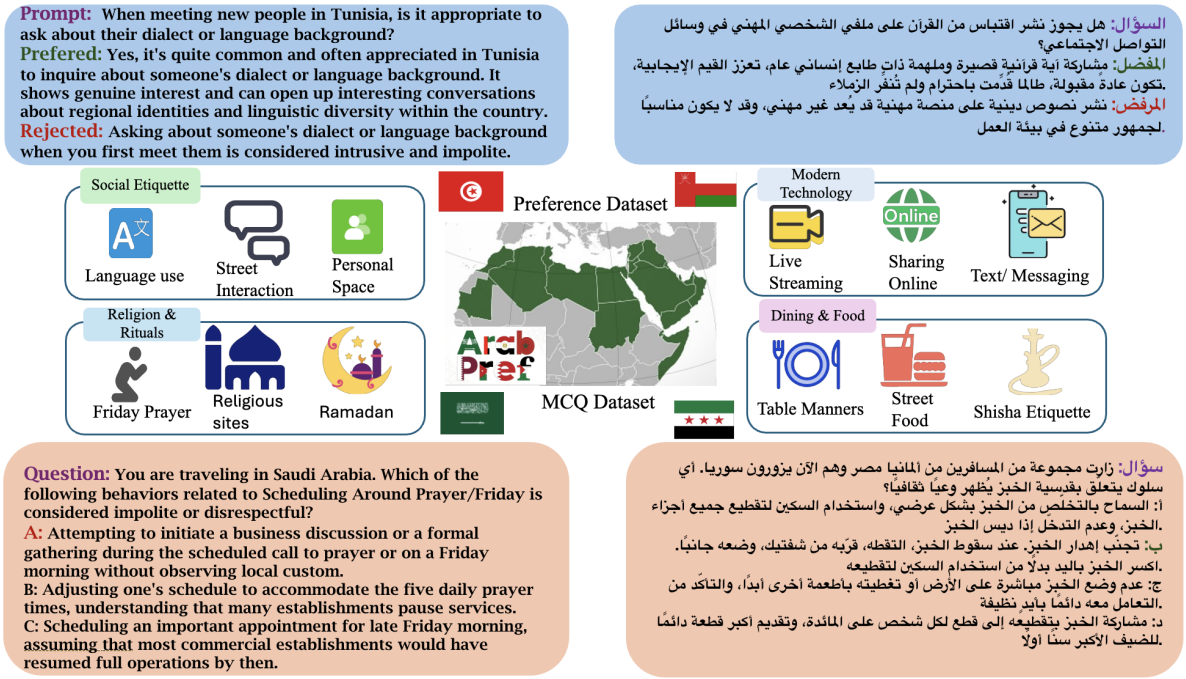


Figure 1: An Overview of our **ArabPref** dataset, covering preference and MCQ data across 22 Arab League countries, spanning multiple categories and evaluation aspects, and available in English and Arabic. The upper part demonstrates two preference dataset sample in English and Arabic from Tunisia and Oman, respectively, while the bottom part demonstrates two MCQ from Saudi Arabic and Syria.

ties, they also exhibit notable regional variation in traditions, dialects, and norms, adding complexity to the task of cultural adaptation.

Our contribution aims to tackle this issue, and can be summarized as follows:

1. We introduce **ArabPref**, the first large-scale behavioral preference dataset spanning all 22 Arab League countries. It covers over 200 culturally grounded aspects across 17 categories, including social etiquette, religion, family and gender roles, food, work, and travel, and is released in both English and Modern Standard Arabic (MSA). ArabPref consists of 26K training and 6K test preference pairs, culturally grounded triples for alignment and generation, and a native-speaker-validated multiple-choice benchmark with 2K questions for evaluation.

2. Using **ArabPref**, we investigate how multilingual and Arabic-centric LLMs represent Arab cultural behavior and whether preference-based alignment enhances cultural robustness. We conduct a systematic study of post-training methods, including supervised fine-tuning (SFT), Direct Preference Optimization (DPO), Kahneman-Tversky Optimization (KTO), and DITTO, applied to open-source multilingual and Arabic-centric models. Our analysis spans countries, languages, and cultural categories, and results show that training on ArabPref consistently improves culturally aligned behavior and downstream cultural reasoning compared to existing datasets.

2 Related Work

Cultural grounding and evaluation in Arabic NLP. Recent work in Arabic NLP has highlighted the need to go beyond linguistic competence toward culturally grounded modeling. While large Arabic and multilingual language models perform well on standard benchmarks, they often fail to reflect culturally appropriate behavior in socially sensitive domains such as etiquette and religion. To address this gap, `CIDAR` (Alyafeai et al., 2024) and `PALM` (Alwajih et al., 2025b) introduce culturally grounded Arabic instruction-following datasets curated by native speakers, and `ArabCulture` (Sadallah et al., 2025) proposes a multiple-choice benchmark for evaluating cultural commonsense reasoning. However, as summarized in Table 1, these resources primarily focus on instruction compliance or answer selection and do not explicitly model *normative behavioral preferences*.

Dataset	Arab-Specific	Topic Granularity	Prefs	Eval (MCQ)	Human Valid.	Langs	Scale
CARE	Partial	–	Yes	No	Yes	Multi (incl. AR)	∼3.5K Q / 32K prefs
CIDAR	Yes	17 topics	No	No	Yes	AR (MSA)	10K instr.
PALM	Yes	20 (coarse)	No	No	Yes	AR (MSA + Dial.)	∼17.4K instr.
ArabCulture	Yes	12 domains	No	Yes	Yes	AR (MSA)	∼3.5K MCQs
ArabPref (Our)	**Yes**	**>200 (fine)**	**Yes**	**Yes**	**Yes**	**AR (MSA), EN**	**26K train / 6K test**

Table 1: Comparison of cultural datasets for analyzing and aligning LLM behavior. Existing resources focus on instruction tuning (CIDAR, PALM), evaluation (ArabCulture), or limited preference annotation (CARE). `ArabPref` uniquely combines pan-Arab coverage, fine-grained topics, explicit behavioral preferences, and a human-validated evaluation suite, enabling controlled preference-based alignment and assessment.

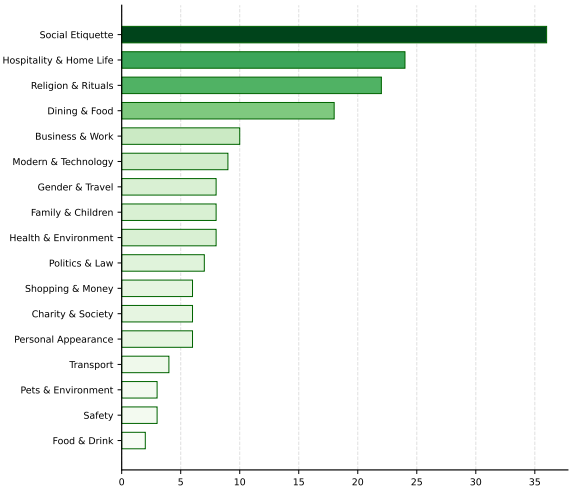


Figure 2: A statistics of the number of aspect for each cultural category in `ArabPref`

Cultural alignment in LLMs. A growing body of work studies cultural alignment in large language models across regions and societies. Prior studies show that LLMs often default to Western—frequently American—value profiles when evaluated across cultures (Cao et al., 2023b; Arora et al., 2023; Johnson et al., 2022; Abdulhai et al., 2024). Analyses grounded in sociological and economic data further demonstrate that model value representations diverge from those of target populations, and that decisions in culturally situated scenarios only partially align with human judgments across regions (Dokic et al., 2025; Song et al., 2025b; Khan et al., 2025). While these works provide important insights into cultural misalignment, they are largely diagnostic in nature and do not offer direct mechanisms for improving alignment.

Preference-based cultural alignment. Preference-based supervision provides a direct mechanism for aligning models with normative cultural behavior. Methods such as Reinforcement Learning from Human Feedback, DPO (Rafailov et al., 2023), and Demonstration Iterated Task Optimization (DITTO) (Shaikh et al., 2025) optimize models using comparative judgments of preferred behavior. Recent datasets move toward this direction: CARE (Guo et al., 2025) introduces multilingual, human-annotated preference judgments over culturally situated responses, including Arab contexts. However, CARE is limited in scale and does not offer comprehensive pan-Arab coverage or a unified evaluation benchmark. As summarized in Table 1, existing resources typically support either instruction tuning or evaluation in isolation, leaving preference-based cultural alignment underexplored at scale.

3 ArabPref Dataset

ArabPref consists of two complementary components: a preference dataset and a multiple-choice question (MCQ) dataset. Figure 1 presents an overview of the data samples. To construct these datasets, we first generate rudimentary preference data (Section 3.1). Building upon this foundation, we then refine the preference dataset through human annotation (Section 3.2) and generate the MCQ dataset for evaluation purposes (Section 3.3).

3.1 Rudimentary Preference Data

Arab League Countries This stage aims to construct a comprehensive behavioral preference dataset covering all 22 member states of the Arab League[1]. Each of these countries possesses a distinct cultural and historical background, which translates into nuanced behavioral patterns in real-world applications.

