Palm: A Culturally Inclusive and Linguistically Diverse Dataset for Arabic LLMs

Anonymous ACL submission

Abstract

As large language models (LLMs) become increasingly integrated into daily life, ensuring their cultural sensitivity and inclusivity is more critical than ever. One significant challenge, however, is the development of comprehensive instruction datasets needed to train and evaluate LLMs effectively. In this paper, we present the results of a year-long community-driven project aimed at creating a dataset that spans *all 22* Arab countries. Named *Palm*, the dataset includes instructions in both Modern Standard Arabic (MSA) and dialectal Arabic (DA), covering a diverse range of 20 topics. Developed by a team of 45 researchers from across the Arab world, *Palm* embodies a broad and inclusive perspective. We also leverage *Palm* to evaluate cultural and dialectal capabilities of several frontier LLMs, uncovering their limitations. *Palm* is publicly available at anonymous-link.

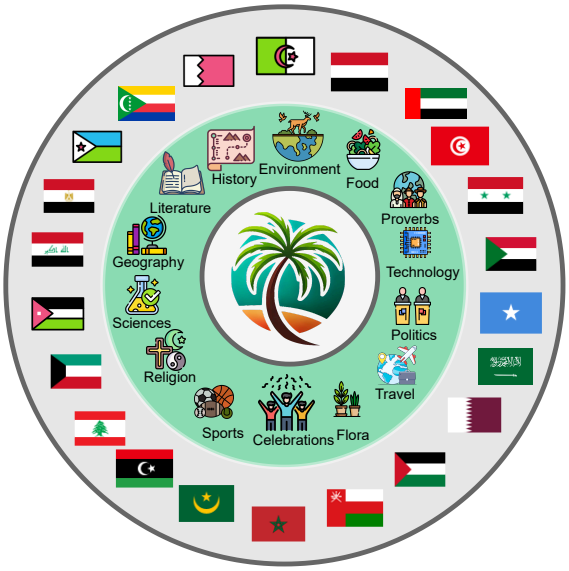


Figure 1: *Palm* covers all 22 Arab countries, each represented by its own flag, across diverse topics.

1 Introduction

Large language models (LLMs) have become pervasive across a wide range of applications. These models are typically trained to predict the next token in a sequence (Radford et al., 2018), followed by a fine-tuning phase where they learn to respond to human prompts using instruction datasets (Ouyang et al., 2022). However, the responses generated by these models are often biased toward the instruction data they were fine-tuned on, which may not reflect the values and cultures of diverse end users (Shankar et al., 2017; Xu et al., 2024). Since creating instruction datasets manually is a labor-intensive and costly process, a common approach to adapting English-trained models (where publicly available datasets exist) to other languages is through translation. However, LLMs trained on translated data often exhibit Western, Anglocentric, and American biases. For example, an Arabic LLM pre-trained on English-to-Arabic translated data suggested having a beer after prayer (Naous et al., 2024), a recommendation that starkly contradicts Arab cultural values, religious practices, and social norms. This example highlights the critical need for building LLMs that are culturally and linguistically aware, which requires the inclusion of more diverse global communities in their development (Adilazuarda et al., 2024).

In this work, we focus on the Arab world and its communities. The Arab world spans a vast geographical region across Africa and Asia, with a population exceeding 450 million people in 22 countries. It is home to a diverse array of local cultures, customs, traditions, political systems, and social practices. The linguistic landscape is equally rich: Arabic exists in three main forms—Classical Arabic, MSA, and Dialectal Arabic (DA), with MSA and DA being the most widely used today. MSA, the standardized form used in formal settings such as literature, media, and official documents, contrasts sharply with DA, which is employed in everyday conversation and varies signifi-

cantly across regions, sometimes classified at the country level (Abdul-Mageed et al., 2024). Several LLMs have been pre-trained for Arabic, including Jasmine (Billah Nagoudi et al., 2023), JAIS (Sengupta et al., 2023), and AceGPT (Huang et al., 2024). These models demonstrate impressive capabilities in generating Arabic across its different forms. However, when it comes to instruction tuning, the datasets used for these models are predominantly machine-generated or machine-translated, introducing errors and biases into the resulting models (Naous et al., 2024). Our work addresses this issue.

More specifically, we introduce *Palm*, the first comprehensive fully human-created Arabic instruction dataset that is both culturally and linguistically diverse and inclusive. *Palm* is the first dataset at the country level to cover *all* 22 Arab countries, spanning 20 culturally relevant topics. What sets *Palm* apart is its inclusion of instructions in both MSA and local dialects, all of which are human-annotated using reliable, country-specific sources. This dataset was developed through a large community-driven project, leveraging local expertise and collective knowledge. *Palm* serves a dual purpose: it can be used for cultural and dialectal instruction tuning of LLMs, as well as for evaluating their cultural competence regarding the Arab world. By addressing the current gap in Arabic representation, *Palm* aims to significantly enhance the inclusion of Arab cultures and dialects in current and future AI systems.

Our contributions in this work are as follows:

1. We introduce ***Palm***, a novel dataset developed through a year-long collaborative community effort, encompassing culturally informed instructions from **all 22 Arab countries**, including both **MSA** and **local dialects**, and covering multiple linguistic forms and topics.

2. We benchmark several open-source and frontier LLMs on ***Palm***, offering an in-depth analysis of model performance on both MSA and dialectal data across different dimensions.

3. We present a comprehensive analysis using three models as evaluative judges, studying alignment between them and providing insights into the reliability of automated evaluations across different models.

4. We conduct a human evaluation to validate the alignment between automatic model evaluations and human judgments, demonstrating the effectiveness of using automated methods for assessing culturally aware and dialect-specific Arabic instructions.

2 Related Work

Arabic Varieties. Arabic, with its rich linguistic diversity, has garnered significant attention in NLP research. This focus has led to the development of various models for encoding (Antoun et al., 2020; Abdul-Mageed et al., 2021; Inoue et al., 2021) and generating (Billah Nagoudi et al., 2023; Sengupta et al., 2023; Huang et al., 2024) Arabic text, achieving impressive results in both understanding and generation tasks (Elmadany et al., 2023; Seelawi et al., 2021; Nagoudi et al., 2023). Arabic primarily consists of two variants: MSA, used in formal settings, and colloquial dialects, which serve as the primary means of daily communication (Abdul-Mageed et al., 2020). These dialects, often mutually unintelligible, present unique challenges for NLP systems (Abdul-Mageed et al., 2024). Despite progress in Arabic NLP, a significant issue persists: the under-representation of colloquial dialects in existing language models. This gap affects NLP systems performance and leads to under-representation of corresponding cultures (AlKhamissi et al., 2024). Consequently, there is a growing need for more comprehensive Arabic language models that can effectively capture the nuances of both MSA and the diverse colloquial dialects, better representing the linguistic and cultural richness of the Arab world.

Cultural Awareness in Language Models. Recent research has employed diverse methodologies to assess the cultural capabilities of language models (Adilazuarda et al., 2024). One approach involves constructing knowledge databases (Keleg and Magdy, 2023; Shi et al., 2024), often leveraging online resources such as Wikipedia (Nguyen et al., 2023; Fung et al., 2024) and web corpora (Nguyen et al., 2023). However, these evaluations may lack sufficient complexity, as much of the web-scraped content used in their creation might have been part of the training data for the evaluated language models (Petroni et al., 2019). Another methodology utilizes socio-cultural surveys like the World Value Survey (Ramezani and Xu, 2023; AlKhamissi et al., 2024). While valuable, these surveys often cover a limited range of cultural concepts. To address these limitations, researchers

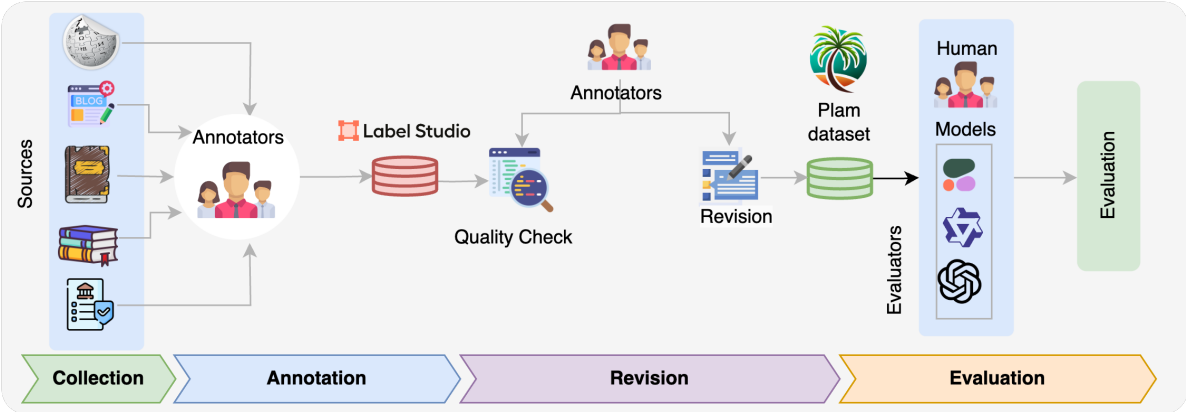


Figure 2: The complete pipeline for *Palm* development. The pipeline begins with the collection of diverse sources, followed by annotation using Label Studio, quality checks, and subsequent revisions. *Palm* is then evaluated by both human and model evaluators, leading to a final assessment phase.

have proposed new datasets and benchmarks specifically designed to evaluate the cultural capabilities of LLMs (Arora et al., 2024; Myung et al., 2024; Singh et al., 2024). Additionally, some studies have focused on measuring Western cultural bias when models are prompted in different languages (Naous et al., 2024).

