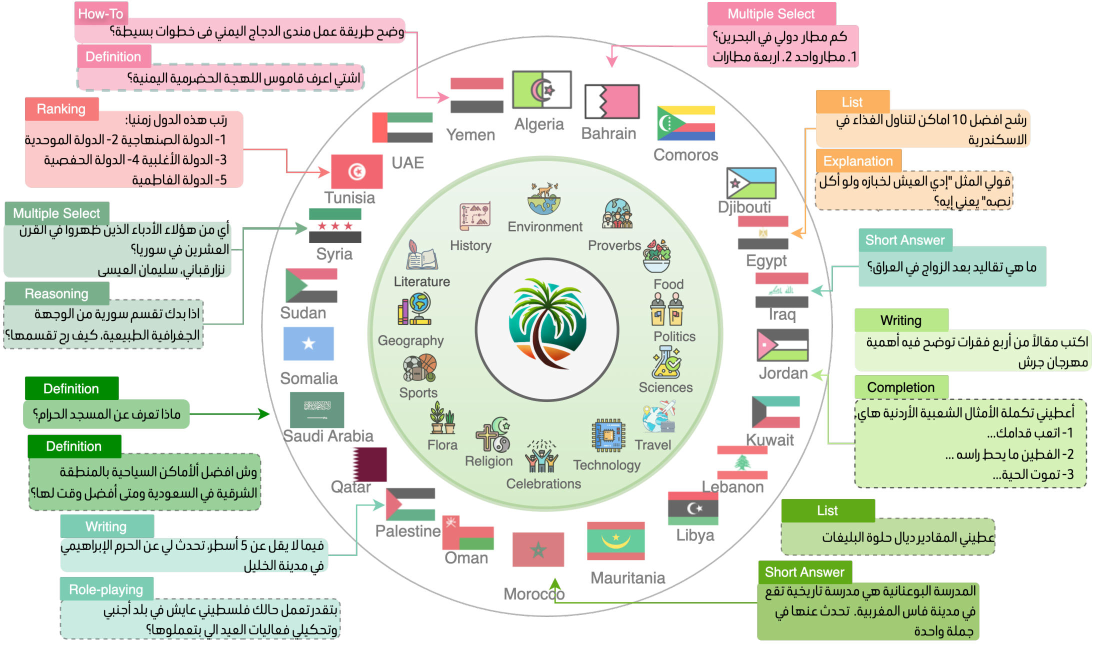
Palm: A Culturally Inclusive and Linguistically Diverse Dataset for Arabic LLMs

Fakhraddin Alwajih[1*], Abdellah El Mekki[1,2*], Samar Mohamed Magdy[1,2], Abdelrahim A. Elmadany[1], Omer Nacar[5], ElMoatez Billah Nagoudi[1,5], Reem Abdel-Salam[7], Hanin Atwany[2], Youssef Nafea[2], Abdulfattah Mohammed Yahya[7], Rahaf Alhamouri[9], Hamzah A. Alsayadi[10], Hiba Zayed[4], Sara Shatnawi[2], Serry Sibaee[5], Yasir Ech-Chammakhy[6], Walid Al-Dhabyani[7], Marwa Mohamed Ali[10], Imen Jarraya[5], Ahmed Oumar El-Shangiti[2], Aisha Alraeesi[2], Mohammed Anwar Al-Ghrawi[11], Abdulrahman S. Al-Batati[5], Elgizouli Mohamed[12], Noha Taha Elgindi[13], Muhammed Saeed[2], Houdaifa Atou[6], Issam Ait Yahia[6], Abdelhak Bouayad[6], Mohammed Machrouh[6], Amal Makouar[6], Dania Alkawi[5], Mukhtar Mohamed[2], Safaa Taher Abdelfadil[2], Amine Ziad Ounnoughene[15], Rouabhia Anfel[16], Rwaa Assi[4], Ahmed Sorkatti[12], Mohamedou Cheikh Tourad[14], Anis Koubaa[17],
Ismail Berrada[6], Mustafa Jarrar[4,18], Shady Shehata[2,3],
Muhammad Abdul-Mageed[1,2,3]

[1] The University of British Columbia, [2] MBZUAI, [3] Invertible AI,
[4] Birzeit University, [5] Prince Sultan University,
[6] UM6P, [7] Cairo University, [8] Prince Sultan University,
[9] JUST, [10] Ain Shams University, [11] Damascus University,
[12] University of Khartoum, [13] Menoufiya University, [14] University of Nouakchott,
[15] National Polytechnic School of Algiers, [16] Full Sail University, [17] Alfaisal University,
[18] Hamad Bin Khalifa University

{fakhr.alwajih, muhammad.mageed}@ubc.ca



How-To
وضح طريقة عمل مندى الدجاج اليمني فى خطوات بسيطة؟

Definition
اشتي اعرف قاموس اللهجة الحضرمية اليمنية؟

Ranking
رتب هذه الدول زمنيا:
1- الدولة الصنهاجية 2- الدولة الموحدية
3- الدولة الأغلبية 4- الدولة الحفصية
5- الدولة الفاطمية

Multiple Select
أي من هؤلاء الأدباء الذين ظهروا في القرن
العشرين في سوريا؟
نزار قباني، سليمان العيسى

Reasoning
اذا بدك تقسم سورية من الوجهة
الجغرافية الطبيعية، كيف رح تقسمها؟

Definition
ماذا تعرف عن المسجد الحرام؟

Definition
وش افضل ألأماكن السياحية بالمنطقة
الشرقية في السعودية ومتى أفضل وقت لها؟

Writing
فيما لا يقل عن 5 أسطر، تحدث لي عن الحرم الإبراهيمي
في مدينة الخليل

Role-playing
بتقدرتعمل حالك فلسطيني عايش في بلد أجنبي
وتحكيلي فعاليات العيد الي بتعملوها؟

Multiple Select
كم مطار دولي في البحرين؟
1. مطارواحد 2. اربعة مطارات

List
رشح افضل 10 اماكن لتناول الغذاء في
الاسكندرية

Explanation
قولي المثل "إدي العيش لخبازه ولو أكل
نصه" يعني إيه؟

Short Answer
ما هي تقاليد بعد الزواج في العراق؟

Writing
اكتب مقالاً من أربع فقرات توضح فيه أهمية
مهرجان جرش

Completion
أعطيني تكملة الأمثال الشعبية الأردنية هاي.
1- اتعب قدامك...
2- الفطين ما يحط راسه ...
3- تموت الحية...

List
عطيني المقادير ديال حلوة البليفات

Short Answer
المدرسة البوعنانية هي مدرسة تاريخية تقع
في مدينة فاس المغربية. تحدث عنها في
جملة واحدة

Figure 1: PALM spans all 22 Arab countries, each represented by its own flag, across 20 diverse areas, including local *celebrations*, *geography*, and *history*. Instructions (input–response pairs) are human-created at the country level. Dashed examples represent local dialects, whereas others use MSA. Only a subset of domains and instruction types is shown here, and we include only example inputs (not responses) due to space constraints. English translations of the Arabic instructions are in Appendix H.

[*] Equal contribution

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32871–32894
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Abstract

As large language models (LLMs) become increasingly integrated into daily life, ensuring their cultural sensitivity and inclusivity is paramount. We introduce PALM, a year-long community-driven project covering *all 22* Arab countries. The dataset includes instructions (input, response pairs) in both Modern Standard Arabic (MSA) and dialectal Arabic (DA), spanning 20 diverse topics. Built by a team of 44 researchers across the Arab world, all of whom are authors of this paper, PALM offers a broad, inclusive perspective. We use PALM to evaluate the cultural and dialectal capabilities of several frontier LLMs, revealing notable limitations. For instance, while closed-source LLMs generally exhibit strong performance, they are not without flaws, and smaller open-source models face greater challenges. Moreover, certain countries (e.g., Egypt, the UAE) appear better represented than others (e.g., Iraq, Mauritania, Yemen). Our annotation guidelines, code, and data for reproducibility are publicly available. More information about PALM is available at our project page: https://github.com/UBC-NLP/palm.