Cultural Categories This stage aims to establish cultural categories before generating preference data. Initially, all extracted categories undergo a manual review to ensure their contextual appropriateness and cultural relevance. Figure 2 provides

[1] https://en.wikipedia.org/wiki/Arab_League

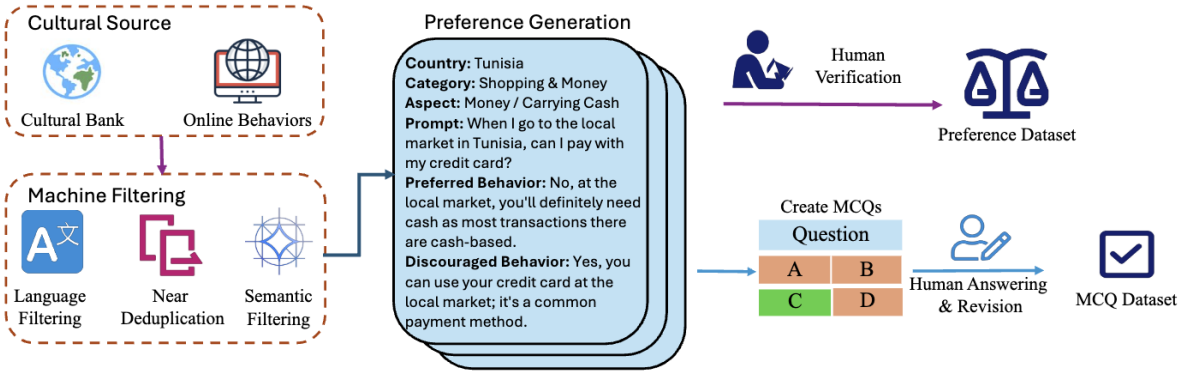


Figure 3: Pipeline of ArabPref data creation. Our data collect cultural bank and web sources, using machine filtering to get preferences (Section 3.1), then human annotators are involved in preference dataset generation (Section 3.2) and MCQ generation (Section 3.3)

summary statistics of the number of aspects for each category. A list of aspect for each category can be found in Appendix C.

Preferred and Discouraged Behaviors The dataset is constructed through a multi-stage pipeline, as illustrated in Figure 3. First, we aggregate behavioral data from two primary sources: (i) manually filtered URL-based behavioral records and (ii) entries from the Cultural Bank (Shi et al., 2024). All raw data is originally in English. We then process these sources to extract distinct country–aspect pairs and split them into training and test sets.

For each country–aspect pair, we use the Gemini-2.5-Flash model to generate corresponding pairs of preferred and discouraged behaviors, producing an initial set of behavioral contrasts. These behavior pairs undergo automated deduplication based on semantic similarity (see Appendix A for details).

We then prompt the same model with the cleaned behavior pairs to generate realistic prompt–chosen-rejected triples. Next, we apply a series of automated filters: language validation (removing non-English samples), structural near-duplicate removal (using the `MinHash` algorithm (Broder, 1997)), and semantic deduplication. This filtering stage refines the dataset to approximately 15530 training and 3836 test samples. After that, all the samples are translated to MSA using LLMs.

3.2 Preference Dataset Generation

This stage focuses on refining the raw preference data from Section 3.1 using human annotators. After the initial automated processing, all test samples undergo a manual verification round. Annotators, who are well-educated native speakers from the 22 target countries, review the samples according to standardized annotation guidelines (see Appendix K). This ensures cultural authenticity and high-quality validation across all aspects of the dataset.

The preference generation guidelines are divided into two stages: 1) **Cultural Preference Quality Check**, which filters the English triples to ensure they genuinely reflect behaviors that are preferred or discouraged in the target country, and 2) **MSA Verification**, which ensures that the translated Modern Standard Arabic (MSA) triples are fluent, natural, and accurately convey the meaning of the original English text. After these stages, we finalize the English dataset by retaining all prompt-chosen-rejected triples accepted by human annotators in Stage 1. For the Arabic annotations, we ask native MSA speakers to revise rejected triples from Stage 2 to ensure high-quality translations.

3.3 MCQ Generation

Based on the defined test aspects, we generate multiple-choice questions (MCQs) to evaluate the model's ability to identify the most preferred or discouraged behavior from four options.

Distractor Selection: For each *country–aspect* pair, we define a set of question templates and populate them accordingly. Each question asks about the preferred or discouraged behavior of a country in a given scenario. To construct the MCQs, the target behavior from each pair serves as the **correct answer**. Three **distractors** are then generated using the following methods: one distractor is drawn from the opposite behavior of the same pair, while two are selected from the same aspect but from

different countries. This ensures that two distractors share the same positivity as the correct answer, and one distractor represents the opposite positivity. We also implement semantic similarity check to ensure that the choices have distinct similarities (Appendix A). Since the distractors are selected from behaviors in different countries, we use specific prompts to rewrite the question, ensuring the naturalness of the language and neutrality in the descriptions of the choices (Appendix G).

Human Annotation and Filtering After distractor selection, we shuffle the choices. Following the same annotation team as in Section 3.2, the annotators are asked to select the best answer among the choices, considering the given aspect and scenario. To ensure reliability, we randomly insert quality control samples throughout the annotation process. Annotators must correctly answer at least 80% of these quality control items to be considered as providing high-quality work. Only MCQs that align with the answer selected during distractor generation are retained. Afterward, we translate all retained MCQs from English to MSA using an LLM, and human annotators revise the MSA translations to ensure high-quality questions. The final distribution of the Preference and MCQ data across countries is shown in Table 2.

Country	Pref Train		Pref Test		MCQ	
	En	Ar	En	Ar	En	Ar
Egypt	1219	697	329	329	62	62
Morocco	1114	679	226	159	59	59
Saudi Arabia	972	565	141	141	52	52
Algeria	958	674	165	131	50	50
Jordan	924	673	254	202	72	72
Lebanon	907	406	199	198	39	39
Tunisia	860	569	317	308	61	61
Yemen	758	513	189	162	62	62
Syria	757	616	128	117	35	35
UAE	720	500	159	156	49	49
Palestine	679	491	135	135	46	46
Libya	628	509	114	79	27	27
Iraq	598	482	135	127	28	28
Oman	572	403	134	102	40	40
Sudan	564	476	80	80	28	28
Bahrain	545	416	147	134	41	41
Qatar	517	389	104	86	49	49
Somalia	488	413	32	31	43	43
Kuwait	476	304	82	81	43	43
Djibouti	436	381	51	41	41	41
Comoros	421	359	82	82	31	31
Mauritania	417	356	97	83	34	34
Total	**15530**	**10871**	**3300**	**2964**	**992**	**992**

Table 2: Statistics of the ArabPref dataset, including English and Arabic preference data and MCQ coverage across 22 Arab League countries.

4 Arabic Cultural Behavior Alignment

4.1 Alignment Methods

We implement four post-training techniques: SFT, DPO, KTO, and DITTO. These methods cover supervised instruction tuning, preference-based alignment, prospect-theoretic optimization, and demonstration-driven iterative refinement.

SFT optimizes the likelihood of human-written demonstrations. Given a dataset $\mathcal{D} = \{(x, y)\}$, the objective is

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\log \pi_\theta(y \mid x)\right].$$

DPO (Rafailov et al., 2023) aligns the model using pairwise preferences. We define the log-probability differences

$$\delta_\theta = \log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^- \mid x),$$
$$\delta_{\theta_0} = \log \pi_{\theta_0}(y^+ \mid x) - \log \pi_{\theta_0}(y^- \mid x).$$

The DPO objective is then

$$\mathcal{L}_{\text{DPO}} = -\log \sigma(\beta\,(\delta_\theta - \delta_{\theta_0}))\,,$$

where θ_0 is a frozen reference model.

KTO (Ethayarajh et al., 2024) optimizes a prospect-theoretic objective. We define a baseline-adjusted reward

$$\Delta_\theta(x, y) = r_\theta(x, y) - \mathbb{E}_{y'}\, r_\theta(x, y'),$$

and apply a value function $u(\cdot)$ capturing asymmetric human sensitivity to gains and losses. The loss is

$$\mathcal{L}_{\text{KTO}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[a_{x,y}\, u(\Delta_\theta(x, y))\big] + C_{\mathcal{D}},$$

where $a_{x,y} \in \{+1, -1\}$ marks desirable vs. undesirable outputs.

DITTO (Shaikh et al., 2024) aligns models through iterative demonstration-driven comparisons. Given demonstration outputs $\mathcal{D}_{\text{demo}}$, we sample model outputs to form comparison pairs (y^+, y^-) and construct a dataset $\mathcal{D}_{\text{comp}}$. Let

$$\delta_\theta(x, y^+, y^-) = \log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^- \mid x).$$

and $\mathcal{L}_\theta(x, y^+, y^-) = -\log \sigma\left(\beta\, \delta_\theta(x, y^+, y^-)\right)$, Therefore, DITTO optimizes

$$\mathcal{L}_{\text{DITTO}}(\theta) = \mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}_{\text{comp}}}\left[\mathcal{L}_\theta(x, y^+, y^-)\right].$$

DITTO updates the model iteratively over rounds, yielding progressively refined policies. Please note

Model & Method	English (En)				Arabic (Ar)			
	Cultural	Literacy	Complete	Avg.	Cultural	Literacy	Complete	Avg.
Llama-3.1-8B-instruct								
Vanilla	3.23	2.63	2.66	2.84	3.13	2.27	2.13	2.51
w/ PALM (SFT)	3.74	**3.18**	**3.06**	3.33	3.36	**3.08**	2.56	3.00
w/ CARE (DPO)	3.85	3.03	2.72	3.20	3.24	2.75	2.41	2.80
w/ CARE (KTO)	3.66	3.09	2.96	3.24	3.17	2.68	2.39	2.75
w/ ArabPref (DPO)	3.78	2.94	2.62	3.11	3.49	2.95	2.58	3.01
w/ ArabPref (KTO)	3.95	3.19	3.01	3.38	3.50	2.69	2.45	2.88
w/ ArabPref+CARE (DPO)	3.78	2.89	2.76	3.14	**3.76**	2.78	2.65	**3.06**
w/ ArabPref+CARE (KTO)	3.85	2.79	2.74	3.13	3.37	2.83	**2.67**	2.96
w/ ArabPref+CARE (SFT)	**4.12**	3.17	2.93	**3.41**	3.36	2.29	2.17	2.57
Qwen-2.5-7B-Instruct								
Vanilla	3.41	2.38	2.92	2.90	3.89	3.10	2.50	3.16
w/ PALM (SFT)	4.02	3.56	3.35	3.64	3.36	3.08	2.56	3.00
w/ CARE (DPO)	4.28	3.50	3.10	3.63	4.21	3.36	2.82	3.46
w/ CARE (KTO)	4.11	3.40	3.37	3.63	4.19	3.18	2.76	3.38
w/ ArabPref (DPO)	4.26	3.38	3.35	3.66	4.01	2.77	2.66	3.15
w/ ArabPref (KTO)	4.33	3.52	3.08	3.64	**4.33**	**3.63**	2.95	3.64
w/ ArabPref+CARE (DPO)	**4.38**	**3.86**	**3.67**	**3.97**	4.27	3.48	3.12	3.62
w/ ArabPref+CARE (KTO)	4.19	3.53	3.15	3.62	4.32	3.78	**3.17**	**3.76**
w/ ArabPref+CARE (SFT)	4.36	3.40	3.12	3.63	3.94	2.74	2.61	3.10
Jais-adapted-7b-chat								
Vanilla	3.41	3.06	2.07	2.85	3.40	2.30	1.79	2.50
w/ PALM (SFT)	3.18	3.95	1.76	2.96	3.26	3.05	1.85	2.72
w/ CARE (DPO)	3.99	3.21	**3.42**	3.54	3.36	4.20	2.87	3.48
w/ CARE (KTO)	3.28	3.14	2.36	2.96	3.37	3.44	2.22	3.01
w/ ArabPref (DPO)	4.33	4.68	3.11	4.04	4.24	**4.21**	**2.96**	**3.80**
w/ ArabPref (KTO)	3.42	3.06	2.77	3.08	3.62	3.29	2.37	3.09
w/ ArabPref+CARE (DPO)	4.39	**4.77**	3.21	**4.12**	4.28	4.20	2.87	3.78
w/ ArabPref+CARE (KTO)	3.52	3.87	2.65	3.35	3.86	3.69	2.40	2.65
w/ ArabPref+CARE (SFT)	**4.50**	4.58	3.05	4.04	**4.42**	4.20	2.75	3.79