Arabic Cultural Awareness in Language Models. Recent research has focused on incorporating Arabic language and culture into LLMs, with several studies evaluating cultural capabilities and developing relevant benchmarks (Myung et al., 2024; Singh et al., 2024; AlKhamissi et al., 2024; Naous et al., 2024; Demidova et al., 2024). AlKhamissi et al. (2024) used the World Values Survey to assess LLM alignment with Arab cultural values, finding that underrepresented groups showed reduced alignment. They also noted improved cultural alignment when fine-tuning English models on Arabic data. Naous et al. (2024) revealed a significant bias towards Western entities in LLMs, attributing this to the prevalence of translated pretraining data rather than original Arabic content. To address limitations in country coverage, topics, and dialects, researchers have introduced new datasets. Mousi et al. (2024) proposed a benchmark covering 180 questions across 9 topics and 3 Arabic regions, while Alyafeai et al. (2024) created a dataset of 10,000 instruction-output pairs representing 17 topics in Modern Standard Arabic (MSA). However, existing work has largely focused on automatic annotation and MSA, leaving gaps in dialectal and country-specific cultural representation. Our work, **Palm**, addresses these gaps by providing the first large-scale, fully human-created collection of Arabic cultural instructions and responses from all Arab countries. Palm covers 20 diverse topics and multiple dialects, with careful consideration for local cultures. Table 1 compares Palm with existing Arabic cultural instruction datasets.

3 *Palm* Dataset

Palm is a fine-grained, manually curated, and culturally aligned dataset that includes 22 Arab countries. It features a diverse set of instructions including samples from MSA and DA, all entirely human-generated. It covers diverse topics such as literature, history, food, local geography, and proverbs, providing a comprehensive view of the culture of Arabic local communities. Figure 1 illustrates the covered 22 countries along with the included topics. In the following, we provide a detailed description of the data creation.

3.1 Annotators and Platform

A diverse, trained team of 45 participants from 15 Arab countries contributed to this community-driven data creation project. At least two annotators per country, all with higher education backgrounds in various disciplines, participated in gathering and labeling the instructions. This approach ensured a rich, multifaceted dataset representative of the region's linguistic diversity.

3.2 Data Labeling Process

To facilitate data creation, we employed Label Studio platform (Tkachenko et al., 2020) as our annotation platform. The annotation project was organized by countries and topics, with each annotator granted access to their corresponding country in the

	Multilingual		Arabic Specific		
	AYA	**BLEnD**	**AraDiCE**	**CIDAR**	***Palm* (ours)**
# of Arab Countries	✗	1 out of 16‡	✗	✗	22
MSA Coverage	✓	✓	✓	✓	✓
# of Arabic Dialects	✗	✗	3⋆	✗	9⋆⋆
Cultural Coverage	limited	limited	limited	✓	✓
Human Collected	✗	✓	✗	✗	✓
Human Revised	✓	✗	✓	✓	✓
# of Arabic Instructions	5K out of 204K	3.6K out of 55K	180 out of 45K	10K (100%)	18K (100%)

Table 1: Comparison of other datasets with *Palm* based on the number of Arabic instructions, number of Arabic countries, covered dialects, cultural coverage, human-collected data, and human-revised data. ⋆ AraDiCE covers three dialect regions: Levantine, Egyptian, and Gulf mostly exploiting available NLP datasets. ⋆⋆*Palm* covers nine different Arabic dialects: Egyptian, Jordanian, Mauritanian, Moroccan, Palestinian, Saudi, Syrian, Tunisian, and Yemeni. ‡ BLEnD evaluates cultural sensitivity across diverse multilingual cultures, including only Algeria for Arabic.

Figure 3: An MSA instruction example from *Palm*.

platform. Annotators selected their assigned country and then chose from a list of topics to create instructions following guidelines in Section 3.3.

To ensure data integrity and annotation quality, we implemented a rigorous weekly review process. This involved assessing the annotations' accuracy, source reliability, instruction diversity, and rate of progress. Additionally, we established a dedicated communication channel through Slack to facilitate real-time collaboration among team members. Figure 2 presents the full pipeline for the creation of *Palm* dataset from collection of information to revision of created instructions.

3.3 Annotation Guidelines

Because our community project involves participants from 15 Arab countries, we have established comprehensive and clear annotation guidelines for *Palm*, serving as the core of our work. The participants were carefully trained using these guidelines to ensure consistency and quality across all aspects of the project. These guidelines provide detailed information about the types of instructions, the different topics and domains covered in our dataset, examples of a wide range of instructions across all domains, and samples of trustworthy resources to ensure the authenticity and quality of the data. We asked the annotators to create instructions for two main categories: *general category*, which covers MSA instructions for general knowledge such as sports, celebrations, and technology; and *country-specific category*, where the annotator provides instructions reflecting their country's culture in multiple topics such as customs, local geography, national history, proverbs, food, and celebrations. The country-specific instructions could be expressed in either MSA or the corresponding dialect. For a more detailed exposition of the annotation guidelines, refer to Appendix A.[1]

3.4 Dataset Analysis

Unlike other Arabic instruction datasets listed in Table 1, *Palm* offers several unique features: it covers a larger number of Arabic-speaking countries and dialects, and is entirely created and revised by humans. More specifically, *Palm*, comprises a total of 17,971 instructions collected from 22 Arab countries. It is meticulously designed to reflect the cultural and linguistic diversity of the Arab world, encompassing a wide range of topics and instruction classes expressed in different Arabic varieties. Figures 3 and 4 show examples for MSA and Dialectal (Moroccan) instructions, respectively. In the following, we provide a detailed quantitative analysis of the dataset, highlighting the distribution of instructions across countries, topics, and resource levels.

[1]Our full annotation guidelines manual is available at anonymous-link.

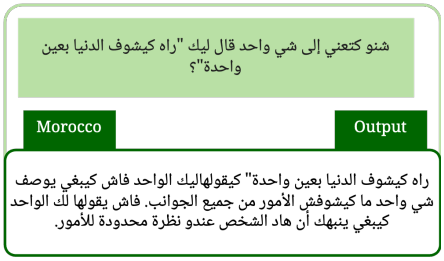


Figure 4: A dialectal (Moroccan) instruction example from *Palm*.

Overall Statistics. We categorize covered countries in *Palm* into *high-resource* (over 500 instructions) and *low-resource* (around 100 instructions) groups for effective data analysis. The high-resource countries contributed significantly to the dataset, each providing an average of over 500 instructions. These countries include Egypt, Jordan, Mauritania, Morocco, Palestine, Saudi Arabia, Sudan, Syria, Tunisia, United Arab Emirates (UAE), and Yemen. Additionally, we included a *General* category, consisting of 1,409 instructions that are not specific to any single country but are general to the Arab world making up 8%. Collectively, these high-resource countries and the General category contributed a total of 15,217 instructions, accounting for approximately 84% of the entire dataset. An important aspect of our dataset is the inclusion of dialectal variations. For the ten high-resource countries, we asked annotators to add approximately 380 examples on average in their respective dialects. This initiative resulted in a total of 4,129 dialectal examples, enriching the dataset with diverse language usage and improving its cultural authenticity and inclusion. Details of the dialectal data distribution are shown in Table D.1 in Appendix D. In contrast, the low-resource countries have fewer instructions, primarily due to limited resources and annotator availability. These countries include Algeria, Bahrain, Comoros, Djibouti, Iraq, Kuwait, Lebanon, Libya, Oman, Qatar, and Somalia. Together, the low-resource countries contribute a total of 1,345 instructions, making up about 8% of the dataset.

***Palm* Arabic Varieties Coverage.** *Palm* is comprised of country-specific data in MSA complemented by country-level dialects, making it linguistically inclusive. Table D.1 provides the distribution of MSA and dialect data for each country, while Figure F.1 showcases selected examples from MSA. Dialects are essential for everyday communication, encompassing both spoken and written forms. To facilitate the study and understanding of textual dialects, the dataset includes numerous examples of dialectal instructions. Figure F.2 illustrates these examples and offers valuable insights into the variations of dialectical expressions across different regions.

Classes. *Palm* offers over 20 diverse classes of instructions and responses, ranging from *multiple choice questions (MCQs)* to *scenario-based inquiries*. Each variety in our dataset is richly represented, to ensure linguistic diversity of the respective countries and topics. Additionally, we include classes such as *critical thinking questions*, *role-playing scenarios*, and *short-answer queries*, providing a comprehensive array of learning materials for models. Overall, the instructions within the dataset are organized into three main categories: i) *informative instructions* which include tasks related to summarizing, listing, explaining, etc.; ii) *interactive instructions* include input/output tasks such as true or false statements, completion questions, and MCQs; and iii) *creative instructions* such as ones for suggesting ideas, writing short notes, and planning. A detailed description of the instructions types is in Table C.1, Appendix C.

4 Evaluation

4.1 Evaluation Description

To demonstrate the efficacy of *Palm*, we employ it as a comprehensive evaluation benchmark for Arabic-aware LLMs. This evaluation serves dual purposes: assessing LLMs' performance across individual countries, topics, and dialects to provide a nuanced measure of Arabic cultural awareness, and offering a robust methodology for future researchers to evaluate model proficiency in handling Arabic culture and dialects. We evaluate 18 Arabic-aware LLMs (listed in Table G.1, Appendix G.1), including GPT-4o, Claude 3.5 Sonnet, Command R+ (CMDR+), QWEN 2.5, JAIS, AceGPT, and LLaMA 3.1, by generating responses (with greedy decoding, temperature=0) to a random sample of *Palm* instructions balanced across all countries, dialects, and topics and assessing the outputs.

The evaluation metrics can be categorized into surface-level evaluation metrics and other metrics that judge the answer and how it relates to the ground truth, including *correctness*, *coherence*, *helpfulness*, and *details*. The description of these metrics is provided in Appendix E.