1 Introduction

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 data they were pre-trained or fine-tuned on, which may not reflect the values and cultures of diverse end users (Shankar et al., 2017; Xu et al., 2024; Naous et al., 2024). LLMs pre-trained on translated data from English 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). However, having a benchmarking tool for cultural and linguistic coverage for LLMs is a crucial phase.

In this work, we focus on the Arab world and its communities which span 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 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 significantly 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), AceGPT (Huang et al., 2024), ALLAM (Bari et al., 2024), Fanar (Team et al., 2025), and NileChat (El Mekki et al., 2025). These models demonstrate powerful capabilities in generating Arabic across its different forms. However, when it comes to instruction tuning, the datasets used for some of these models (such as `JAIS` and `AceGPT`) are predominantly machine-generated or machine-translated, resulting in a set of instructions that are not related to Arab culture. In addition, most of these models lack evaluation on Arabic country-specific cultural awareness for all Arab countries, as most of them were evaluated on general NLP tasks but lack evaluation on specific Arabic countries' cultures and dialects. Our work addresses this need by providing a large dataset of Arabic instructions to ensure better cultural representation of Arab communities.

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.

We offer the following contributions:

1. We present **PALM**, a novel dataset developed

through a year-long collaborative community effort. It includes culturally informed instructions from **all** 22 **Arab countries** in both **MSA** and **local dialects**, spanning multiple linguistic forms and topics.

2. We benchmark several open-source and frontier LLMs on **PALM**, providing a detailed analysis of model performance on both MSA and dialectal data across various dimensions.

3. We offer a comprehensive analysis using three models as evaluative judges, examining their alignment and highlighting the reliability of automated evaluations.

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

2 Related Work

Arabic Varieties. Arabic, with its rich linguistic diversity, has attracted increasing attention in NLP. This focus has propelled the development of 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; Team et al., 2025) Arabic text, yielding powerful results in both understanding and generation tasks (Elmadany et al., 2023; Seelawi et al., 2021; Nagoudi et al., 2023). Nonetheless, a critical gap persists: the underrepresentation of Arabic dialects in current language models, which affects both performance and cultural inclusion (AlKhamissi et al., 2024). Consequently, there is a pressing need for more comprehensive Arabic language models that can capture both Modern Standard Arabic (MSA) and diverse colloquial dialects, reflecting the linguistic and cultural richness of the Arab world.

Cultural Awareness in Language Models. Recent research employs diverse methodologies to assess the cultural capabilities of language models (Adilazuarda et al., 2024). One approach constructs knowledge databases (Keleg and Magdy, 2023; Shi et al., 2024), often drawing on online resources such as Wikipedia (Nguyen et al., 2023; Fung et al., 2024) and web corpora (Nguyen et al., 2023). However, these evaluations can be overly simplistic, since much of the web-scraped content might already appear in the models' training data (Petroni et al., 2019). Another methodology uses socio-cultural surveys like the World Value Survey (Ramezani and Xu, 2023; AlKhamissi et al., 2024), which, while valuable, often cover only a narrow range of cultural concepts. To overcome these limitations, researchers have introduced new datasets and benchmarks tailored to evaluating the cultural capabilities of LLMs (Arora et al., 2024; Myung et al., 2024a; Singh et al., 2024a).

Arabic Cultural Awareness in Language Models. Recent studies have also increasingly focused on integrating Arabic language and culture into LLMs, alongside the development of relevant benchmarks (Myung et al., 2024a; Singh et al., 2024a; AlKhamissi et al., 2024; Naous et al., 2024; Demidova et al., 2024; Alwajih et al., 2024, 2025). For instance, AlKhamissi et al. (2024) leveraged the World Values Survey to assess LLM alignment with Arab cultural values, revealing reduced alignment for underrepresented groups. Similarly, Naous et al. (2024) identified a significant Western bias in LLMs, attributing it to the prevalence of translated rather than original Arabic data in pre-training corpora. Researchers have also introduced new datasets. For example, Mousi et al. (2024a) proposed a benchmark comprising 180 questions spanning nine topics and three Arabic regions, and Alyafeai et al. (2024a) introduced a localized dataset of 10,000 instructions covering 17 topics in MSA. However, much of this work relies on automatic annotation and remains centered on MSA, leaving gaps in dialectal and country-specific cultural representation. Moreover, most available benchmarks are limited in size and instruction formats, often focusing on multiple-choice questions (Mousi et al., 2024a; Team et al., 2025).

Our work, PALM, closes these gaps by presenting the first large-scale, fully human-curated collection of Arabic cultural input-output instruction pairs from all Arab countries. It encompasses multiple instruction types expressed in MSA and diverse Arabic varieties. Table 1 compares PALM to existing datasets involving any level of Arabic instructions.

3 PALM Dataset

As stated earlier, PALM is a manually curated, culturally aligned dataset covering all 22 Arab countries. It features a diverse set of instructions (input, output pairs) from both MSA and *ten* different Ara-

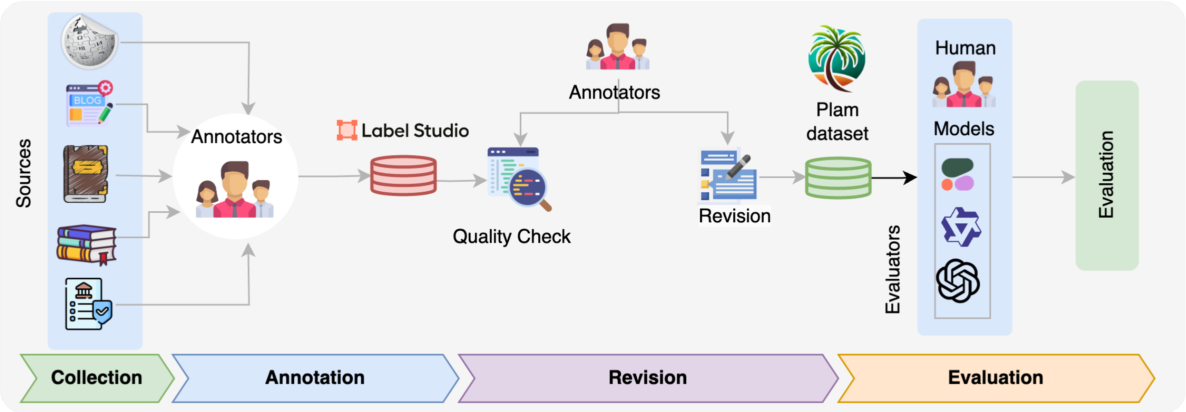


Figure 2: The complete pipeline for PALM creation, beginning with data collection from diverse sources, followed by annotation in Label Studio, quality checks, and subsequent revisions. Finally, PALM undergoes both human and model evaluations, culminating in a final assessment phase.

bic dialects, all entirely human-produced. PALM comprises 20 different **topical areas**, such as *celebrations*, *history*, *geography*, *literature*, *politics*, *proverbs*, and *sports*, crafted at the local, country-specific, level or at the level of the whole Arab world (e.g., *technology*). As such, PALM represents a comprehensive view of the culture of Arabic local communities. Figure 1 illustrates the composition of PALM across different countries and areas. We now describe how we created PALM.

3.1 Team Structure

PALM is a community project involving 44 trained native speakers, all of whom are authors of this work. We aimed to incorporate local knowledge from every Arab country and succeeded for 15 out of 22. For each of these 15 countries, we assigned at least two annotators. For the remaining seven, two annotators from neighboring countries were chosen to ensure cultural familiarity.[1] Each team member holds at least a bachelor's degree, with most having advanced degrees. Within each country, members hail from different regions, promoting inclusive cultural coverage and broad dialectal variety. To our knowledge, PALM stands among the most comprehensive datasets of its kind in the Arab world, both culturally and linguistically.

3.2 Annotation Guidelines

We developed our annotation guidelines iteratively over a period of about three months. The first version of the guidelines, created by four senior team members in consultation with a wider pool of participants, introduced the main objectives of the project, the topical areas from which the data will be created, and several categories of instruction types specific to each country (e.g., various types of open-ended requests and questions) illustrated with rich sets of examples. We also included samples from trustworthy **information sources** (see Appendix A) that can be treated as references while creating the data. This initial version of the guidelines was then shared with the team members who were asked to build a pilot dataset based on these guidelines. After a series of meetings, we further improved the guidelines, reaching an extensive version totaling 100 pages. This final version was then used to train all team members in order to ensure consistency across all aspects of the project.