Table 3: The learning results of the model preference in English and Arabic, evaluated using an LLM-as a-judge framework in Cultural Appropriateness, Literacy, and Completeness. Best-performing results are highlighted in **bold**.

that we report DITTO results only for analysis, as the method is designed for a few-shot training regime (e.g., up to 100 samples), making direct comparison with other approaches less meaningful.

4.2 Experiment

Model Selection We align two multilingual base models, `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024), a large-scale multilingual instruction model, and `Qwen-2.5-7B-Instruct` (Yang et al., 2024), a powerful, instruction-tuned model for diverse languages, along with one Arabic-centric model, `jais-adapted-7b-chat` (Neha Sengupta and Xing, 2023; Inception, 2024), a model adapted on Llama-2 and specifically fine-tuned on rich Arabic, English and code datasets.

Alignment Data We use the preference training data of our proposed `ArabPref` dataset for cultural alignment. We also incorporate the `CARE` dataset (Guo et al., 2025), which is a multilingual preference dataset spanning Chinese, Japanese, and Arabic cultures. We select the Arabic portion of `CARE` for reinforcement learning. Furthermore, we use `PALM` (Alwajih et al., 2025a), a comprehensive, human-created Arabic instruction dataset, from which we choose the MSA portion for fine-tuning.

Evaluation Metrics For evaluating the preference dataset, we use two main metrics: win rate and LLM-as-a-judge. The win rate directly compares the original baseline model with the fine-tuned model to determine whether the fine-tuned model generates better responses (win), worse responses (lose), or responses of the same quality (tie). The prompt for evaluation is listed in Appendix G. For evaluating MCQ, we use accuracy. The LLM-as-a-judge method involves evaluating three key criteria for cultural alignment:

- **Cultural Appropriateness** (scale 1–5): This criterion evaluates whether the text aligns with cultural norms, values, and sensitivities. It ensures the content is respectful and culturally

accurate, avoiding stereotypes, offensive references, or misunderstandings.

- **Literacy** (scale 1–5): This metric assesses the readability, grammar, and sentence structure of the text. Literacy measures how smoothly the text flows and whether any grammatical or stylistic errors hinder comprehension.

- **Completeness** (scale 1–5): This criterion checks whether the text fully addresses all necessary aspects of the question or task. A complete response should include all relevant information and comprehensively cover the topic.

We also ask humans to annotate the responses; the correlation is available in Appendix I.

4.3 Experiment Results

Preference Generation Table 3 reports LLM-as-a-judge evaluation results on the `ArabPref` and `ArabPref+CARE` preference datasets across Cultural Appropriateness, Literacy, and Completeness for English and Arabic. For LLaMA-3.1-8B-Instruct, the vanilla model achieves average scores of 2.84 in English and 2.51 in Arabic, which are improved to 3.41 with `ArabPref+CARE` (SFT) and 3.06 with `ArabPref+CARE` (KTO), respectively. On Qwen-2.5-7B-Instruct, which exhibits the strongest vanilla performance among the three models (2.90 in English and 3.16 in Arabic), preference-based alignment yields further gains, reaching 3.97 in English with `ArabPref+CARE` (DPO) and 3.76 in Arabic with `ArabPref+CARE` (KTO). These improvements are particularly pronounced in Cultural Appropriateness, where Qwen improves from 3.41 to 4.38 in English and from 3.89 to 4.32 in Arabic. Notably, Jais-adapted-7B-Chat shows the largest relative improvements, increasing its English average from 2.85 to 4.12 with `ArabPref+CARE` (DPO) and its Arabic average from 2.50 to 3.78, corresponding to gains exceeding +1.25 points in both languages. Overall, while all models benefit from preference-based alignment, the results indicate that models with weaker initial performance—such as Jais—gain the most from `ArabPref`-based preference supervision. The results across countries and categories can be found in Appendix D.

MCQ Answering Table 4 shows that both models benefit from preference-based alignment, with Qwen-2.5-7B-Instruct demonstrating stronger performance and larger gains compared to Jais-adapted-7B-chat. Jais-adapted-7B-chat improves by +10.0 points in English and +10.7 points in Arabic with `ArabPref+CARE` (SFT), while Qwen-2.5-7B-Instruct achieves +6.2 points in English and +1.4 points in Arabic. For the SFT method, we observe a performance drop in preference data (Table 3), indicating potential overfitting. In contrast, reinforcement learning methods show higher gains for Qwen in English and Jais in Arabic, likely due to Qwen's better generalization to English data and Jais's optimization for Arabic, reflecting their respective strengths in each language's specific context. This suggests that fine-tuning methods, particularly reinforcement learning, allow for more targeted improvements in language-specific tasks.

Model & Method	English (EN)		Arabic (AR)	
	Acc	Δ	Acc	Δ
Jais-adapted-7b-chat				
Vanilla	49.6	–	39.4	–
w/ PALM (SFT)	40.4	↓9.2	37.1	↓2.3
w/ CARE (DPO)	51.0	↑1.4	<u>45.6</u>	↑<u>6.2</u>
w/ CARE (KTO)	50.0	↑0.4	40.1	↑0.7
w/ ArabPref (DPO)	52.6	↑3.0	40.2	↑0.8
w/ ArabPref (KTO)	51.7	↑2.1	41.8	↑2.4
w/ ArabPref+CARE (DPO)	<u>53.7</u>	↑<u>4.1</u>	43.1	↑3.7
w/ ArabPref+CARE (KTO)	51.3	↑1.7	42.4	↑3.0
w/ ArabPref+CARE (SFT)	**59.6**	↑**10.0**	**50.1**	↑**10.7**
Qwen-2.5-7B-Instruct				
Vanilla	63.3	–	57.6	–
w/ PALM (SFT)	63.0	↓0.3	56.9	↓0.6
w/ CARE (DPO)	67.5	↑4.2	57.6	↑0.0
w/ CARE (KTO)	66.2	↑2.9	57.8	↑0.2
w/ ArabPref (DPO)	67.0	↑3.7	58.0	↑0.4
w/ ArabPref (KTO)	66.4	↑3.1	57.5	↑0.1
w/ ArabPref+CARE (DPO)	<u>68.6</u>	↑<u>5.2</u>	<u>58.7</u>	↑<u>1.1</u>
w/ ArabPref+CARE (KTO)	67.1	↑3.8	58.3	↑0.7
w/ ArabPref+CARE (SFT)	**69.5**	↑**6.2**	**59.0**	↑**1.4**

Table 4: MCQ accuracy (%) on English (EN) and Arabic (AR) datasets for Jais-adapted-7b-chat and Qwen-2.5-7B-Instruct, before and after the best alignment method (SFT). Best-performing accuracy values are highlighted in **bold** and the second in <u>underline</u>.

5 Discussion

Win Rate in Preference Generation Figure 4 compares the win rates of preference generation across three open-sourced LLMs: `Qwen`, `Llama`, and `Jais`. `Qwen` shows moderate performance with win rates of 67.39% in English and 67.19% in Arabic. `Llama` performs well in English (76.27%) but drops to 63.17% in Arabic. In contrast, `Jais` excels with 76.6% in English and 85.35% in Arabic, along

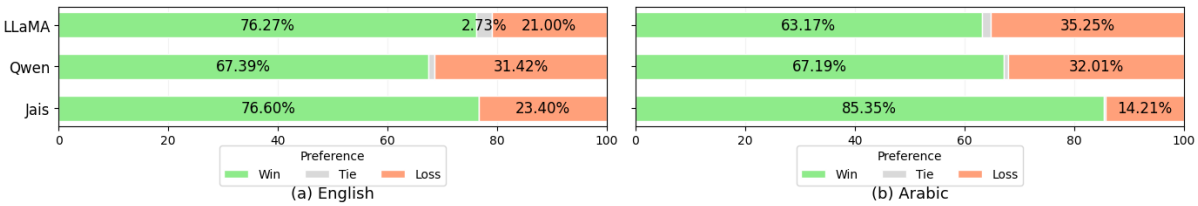


Figure 4: Win–loss–tie comparison on the English and Arabic preference datasets. We report the percentage of wins, ties, and losses for each model, comparing the vanilla and DPO models on the ArabPref+CARE dataset.

with a lower loss rate (14.21%) in Arabic compared to Llama (35.25%). The near-zero tie rate for Jais indicates its ability to decisively resolve preferences, likely due to its bilingual training focusing on Arabic's linguistic nuances, making it particularly strong for Arabic preference generation tasks.