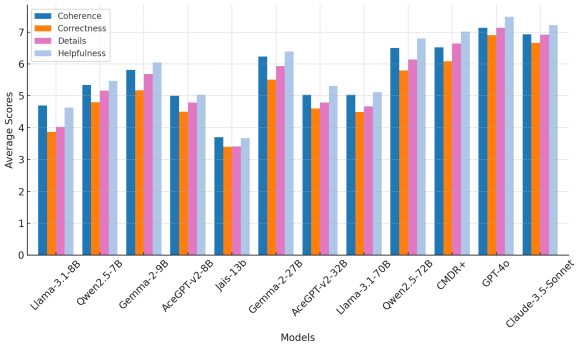


Figure 5: Performance comparison of evaluated LLMs across the four metrics: *correctness*,*coherence*, *helpfulness*, and *details*.

4.2 Surface-Level Evaluation

Following Arora et al. (2024), we employ surface-level attributes to automatically evaluate the generated answers. While these metrics do not assess the *correctness* of the responses, they serve to evaluate three key aspects: **(i) language consistency** between the instruction and generated answer; **(ii) preservation of prompt given dialect (dialectal consistency)** in generated answer; and **(iii) presence of sequence repetitions** in the generated answer. This approach allows us to examine the model's ability to maintain linguistic and cultural fidelity in its responses without directly assessing factual accuracy.

4.3 LLM-as-Judge Evaluation

To assess the *correctness* of the generated responses, we employ the LLM-as-Judge methodology (Zheng et al., 2023). We select a trio of high-performing models in Arabic language tasks: GPT-4o, CMDR+, and QWEN 2.5 72B. We prompt these models to evaluate the quality of generated answers on a scale of 1 to 10 through the built-in evaluation pipeline in LangChain, [2] considering both the given instruction and the ground truth. The specific prompt used for this evaluation is detailed in Figure E.1. For each generated response, we calculate the mean score across the three judges to determine the final rating. To ensure the reliability of our evaluation method, we compute the Intra-class Correlation Coefficient (ICC) among the three judges. The resulting **ICC value of 0.68** indicates a good correlation and robust scoring consistency across the judges. Table G.2 shows the statistics on the subset of data used for this LLM-as-Judge evaluation, totaling **1,926** samples.

[2] https://langchain.com/

In addition to the *correctness* metric, we replicate the same methodology to evaluate the generated answers using three other metrics: *coherence*, *helpfulness*, and *details*.

4.4 Human Evaluation

To ensure the reliability and validity of our automatic evaluation, we conduct a human evaluation on a subset of the generated responses that includes diverse examples across different countries, topics, and both MSA and dialectal responses. We enlisted a group of evaluators from the same data creation team, representing various Arab countries (with their corresponding cultures and dialects). The evaluators were tasked with assessing the selected generated responses from five LLMs based solely on the *correctness* criterion. For quality assurance, each generated response underwent evaluation by a minimum of three independent human evaluators. We subsequently calculate the ICC across all evaluators. The analysis yielded **an average ICC of 0.67** for the human evaluation process, indicating high inter-rater agreement. This subset of data used for evaluation was sampled from the same subset of the LLM-as-Judge evaluation including 92 MSA samples, and 20 dialectal samples for Egypt, Morocco, Syria and Yemen countries.

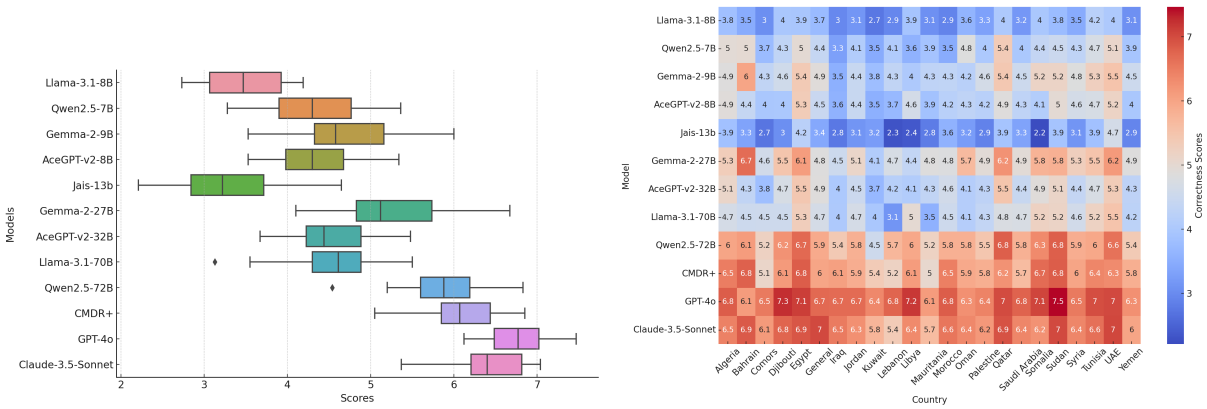
It is worth noting that after computing the correlations between the human evaluation scores and the LLM-as-judge *correctness* scores, we found a strong correlation. The **Pearson correlation coefficient is 0.76**, and the **ICC is 0.78**, lending strong credibility to the LLM-as-Judge results presented in the following section.

5 Results and Discussion

In this section, we analyze results of a subset of 12 LLMs that yield notable insights. Complete results are available in Appendix G.2. Section G.2.1 in the Appendix presents the results of the other models and shows that smaller models generally underperform in providing correct answers to the Arabic cultural questions.

5.1 Surface-Level Results

Table 2 presents the results for the surface-level attributes, namely repetitions, language consistency, and dialectal consistency. The results show that the leading models such as Qwen 2.5, GPT-4o, and Claude-3.5-Sonnet have a negligible rate of repetitions when they are prompted with Arabic instructions, while models such as Gemma-2 and



(a) (b)

Figure 6: LLM-as-judge average correctness scores across Arabic countries.

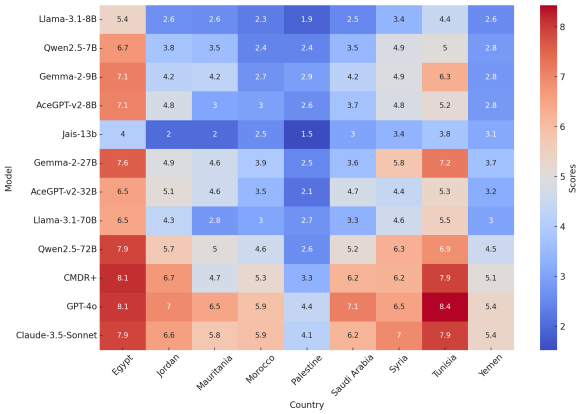


Figure 7: Performance of various models using Correctness score across Arabic dialects.

Llama 3.1 have higher repetition rates where 42% of Gemma-2-9b generated answers have repetitions. For language consistency rate, all models share the same trend; however, the dialectal consistency rate is less than approximately 10% for all LLMs, showing that when Arabic-aware LLMs are prompted with dialectal Arabic instructions, the answers are generated approximately 90% of the time in MSA.

Model	Repetitions	Language Consistency	Dialectal Consistency
AceGPT-v2-32B	0.36	90.45	2.78
AceGPT-v2-8B	0.57	90.71	3.33
Llama-3.1-70B	2.75	91.17	10.56
Llama-3.1-8B	8.20	91.28	10.56
Qwen2.5-72B	0.00	91.33	8.33
Qwen2.5-7B	0.93	90.81	3.89
CMDR+	1.04	90.81	6.67
gemma-2-27b	30.22	90.29	7.78
gemma-2-9b	42.47	90.76	6.11
GPT-4o	0.00	91.07	8.33
jais-13b	0.73	90.71	2.22
Claude-3.5-Sonnet	0.00	91.02	9.44

Table 2: Surface-level attributes rates (%) results

5.2 LLM-as-Judge Results

Overall Results. Figure 5 presents a comparison of the results for the four evaluated metrics using LLM-as-Judge: *coherence*, *correctness*, *details*, and *helpfulness*. Across all models, there is a general trend of higher scores in *coherence* and *helpfulness* compared to *correctness* and *details*. Claude-3.5-sonnet, GPT-4o, and Command R+ consistently achieve the highest scores across all four metrics. Specifically, GPT-4o achieves an average score of 7.14 in *coherence* and 7.48 in *helpfulness*, while Claude-3.5-Sonnet follows closely with scores of 6.93 and 7.22, respectively. Their performance in *correctness* and *details* is similarly strong, highlighting their well-rounded capabilities. Interestingly, even smaller models like Gemma-2-7B show competitive performance in *coherence* and *helpfulness*, though they tend to lag in *correctness* and details. The chart also reveals that as model size increases, there is typically an improvement across all metrics, with the most pronounced gains observed in *correctness* and *details*. This trend underscores the impact of model scale on performance across various aspects of language understanding and generation.

Per-Country Results. Focusing on the *correctness* metric, Figure 6a and Figure 6b describe the *correctness* scores per country. The Figures show that more advanced models like GPT-4o and Claude-3.5-Sonnet consistently demonstrate higher *correctness* scores, often exceeding 6.0, suggesting better cultural understanding. For instance, GPT-4o achieves scores of 7.5 for Syria and 7.3 for Djibouti. In contrast, earlier or smaller models such as Llama-3.1-8B and Jais-13b show lower

correctness scores, frequently below 4.0, with Jais-13b scoring as low as 2.2 for Saudi Arabia. The variation in scores across countries for each model indicates differing levels of cultural competence depending on the specific Arab nation, highlighting the complexities of capturing diverse cultural nuances within the Arabic-speaking world. For example, Claude-3.5-Sonnet scores 7.0 for Yemen and 5.8 for Lebanon, suggesting it may have a better grasp of Yemeni cultural contexts compared to Lebanese ones.