In our guidelines, we asked the annotators to create instructions for two main categories: ***general category***, which covers MSA instructions for general knowledge such as *science* and *technology*; and ***country-specific category***, where the annotators provide instructions reflecting their country's culture in multiple topics as mentioned earlier, including *local celebrations*, *customs*, *local geography, national history, proverbs,* and *food*. The country-specific instructions could be expressed in either MSA or the dialect corresponding to the respective country. For more details about our annotation guidelines, refer to Appendix A.[2]

3.3 Platform and Quality Assurance

We used Label Studio (Tkachenko et al., 2020) as our annotation platform, forming country-specific sub-teams. Each annotator accessed their respective country's sections and created instructions in

[1] Countries without local team members are Bahrain, Comoros, Djibouti, Iraq, Libya, Qatar, and Somalia.

[2] Our full annotation guidelines manual is available at this link.

	Multilingual		Arabic Specific		
	AYA	BLEnD	AraDiCE	CIDAR	PALM (ours)
# Arab countries	-	1	-	-	22
# Arabic dialects	-	-	6	-	10
Cultural coverage?	limited	✓	limited	✓	✓
Human collected?	mixed	✓	✗	✗	✓
Human revised?	✓	✗	✓	✓	✓
From scratch?	mixed	✓	✗	✗	✓
Open classes?	✓	✗	✗	✗	✓
# Arabic instructions	5K out of 204K	3.6K out of 55K	180 out of 45k	10K (100%)	18K (100%)

Table 1: PALM in comparison. PALM is specifically designed to capture country-specific knowledge from all Arab countries, exceeding existing datasets in both geographic scope and dialectal diversity. Collected entirely by human annotators from scratch (unlike AraDice, which involves translation and data retargeting, and CIDAR, which relies on localization), PALM is also the only Arabic dataset based on open-ended instructional prompts (e.g., writing instructions, role-playing, reasoning) rather than solely QA. Further details on how PALM compares to other datasets are provided in Appendix F.

relevant topical areas, following our carefully designed annotation guidelines. We also implemented a structured review process to ensure data quality and annotation consistency. Weekly meetings addressed annotation accuracy, source reliability, instruction diversity, and progress. A dedicated Slack channel enabled real-time collaboration. Figure 2 illustrates the pipeline for constructing PALM, from data collection to final instruction revision. After completing the dataset, we conducted a comprehensive review in which team members cross-reviewed each other's contributions, ensuring each sample was examined by at least two reviewers. Section E in the appendix analyzes the impact of this revision process and highlights the resulting improvements in data quality.

3.4 Dataset Analysis

Unlike other Arabic instruction datasets listed in Table 1, PALM offers several unique features: it covers *all* 22 Arab countries, a larger number of dialects, and is entirely created and reviewed by humans. Totaling $17,411$ instruction pairs, PALM is meticulously designed to reflect the cultural and linguistic richness of Arab countries. In order to further characterize the dataset and showcase the various types of instructions developed, we provide a detailed quantitative analysis here.

Overall Statistics. To facilitate analysis, we divide the countries in PALM into two categories: *high-resource* (more than 500 instructions) and *low-resource* (around 100 instructions). High-resource countries include Egypt, Jordan, Mauritania, Morocco, Palestine, Saudi Arabia, Sudan, Syria, Tunisia, UAE, and Yemen. We also incorporate a *General* category containing $1,109$ instructions that are not tied to a single country but rather pertain to the Arab world at large. In total, the high-resource countries and the General category account for $16,066$ instructions, representing roughly 92% of the dataset.

A key feature of our dataset is its inclusion of dialects. For the ten out of the eleven high-resource countries, we asked annotators to provide around 380 examples in their respective local dialects, resulting in $4,211$ dialectal instruction pairs. This addition substantially broadens the dataset's linguistic diversity and enhances its cultural authenticity. Details of the dialectal distribution are presented in Table D.1 in Appendix D.

In contrast, the low-resource countries—Algeria, Bahrain, Comoros, Djibouti, Iraq, Kuwait, Lebanon, Libya, Oman, Qatar, and Somalia—have fewer instructions, largely due to limited annotator availability and resources. Collectively, these countries contribute $1,345$ instructions, amounting to approximately 8% of the dataset.

Instruction Types. Instructions in PALM span a range of diverse categories. During our meetings, we frequently discussed ways to introduce richer components such as role-playing and reasoning, in addition to various information requests and question prompts. One method for characterizing the data after its creation is to extract the *verbs* and subsequent *nouns* from each instruction, allowing us to identify examples such as "*summarize* the follow-

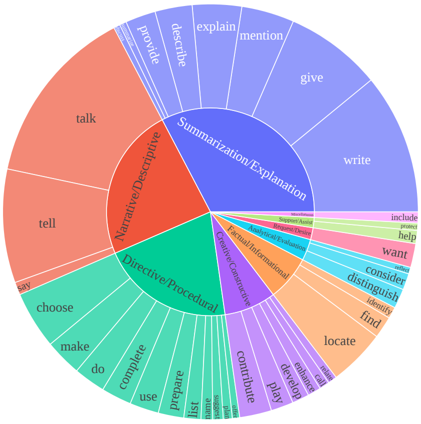


Figure 3: Distribution of instruction types based on verb usage in PALM, illustrating the frequency and categorization of verb–noun pairs and offering insights into the dataset's instructional diversity. Appendix C.1 presents a drill-down of this analysis.

ing *article*". To implement this, we first translate all instructions into English and then use `GPT4-o1` to extract the relevant verbs and nouns. We retain only those verbs appearing at least 25 times and nouns that co-occur with each verb at least five times. We next employ `GPT-o1` again to cluster these verb-noun pairs according to different instruction types and assign a representative name to each cluster, yielding the following: *(1) Summarization and Explanation* (requests for summaries or definitions), (2) *Directive and Procedural* (guidance on specific actions), *(3) Factual and Informational Queries* (requests for factual information), *(4) Creative and Constructive Generation* (content creation tasks), *(5) Analytical and Evaluation-Based* (critical thinking tasks), and *(6) Narrative and Descriptive Tasks* (storytelling or descriptive content). Figure 3 depicts the distribution of these instruction types in *Palm*. For a more detailed breakdown of each theme and its corresponding verbs and nouns, refer to Figures C.1 in Appendix C.

Dataset Splits. The PALM dataset is organized into three distinct splits: training, public test, and private test sets, each designed to serve a specific purpose. ***Training Set:*** This set, comprising $13,559$ instructions, will be publicly released as training data for models designed to achieve cultural and linguistic alignment with Arabic communities. It provides researchers with a valuable resource for developing models that are more attuned to Arabic cultural contexts. ***Public Test:*** This set, hereinafter referred to as the *test* set, will be publicly released as a benchmarking dataset. It enables researchers to evaluate their models' performance on culturally-specific instructions from the Arab world during the development phase. The test set comprises $1,926$ instructions. ***Private Test:*** This set will remain private and be accessible exclusively through a leaderboard, ensuring a fair comparison of different models and approaches by preventing data leakage. As with the public test set, it consists of $1,926$ instructions.

4 Evaluation

4.1 Evaluation Setup

To demonstrate the efficacy of PALM, we employ the test set as an evaluation benchmark. This evaluation serves dual purposes: (i) assessing current LLMs' performance across individual countries, topics, and dialects to provide a nuanced measure of Arabic cultural awareness, and (ii) offering a robust methodology for future researchers to evaluate model proficiency in handling Arabic culture and dialects. We evaluate 18 Arabic-aware LLMs[3], 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)* for the test set of PALM instructions and assessing the outputs.[4] Evaluation metrics can be categorized into *surface-level* and *LLM-as-Judge* metrics (Zheng et al., 2023).

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 responses, they enable us to measure three key aspects: *(i) language consistency* between the instruction and the generated answer; *(ii) preservation of the prompt's dialect (dialectal consistency)* in the generated answer; and *(iii) presence of sequence repetitions* within the generated answer. This approach allows us to examine the model's capacity to maintain linguistic and cultural fidelity without directly assessing factual accuracy.