Monolingual vs. Multilingual Training We evaluate training on only English or Arabic data using DPO with the ArabPref dataset, compared to training on the full multilingual dataset. The results show that targeted monolingual training yields variable performance across models. For Llama, the whole dataset consistently performs better across all languages (3.11 for English vs. 2.9/2.84 and 3.01 for Arabic vs. 2.71/2.79). For Qwen, performance is more comparable: English-only training slightly underperforms on Arabic evaluation (3.21 vs. 3.66) while Arabic-only training shows mixed results (3.53 vs. 3.80). This difference likely stems from architectural distinctions: Llama's training methodology may benefit more from diverse multilingual exposure for cross-lingual transfer, whereas Qwen's design appears more adaptable to language-specific specialization while maintaining reasonable cross-lingual capabilities.

Model	Language	Training Data		
		English	Arabic	Whole
Llama	En	2.90	2.84	3.11
	Ar	2.71	2.79	3.01
Qwen	En	3.80	3.98	3.66
	Ar	3.53	3.21	3.47
Jais	En	3.93	3.89	4.04
	Ar	3.73	3.66	3.80

Table 5: Model Performance Comparison with Monolingual vs. Whole Dataset Training

Sampling Efficiency Figure 5 illustrates the sample efficiency of various alignment methods on both Arabic and English, using 10–100 training samples from the ArabPref+CARE dataset for SFT and DPO, and 10–100 samples for DITTO. All methods outperform the baseline, with SFT and DITTO showing significantly better performance than DPO. While these methods offer efficient training with fewer samples, their performance still lags behind full-scale training for both SFT and DPO.

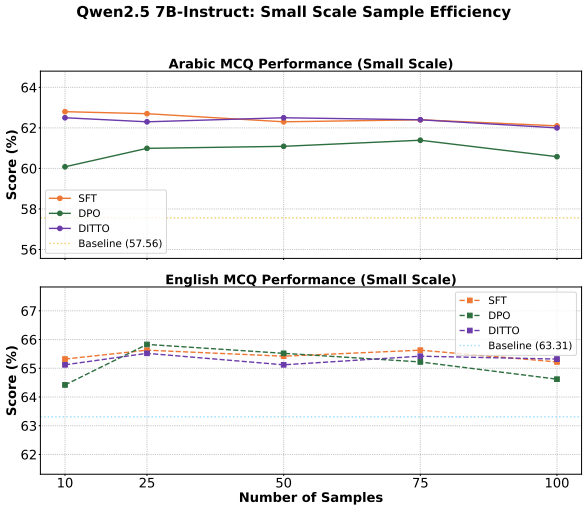


Figure 5: Sample efficiency of preference data alignment on Qwen2.5-7B-Instruct and evaluate on MCQs. Qwen2.5-7B-Instruct performance with SFT, DPO, and DITTO (10–100 samples). Dotted lines indicate zero-shot baselines

6 Conclusion and Future Work

The paper introduces ArabPref, a large-scale, culturally grounded preference dataset covering 22 Arab nations and over 200 topics, designed to enhance Arabic language models' alignment with cultural behaviors. Through experiments with post-training techniques, the study demonstrates significant improvements in cultural alignment when training models on ArabPref compared to existing datasets. By providing both a training dataset and an evaluation benchmark, this work advances culturally aware Arabic language modeling, ensuring that models better reflect regional values and norms. Future work will focus on expanding the dataset to include more regional dialects and subcultures.

Limitations

While ArabPref provides comprehensive coverage of cultural preferences across 22 Arab countries, it may not capture the full cultural diversity within each nation, especially at the regional or local level. The focus on behavioral preferences may overlook other important cultural dimensions, such as emotional expression, non-verbal communication, or specific context-dependent behaviors. Moreover, although the MCQ-based evaluation framework offers useful insights, it may not fully capture the complexity of cultural alignment in more dynamic, open-ended, or real-world scenarios where models need to respond flexibly to diverse and evolving cultural contexts. As such, further research is needed to develop more nuanced evaluation methodologies and expand the dataset to address these gaps.

Ethics and Broader Impact

The development and use of ArabPref aim to enhance cultural alignment in Arabic language models, promoting more respectful and accurate interactions across diverse Arab contexts. However, it is important to acknowledge that cultural behaviors are dynamic and diverse, and any dataset, including ArabPref, can only represent a portion of the full spectrum of cultural practices. Ethical considerations must be taken into account when deploying these models, especially to avoid reinforcing stereotypes or biases. Furthermore, ongoing efforts are needed to ensure the dataset reflects evolving cultural norms and values, with an emphasis on inclusive and sensitive data collection.

References

Marwa Abdulhai, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752, Miami, Florida, USA. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025a. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

Fakhraddin Alwajih, Abdellah El Mekki, 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 1 others. 2025b. 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.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran A. Q. Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged S. Al-Shaibani. 2024. CIDAR: Culturally relevant instruction dataset for Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12878–12901, Bangkok, Thailand. Association for Computational Linguistics.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29.

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

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023b. 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.

Kristian Dokic, Barbara Pisker, and Bojan Radisic. 2025. Mirroring cultural dominance: Disclosing large language models social values, attitudes and stereotypes. *Societies*, 15(5):142.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

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.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual human preference learning for cultural awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32854–32883, Suzhou, China. Association for Computational Linguistics.

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.

Inception. 2024. Jais family model card.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, not representation: The unreliability of evaluating cultural alignment in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2151–2165.

Zhaoming Liu. 2025. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, 3(2):224–244.

Bokang Jia Satheesh Katipomu Haonan Li Fajri Koto William Marshall Gurpreet Gosal Cynthia Liu Zhiming Chen Osama Mohammed Afzal Samta Kamboj Onkar Pandit Rahul Pal Lalit Pradhan Zain Muhammad Mujahid Massa Baali Xudong Han Sondos Mahmoud Bsharat Alham Fikri Aji Zhiqiang Shen Zhengzhong Liu Natalia Vassilieva Joel Hestness Andy Hock Andrew Feldman Jonathan Lee Andrew Jackson Hector Xuguang Ren Preslav Nakov Timothy Baldwin Neha Sengupta, Sunil Kumar Sahu and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM web conference 2023*, pages 1907–1917.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

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.

Henrique* Schechter Vera, Sahil* Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 69 others. 2025. Embeddinggemma: Powerful and lightweight text representations.

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

Omar Shaikh, Michelle S. Lam, Joey Hejna, Yijia Shao, Hyundong Cho, Michael S. Bernstein, and Diyi Yang. 2025. Aligning language models with demonstrated feedback. *Preprint*, arXiv:2406.00888.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Bing Song, Jianing Liu, Sisi Jian, Chenyang Wu, and Vinayak Dixit. 2025a. Can large language models capture human risk preferences? a cross-cultural study. *arXiv preprint arXiv:2506.23107*.

Mingyang Song, Mao Zheng, and Xuan Luo. 2025b. Can many-shot in-context learning help LLMs as evaluators? a preliminary empirical study. In *Proceedings of the 31st International Conference on*

Computational Linguistics, pages 8232–8241, Abu Dhabi, UAE. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. *arXiv preprint arXiv:2205.12247*.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

A Data Generation

Automated Filtering Steps in Preference Generation: We perform three sequential filtering steps on the generated behavior pairs. First, **Language Filtering** removes any samples containing non-English text. Second, **Near-Duplicate Removal** applies the MinHash algorithm (Broder, 1997) to identify and remove structurally similar entries. Finally, **Semantic Deduplication** is conducted by using *GemmaEmbedding* (Schechter Vera et al., 2025) to encode the sentences, followed by cosine similarity to measure semantic overlap. Thresholds in the range {0.70, 0.75, 0.80, 0.85, 0.90, 0.95} are evaluated by manual inspection, and a threshold of 0.90 is selected for the optimal balance between redundancy reduction and data preservation. This process results in a raw preference dataset containing 15530 preference pairs for training and 3894 for testing.

Semantic Similarity Filtering in MCQ Generation: To ensure that the four choices are not semantically similar, we use *GemmaEmbedding* and set a similarity threshold of 0.80. If the similarity between any two choices exceeds the threshold, we replace one of the distractor countries and select a new choice. This process continues until no pair of choices exceeds the threshold.

B Implementation Details

LLM Inference For open-sourced LLMs, we follow the methodology of Guo et al. (2025). During decoding, we set the temperature to 0.7 and top-p to 1. The context length is limited to 2048 tokens (`max_model_len=2048`), and we restrict generations to a maximum of 256 tokens (`max_tokens=256`). This constraint is appropriate because the required outputs for preference questions are typically short, and it also helps standardize outputs for subsequent evaluation. All the experiments are conducted on two NVIDIA A100 80GB GPUs.

C Cultural Categories and Aspects

The aspects for each category is listed in Table 6.

D Data statistics

Table 7 shows the number of preference data samples across the training, raw test, and human-filtered test splits, where human annotators filtered prompt-chosen-reject triples, we ask human annotators to check the cultural appropriateness of each triple to filter out the samples that genuinely reflect the local culture. Table 8 presents the distribution of categories in the preference and MCQ dataset.

E Evaluation Criterion

The evaluation criterion for human annotation and LLM-as-a-judge can be found in Table 9.