Per-Dialect Results. Figure 7 illustrates the performance of various models using *correctness* scores across several Arabic dialects evaluating the performance of LLMs when prompted with dialectal instructions. Larger models such as GPT-4o, Claude-3.5-Sonnet, and CMDR+ consistently achieve higher scores (often above 7.0) across most countries. For instance, GPT-4o scores 8.1 for Egypt and 8.4 for Tunisia. In contrast, smaller models like Llama-3.1-8B and Jais-13b generally have lower scores, frequently below 4.0. Performance varies across countries, with some dialects appearing more challenging than others. Morocco and Palestine often have lower scores across models, while Egypt and Tunisia tend to have higher scores.

Per-Topic Results. Figure G.2 in Appendix G.2 illustrates the performance of various models across different topics. GPT-4o and Claude-3.5-Sonnet consistently demonstrate superior performance across most topics, with scores frequently above 6.0. For instance, GPT-4o achieves the highest score of 7.4 in the History category, while Claude-3.5-Sonnet scores 7.0 in both History and Proverbs categories. In contrast, models like Llama-3.1-8B and Jais-13b generally show lower performance, with scores often below 4.0 across various topics. The Food category appears challenging for most models, with generally lower scores compared to other topics. Notably, some models exhibit strengths in specific areas. For example, Qwen2.5-72B performs well in Celebrations (6.3) and Science (6.4), while Gemma-2-27B shows strong results in Flora (5.9) and Science (5.9).

Results for the other metrics are presented in Appendix G.2, namely *coherence* (Figure G.1a, Figure G.1b), *details* (Figure G.1c, FigureG.1d), and *helpfulness* (Figure G.1e, Figure G.1f).

5.3 Human Evaluation Results

Table 3 presents the average human-evaluation *correctness* scores for dialectal Arabic instructions in four Arab countries. AceGPT-v2-32B consistently performed well, achieving the highest scores for Egypt (6.47) and Morocco (4.55). Claude-3-5-Sonnet showed strong performance for Morocco (6.23) and Syria (4.65). Llama-3.1-8B and Qwen2.5-72B had middle-point results, while Jais-13b generally scored lowest except for Yemen. Notably, scores varied significantly between countries for each model, suggesting that performance on dialectal Arabic instructions is highly dependent on the specific country and dialect being evaluated.

Country	AceGPT-v2-32B	Llama-3.1-8B	Qwen2.5-72B	Claude-3-5-Sonnet	Jais-13b
Egypt	6.47	4.26	4.71	4.15	4.08
Morocco	4.55	2.87	4.44	6.23	3.10
Syria	3.27	3.40	4.03	4.65	2.27
Yemen	2.13	1.85	2.58	4.28	2.90

Table 3: Average human-evaluation correctness scores for dialectal Arabic instructions per country.

A similar trend is noticed for G.4, where Claude-3-5-Sonnet achieves the highest score for several countries, notably Syria (7.25) and Egypt (6.83). AceGPT-v2-32B and Qwen 2.5-72B also performs well in multiple regions. Performance varies significantly by country.

6 Conclusion

In this work, we introduced *Palm*, a culturally inclusive and linguistically diverse dataset that covers all 22 Arab countries. *Palm* is designed to enhance the cultural capabilities and facilitate benchmarking of Arabic LLMs. Through a year-long, community-driven effort involving 45 researchers from across 15 different Arab countries, *Palm* offers a comprehensive set of instructions that cover both MSA and various regional dialects. Our evaluations using *Palm* demonstrate the importance of culturally tailored datasets in assessing LLMs, highlighting the gaps in existing models when it comes to understanding and generating culturally relevant and dialect-specific responses. *Palm* not only improves the representation of diverse Arab cultures in technology but also provides a benchmark for future work in culturally sensitive and inclusive NLP. By making *Palm* publicly available, we aim to foster continued research and development in the field, ultimately contributing to the creation of more culturally aware language technologies.

7 Limitations

While *Palm* provides a valuable resource for culturally aware and dialectally diverse Arabic LLM training, it has some limitations. For low-resource countries, annotations were often provided by non-native annotators. This can result in more general content that may lack the depth and cultural subtleties that native speakers could offer, affecting the richness of cultural representation.

The diversity of dialects within a single country also poses challenges. Many Arab countries have multiple regional dialects, which require a larger, more geographically diverse team to cover comprehensively. Due to project scale, some dialectal variations may not be fully represented, limiting the dataset's ability to capture the complete linguistic diversity.

Last, while automatic evaluation using LLMs allows for scalable assessment, it has limitations with dialects and Arabic cultural subtleties. LLMs may struggle to accurately evaluate culturally specific or dialectal content, introducing potential biases. Thus, human evaluation remains essential alongside automated methods to ensure reliable results.

8 Ethics Statement

In developing *Palm*, we prioritized ethical considerations, particularly in ensuring cultural sensitivity and inclusivity. All annotations were created by informed participants, respecting local norms and values while aiming to accurately represent diverse Arab cultures. We also ensured data privacy and consent throughout the project, with clear guidelines for annotators to follow. While *Palm* aims to reduce biases in Arabic LLMs, we acknowledge the risk of unintentional cultural bias, especially in regions where native annotators were unavailable. We encourage ongoing community involvement to address these gaps and ensure continuous improvement of the dataset.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. 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, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876, Online. Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. *Preprint*, arXiv:2403.15412.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged 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 and virtual meeting. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2024. Calmqa: Exploring culturally specific long-form question answering across 23 languages. *Preprint*, arXiv:2406.17761.

El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim Elmadany, Alcides Inciarte, and Md Tawkat Islam Khondaker. 2023. JASMINE: Arabic GPT models for few-shot learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16721–16744, Singapore. Association for Computational Linguistics.

Anastasiia Demidova, Hanin Atwany, Nour Rabih, Sanad Sha'ban, and Muhammad Abdul-Mageed. 2024. John vs. ahmed: Debate-induced bias in multilingual LLMs. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 193–209, Bangkok, Thailand. Association for Computational Linguistics.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A challenging benchmark for Arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *Preprint*, arXiv:2402.09369.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. Acegpt, localizing large language models in arabic. *Preprint*, arXiv:2309.12053.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2024. AraDiCE: Benchmarks for dialectal and cultural capabilities in llms.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Preprint*, arXiv:2406.09948.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. 2023. Dolphin: A challenging and diverse benchmark for Arabic NLG. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1404–1422, Singapore. Association for Computational Linguistics.

Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

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*, volume 21 of *WWW '23*, page 1907–1917. ACM.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). 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, Alham Fikri Aji, Zhiqiang

Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *Preprint*, arXiv:1711.08536.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua yu, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *Preprint*, arXiv:2404.15238.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label studio: Data labeling software. *Open source software available from https://github. com/heartexlabs/label-studio*, 2022.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. *Preprint*, arXiv:2404.00929.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Appendices

We provide an addition organized as follows:

- Annotation Guidelines A.
- Details of Instruction Dataset Topics and Categories B.
- Diverse Instruction Formats and Linguistic Variations C.
- Statistics D.
- Evaluation Prompt E.
- Selected Examples F.

A Annotation Guidelines

The key guidelines include the following criteria:

- Use Trustworthy Sources: Annotators are instructed to use only reliable sources, which may include but are not limited to: Wikipedia, Online encyclopedias, Books, Governmental websites, and Specialized websites
- Maintain Objectivity: For certain topics and domains, annotators are asked to provide answers that are objective and based on factual information and established knowledge.
- Avoid Personal Opinions: In domains such as politics and religion, annotators should focus solely on presenting information without incorporating personal beliefs or interpretations.
- Encourage Creativity: Annotators are encouraged to be creative by generating a diverse range of instructions across all domains.

B *Palm* Topics and Domains

Palm focuses on a variety of topics and domains across Arab countries, capturing essential details for understanding daily life and cultural aspects. It is divided into three main categories:

- General domain: refers to a wide range of topics that encompass general knowledge, such as sports, mathematics, food, flora,..etc. These subjects/topics contain information that is not limited to or specific to any particular country or culture, making them relevant.
- Country-specific domain: refers to various topics that focus on knowledge and information about Arab countries. It highlights important details about daily life, cultural practices, traditions, and social norms in these nations.
- Hybrid Domains: These topics combine both country-specific instructions and general knowledge, with a primary focus on Arab countries. While they emphasize regional contexts, they also include general information that can apply to broader themes or encompass Arab nations.

B.1 General domains

As discussed in the previous subsection, general domains includes a wide range of topics that encompass general knowledge applicable across various contexts, such as science, sports, and technology. This includes: Science: Biology, Physics, Chemistry, Math, and Astronomy.

B.2 Hybrid domains

We refer to hybrid domains as topics that span both general and country-specific domains. These topics include sports, technology, politics, religion, travel, food, flora, and environmental life. In the context of the general domain, these topics encompass instructions that provide general knowledge about each subject. In contrast, when discussing them within the country-specific domain, they refer to a set of instructions that capture cultural preferences and relevant information specific to that country.

B.3 Country specific domains

As discussed in the previous subsection, country-specific domains refer to topics that focus exclusively on knowledge and information relevant to particular Arab countries, highlighting unique cultural practices, traditions, and social norms. This includes: Language and Proverbs. A brief description of each topic included in the dataset is provided next:

Language: Language variation in the Arab region is significant, as it includes different dialects that reflect the cultural diversity of the area. We created instructions in both variations MSA and various dialects, which are further categorized by region. To enhance the dataset, the language domain has been added, incorporating these dialectal variations to capture the nuances of regional communication effectively. This domain

covers a wide range of instructions as: i) inquiries about the meanings of specific dialectal words or phrases in certain dialect, ii) guidance on how to use certain dialectal words or phrases, iii) rewriting sentences from dialectal Arabic to MSA and vice versa, and iv) constructing sentences using a certain dialectal word.