4.3 LLM-as-Judge Evaluation

For the LLM-as-Judge assessment (Zheng et al., 2023), we focus on the *correctness* metric. We select three models with strong performance in Arabic tasks—`GPT-4o`, `CMDR+`, and `QWEN 2.5`

[3] The full list of models is in Table I.1 in Appendix.

[4] Models such as Fanar (Team et al., 2025) and Allam (Bari et al., 2024) are reported to provide encouraging performance but were not available for evaluation at the time of submission.

72B—and prompt them to rate answer correctness on a scale of 1 to 10, using LangChain's built-in evaluation pipeline.[5] Each rating considers both the instruction and the ground truth, as shown in Figure G.1. We then compute a mean correctness score for each generated response, capturing factual accuracy relative to the provided ground truth and ensuring no major errors. To gauge the reliability of this method, we calculate the Intraclass Correlation Coefficient (ICC) among the three models, yielding an **ICC of 0.68**, which indicates good scoring consistency. Table I.2 summarizes the **1,926** samples in the test set used for this evaluation.[6]

4.4 Human Evaluation

To validate our automatic evaluation, we conducted a human evaluation on a subset of generated responses, encompassing diverse examples from multiple countries, topics, and both MSA and dialects. We enlisted evaluators from the same data creation team, representing various Arab countries. These evaluators assessed responses from five LLMs based solely on the *correctness* criterion. Each response was evaluated by at least three independent human evaluators. We then computed the intraclass correlation coefficient (ICC) across all evaluators, yielding an **average ICC of 0.67**, indicating high inter-rater agreement.

The evaluation subset was sampled from the same test set used in the LLM-as-Judge evaluation and included **92 MSA samples** plus **20 dialectal samples** each for Egypt, Morocco, Syria, and Yemen. We further computed the correlation between the human evaluation scores and the LLM-as-Judge *correctness* scores, observing a strong relationship. The **Pearson correlation coefficient is 0.76 (p-value < 0.05)** and the **ICC is 0.78**, lending strong credibility to the LLM-as-Judge results presented in the following section.

5 Results and Discussion

Here, we analyze results of a subset of 12 LLMs that yield notable insights.[7]

5.1 Surface-Level Results

Table 2 shows the results for the surface-level attributes, namely repetitions, language consistency, and dialectal consistency. The top-performing models (e.g., Qwen 2.5, GPT-4o, and Claude-3.5-Sonnet) display negligible repetition rates when prompted with Arabic instructions, whereas Gemma-2 and Llama 3.1 exhibit higher repetition rates. For instance, 42% of responses generated by Gemma-2-9b contain repetitions. Regarding language consistency, all models follow a similar trend. However, dialectal consistency remains below roughly 10% across all LLMs, indicating that even when these Arabic-aware models are prompted with dialectal Arabic instructions, they tend to produce answers in MSA about 90% of the time.

Model	A1	A2	A3
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: Results of surface-level attributes (%). **A1:** repetitions, **A2:** lang consistency, **A3:** dialect consistency.

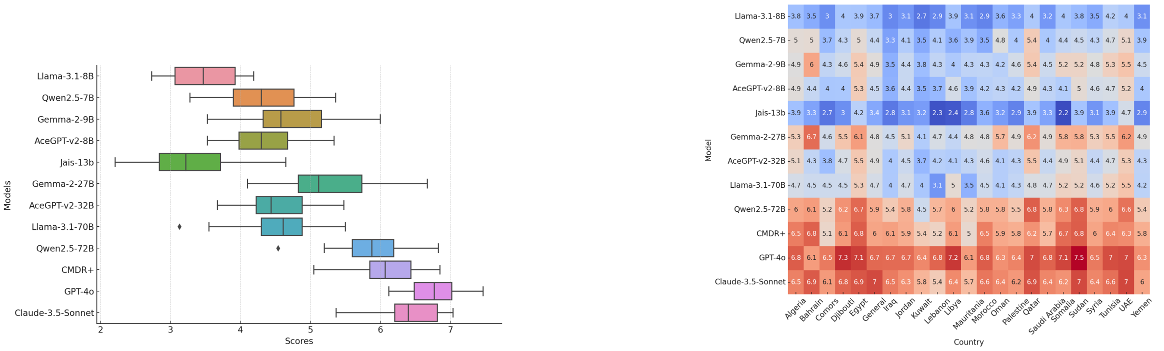
5.2 LLM-as-Judge Results

Overall Results. Figure 4a presents the results of our *correctness* evaluation using LLM-as-Judge. Overall, *GPT-4o and Claude-3.5-Sonnet exhibit the highest performance*, achieving median scores above 6.0. These larger models are followed closely by CMDR+ and Qwen2.5-72B, which display robust performance with median scores ranging from 5.8 to 6.0. *Mid-range models*—including AceGPT-v2-32B, Gemma-2-27B, and Llama-3.1-70B—cluster around median scores of 4.5 to 5.0, indicating moderate competence in Arabic evaluation. Conversely, *lower-capacity models* such as Jais-13b and Llama-3.1-8B demonstrate comparatively weaker performance, with median scores between 3.0 and 4.0. Notably, the boxplot whiskers suggest considerable variance across all models, indicating that performance can be inconsistent depending on the

[5] https://langchain.com/

[6] We also apply the same methodology to evaluate *coherence*, *detail*, and *helpfulness*, with descriptions provided in Appendix G.

[7] Complete results are available in Appendix I.2, with Section I.2.1 showing that smaller models generally underperform in providing correct answers to the Arabic cultural questions.



(a)

(b)

Figure 4: LLM-as-judge average correctness scores across Arabic countries.

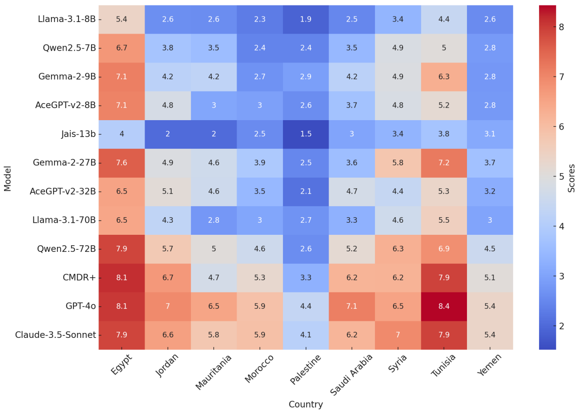


Figure 5: Correctness scores across Arabic dialects.

specific test cases. An interesting, albeit unsurprising, finding is the overall correlation between model size and performance, with larger models (>70B parameters) consistently outperforming their smaller counterparts. We report the results for the *coherence*, *detail*, and *helpfulness* metrics in Figure I.1, located in Appendix I.2.

Per-Country Results. Figure 4b presents the *correctness* scores by country. The figure shows that more advanced models like `GPT-4o` and `Claude-3.5-Sonnet` consistently exhibit higher scores—often exceeding 6.0—indicating better cultural understanding. For example, `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`, display lower *correctness* scores, frequently below 4.0, with `Jais-13b` scoring as low as 2.2 for Saudi Arabia. The variation in scores across different countries underscores each model's varying degree of cultural competence, highlighting the complexity of capturing diverse cultural nuances within the Arabic-speaking world. For instance, `Claude-3.5-Sonnet` scores 7.0 for Yemen but 5.8 for Lebanon, suggesting a stronger grasp of Yemeni contexts relative to Lebanese ones.

Per-Dialect Results. Figure 5 shows the *correctness* scores of various models when prompted with dialectal instructions across multiple Arabic dialects. Larger models, such as `GPT-4o`, `Claude-3.5-Sonnet`, and `CMDR+`, consistently exhibit higher scores—frequently above 7.0—in most countries. For example, `GPT-4o` achieves 8.1 for Egypt and 8.4 for Tunisia. In contrast, smaller models (e.g., `Llama-3.1-8B` and `Jais-13b`) generally perform below 4.0. Performance also varies by country, with some dialects presenting more difficulty than others. Morocco and Palestine often yield lower scores for multiple models, whereas Egypt and Tunisia tend to yield higher ones.[8]

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 substantially between countries for each model, suggesting that *performance on dialectal Arabic instructions is highly dependent on the specific country and dialect being evaluated.*

A similar trend is noticed for Table I.4, where

[8] Results per topic are presented in Appendix I.2. LLMs demonstrated higher performance in topics such as celebrations, history, and travel, while achieving lower scores in categories like sports and food.