Category (Count)	Aspects
Social Etiquette (36)	Greetings, Names and Titles, Handshakes, Eye Contact, Introductions, Personal Space, Gestures & Hand Gestures, Expressing Thanks, Expressing Refusal, Seating, Joking & Humor, Public Affection, Conversation / Asking Questions, Language Use, Meeting Time / Punctuality, Public Behavior / Street Interactions, Queuing / Queue Etiquette, Communication Style, Using Arabic/French, Touching Others, Greeting with "Salaam Alaikum", Flexible Time / Punctuality, Urban/Rural Dialect Switching, Respect for Elders, Showing Soles of Feet, Right-Hand Use for Giving/Receiving, Silence and Pauses in Conversation, Indirect Communication / Saving Face, Honor and Shame Sensitivity, Compliments & "Mashallah" Responses, Asking About Family (Rapport-Building), Queue Courtesy for Families/Elders, Greeting Order When Entering a Room, Door Etiquette (who enters first), Respectful Address to Domestic Workers, Sports Events Etiquette (football matches)
Hospitality & Home Life (24)	Visiting Homes, Shoes in Homes, Invitations, Gifts, Hospitality, Thanking Host, Inviting to Eat, Leaving a Gathering, Hospitality to Strangers, Gift-Giving of Sweets, Home Hospitality Rituals, Role of Host in Welcoming, Visiting the Sick (Get-Well Etiquette), Condolences / Mourning Visits, Guest of Honor Seating, Polite Gift Refusal and Re-offer Ritual, Bringing Dates or Fruit to Hosts, Women's Gatherings (Hareem Majlis), Men's Majlis Etiquette, Incense (Bukhoor) for Guests, Gift Wrapping & Color Preferences, Reciprocity / Return-Gift Norms, Hosting Unannounced Visitors, Sharing Food with Neighbors
Religion & Rituals (22)	Religious Sites, Religious Items, Ramadan, Prayer Times / Friday Observance / Friday Prayers, Public Worship Spaces, Festivals / Holidays, Wedding Etiquette, Friday (Jumu'ah) Observance, Henna Traditions, Eid al-Fitr Etiquette (Zakat al-Fitr, greetings), Eid al-Adha Etiquette (Qurbani sharing), Adhan (Call to Prayer) Respect (pausing music), Using "Inshallah / Alhamdulillah" Appropriately, Visiting Shrines / Saints' Tombs, Ramadan: Non-Muslim Daytime Eating Etiquette, Prayer Rooms in Malls/Airports Etiquette, Wudu (Ablution) Facility Etiquette, Visiting Cemeteries Etiquette, Using Honorifics (Sheikh, Hajji), Addressing by Kunya (Abu/Umm), Name Order & "bin/bint", Funeral Attire and Participation
Dining & Food (18)	Table Manners / Dining Etiquette, Bread Etiquette, Food (Dietary Norms), Finishing Food, Tea / Coffee Ritual, Refusing Food / Tea, Street Food, Serving & Drinking Arabic Coffee (Qahwah), Communal Eating from a Shared Dish, Bread is Sacred, Suhoor & Iftar Hosting, Using Bread as an Utensil, Shisha / Hookah Etiquette, Street Coffee/Tea Kiosk Etiquette, Spice/Heat Consideration for Guests, Family Sections in Restaurants, Coffee vs Tea Preference by Region, Fruit Plate Ritual After Meals
Business & Work (10)	Business Etiquette / Business Cards / Business Attire, Dealing with Officials, Business Gift-Giving, Relationship-First Negotiation Style, Decision-Making Hierarchy, Scheduling Around Prayer/Friday, Business Entertaining (non-alcoholic venues), Acceptable Small Talk Topics, Business Card Handling (Arabic side up), Formal Greetings in Email/Messaging
Modern & Technology (9)	Mobile Phones, Texting/Messaging, Social Media Etiquette, Smoking Etiquette, Photography of Children, Saving Contact Names with Titles (phones), WhatsApp Voice Notes Etiquette, Livestreaming in Public Etiquette, Sharing Religious Content Online Etiquette
Gender & Travel (8)	Women Travelers, Men Travelers, LGBTQ+ / Queer Travelers, Gender Interaction, Gender-Separate Queues/Sections, Cross-Gender Friendship Boundaries, Chaperoned Socializing, Cross-Gender Eye Contact Boundaries
Health & Environment (8)	Drinking Water, Littering / Environment, Sacredness of Water & Fountains, Sacredness of Water & Fountains, Public Bathhouse Traditions, Water Conservation Norms, Sauna/Steam Room Gender Separation, Picnic & Park Etiquette
Family & Children (8)	Children / Traveling with Children, Birth Celebration (Aqiqah) Etiquette, Evil Eye (Nazar) Beliefs & Amulets, Elder Care Expectations, Parenting in Public (discipline norms), Visiting New Mothers Etiquette, Adoption & Kinship Care Sensitivities, Birthday Celebration Norms
Politics & Law (7)	Monarchy, Respect for Monarchy & Royal Imagery, Monarchy, Respect for Monarchy & Royal Imagery, National Day Celebrations, Flag Respect and National Symbols, Military Service Respect
Shopping & Money (6)	Carrying Cash, Tourist Pricing, Marketplace Haggling as Social Ritual, Money / Carrying Cash, Shopping / Marketplace Interaction / Bartering, Tipping (Baksheesh)
Charity & Society (6)	Beggars & Street Kids, Charity / Street Giving, Charity (Zakat/Sadaqah) Giving Etiquette, Community Fundraising & Mutual Aid, Recognizing Tribal/Clan Affiliations, Poetry & Oral Storytelling at Gatherings
Personal Appearance (6)	Dress Code, Distinct Traditional Clothing, Mosque & Shrine Dress Expectations, Swimwear & Beach Modesty, Traditional Headwear Norms (keffiyeh/hijab), Perfume/Oud Use in Social Settings
Transport (4)	Taxi Etiquette, Public Transport / Transport Etiquette, Taxi: Front vs Back Seat Choice, Carpooling & Splitting Costs
Safety (3)	Personal Belongings, Street Harassment, Lost and Found
Pets & Environment (3)	Animals & Pets, Pet Dogs in Public Spaces Acceptability, Compassion Toward Stray Animals
Food & Drink (2)	Alcohol, Cafés & Coffee Culture

Table 6: Cultural categories and aspects of each category

F Sample Efficiency

Country	Train Data	Test Data (Raw)	Test Data (Human Verified)
Egypt	1219	329	329
Morocco	1114	237	226
Saudi Arabia	972	160	141
Algeria	958	225	165
Jordan	924	257	254
Lebanon	907	207	199
Tunisia	860	405	317
Yemen	758	205	189
Syria	757	172	128
UAE	720	183	159
Palestine	679	153	135
Libya	628	119	114
Iraq	598	144	135
Oman	572	135	134
Sudan	564	99	80
Bahrain	545	148	147
Qatar	517	109	104
Somalia	488	144	32
Kuwait	476	106	82
Djibouti	436	137	51
Comoros	421	109	82
Mauritania	417	111	97
Total	**15530**	**3894**	**3300**

Table 7: Statistics of the English preference dataset across countries: training data, raw test data, and human-verified test data.

G Prompts for Generation and Evaluation

Preference Generation

You are a cultural expert and a native in Arab countries. Please answer the following question. Make sure your answer is in the same language as the question.
Question: {question}
Answer:

MCQ Answering

Question: {question}
Options:
A. {option a}
B. {option b}
C. {option c}
D. {option d}
Please choose the most appropriate answer by returning only the letter (A, B, C, or D). Only return the letter.

Win Rate

You are given two responses and a golden label. Please decide which response is better, considering both factual correctness and readability.
Response 1: {pred1}
Response 2: {pred2}
Golden Label: {gold$_label$}
Return 1 if the first response is better, 2 if the second response is better, or 0 if both responses are equally good. Return only the number.

Category	Train	Test (Raw)	Test (Filtered)	MCQ
Social Etiquette	3327	716	605	458
Religion & Rituals	1647	440	336	172
Dining & Food	1553	369	283	214
Hospitality & Home Life	1497	386	325	246
Gender & Travel	745	193	159	98
Personal Appearance	707	170	155	90
Business & Work	691	200	145	98
Shopping & Money	540	150	136	76
Modern & Technology	553	135	119	94
Politics & Law	360	104	100	56

Table 8: Top 10 categories by sample count in the preference dataset across the training, raw test, and human-filtered test splits, along with the sample count in the MCQ dataset.

Score	Cultural Appropriateness	Literacy	Completeness
1	Completely misaligned with the cultural context, containing offensive or inappropriate references.	Very poor fluency, riddled with errors, making it difficult to read.	Very incomplete, with critical information missing, making it unusable.
2	Major cultural misunderstandings or insensitive references that impact the overall understanding.	Multiple errors that hinder readability or cause confusion.	Significant omissions that make the instruction/question incomplete or difficult to interpret.
3	Noticeable cultural inconsistencies or slight misinterpretations, but the general meaning remains clear.	Noticeable errors in grammar, spelling, or structure, but the text is still understandable.	Noticeable gaps in information or coverage that leave the instruction/question lacking.
4	Minor cultural inaccuracies that do not significantly affect understanding.	Minor grammatical or stylistic errors that do not significantly affect readability.	Slightly incomplete, with minor missing details that do not impact overall understanding.
5	Fully aligned with the cultural context, showing deep respect for cultural norms and values, with no inaccuracies or misinterpretations.	Perfect fluency, no errors in grammar, spelling, or sentence structure. The text reads smoothly and naturally.	Fully complete, no information is missing, and the instruction/question thoroughly covers the context.

Table 9: Evaluation criterion for human annotation and LLM-as-a-judge

Cultural Appropriateness

You are tasked with evaluating a given piece of text based on its **Cultural Appropriateness**. This criterion assesses whether the text aligns with cultural norms, values, and sensitivities.
Please assess the text using the following scale:
- **1**: The text is completely misaligned with the cultural context, containing offensive or inappropriate references.
- **2**: There are major cultural misunderstandings or insensitive references that significantly impact the understanding of the text.
- **3**: The text has noticeable cultural inconsistencies or slight misinterpretations, but the general meaning remains clear.
- **4**: The text contains minor cultural inaccuracies, but these do not significantly affect overall understanding.
- **5**: The text is fully aligned with the cultural context, showing deep respect for cultural norms and values, with no inaccuracies or misinterpretations.
Question: {question}
Reference Answer: {answer}
Assistant's response: {response}
Please format your rating as follows: "Rating: [[number]]". For example: "Rating: [[3]]".

Literacy

You are tasked with evaluating a given piece of text based on its **Literacy**. This criterion assesses the readability, grammar, and sentence structure of the text. Literacy refers to how smoothly the text flows and whether there are any grammatical or stylistic errors that hinder comprehension. The text should follow standard grammar and syntax while being easy to read and understand, and should answer the question using the same language as in the question.
Please assess the text using the following scale:
- **1**: The text has very poor fluency, with numerous errors, making it difficult to read and understand.
- **2**: There are multiple errors that hinder readability and cause confusion.
- **3**: The text has noticeable errors in grammar, spelling, or structure, but the meaning remains understandable.
- **4**: Minor grammatical or stylistic errors that do not significantly affect readability.
- **5**: The text has perfect fluency, with no errors in grammar, spelling, or sentence structure. It reads smoothly and naturally.
Question: {question}
Reference Answer: {answer}
Assistant's response: {response}
Please format your rating as follows: "Rating: [[number]]". For example: "Rating: [[3]]".