Proverbs: Proverbs are important features in the Arab world, reflecting the values, beliefs, and traditions of different communities. They convey life lessons and social norms, often passed down through generations. Including proverbs in the dataset helps capture this cultural heritage and enriches our understanding of local traditions. This domain includes a wide range of instructions as: i) the meanings of specific dialectal proverbs, ii) examples of dialectal proverbs related to particular topics, iii) guidance on how to respond to proverbs in conversation, iv) questions about when to use certain dialectal proverbs, and v) completing specific dialectal proverbs.

Sports: Sports dominant a hug part in the Arab world, reflecting cultural identity and community spirit. Each country has its own unique sporting traditions, from camel racing and falconry in the Gulf region for instance to the widespread for football across all Arab nations. Including sports in the dataset helps capture the diversity of cultural practices. This domain includes a wide range of instructions as i) inquiries about general sports rules, ii) requests for general sports information, such as when a specific event was held or who won the World Cup, and iii) inquiries about sports events information specific to Arab countries, including questions like who won the last match, which team is the best, who are the top nominated players for a particular year, and what are the most popular sports in a specific country, history of a certain sport.

Politics: Political domain is multidisciplinary in nature, drawing from fields such as political science, international relations, public administration, law, economics of all Arab countries. In this domain, the annotators instructed to provide factual and objective information rather than subjective opinions. The set of instructions includes: i) general information about politics and its subdomains, such as voting systems and ideologies,...etc and ii) country-specific inquiries that cover topics like political parties, their establishment dates, relevant laws, forms of government and their structures, as well as information about notable political figures.

Religion: Religion plays a central role in the Arab world, where all major monotheistic faiths (Judaism, Christianity, and Islam) originated. This rich religious heritage influences cultural practices, social norms, and community values across the region. To enhance our dataset, we asked the annotators to create instructions that provide factual and objective information about these three religions, without focusing on specific rituals. The instructions cover: i) Prominent Religious Figures: Names and details about significant figures and their roles in each religion. ii) Significant Religious Sites: Important locations associated with each faith and their historical significance. iii) Religious Institutions: An overview of different religious institutions and their roles within the community.

Food: Arab cuisine is known for its rich use of aromatic and flavorful spices, which is a defining feature of Middle Eastern cooking. Each meal includes a variety of dishes that reflect centuries-old traditions shaped by the trade of ingredients and spices among Arab communities. While there are many similarities across regions, each area has its own unique culinary traditions. The set of instructions includes: i) general inquiries cover questions about the ingredients of specific dishes, nutritional breakdowns for certain meals, BMI calculations, general information about nutrition, and food safety practices. ii) country-specific topics focus on ingredients and cooking methods for particular dishes, providing historical and nutritional information about those dishes. In addition, it also includes details about famous meals, recommendations on what to eat at each meal, names of popular dishes and beverages, and suggestions for recipes using specific ingredients.

Celebration: Celebrations in Arab culture are unique and vibrant, characterized by a strong emphasis on food, music, and the joy of bringing family and friends together. These gatherings reflect the community-oriented spirit of the culture,

where happiness and hospitality play central roles. This domain includes a wide range of instructions, such as: i) Names of Famous Celebrations: National, historic, or religious celebrations, along with their locations and timings, and how each country or city celebrates them. ii) Stories Behind Each Celebration: Details about the traditions associated with each celebration, including food, clothing, and music. iii) Names of Famous Festivals: Information about notable festivals and their specific times.

History: Arab countries have a rich history that spans centuries, beginning with ancient civilizations such as Babylon in Iraq and ancient Egypt, and continuing to the present day. The civilizations that emerged in this region have significantly impacted the world, contributing to advancements in science, mathematics, literature, and art. This deep historical legacy shapes the cultural practices and celebrations observed today, reflecting the diverse traditions and values that have evolved over time. Including historical context in the instruction dataset is crucial, as it helps to understand the significance of cultural practices. This domain includes a wide range of instructions, such as: i) Ancient Civilizations: by Providing the names of different ancient civilizations, detailing their stories, including how they prospered and the reasons for their decline; ii) Wars and Conflicts: by exploring the narratives behind significant wars, including those related to independence and Islamic conquests; iii) Historical Events: by highlight specific events that occurred in particular centuries and their significance; iv) Rulers and Leaders: by discuss various rulers, their stories, and the impact they had on their countries.

Travel: As discussed, the Arab region has a rich history, making it an attractive destination for tourists eager to explore various ancient sites spanning centuries. Additionally, the diverse geography provides excellent opportunities for vacationing. To enhance our dataset, we have included travel instructions that cover: i) Best Places and Historical Landmarks: Identify notable historical landmarks in each city. Types of Tours Available: Outline the different types of tours, including religious tours, sightseeing, and wellness retreats. ii) Visit Planning: Provide guidance on how to plan a day or week-long visit to a specific country, including recommendations on what to eat, where to stay, and where to shop. iii) Historical Sites Access: List different historical sites and provide directions on how to reach them. iv) Shopping and Leisure: Highlight shopping areas and beaches specific to each city. v) Advice and Suggestions: Offer recommendations on what to visit and activities to engage in during the trip.

Flora and Environment Life: The flora and environmental life of the Arab region are integral to its cultural identity and heritage. The diverse ecosystems, ranging from arid deserts to lush coastal areas, support a wide variety of plant and animal species that have shaped local traditions and practices. Including instructions that cover flora and environmental life in the dataset is essential for fostering cultural awareness and understanding the intricate connections between nature and local traditions This domain includes a wide range of instructions, such as: i) Wildlife and national parks: questions about different types of wildlife, including distinct animal species, as well as information on national parks and reserves; ii) Environmental challenges: inquiries regarding environmental issues, such as the impacts of climate change; iii) Types of plants: information about various types of plants, including herbal, poisonous, and agricultural species, along with their habitats; iv) planting suggestions: recommendations on how to effectively plant specific species; v) Endangered and non-endangered animals: questions about endangered species, their habitats, and where they can be found.

Local Geography: Local geography significantly shapes the cultural landscape of the Arab region, influencing everything from settlement patterns to social customs. The diverse terrains ranging from mountains and deserts to rivers and coastal areas affect climate, agriculture, and resource availability, which in turn impact local lifestyles and traditions. By including instructions that cover local geography in the dataset, we can enhance cultural awareness and highlight how landscapes shape cultural preferences and practices. This domain includes a wide range of instructions, such as: i) length of the coastal strip of a specific city/country; ii) terrain diversity; iii) surrounding countries; iv) water resources, and rivers; v) economic impact regarding the geographic location of a country; vi) prominent geographical features for a specific coun-

try; vii) geographic information about cities.

C Diverse Instruction Formats and Linguistic Variations

The instruction datasets cover a diverse set of instruction formats. In summary, these instruction formats include various types of tasks such as:

- **Informative Instructions:** This category includes instructions such as summarizing, listing, and explaining/discussing/defining/describing topics.
- **Interactive Instructions:** This category consists of instructions like true or false questions, completion tasks, question and answer (QA) formats, and multiple choice questions (MCQs).
- **Creative Instructions:** This category includes instructions that involve suggesting, writing, and planning.

Detailed information is indicated in the table C.1 with examples for each case.

This variety of instruction formats makes the datasets flexible and able to meet different user needs, improving engagement and making the content more effective.

D Statistics

D.1 Comparative Analysis of Token Length Distributions Across Models

Figure D.1 presents a comparative analysis of instruction and output token lengths for three LLMs tokenizers: Llama 3.1, Qwen 2.5, and Gemma 2. The figure includes histograms with density curves for instruction and output lengths (top panels) and boxplots for the same data (bottom panels).

In the Instruction Lengths Figure D.1(a), most instructions across all models range between 10 and 20 tokens. Notably, Gemma 2 exhibits a higher concentration of shorter instructions (5–15 tokens), while Llama 3.1 and Qwen 2.5 tend toward slightly longer instructions, with frequencies gradually declining beyond 20 tokens.

For the Output Lengths Figure D.1(b), all models display a peak around 50 tokens, with distributions extending up to 250 tokens. Llama 3.1 tends to generate longer outputs overall, evident from a more pronounced tail toward higher token counts compared to the other models.

The Instruction Length Boxplots Figure D.1(c) show that Llama 3.1 and Qwen 2.5 have similar distributions, with median instruction lengths around 15 tokens and comparable variability. Gemma 2 has a slightly shorter median length and a narrower spread, indicating less variation in instruction lengths.

In the Output Length Boxplots Figure D.1(d), Llama 3.1 again produces the longest outputs, with a median around 90 tokens and outliers extending beyond 250 tokens. Qwen 2.5 and Gemma 2 have median output lengths around 70–80 tokens, with fewer extreme outliers.

Overall, this analysis demonstrates that Llama 3.1 generates longer outputs compared to Qwen 2.5 and Gemma 2, while Gemma 2 often produces shorter instructions. These variations highlight differences in how the models handle input-output lengths.

Additionally, Table D.2 presents the average character lengths for instructions and outputs across various countries in the dataset. Countries like Tunisia, UAE, and Jordan have longer average instruction lengths, while Lebanon and Bahrain feature shorter instructions. For output lengths, Egypt and Somalia have the highest averages, while Qatar and Syria have shorter outputs. This table provides a detailed view of the character length variations across the dataset.