Country	M1	M2	M3	M4	M5
Egypt	6.47	4.26	4.71	4.15	4.08
Moroc	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: Avg human eval correctness for dialect instructions per country. **M1:** AceGPT-v2-32B, **M2:** Llama-3.1-8B, **M3:** Qwen2.5-72B, **M4:** Claude-3-5-Sonnet, **M5:** Jais-13b.

`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 perform 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 44 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.

Limitations

While PALM serves as a valuable resource for training and benchmarking culturally aware, dialectally diverse Arabic LLMs, it has some limitations. In low-resource countries, content was often contributed by annotators from neighboring countries rather than local speakers. Although these annotators share certain cultural similarities, the resulting instructions may lack the depth and nuances that native, local speakers would bring.

Moreover, many Arab countries have multiple regional dialects, which require larger, geographically diverse teams to fully represent. Due to the scale of the project, some dialectal variations may not be covered in detail, limiting the dataset's ability to capture every linguistic nuance.

Lastly, although automatic evaluations using LLMs facilitate scalable assessment, they can fall short when dealing with dialects and subtle cultural elements. These models may misjudge culturally specific or dialectal content, introducing biases. Consequently, human evaluation remains essential alongside automated methods to ensure reliable results.

Ethics Statement

In developing PALM, we emphasized cultural sensitivity, inclusivity, and ethical responsibility. All annotations were created by informed participants, each of whom is a co-author on this paper, ensuring that every contributor receives full credit for their work. We adhered strictly to publicly available and reputable sources, refraining from using any private or sensitive data. Clear guidelines were also provided to respect local norms, maintain data privacy, and secure participant consent.

Although PALM aims to mitigate biases in Arabic LLMs, unintentional cultural bias may still occur—particularly in regions lacking direct local representation. We encourage ongoing community involvement to address these gaps, ensuring continual refinement and improvement of the dataset.

Reproducibility. Our test data, prompts, and code necessary to produce all results reported in this work are publicly available at PALM. Our private test set will also be available through a leaderboard.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[9] and UBC Advanced Research Computing-Sockeye.[10]

[9] https://alliancecan.ca

[10] https://arc.ubc.ca/ubc-arc-sockeye

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 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.

Fakhraddin Alwajih, Samar Mohamed Magdy, Abdellah El Mekki, Omer Nacar, Youssef Nafea, Safaa Taher Abdelfadil, Abdulfattah Mohammed Yahya, Hamzah Luqman, Nada Almarwani, Samah Aloufi, Baraah Qawasmeh, Houdaifa Atou, Serry Sibaee, Hamzah A. Alsayadi, Walid Al-Dhabyani, Maged S. Al-shaibani, Aya El aatar, Nour Qandos, Rahaf Alhamouri, Samar Ahmad, Razan Khassib, Lina Hamad, Mohammed Anwar AL-Ghrawi, Fatimah Alshamari, Cheikh Malainine, Doaa Qawasmeh, Aminetou Yacoub, Tfeil moilid, Ruwa AbuHweidi, Ahmed Aboeitta, Vatimetou Mohamed Lemin, Reem Abdel-Salam, Ahlam Bashiti, Adel Ammar, Aisha Alansari, Ahmed Ashraf, Nora Alturayeif, Sara Shatnawi, Alcides Alcoba Inciarte, AbdelRahim A. Elmadany, Mohamedou cheikh tourad, Ismail Berrada, Mustafa Jarrar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. Pearl: A multimodal culturally-aware arabic instruction dataset. *Preprint*, arXiv:2505.21979.

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

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. 2024a. 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.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran AQ Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, et al. 2024b. Cidar: Culturally relevant instruction dataset for arabic. *arXiv preprint arXiv:2402.03177*.

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.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2024. Allam: Large language models for arabic and english. *Preprint*, arXiv:2407.15390.

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.

Abdellah El Mekki, Houdaifa Atou, Omer Nacar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. Nilechat: Towards linguistically diverse and culturally aware llms for local communities. *Preprint*, arXiv:2505.18383.

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. 2024a. AraDiCE: Benchmarks for dialectal and cultural capabilities in llms.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2024b. Aradice: Benchmarks for dialectal and cultural capabilities in llms. *arXiv preprint arXiv:2409.11404*.

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. 2024a. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Preprint*, arXiv:2406.09948.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024b. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv 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. 2024a. 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.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024b. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

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.
- Revision Process E.
- Comparative Analysis of Arabic Instructional Datasets F.
- Evaluation Prompt G.
- Selected Examples H.

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.

Information Sources. To ensure high data quality, we underscore the importance of consulting reliable and authoritative sources during instruction creation. Annotators were consistently advised to perform thorough verification of these sources. Exemplary sources include Wikipedia and other reputable online encyclopedias, academic books, governmental websites, and specialized platforms (e.g., health organization websites offering medical information). We explicitly cautioned against relying on single-individual sources, such as personal posts or social media content, unless the individual is a widely recognized expert in the respective field. Notably, for domains such as travel, culinary arts, and culturally specific celebrations, the most valuable insights often derive from online discussions and forums. In these cases, we leveraged annotators' local cultural knowledge and judgment to ensure the trustworthiness and relevance of the data.

Our full annotation guidelines manual is available at https://github.com/UBC-NLP/palm/blob/main/guidelines.md.

B *Palm* Topics and Domains

Palm categorizes its instructions into three main domains: **General**, **Hybrid**, and **Country-Specific**. The **General Domain** covers topics with universally applicable knowledge, such as science, sports, and technology. The **Hybrid Domain** consists of topics that include both general and country-specific knowledge, bridging regional cultural insights with broader themes. Lastly, the **Country-Specific Domain** focuses exclusively on Arab nations, highlighting their traditions, social norms, and linguistic nuances. Table B.1 provides a breakdown of the key instruction domains within *Palm*.

C Diverse Instruction Formats and Linguistic Variations

Figure C.1 presents sunburst charts for every instruction type category in *Palm*. Each subfigure highlights a unique theme, derived from verb usage and the subsequent noun, that sheds light on the diversity of instructional approaches within the dataset. All of these subcharts are drilldowns from Figure 3.

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

Domain	Definition
Science	Covers various scientific fields, including biology, physics, chemistry, mathematics, and astronomy. Instructions range from fundamental concepts to applied sciences and technological advancements.
Food	Covers general knowledge about ingredients, nutrition, and food safety, as well as country-specific dishes, traditional recipes, and meal customs in Arab countries.
Sports	Includes general sports rules and history, as well as country-specific sporting traditions, major tournaments, and notable athletes in the Arab world.
Politics	Covers both general political concepts (e.g., voting systems, ideologies) and country-specific topics like political parties, government structures, and notable leaders.
Religion	Explores the major monotheistic religions (Islam, Christianity, Judaism), focusing on historical sites, religious figures, and institutions while avoiding specific rituals.
History	Encompasses ancient civilizations, historical events, wars, and influential leaders, highlighting their impact on Arab culture and heritage.
Travel	Provides information on notable historical landmarks, best travel destinations, itineraries, and cultural tourism across Arab countries.
Flora & Environment	Discusses wildlife, national parks, climate change, agricultural practices, and native plant species in different Arab regions.
Local Geography	Focuses on terrain diversity, water resources, economic impact, and geographical landmarks of specific Arab countries.
Celebrations	Highlights national, historical, and religious festivals, their cultural significance, associated traditions, and unique practices in different Arab communities.
Language	Examines Modern Standard Arabic (MSA) and dialectal variations, including translation tasks, word usage, and sentence restructuring between dialects and MSA.
Proverbs	Captures the cultural relevance of Arabic proverbs, their meanings, usage, and context in everyday conversations.

Table B.1: Instruction areas/domains in *Palm* categorized by their relevance to general, hybrid, and country-specific knowledge.

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

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.