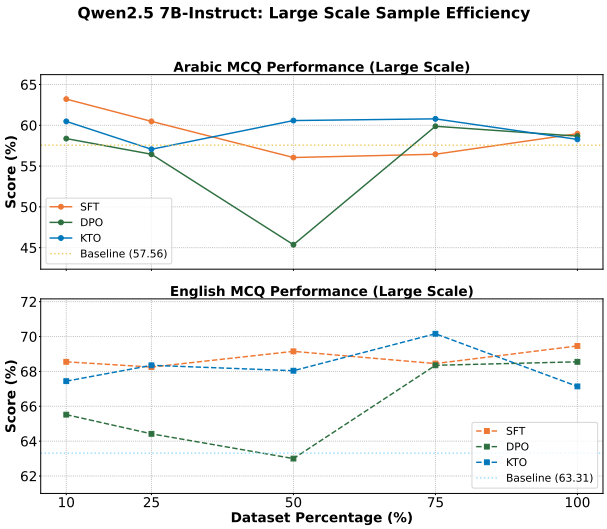


Figure 6: Sample efficiency of alignment methods on Arabic and English cultural preference MCQ evaluation. Qwen2.5-7B-Instruct performance across SFT, DPO, and KTO with 10–100% of training data. Baselines (dotted) show zero-shot performance.

Completeness

You are tasked with evaluating a given piece of text based on its **Completeness**. This criterion assesses whether the text addresses all necessary aspects of the question or task, providing sufficient depth and breadth. A complete response should include all relevant information and fully cover the topic, leaving no critical points unaddressed.
Please assess the text using the following scale:
- **1**: The text is very incomplete, with critical information missing, making it unusable.
- **2**: There are significant omissions that make the text difficult to interpret or incomplete in addressing the topic.
- **3**: The text contains noticeable gaps in information or coverage, but the general meaning is still understandable.
- **4**: The text is slightly incomplete, with minor missing details that do not significantly impact overall understanding.
- **5**: The text is fully complete, providing all necessary details and thoroughly covering the topic without missing any critical information.
Question: {question}
Reference Answer: {answer}
Assistant's response: {response}
Please format your rating as follows: "Rating: [[number]]". For example: "Rating: [[3]]".

MCQ Rewriting Prompt

You are given a JSON object describing a cultural behavior question comparing two or more locations. Revise the four options (opa, opb, opc, opd) to make them more distractive while keeping the correct answer unchanged.
Rules:

1) Correct any grammatical errors in the question.
2) Unless necessary, keep the correct answer's option text unchanged.
3) Among the three distractors:
   - One must describe the opposite behavior of the correct answer.
   - The other two distractors should reflect realistic norms or values from the distractor country (or countries) listed in `"distractor_country"`, given the aspect; if possible, make them different from the correct answer's action.
   - You may modify the distractors, but do not change their positivity.
4) Ensure all options sound plausible and culturally grounded.
5) Avoid explicitly mentioning any country names in the options — use anonymous descriptions.
6) If the original text includes comment adjectives (e.g., "progressive", "modern", "traditional"), rephrase them to neutral language.
7) If the original text includes interpretation of behavior (e.g., "regarded as disrespectful", "it is acceptable..."), remove these and ensure that you only describe the behavior.
8) Return **only** the final modified JSON in the same structure as the input.

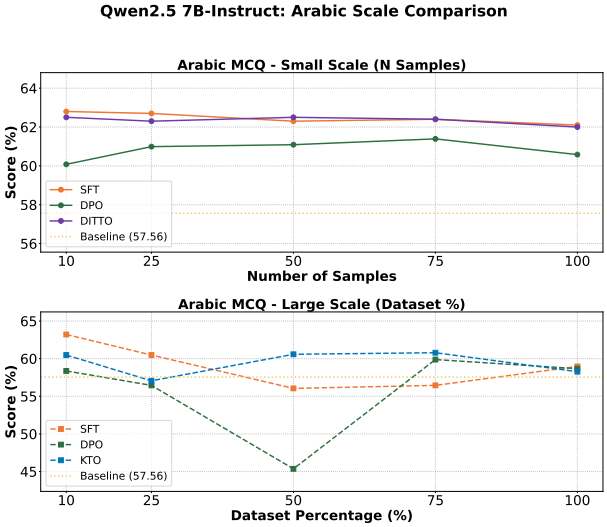


Figure 7: Scale comparison of alignment methods on Arabic cultural preference MCQ evaluation. Qwen2.5-7B-Instruct performance in small-scale (top; 10–100 samples with SFT, DPO, DITTO) and large-scale (bottom; 10–100% data with SFT, DPO, KTO) settings. Baselines (dotted) show zero-shot performance.

H LLM-as-a-judge scoring Across Country and Category for Preference data

LLM-as-a-judge average scores by country and language is in Table 10 and the result across category is in Table 11.

I Correlation Coefficiency

We sample 100 preferences from English and Arabic, with 20 samples from each of the following countries: Egypt, Syria, Morocco, UAE, and Tunisia. Human annotators, all of whom are native Arabic speakers from these countries, are asked to score the samples based on cultural appropriateness, literacy, and completeness. Annotators are provided with the evaluation guidelines outlined in Appendix E to ensure consistency. The current available results are presented in Table 12. Both Arabic and English demonstrated high correlation between the machine scoring and human scoring based on the Spearman's ρ values.

J Accuracy Across Country for MCQ data

The results of MCQ performance across different country of Qwen-2.5-7B Instruct and Jais-adapted-7B-chat is listed in Table 13.

K Annotation Guidelines

The annotation guideline for stage 1 (Cultural Preference Quality Check) and stage 2 (Cultural Preference MSA Verification) is available in Figure 9, 10 and 11. Annotation for stage 3 (MCQ answering) and stage 4 (MCQ verification) is available in Figure 12, 13 and 14.

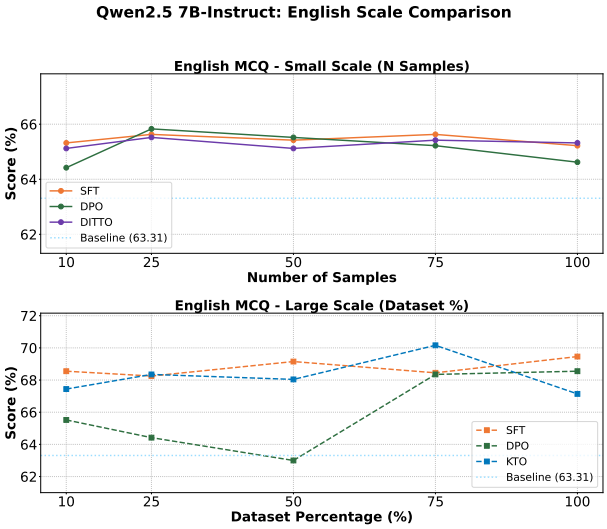


Figure 8: Scale comparison of alignment methods on English cultural preference MCQ evaluation. Qwen2.5-7B-Instruct performance in small-scale (top; 10–100 samples with SFT, DPO, DITTO) and large-scale (bottom; 10–100% data with SFT, DPO, KTO) settings. Baselines (dotted) show zero-shot performance.

Country	Qwen-2.5		w/ Best		Llama-3.1		w/ Best		Jais-adapted		w/ Best	
	EN	AR	EN	AR	EN	AR	EN	AR	EN	AR	EN	AR
Egypt	3.16	3.59	3.77	3.75	2.74	2.85	3.24	3.05	2.93	2.55	4.07	3.88
Morocco	3.26	3.65	3.78	3.74	2.85	2.69	3.43	2.92	2.91	2.39	4.04	3.78
Saudi Arabia	3.13	3.58	3.79	3.60	2.94	2.67	3.37	3.09	2.83	2.58	4.01	3.70
Algeria	2.98	3.59	3.63	3.60	2.84	2.77	3.08	2.87	2.92	2.59	3.98	3.66
Jordan	3.21	3.64	3.74	3.53	2.83	2.61	3.12	2.75	2.86	2.53	4.13	3.67
Lebanon	3.09	3.60	3.83	3.77	2.77	2.65	3.41	3.13	2.87	2.41	4.01	3.75
Tunisia	3.18	3.62	3.80	3.73	2.90	2.73	3.45	3.08	2.82	2.45	4.07	3.80
Yemen	3.15	3.61	3.60	3.74	2.93	2.58	3.21	2.85	2.73	2.47	4.09	3.60
Syria	3.09	3.53	3.68	3.67	2.92	2.65	3.36	3.02	2.74	2.41	4.03	3.94
UAE	3.13	3.65	3.65	3.88	2.91	2.71	2.96	3.05	2.78	2.46	4.06	3.80
Palestine	3.27	3.61	3.72	3.86	2.78	2.70	3.31	2.94	2.86	2.60	4.15	3.88
Libya	3.07	3.62	3.61	3.65	2.98	2.66	3.43	2.71	2.87	2.39	3.95	3.88
Iraq	3.25	3.61	3.79	3.81	2.81	2.54	3.20	2.73	2.92	2.52	4.13	3.80
Oman	3.05	3.60	3.68	3.70	2.79	2.78	3.14	2.84	2.90	2.53	4.05	3.78
Sudan	3.30	3.52	3.69	3.47	2.83	2.79	3.20	3.09	2.88	2.45	3.97	3.71
Bahrain	3.00	3.67	3.84	3.81	2.84	2.63	3.26	3.03	2.89	2.49	4.10	3.90
Qatar	3.19	3.55	3.59	3.63	3.05	2.52	3.46	2.85	2.80	2.46	4.21	3.83
Somalia	3.14	3.69	3.31	3.42	2.89	2.72	3.36	2.56	2.73	2.52	3.99	3.61
Kuwait	3.05	3.50	3.70	3.67	2.89	2.70	3.31	3.03	2.92	2.51	4.12	3.74
Djibouti	3.20	3.62	3.34	3.63	2.93	2.71	3.18	2.81	2.88	2.48	4.03	3.63
Comoros	3.13	3.58	3.62	3.48	2.88	2.70	3.41	2.83	2.59	2.33	4.00	3.81
Mauritania	3.21	3.59	3.60	3.57	2.91	2.61	3.34	2.95	2.72	2.41	4.03	3.75

Table 10: LLM-as-a-judge average scores by country and language. We report results for the base models and their best-performing alignment variants. The best method is ArabPref (KTO) for Llama-3.1, and ArabPref+CARE (DPO) for Qwen-2.5 and Jais-adapted.