D.2 Lexical Onset Analysis of Prompts

Figure D.2 presents a bar plot illustrating the distribution of first words, defined as space-delimited strings, in instructions within the **Palm** dataset. The most frequent initial word is "ما" ("what"), appearing **4,265** times, followed by "كيف" ("how") with **877** occurrences and "من" ("who/from") with **704** instances. This indicates a strong emphasis on informational and definitional queries. Additionally, verbs such as "اذكر" ("provide"/"list") and "اكتب" ("write") are prevalent, appearing **516** and **348** times respectively, suggesting a focus on task-oriented instructions. The presence of dialectal variations like "شو" ("what") and "شنه" ("what") alongside standard forms underscores the dataset's comprehensive coverage of both MSA and colloquial dialects. The 'Others' category, comprising **2,589** instances, reflects the dataset's diversity in addressing various user queries. Overall, the distribution reveals that *Palm* facilitates information retrieval and explanatory responses, essential for

Category	Sub-Category	Example
Informative	Summarizing	Instruction: أذكر ملخصا عن التركيبة السكانية للحسكة
	Listing	Instruction: أذكر قائمة أغلى اللاعبين فى الدوري السعودي للمحترفين؟
	Explaining Discussing Defining Describing	Instruction: أشرحلي ذ المثل: اشوف الشيباني التاكى الماشاف آفگراش الواگف
Interactive	True or False questions	Instruction: هل العبارة التالية صح أم خطأ؟ Input: أول قمر اصطناعي كويتي أطلق في ٣٠ يونيو ٢٠٢٠ .
	Completion tasks	Instruction: أكمل العبارة التالية: Input: عُرفت..... سابقاً باسم جزيرة الصوة في الشمال الشرقي من إمارة أبوظبي.
	Question Answer (QA) formats	Instruction: ماذا تعرف عن هيكل المحكمة الوطنية في الصومال؟
	Multiple choice questions (MCQs)	Instruction: أختر الإجابة الصحيحة للعبارة التالية: أقيم المعرض الدولي للصيد والفروسية ٢٠٢٢ في إمارة Input: ١. الشارقة ٢.دبي ٣أبوظبي
Creative	Suggesting	Instruction: أقترح ٨ أفكار مشاريع صديقة للبيئة في مصر
	Writing	Instruction: أكتب نبذة عن مدينة سبتة.
	Planning	Instruction: صمم خطة ليوم للإقامة في مدينة القاهرة

Table C.1: The diverse categories of instruction formats with their sub-categories and corresponding examples.

Topic	Celeb.	Env.	Flora	Food	General	Hist.	Lang.	Lit.	Geog.	Politics	Proverbs	Religion	Sports	Tech	Travel	Science	Dialect	Total
Egypt	118	90	116	118	—	105	22	140	177	71	109	109	117	5	135	24	438	1480
Jordan	109	122	100	112	—	74	125	113	98	83	100	98	186	79	114	—	500	1513
Mauritania	69	28	13	109	—	285	105	78	343	40	112	39	37	—	38	1	294	1298
Morocco	32	70	33	256	—	247	139	45	87	206	220	111	245	42	188	9	717	1939
Palestine	46	102	143	181	—	103	202	130	137	60	178	114	109	—	13	—	525	1518
Saudi Arabia	109	200	61	42	—	104	10	140	163	17	39	61	142	100	111	—	296	1299
Sudan	18	—	—	28	—	450	—	40	285	124	—	27	17	—	8	—	—	997
Syria	31	72	201	131	—	202	—	198	65	55	240	100	38	4	202	26	551	1591
Tunisia	12	16	13	40	—	48	51	13	48	11	16	43	50	92	14	160	29	787
UAE	50	21	17	80	—	127	—	132	152	133	—	23	35	34	191	27	26	1049
Yemen	40	16	53	50	—	172	251	119	114	92	262	43	235	25	46	114	592	1746
General	10	46	42	42	—	52	363	45	60	32	—	12	80	13	12	300	—	1409
Algeria	6	—	—	20	—	50	—	1	48	17	—	2	25	—	3	—	—	172
Bahrain	—	—	—	—	100	—	—	—	—	—	—	—	—	—	—	—	—	100
Comoros	7	4	10	8	13	6	5	7	6	3	—	5	8	4	15	—	—	101
Djibouti	2	14	—	8	18	23	3	—	20	7	—	3	—	2	—	—	—	100
Iraq	13	11	—	17	—	11	21	5	—	8	—	—	8	—	19	—	—	113
Kuwait	15	—	—	21	—	16	—	16	16	—	—	16	16	16	15	—	—	147
Lebanon	—	—	—	—	100	—	—	—	—	—	—	—	—	—	—	—	—	100
Libya	—	—	—	1	100	—	—	—	—	—	—	—	—	—	—	—	—	101
Oman	—	—	—	—	100	—	—	—	—	—	—	—	—	—	—	—	—	100
Qatar	—	—	10	10	38	61	—	12	11	—	—	—	28	—	40	—	—	210
Somalia	2	7	—	9	31	6	—	—	13	26	—	—	7	—	—	—	—	101

Table D.1: The overall statistics of instructions number per country, per topic, and per dialect. Celb.: Celebrations, Env: Environment, Hist.: History, Lit.: Literature, Geog.: Geography.

Country	Instruction	Response
Egypt	52.34	1,444.13
Jordan	58.26	334.34
Mauritania	48.00	505.51
Morocco	56.10	816.22
Palestine	54.78	460.43
Saudi Arabia	52.55	444.23
Somalia	34.61	1,639.81
Sudan	39.85	528.33
Syria	45.76	246.92
Tunisia	73.26	957.76
UAE	72.27	280.12
Yemen	48.55	624.52
Algeria	53.13	1,152.94
Bahrain	33.65	1,039.91
Comoros	51.95	479.75
Djibouti	39.43	760.79
Iraq	34.38	1,010.91
Kuwait	56.39	479.41
Lebanon	27.94	336.42
Libya	30.11	613.85
Qatar	59.86	122.10
Oman	31.70	381.61

Table D.2: Average character length for instructions and responses by country.

training LLMs to handle a wide range of culturally nuanced and linguistically diverse inquiries effectively.

E Evaluation Prompt and Metrics

The evaluation metrics are defined as follows:

1. **Correctness**: Measures the factual accuracy of the response in relation to the instruction. A correct response should provide accurate information without errors or misconceptions.

2. **Coherence**: Evaluates the logical consistency and clarity of the response. A coherent response should be well-structured, logically organized, and easy to understand.

3. **Helpfulness**: Determines the utility of the response to the user. A helpful response should provide valuable information that satisfies the user's needs.

4. **Details**: Measures the depth and comprehensiveness of the response. A detailed response should provide sufficient elaboration and cover relevant aspects of the topic.

F Selected Examples

G Evaluation

G.1 Evaluated LLMs

G.2 LLM-as-Judge Results

G.2.1 Ablation Study on Model Size and Performance

In our ablation study, we observed a notable disparity in performance between smaller and larger LLMs. As indicated in Table G.3, the smaller models—those with fewer than 4 billion parameters—achieved an average correctness score of 2.40, whereas the larger models attained a significantly higher average score of 4.01 on the correctness metric. This suggests that smaller LLMs may lack the capacity to effectively handle the complexity of the culturally rich and dialectal content in our dataset. Consequently, we did not include these smaller models in our primary comparisons with larger LLMs. The correctness scores were

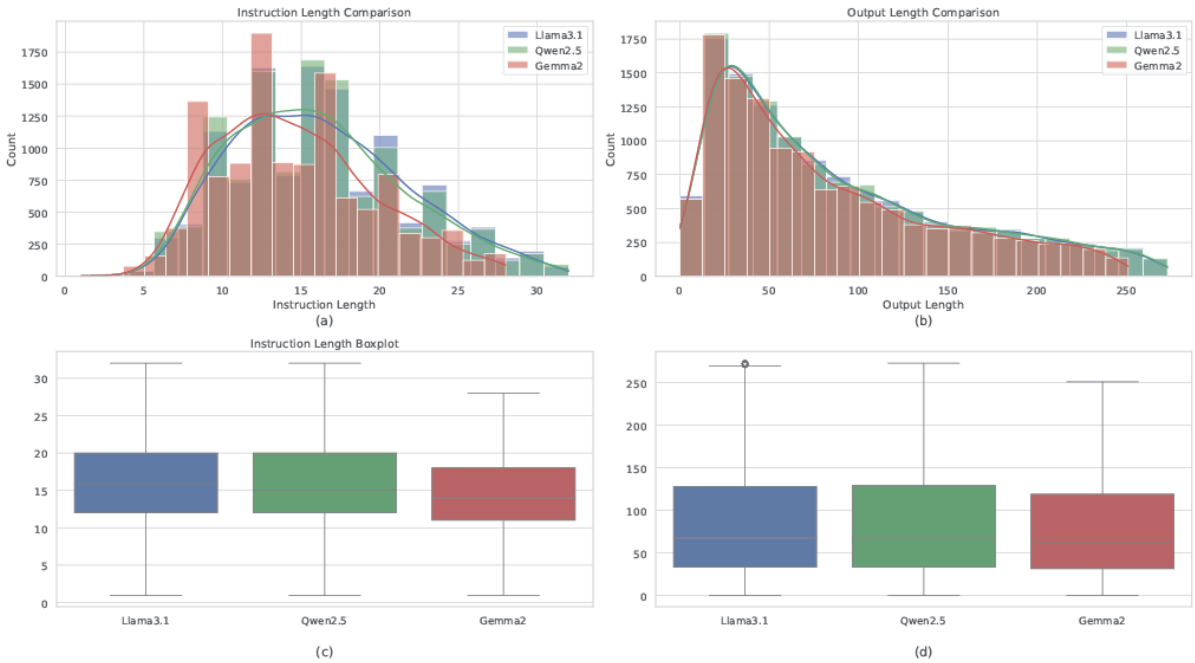


Figure D.1: Token Lengths Comparison.