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

E Revision Process

To quantitatively assess the changes made during the review process, we employed Python's `difflib.SequenceMatcher` to compute similarity ratios between the original and revised versions of the instructions, inputs, and outputs. Specifically, we defined the difference score as $1-$similarity and observed mean differences of 0.012 for instructions, 0.001 for inputs, and 0.017 for outputs. Notably, only 22% of the samples underwent any modifications, reflecting the effectiveness of our rigorous annotator training and the weekly meetings held to address emerging issues early in the project. This proactive approach ensured clarity and consistency throughout the annotation workflow.

We also examined a sample of 200 instructions to characterize the types of revisions made during the second phase of data review. These revisions primarily fell into three categories. First, **grammar and mechanics** revisions included punctuation adjustments (such as adding missing marks or removing extraneous ones) and grammatical corrections (e.g., fixing subject–verb agreement, verb tenses, and pronoun usage). Second, **question revisions** involved rephrasing or clarifying questions to preserve their core meaning. Third, **answer revisions** comprised either summarizing responses to highlight key points or expanding them by adding further details and context.

F Comparative Analysis of Arabic Instructional Datasets

Recent efforts in developing instructional datasets for Arabic language processing have produced a variety of resources, each with distinct strengths and limitations. Multilingual datasets such as AYA (Singh et al., 2024b) and BLEnD (Myung et al., 2024b) have contributed valuable resources by including Arabic instructions; however, their focus on multiple languages means that Arabic-specific nuances are underrepresented. For example, while AYA provides 5K Arabic instructions out of a total of 204K, BLEnD offers only 3.6K Arabic instructions among 55K entries, and BLEnD's coverage is limited to just one Arab country. In contrast, Arabic-specific datasets like AraDiCE (Mousi et al., 2024b) and CIDAR (Alyafeai et al., 2024b) have been developed to capture more localized content. AraDiCE, which spans six dialects, often relies on translation and data retargeting methods, and its limited number of native instructions (180 out of 45K) may not fully capture the linguistic diversity. Similarly, although CIDAR contains a full set of 10K Arabic instructions with human revisions, it lacks the breadth in geographic and dialectal diversity.

Our dataset, PALM, addresses these gaps by providing a more comprehensive resource tailored to the Arabic language. It uniquely covers 22 Arab countries and incorporates 10 Arabic dialects, ensuring broader cultural and regional representation. Importantly, PALM is the only dataset in this comparison that is built entirely from scratch through human collection and revision, rather than relying on machine translation or localization. Moreover, by focusing on open-ended instructional prompts—including tasks such as writing, role-playing, and reasoning—PALM offers richer linguistic expressions and a more authentic reflection of native language use. This meticulous design aims to better support the development and evaluation of Arabic language models in a variety of real-world applications.

G Evaluation Prompt and Metrics

The evaluation metrics are defined as follows (on a scale of 1 to 10):

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.

H Selected Examples

Table H.1 presents a collection of combined Arabic and English examples of instructions and outputs. These examples, originally taken from Figure 1, have been translated into English with the answers shortened to save space.

I Evaluation

I.1 Evaluated LLMs

Table I.1 enumerates the LLMs employed to generate evaluation answers for PALM. These models were selected from a curated list of Arabic-aware systems, with each entry including its size (in billions of parameters) and release date. Note that we used the instruct version for all LLMs.

I.2 LLM-as-Judge Results

Overall Results for Different Metrics. Figure I.1 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

Country	Topic	Instruction	Output
Syria	literature	أي من هؤلاء الأدباء الذين ظهروا في القرن العشرين في سوريا ؟ نزار قباني، سليمان العيسى Which of these writers emerged in 20th-century Syria? Nizar Qabbani, Suleiman al-Issa	نزار قباني و سليمان العيسى Nizar Qabbani and Suleiman al-Issa
Tunisia	history	رتب هذه الدول زمنيا: - الدولة الصنهاجية - الدولة الموحدية - الدولة الأغلبية - الدولة الحفصية - الدولة الفاطمية Arrange these states in chronological order: - The Sanhaji state - The Almohad state - The Aghlabid state - The Hafsid state - The Fatimid state	بالتأكيد. هذا هو الترتيب الزمني الصحيح للدول المذكورة التي تعاقبت... Certainly. This is the correct chronological order of the mentioned states that succeeded each other...
Saudi Arabia	travel	ماذا تعرف عن المسجد الحرام؟ What do you know about Masjid al-Haram?	على الرحب والسعة، المسجد الحرام أحد أهم وأشهر معالم السياحة... You're very welcome, Masjid al-Haram is one of the most important and renowned tourist landmarks...
Palestine	travel	فيما لا يقل عن ٥ أسطر، تحدث لي عن الحرم الإبراهيمي في مدينة الخليل. In at least 5 lines, tell me about the Ibrahimi Mosque in the city of Hebron.	المسجد الإبراهيميّ، أو الحرم الإبراهيميّ الشريف، وهو عند اليهود باسم ... The Ibrahimi Mosque, or the Noble Ibrahimi Sanctuary, is known among the Jews as...
Morocco	travel	المدرسة البوعنانية هي مدرسة تاريخية تقع في مدينة فاس المغربية. تحدث عنها في جملة واحدة The Bouanania School is a historic school located in the city of Fez, Morocco. Talk about it in one sentence.	المدرسة البوعنانية هي مدرسة تاريخية تقع في مدينة فاس المغربية....
Jordan	celebrations	اكتب مقالاً من أربع فقرات توضح فيه أهمية مهرجان جرش Write an essay consisting of four paragraphs that explains the significance of the Jerash Festival.	يلعب المهرجان الذي يستند على الإرث التاريخي والإبداعي العريق دورًا... The festival, which is based on a rich historical and creative heritage, plays a role...
Egypt	travel	رشح أفضل ١٠ أماكن لتناول الغذاء في الإسكندرية. Recommend the top 10 places to eat in Alexandria.	بكل تأكيد! إليك تقرير عن أفضل ١٠ مطاعم في الإسكندرية [٥٠٠ام] أفضل مطاعم... Certainly! Here is a report on the top 10 restaurants in Alexandria Top restaurants...
Algeria	travel	اقترح افضل الشواطئ في تلمسان التي يجب زيارتها. Suggest the best beaches in Tlemcen to visit.	بما أن تلمسان مدينة ساحلية تطل على البحر الابيض المتوسط،... Since Tlemcen is a coastal city overlooking the Mediterranean Sea,...
Yemen	food	وضح طريقة عمل مندى الدجاج فى خطوات بسيطة؟ Explain how to prepare chicken mandi in simple steps?	طريقة عمل مندى الدجاج فى خطوات بسيطة، لتقديم وجبة مغذية... The method for preparing chicken mandi in simple steps, to serve a nutritious meal...

Table H.1: Examples of Arabic and English instructions and outputs. All examples are derived from Figure 1, translated to English, and the answers have been abbreviated for space.

	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 I.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.

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-Topic Results. Figure I.3 in Appendix I.2 presents the performance of various models across different topics. `GPT-4o` and `Claude-3.5-Sonnet` consistently exhibit superior performance, with scores frequently above 6.0. For instance, `GPT-4o` achieves a top score of 7.4 in the *History* category, while `Claude-3.5-Sonnet` scores 7.0 in both *History* and *Proverbs*. In contrast, models such as `Llama-3.1-8B` and `Jais-13b` generally perform worse, often scoring below 4.0 in multiple topics. The *Food* category appears particularly challenging, displaying lower scores compared to other areas. Some models show particular strengths in specific domains. For example, `Qwen2.5-72B` scores 6.3 in *Celebrations* and 6.4 in *Science*, while `Gemma-2-27B` earns 5.9 in both *Flora* and *Science*.

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

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 I.2: Number of samples per country for automatic evaluations.