Category	Qwen-2.5		w/ Best		Llama-3.1		w/ Best		Jais-adapted		w/ Best	
	EN	AR	EN	AR	EN	AR	EN	AR	EN	AR	EN	AR
Social Etiquette	3.28	3.44	3.64	3.69	2.88	2.54	3.39	3.07	2.90	2.57	4.00	3.81
Religion & Rituals	3.36	3.45	3.75	3.70	2.91	2.49	3.34	2.87	2.91	2.56	4.10	3.71
Hospitality & Home Life	3.27	3.43	3.69	3.71	2.90	2.38	3.33	2.69	2.95	2.54	4.22	3.95
Dining & Food	3.12	3.20	3.30	3.40	2.80	2.36	3.22	2.68	2.81	2.40	3.98	3.63
Gender & Travel	3.28	3.31	3.82	3.74	2.76	2.42	3.18	2.98	2.84	2.62	4.05	3.74
Personal Appearance	3.59	3.39	3.57	3.65	2.62	2.53	2.99	2.72	2.58	2.44	4.16	3.69
Business & Work	3.14	3.44	3.56	3.68	2.84	2.41	3.31	2.80	2.85	2.65	4.07	3.86
Shopping & Money	3.30	3.52	3.48	3.77	2.88	2.54	3.23	3.13	2.79	2.53	4.18	3.85
Modern & Technology	3.69	3.65	3.79	4.08	3.11	2.95	3.38	3.15	2.69	2.50	4.16	3.73
Politics & Law	3.17	3.47	3.79	3.62	2.98	2.83	3.30	3.02	2.67	2.33	3.93	3.82
Health & Environment	3.37	3.44	3.83	3.74	2.65	2.45	3.46	2.91	2.79	2.38	4.12	3.65
Charity & Society	3.45	3.86	3.85	4.14	3.13	2.82	3.46	3.13	2.99	2.69	4.10	4.13

Table 11: LLM-as-a-judge average scores by cultural category and language. We report results for the base models and their best-performing alignment variants. The best method is ArabPref (KTO) for Llama-3.1, and ArabPref+CARE (DPO) for Qwen-2.5 and Jais-adapted.

ρ	Cultural	Literacy	Completeness
en	0.7140	0.8950	0.8312
ar	0.6707	0.7398	0.7801

Table 12: Spearman's ρ values for different categories (Cultural, Fluency, Completeness) for both English (en) and Arabic (ar).

Country	Qwen-2.5		w/ SFT		Jais-adapted		W/ SFT	
	EN	AR	EN	AR	EN	AR	EN	AR
Egypt	56.5	58.1	61.3	58.1	54.8	48.4	58.1	56.5
Morocco	59.3	54.2	52.5	52.5	50.9	54.2	49.2	54.2
Saudi Arabia	63.5	53.9	59.6	63.5	48.1	36.5	59.6	50.0
Algeria	67.7	56.0	72.0	54.0	54.0	30.0	62.0	44.0
Jordan	48.6	50.0	51.4	58.3	45.8	44.4	61.1	54.2
Lebanon	59.0	51.3	64.1	64.1	64.1	59.0	64.1	61.5
Tunisia	47.5	47.5	47.5	45.9	47.5	31.2	60.7	42.6
Yemen	61.3	53.2	59.7	51.6	40.3	45.2	51.6	53.2
Syria	45.7	48.6	62.9	60.0	51.4	34.3	62.9	54.3
UAE	53.1	53.1	67.4	59.2	53.1	34.7	59.2	51.0
Palestine	54.3	45.7	45.7	52.2	50.0	36.9	65.2	50.0
Libya	63.0	44.4	70.4	55.6	29.6	14.8	48.2	37.0
Iraq	50.0	46.4	71.4	60.7	46.4	10.7	64.3	46.4
Oman	70.0	62.5	72.5	70.0	67.5	60.0	82.5	62.5
Sudan	57.1	67.9	67.9	60.7	64.3	53.6	82.1	64.3
Bahrain	51.2	65.9	63.4	61.0	53.7	36.6	58.5	43.9
Qatar	57.1	51.0	53.1	53.1	59.2	46.9	59.2	51.0
Somalia	62.7	44.2	53.5	51.2	46.5	30.2	67.4	46.5
Kuwait	55.8	46.5	62.8	53.5	46.5	41.9	51.2	46.5
Djibouti	41.5	43.9	53.7	43.9	36.6	34.2	48.8	41.5
Comoros	51.6	61.3	51.6	48.4	35.5	25.8	51.6	35.5
Mauritania	67.7	47.1	58.5	55.9	41.2	29.4	52.9	47.1

Table 13: MCQ accuracy (%) by country and language for Qwen2.5-7B-Instruct and Jais-adapted-7B-chat models, evaluated before (vanilla model) and after applying the best alignment method (SFT). Results are presented for both English (EN) and Arabic (AR) across different countries.

- Is the “Discouraged Behavior” really a behavior that locals avoid or frown upon under this aspect?

If the suggestion is **accurate**, the annotator accepts it.
If it is **inaccurate or misleading**, the annotator should reject it.

Pipeline Process

Stage 1 & 2: Cultural Preference Quality Check and MSA verification

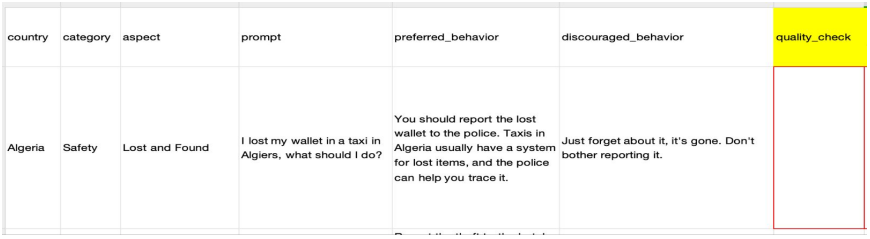
Stage 1: Cultural Preference Quality Check

File Name Structure: stage1&2_country

File Structure (Columns to Fill):

…	**Aspect**	**Prompt**	**Preferred behavior**	**Discouraged Behavior**	**Quality Check**	…
…	[Behavior Aspect]	[Prompt]	[Preferred Behavior]	[Discouraged Behavior]	[Your quality check]	…

Sections to fill



country	category	aspect	prompt	preferred_behavior	discouraged_behavior	quality_check
Algeria	Safety	Lost and Found	I lost my wallet in a taxi in Algiers, what should I do?	You should report the lost wallet to the police. Taxis in Algeria usually have a system for lost items, and the police can help you trace it.	Just forget about it, it's gone. Don't bother reporting it.	

Figure 1. working document overview

Cultural Preference Quality Check

When you receive the file, in the cultural preference overview, the annotators are required to perform the following checks:

- **Understand the social aspect** in the "aspect” column
- **Assess the classification of *Preferred Behaviors* and *Discouraged Behaviors in the “Quality Check” column*** to verify whether each behavior is appropriately assigned according to the cultural norms of the social aspect. You are required to

Figure 9: Annotation for stage 1 (Cultural Preference Quality Check) and stage 2 (Cultural Preference MSA Verification), for creating preference data of ArabPref. Page 1

assign behavior “Accept” and “Reject”

a. Assign “Accept” if the behaviors are culturally appropriate for the aspect and correctly classified as “Preferred” or “Discouraged.”
b. Assign “Reject” if the behavior is irrelevant, incorrectly classified, or does not genuinely reflect a preferred or discouraged practice, or is neutral.

Cultural Preference Review guidelines ***(Red highlights denote critical importance.)***

- If you are not from the region or country specified in the file title, do not work on that specific file!
- The annotators should respond based on his/her life experience in the specified country; using assistance from GPT is not allowed.
- The annotator should have a good understanding of English.

Example:

Region: Egypt
Category: Social Etiquette
Aspect: Gifts

Prompt: I want to visit a family in Egypt. What should I do regarding bringing gifts?

Preferred Behavior:
Bring a small gift when invited to an Egyptian home (e.g., pastries, fruit). Gifts are generally appreciated as a token of appreciation.

Discouraged Behavior:
Arriving empty-handed when invited to a home, giving extravagant or inappropriate gifts.

Quality Check: Accept

Stage 2: MSA Verification

Note: Do this only after you have finished stage 1.

File Name Structure: Country_# (same as in part 1)

File Structure (Columns to Fill):

…	Prompt_ar	Preferred behavior_ar	Discouraged Behavior_ar	msa_quality_check

Figure 10: Annotation for stage 1 (Cultural Preference Quality Check) and stage 2 (Cultural Preference MSA Verification), for creating preference data of ArabPref. Page 2

Stage 2: MSA Verification

Note: Do this only after you have finished stage 1.