	LLM	Size	Release Date
Closed LLMs	Claude-3.5-Sonnet	-	Jun. 2024
	GPT-4o	-	Aug. 2024
Open LLMs	Command R+	104B	Aug. 2024
	Qwen2.5-72B	72B	Sep. 2024
	Llama-3.1-70B	70B	Jul. 2024
	AceGPT-v2-32B	32B	Jun. 2024
	gemma-2-27b	27B	Jul. 2024
	gemma-2-9b	8B	Jul. 2024
	Llama-3.1-8B	8B	Jul. 2024
	AceGPT-v2-8B	8B	Jun. 2024
	Qwen2.5-7B	7B	Sep. 2024
	jais-13b	13B	Aug. 2023
	Phi-3.5-mini	3.8B	Aug. 2024
	Qwen2.5-3B	3B	Sep. 2024
	Llama-3.2-3B	3B	Sep. 2024
	gemma-2-2b	2B	Jul. 2024
	Qwen2.5-1.5B	1.5B	Sep. 2024
	Llama-3.2-1B	1B	Sep. 2024

Table G.1: The LLMs used to generate answers for evaluation of PaLM were selected from a list of Arabic-aware models. Each LLM with its corresponding size in Billion parameters and release date. We used the instruct version for all LLMs.

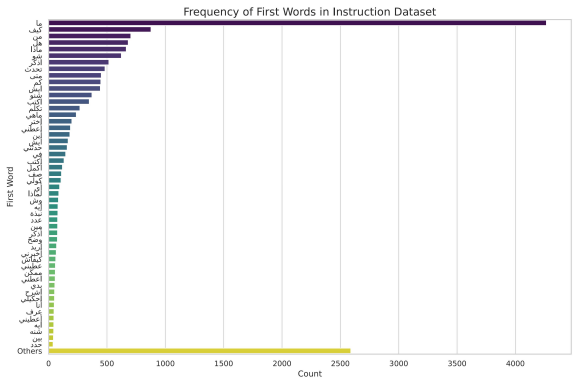


Figure D.2: Frequency of first words.

evaluated using the `Command R+` model as the evaluator, providing a consistent benchmark across all assessments.

G.3 Human Evaluation Results

```
================================ System Message ================================

You are a helpful assistant.

================================ Human Message =================================

[Instruction]
Please act as an impartial judge and evaluate the quality of the response provided
by an AI assistant to the user question displayed below. {criteria}[Ground truth]
{reference}
Begin your evaluation by providing a short explanation. Be as objective as possible.
After providing your explanation, you must rate the response on a scale of 1 to 10
by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

[Question]
{input}

[The Start of Assistant's Answer]
{prediction}
[The End of Assistant's Answer]
```

Figure E.1: Prompt used in our evaluation.

Country	Count
Egypt	146
Jordan	140
Mauritania	59
Morocco	75
Palestine	121
Saudi Arabia	133
Sudan	98
Syria	164
Tunisia	191
UAE	137
Yemen	250
General	20
Algeria	36
Bahrain	5
Comoros	66
Djibouti	56
Iraq	46
Kuwait	75
Lebanon	10
Libya	11
Oman	10
Qatar	40
Somalia	37

Table G.2: Number of samples per country for automatic evaluations.

Country	Phi-3.5-Mini	Gemma-2-2B	Qwen2.5-1.5B	Qwen2.5-3B	LLaMA-3.2-1B	LLaMA-3.2-3B
Algeria	3.66	2.92	2.31	3.42	1.67	2.60
Bahrain	3.00	4.00	2.20	3.40	1.40	2.20
Comoros	1.79	2.70	1.80	2.23	1.85	2.02
Djibouti	3.00	2.89	2.55	2.66	1.56	2.18
Egypt	3.23	3.04	2.72	3.15	1.72	2.28
General	2.45	2.65	2.47	2.55	1.55	2.10
Iraq	2.43	2.74	1.33	1.78	1.64	1.74
Jordan	3.08	2.71	2.31	2.84	1.67	2.18
Kuwait	2.37	2.21	1.54	2.17	1.63	1.73
Lebanon	1.90	1.70	2.20	2.70	1.40	1.50
Libya	3.09	4.18	2.18	2.00	1.55	1.82
Mauritania	2.86	3.04	2.18	2.75	1.69	1.92
Morocco	2.14	2.48	1.73	2.38	1.42	1.96
Oman	3.10	3.30	2.40	3.30	1.60	2.30
Palestine	2.66	2.61	2.18	2.37	1.50	1.97
Qatar	2.49	3.28	1.72	2.60	1.70	2.67
Saudi Arabia	2.41	2.97	2.39	2.63	1.69	1.96
Somalia	2.89	4.51	2.08	2.59	1.92	2.54
Sudan	3.06	3.32	2.24	2.64	1.66	2.30
Syria	2.83	2.54	2.19	2.51	1.61	2.20
Tunisia	3.82	3.24	2.67	3.51	1.84	2.60
UAE	3.27	3.27	2.47	3.14	1.79	2.41
Yemen	2.93	3.02	1.88	2.69	1.69	2.01

Table G.3: Average correctness scores of Small LLMs models across countries using CMDR+.

Country	AceGPT-v2-32B	Llama-3.1-8B	Qwen2.5-72B	Claude-3-5-Sonnet	Jais-13b
Algeria	6.58	5.50	5.33	5.17	6.67
Bahrain	3.25	3.00	3.33	5.00	2.58
Comoros	5.67	5.00	2.17	4.17	4.08
Djibouti	5.33	4.25	5.42	4.08	3.67
Egypt	5.83	5.00	5.17	6.83	4.33
General	5.75	1.92	3.92	4.58	4.45
Iraq	3.83	3.42	4.08	5.75	3.67
Jordan	2.92	2.33	4.25	5.00	4.67
Kuwait	2.83	3.83	2.50	3.58	3.09
Lebanon	4.58	2.42	4.67	5.75	2.92
Libya	3.92	2.58	5.33	5.08	2.33
Mauritania	1.67	2.00	2.75	3.27	1.67
Morocco	3.50	2.33	3.17	4.92	4.92
Oman	4.00	3.00	5.50	4.92	3.50
Palestine	3.25	2.83	3.75	5.17	3.92
Qatar	2.92	2.25	3.75	3.08	3.17
Saudi Arabia	1.09	3.42	5.00	4.25	1.75
Somalia	3.83	4.75	5.17	5.50	3.50
Sudan	4.33	4.17	5.92	4.42	3.58
Syria	5.67	5.50	6.40	7.25	4.17
Tunisia	4.92	6.58	4.08	4.36	3.18
UAE	3.83	2.75	4.83	4.58	2.42
Yemen	5.17	4.75	4.67	5.58	3.75

Table G.4: Average human-evaluation correctness scores for MSA instructions per country.

اذكر أبرز أعمال الشاعر الفلسطيني محمود درويش

Output

Palestine

بكل سرور!

1. ذاكرة النسيان
2. يوميات الحزن العادي
3. لا أريد أن تنتهي هذه الحكاية
4. المزيد من الردود
5. ارتباك العائد
6. لماذا تركت الحصان وحيداً
7. تأثير الفراشة

اكتب نبذة عن الحيوانات النادرة في جزيرة سقطرى

Output

Yemen

في أحدث دراسة جرت في 22 موقعا هاما للطيور في أرخبيل سقطرى، سجلت 291 نوعاً من الطيور، 44 نوعا منها مقيم ومتكاثر في الأرخبيل، كما كشفت دراسة أعدتها منظمة حماية الطيور الدولية، بالتعاون مع مجلس حماية البيئة، عن احتواء أرخبيل سقطرى 13 نوعا مستوطنا من الطيور لا توجد في أي مكان آخر من العالم.

هل القيروان عاصمة تونس؟

1ـ نعم

2 ـ لا

Output

Tunisia

الإجابة:

لا

متى تأسست المكتبة الوطنية الصومالية في مقديشو؟

1. 1970 م
2. 1975 م
3. 1980 م
4. 1985 م

Output

Somalia

الإجابة:

2. 1975

كم عدد المآذن في مسجد الشيخ زايد الكبير؟

Output

United Arab Emirates

عدد المآذن في المسيد أربع مآذن وكل مأذنة طولها 106 أمتار تقريبًا.

كم تمثل الغابات في موريتانيا؟

Output

Mauritania

بكل سرور، إن تقديرات منظمة الأمم المتحدة للأغذية والزراعة (الفاو)، ذكرت أن "الغابات لا تمثل في موريتانيا إلا 0.3 في المئة من إجمالي مساحة البلاد". وأضافت الوزيرة في تصريح لوسائل اعلام محلية أن "الفترة ما بين عامي 1981 و2020 شهدت وتيرة تقلص المساحات الغابية في عموم التراب الوطني مستوى 42 في المئة، إضافة إلى انجراف الغلاف النباتي وفقدانه لدوره في ظرف 10 سنوات".

حدد العبارة الصحيحة:

•القرويين هي أقدم جامعة في العالم تقع في مدينة القاهرة.

•جامعة القرويين هي أقدم جامعة في العالم تقع في مدينة فاس.

Output

Morocco

العبارة الصحيحة هي:

جامعة القرويين هي أقدم جامعة في العالم وتقع في مدينة فاس، بالمغرب. تُعدّ وفقًا لليونسكو وكتاب غينيس للأرقام القياسية أقدم مؤسسة تعليم عالٍ وأول جامعة تمنح إجازة في الطب في العالم، وما زالت تُدرِّس حتى الآن دون انقطاع. ويشير إليها المؤرخون بأنها أقدم جامعة في العالم.

ما هي عملة ليبيا؟

1. الريال الليبي
2. الدرهم الليبي
3. الدينار الليبي

Output

Libya

الإجابة:

3. الدينار الليبي

Figure F.1: MSA examples of instructions from *Palm*.