I.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 I.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

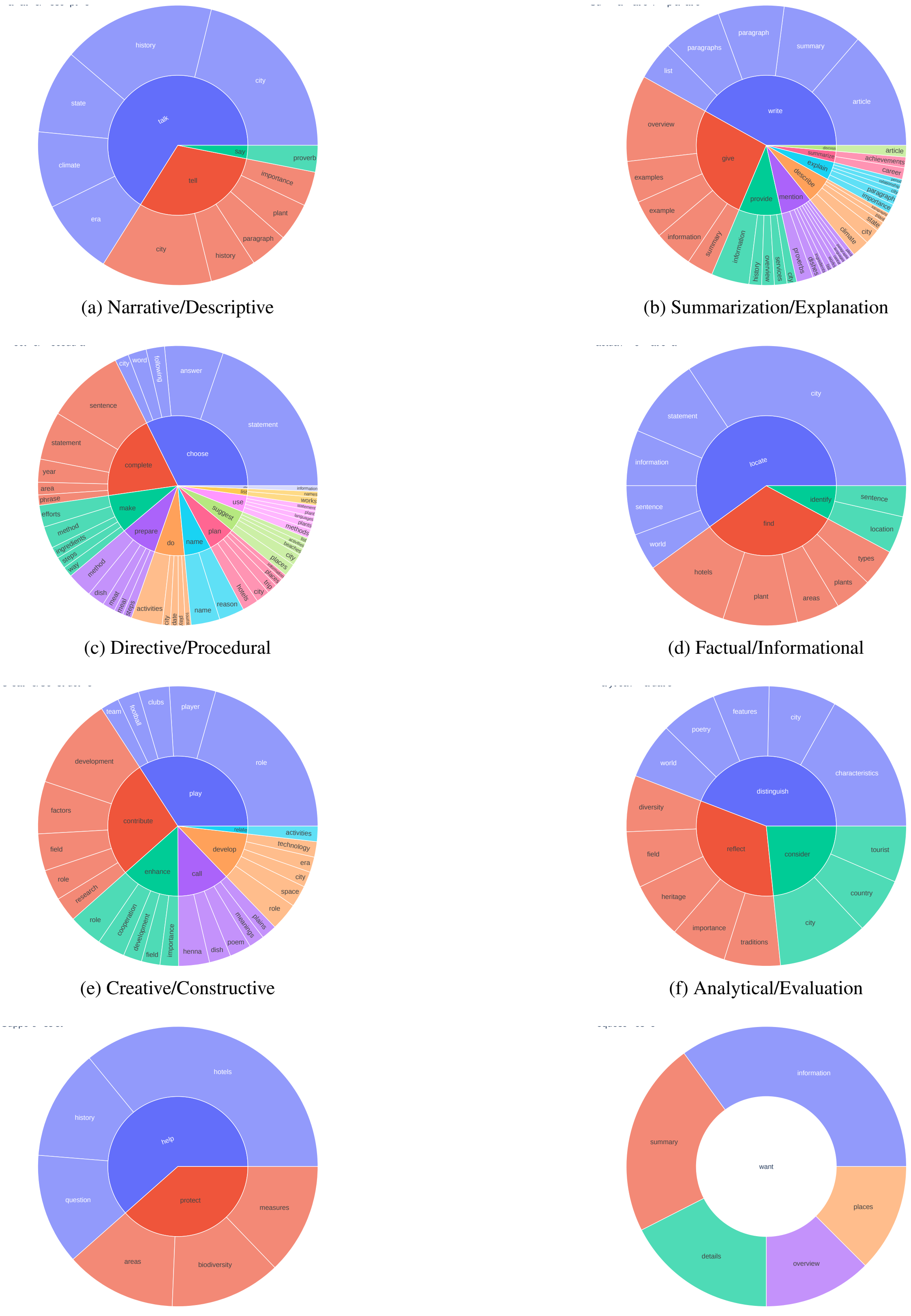
Country	Phi-3.5-Mini	Gemma-2-2B	Qwen2.5-1.5B	Qwen2.5-3B	LLaMA-3.2-1B	LLaMA-3.2-3B
Algeria	3.66	2.92	2.31	3.42	1.67	2.60
Bahrain	3.00	4.00	2.20	3.40	1.40	2.20
Comoros	1.79	2.70	1.80	2.23	1.85	2.02
Djibouti	3.00	2.89	2.55	2.66	1.56	2.18
Egypt	3.23	3.04	2.72	3.15	1.72	2.28
General	2.45	2.65	2.47	2.55	1.55	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 I.3: Average correctness scores of Small LLMs models across countries using CMDR+.

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

I.3 Human Evaluation Results

Table I.4 presents average human-evaluation correctness scores for MSA instructions per country.



(a) Narrative/Descriptive

(b) Summarization/Explanation

(c) Directive/Procedural

(d) Factual/Informational

(e) Creative/Constructive

(f) Analytical/Evaluation

(g) Support/Assist

(h) Request/Desire

Figure C.1: Sunburst charts for each instruction type category in *Palm*. Each subfigure represents a distinct theme derived from verb usage, providing insights into the dataset's instructional diversity.

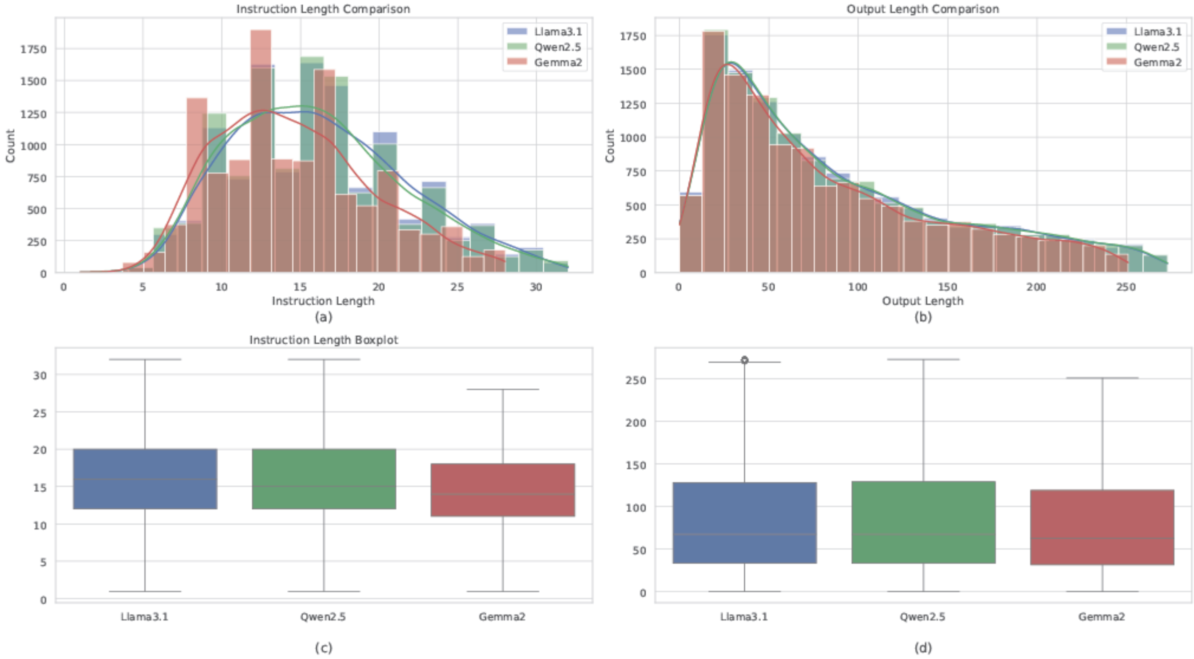


Figure D.1: Token Lengths Comparison.