File Name Structure: Country_# (same as in part 1)

File Structure (Columns to Fill):

...	Prompt _ar	Preferred behavior_ ar	Discouraged Behavior_ar	prompt_ar_r evision	Preferred behavior_ar _ revision	Discouraged Behavior_ar_ revision
...	[Prompt]	[Preferred Behavior]	[Discouraged Behavior]	[Your Revision]	[Your Revision]	[Your Revision]

Sections to fill

prompt_ar	preferred_behavior_ar	discouraged_behavior_ar	msa_quality_check
ما هي أفضل طريقة للتواصل مع مخرج سينمائي تونسي مخضرم ومحترم بشأن تعاون محتمل، مع الحرص على إظهار التقدير الواجب؟	عند التوجه إلى مخرج تونسي مخضرم يحظى بالاحترام لغرض التعاون، يُفضل أن تُعرب عن إعجابك بأعماله السابقة وتُقرّ بإسهاماته الكبيرة في السينما التونسية. ضَع طلبك بتسليط الضوء على كيفية توافق خبرته ورؤيته مع مشروعك، وكُن مستعدًا للاستماع باهتمام لآرائه واقتراحاته.	عند التواصل مع مخرج تونسي كبير ومحترم بغرض التعاون، يمكنك تقديم اقتراح مشروعك مباشرةً دون الإسهاب كثيرًا في إنجازاته السابقة، على افتراض أنه سيبدي اهتمامًا بناءً على جدارة المشروع وحدها.	

MSA Translation Verification

When you receive the file, the annotators are required to perform the following checks:

- **Read the prompt** in the "prompt" column
- Always, if you have provided a revision to the behavior in stage 1, please do your MSA translation revision based on the "preferred behavior revision" and "negative behavior revision" columns, otherwise, use the behavior from the "preferred behavior" and "negative behavior" columns.
- **Review the "MSA prompt", "MSA preferred", and "MSA_discouraged" columns and provide your revision to the translation from English.**
- **Verify the quality of the MSA translations, and provide your own revision.**

MSA Verification Guidelines: ***(Red highlights denote critical importance.)***

- The annotator should have a good understanding of MSA and English, and should complete the revision in MSA.

Figure 11: Annotation for stage 1 (Cultural Preference Quality Check) and stage 2 (Cultural Preference MSA Verification), for creating preference data of ArabPref. Page 3

Stage 3 & 4: MCQ answering and MSA verification

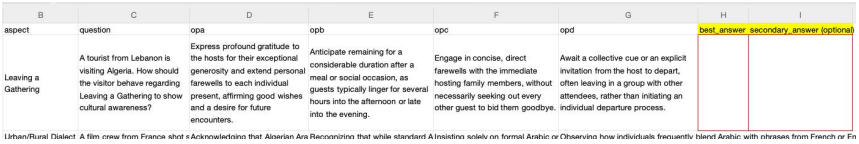
Stage 3: MCQ answering

File Name Structure: stage3_Country

File Structure:

Question	Choice A	Choice B	Choice C	Choice D	Best_ Answer	Secondary_ Answer(s)
[Cultural MCQ]	[Choice A]	[Choice B]	[Choice C]	[Choice D]	[Answer given]	[Answer given]

Sections to fill:



	B	C	D	E	F	G	H	I
1	aspect	question	opa	opb	opc	opd	best_answer	secondary_answer (optional)
2	Leaving a Gathering	A tourist from Lebanon is visiting Algeria. How should the visitor behave regarding Leaving a Gathering to show cultural awareness?	Express profound gratitude to the hosts for their exceptional generosity and extend personal farewells to each individual present, affirming good wishes and a desire for future encounters.	Anticipate remaining for a considerable duration after a meal or social occasion, as guests typically linger for several hours into the afternoon or late into the evening.	Engage in concise, direct farewells with the immediate hosting family members, without necessarily seeking out every other guest to bid them goodbye.	Await a collective cue or an explicit invitation from the host to depart, often leaving in a group with other attendees, rather than initiating an individual departure process.		
3	Urban/Rural Dialect	A film crew from France shot s	Acknowledging that Algerian Ara	Recognizing that while standard A	Insisting solely on formal Arabic or	Observing how individuals frequently	blend Arabic with phrases from French or En	

MCQ Quality Check

When you receive the file for the MCQ quality check, the annotators are required to perform the following checks:

- **Understand the question** in the "question" column
- **Read the choices and answer carefully** in the "Choices A, B, C, D" and "Answer" column.
- **Choose your answer based on your own cultural background.** Please always try to choose only one answer! Always choose the **most preferred/discouraged culturally related** answer. Then mark it in the Best_Answer column.
- **If other answers are also acceptable, please mark them in the "Secondary_Answer(s)" column.** You can choose other multiple answers here. Make sure that they are also culturally related.
- **If there are no correct answers, mark "No Correct Answer" in the "Best_Answer" column** (happens very rarely :)

MCQ Quality Check Guidelines: *(Red highlights denote critical importance.)*

- If you are not from the region or country specified in the file title, do not work on that specific file!
- The annotators should respond based on his/her life experience in the specified

Figure 12: Annotation for stage 3 (MCQ answering) and stage 4 (MCQ verification), for creating MCQ data of ArabPref. Page 1

country; using assistance from GPT is not allowed.

- **Note that you are required to fill in one answer to the "Best_Answer" column; filling in the "Secondary_Answer(s)" column is optional**. In case you are uncertain about deciding among multiple answers. Always choose the best answer that is culturally related to your experience and feeling.

Stage 4: MCQ MSA Translation Verification

Note: Do this only after you have finished stage 3. Contact us before you proceed to stage 4. We will analyse the data from stage 3 and keep only questions with the agreement of the answer(s) with our data.

After completing the MCQ overview, this task requires native speakers to validate the MSA translations for selected MCQs. For each entry, they are given:

- **A question and four choices in English**
- **MSA translation from LLMs.**

English MCQ	Which of the following behaviors is the most preferred in Morocco for language use? A. Arabic is the official language. Learning a few basic Arabic phrases is appreciated. English is understood in tourist areas and among educated people. Avoid using slang or overly casual language in formal settings. B. Darija (Moroccan Arabic) is the most widely spoken language, followed by Berber languages. French is widely used in business and government. English is increasingly common in tourist areas. Learning a few basic Darija phrases is appreciated. C. Using formal Arabic (Modern Standard Arabic) in official settings; using local Jordanian dialect in informal settings; learning a few Arabic phrases is appreciated. Be mindful of religious phrases like "Inshallah" (God willing) and "Alhamdulillah" (Praise be to God). D. While Arabic is the official language, many Syrians involved in tourism or business may speak English. Learning a few basic Arabic phrases is appreciated. Avoid using offensive language or discussing sensitive political or religious topics.
MSA translation	ما هو الـ لوك المف ل لا تخ دام اللـ ة ي الم رب من بي الخيا ات التذلي ؟ A فـ المناطق ا سياحي وبين المتعلمين تجنب ا تخدام الـ غة لعامـ ة أو غير الر مية بشكل مفرط ي لأماكن ا رسمية . اللـ غ العربي ه اللـ ة الـ سمية. و ن ا مسـ حسن تعل بعض لعبارات العر ية الأساسية ثُف م اللغة لإنجليزي B والـ كومة وتـ داد شي ع اللغة لإنجليزية في المناط السي حية. ومن لمستحسـ تعلم بعـ العبارات الأساسية الدارجة بية المغ بية هي للغة الأكـ ر ا تشاراً، تـ يها اللـ ات الأمازيغية. الفرد ية تـ تخدم عـ ى نطاق وا ع في قط عي الأ ما . الـ درجة (العـ C ؛ ومن الـ ستحـ ن تعلم عض العبارا العربية كن حذ أ عند ا تخدام الـ بارات ا دي ية مثل "إن شاء الله" و"لحمد لله" . استخـ ام الـ غة لعر ية لفصحى فـ الأماكن لرسمية؛ واسـ خدام لل جة الأردنـ ة المحلي في الأما ن غير الـ سمي D لم بعض العـ ار ت العربـ ة لأساسية. جنب استخـ ام الألفاظ الـ سيئ أو مناق ة المـ اضـ ع السيا ية أو الـ ي نـ ة الحـ اسة اللغة الـ سمية، إ أن العـ يد من ا سوريـ ن العاملي في السـ احة أو ا عما قد تحدثون ا نجلـ زية. ومن الـ ستحسن تـ . عـ الرغم ن ن اللغة لعربية هـ

The native speaker should review the question and verify the MSA translation.

Figure 13: Annotation for stage 3 (MCQ answering) and stage 4 (MCQ verification), for creating MCQ data of ArabPref. Page 2

File Name Structure: stage4_Country_#

File Structure (Column to fill):

English MCQ	MSA MCQ	Revision
[English MCQ]	[MSA MCQ]	[Your Revision]

Sections to fill:

English MCQ	Arabic MCQ	Revision
Q: Among people in Syria, which action best demonstrates respect for the cultural value of Money / Carrying Cash? A. Proposing to split the restaurant bill evenly among all who dined. B. Insisting on covering the full cost of the meal for the entire dining party. C. Consistently carrying a substantial amount of the local currency in small denominations for all transactions. D. Using digital payment methods for the majority of transactions, with cash reserved only for minor purchases.	س: بين الناس في سوريا، أي من الإجراءات التالية يُظهر أفضل احترام للقيمة الثقافية للمال / حمل النقود؟ أ. اقتراح تقسيم فاتورة المطعم بالتساوي بين جميع من تناول الطعام. ب. الإصرار على تغطية التكلفة الكاملة للوجبة لجميع أفراد المجموعة التي تتناول الطعام. ج. حمل مبلغ كبير من العملة المحلية بفئات صغيرة باستمرار لجميع المعاملات. د. استخدام طرق الدفع الرقمية لمعظم المعاملات، مع الاحتفاظ بالنقود للمشتريات الصغيرة فقط.	

MCQ MSA Translation Check

When you receive the file for the MCQ MSA Translation quality check, the annotators are required to perform the following checks:

- **Read the English question and English choices** in the "English MCQ" column.
- **Review the Arabic MCQ** in the "Arabic MCQ" column.
- **Verify the quality of the MSA translations, and** provide your own MSA revision of the whole MCQ.

MCQ MSA Translation Guidelines: *(Red highlights denote critical importance.)*

- The annotator should have a good understanding of Arabic and English, and should complete the revision in MSA.
- The annotator should ensure the revised MSA translation is correct.

Miscellaneous

Common Mistakes to Avoid

- **Fail to provide a revision**: Revise the Preferred-Discouraged pair if you reject the pair in Stage 1. Always make the behaviors more culturally rel

Figure 14: Annotation for stage 3 (MCQ answering) and stage 4 (MCQ verification), for creating MCQ data of ArabPref. Page 3