مين هنن بعض الشعراء والأدباء اللي برزوا بالأدب السوري بالقرن العشرين؟

Output

Syria

من الشعراء والأدباء اللي برزوا بهديك الفترة فخري البارودي وشفيق جبري، وميخائيل نعيمة وعبد الله يوركي حلاق وعمر بهاء الدين الأميري وغيرهم.

مين هو الأديب اللي كان بالعصر العباسي وينتمي لبلاد الشام؟

Output

Syria

الأديب اللي كان موجود بالعصر العباسي وينتمي إلى بلاد الشام هو أبو تمام .

إيه معني " اللي بيته من ازاز ميحدفش الناس بالطوب"

Output

Egypt

معناه متشوفش عيوب غيرك وتتريق عليهم وتركز في غلطاتهم وأنت ممكن يكون عندك نفس الغلطات أو العيوب بس أنت بنفسك متبقاش شايفها

مين هما اللي بيكتبوا باللغة العربية الفصحى في مصر؟

Output

Egypt

اللي بيكتبوا باللغة العربية الفصحى في مصر عادة هما الناس المثقفة والكُتّاب والصحفيين والأساتذة والمسؤولين الحكوميين. لأن الفصحى هي اللغة الرسمية للكتابة والتوثيق في المؤسسات والإعلام والتعليم. أما في الحياة اليومية والكلام العادي، بنستخدم اللهجة المصرية العامية

بدي تعطيني أمثلة على فرق الفن الشعبي الفلسطيني

Output

Palestine

تكرم عينكإفي عندك كتير فرق ومنهم:1- فرقة الفنون الشعبية2- فرقة العاشقين3- فرقة الأصايل4- فرقة صمود5- فرقة جفرا6- فرقة المجد النصراوية7- فرقة ديرتنا8- فرقة تغاريد9- فرقة عشاق الأقصى10- فرقة فنونيات

شو طريقة زراعة الدحنونة؟

Output

Jordan

انثر بذور الدحنون بالتراب و بتكون أمورك تمام وبرعاية بسيطة بتطلع لحالها، أو بتقدر تزرعها بحوض وتحط طبقة تراب وبعدها البذور تبعيت الدحنونة وبعدين شوية تراب فوقها، المهم انه تضل التربة تبعيتها رطبة.

وشي الأماكن المقدسة الموجودة في السعودية؟

Output

Suadi Arabia

الأماكن المقدسة في الاسلام هم ثلاثة، المسجد الحرام والمسجد النبوي، والمسجد الأقصى. السعودية فيها المسجد الحرام في مكة، والنبوي في المدينة.

واخا تقول ليا شكون هي النكافة؟

Output

Morocco

النكافة هي المرة لي كتقابل وكتزين العروسة فليلة العرس، هي كتكون مكلفة بكلشي من الكساوي ديال العروسة حتى نالمكاياح ديالها. و باش يطلع ليك العرس مزيان ضروري تعرفي تعزلي النكافة، و تقلبي على أحسن ماكاين.

Figure F.2: Dialectal examples of instructions from *Palm*.

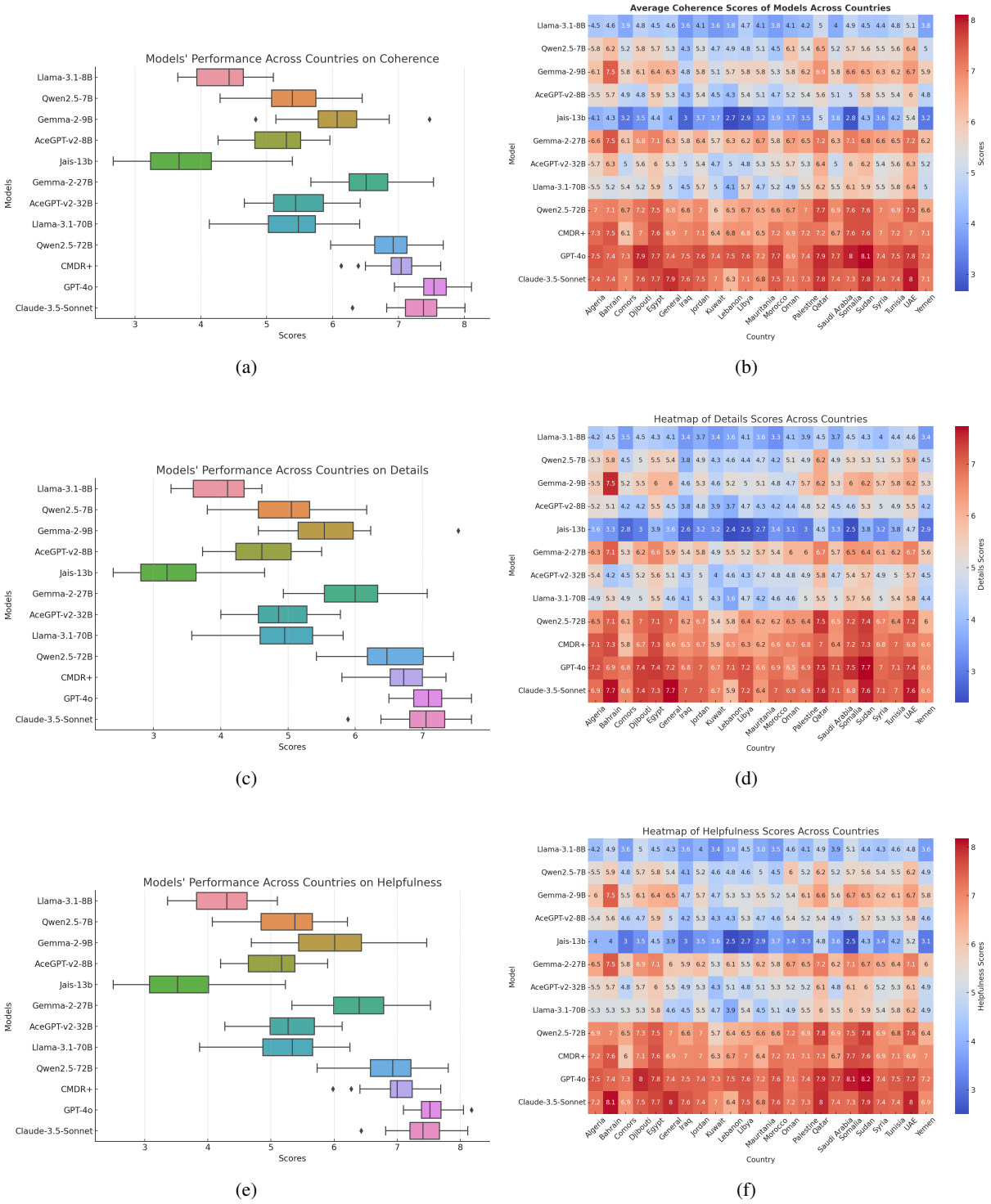


Figure G.1: Comparative analysis of the models across evaluation metrics: Coherence, Details, and Helpfulness.

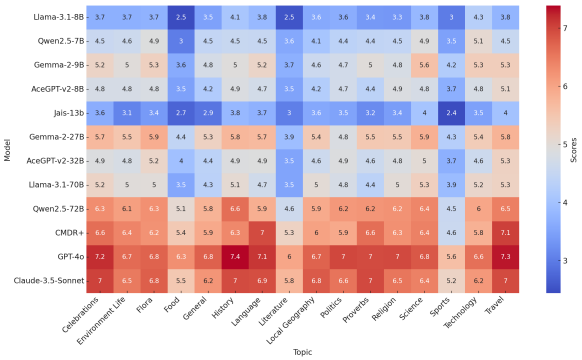
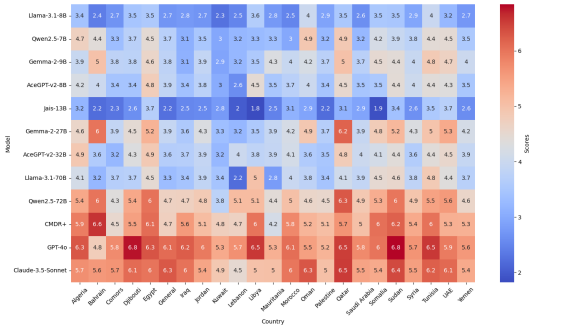
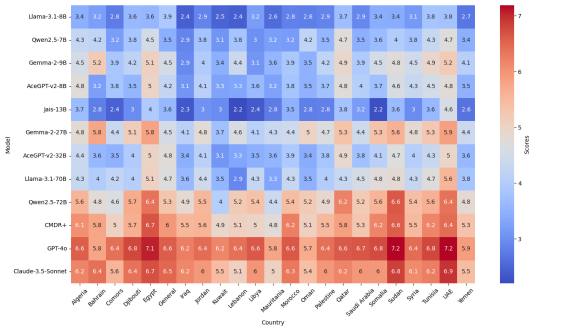


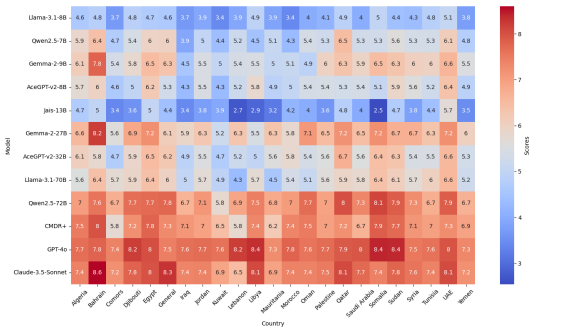
Figure G.2: Performance of various models using Correctness score across different topics.



(a) CMDR+



(b) GPT-4o



(c) Qwen2.5-72B

Figure G.3: LLM-as-judge correctness scores across Arabic countries.