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

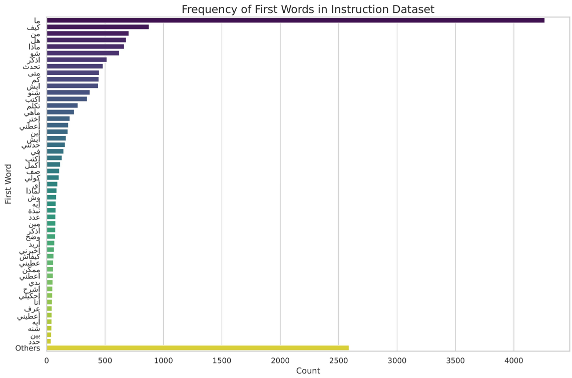
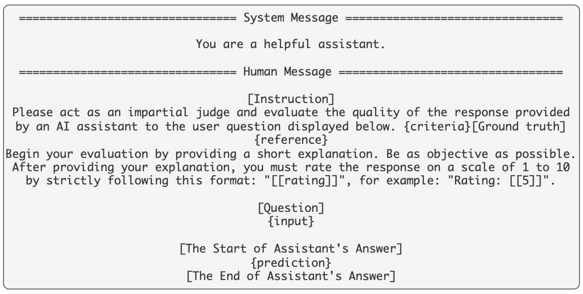


Figure D.2: Frequency of first words.



```
================================ 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 G.1: Prompt used in our evaluation.

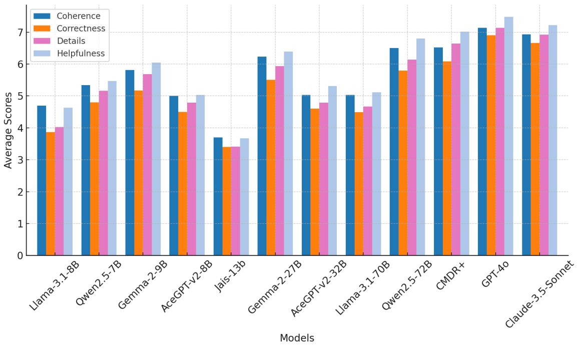
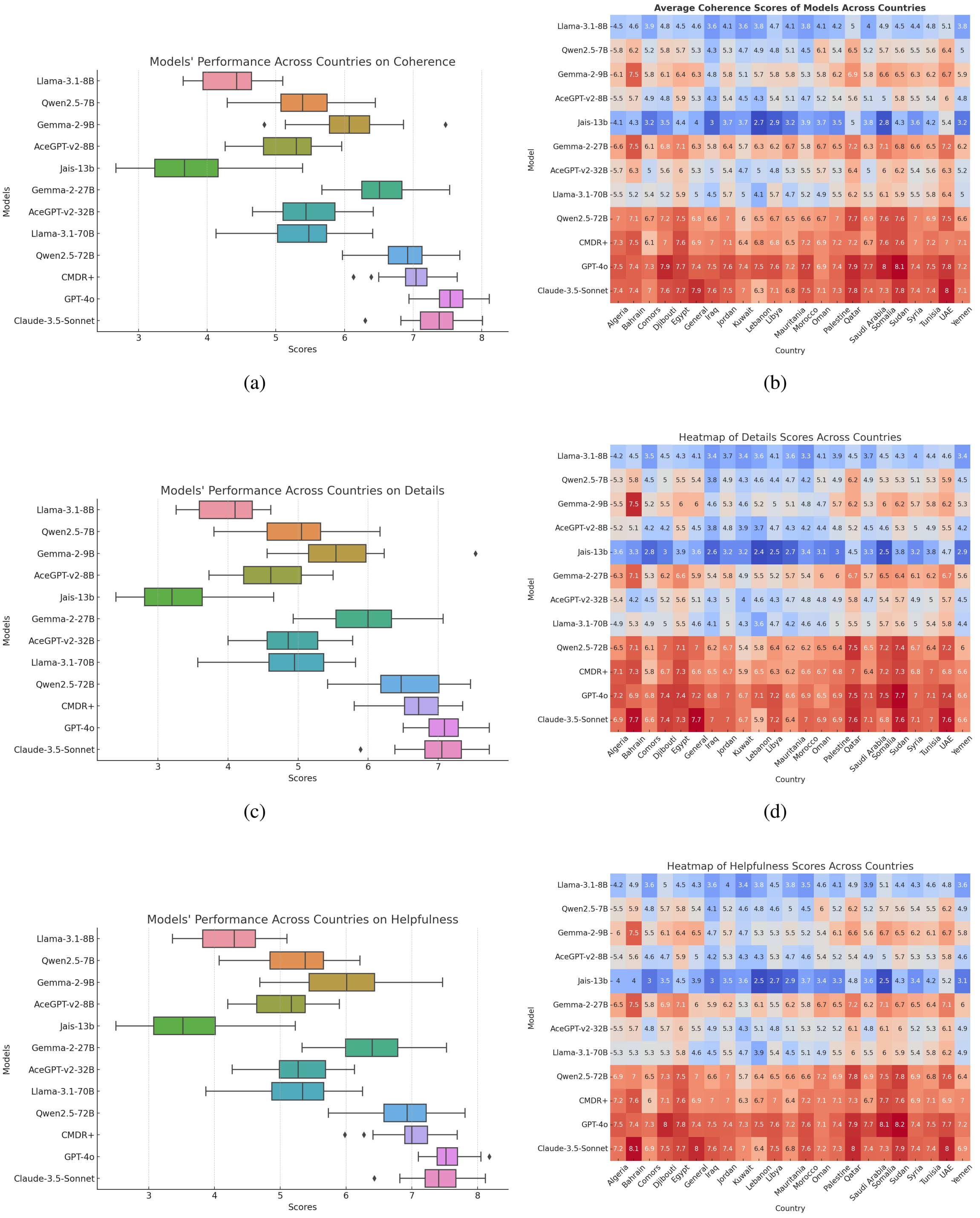


Figure I.1: Performance comparison of 11 evaluated LLMs across the four metrics: *correctness*, *coherence*, *helpfulness*, and *details*.



(a)

(b)

(c)

(d)

(e)

(f)

Figure I.2: Comparative analysis of the models across evaluation metrics: Coherence, Details, and Helpfulness.

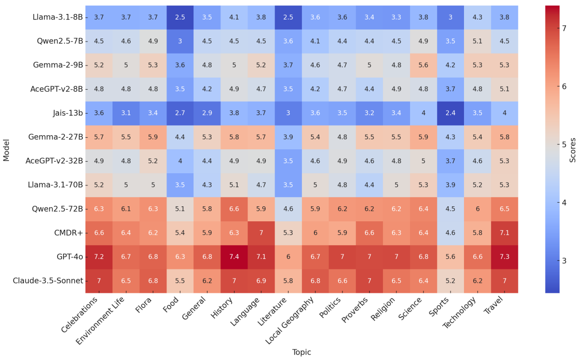
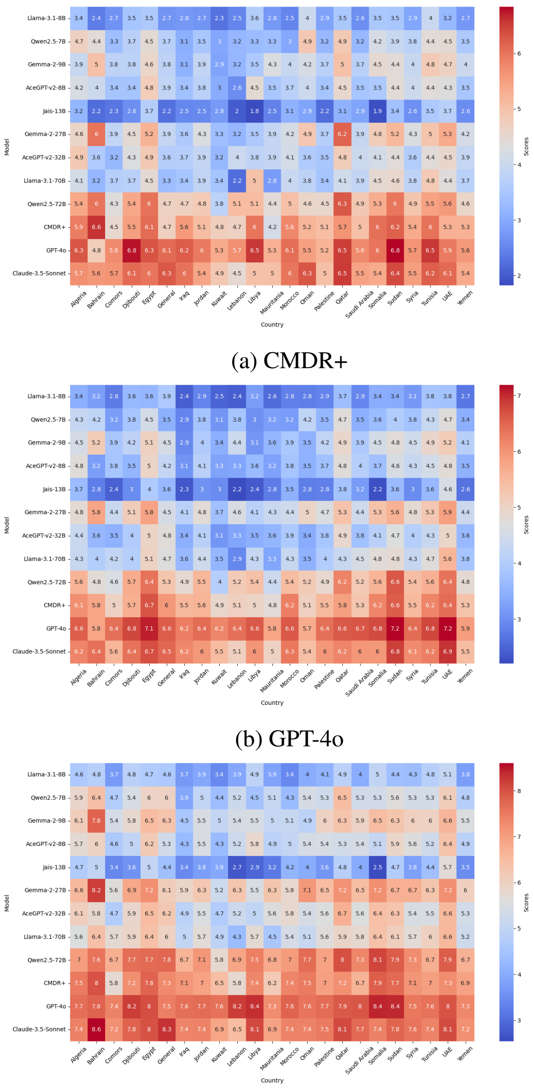


Figure I.3: Performance of various models using Correctness score across different topics.



(a) CMDR+

(b) GPT-4o

(c) Qwen2.5-72B

Figure I.4: LLM-as-judge correctness scores across Arabic countries.