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, Firoj Alam
Qatar Computing Research Institute, Qatar, [†]University of New Brunswick, Canada
[*]American University of Sharjah, UAE
{bmousi, ndurrani, fialam}@hbku.edu.qa

Abstract

Arabic, with its rich diversity of dialects, remains significantly underrepresented in Large Language Models, particularly in dialectal variations. We address this gap by introducing seven synthetic datasets in dialects alongside Modern Standard Arabic (MSA), created using Machine Translation (MT) combined with human post-editing. We present **AraDiCE**, a benchmark for **Ara**bic **Di**alect and **C**ultural **E**valuation. We evaluate LLMs on dialect comprehension and generation, focusing specifically on low-resource Arabic dialects. Additionally, we introduce the first-ever fine-grained benchmark designed to evaluate cultural awareness across the Gulf, Egypt, and Levant regions, providing a novel dimension to LLM evaluation. Our findings demonstrate that while Arabic-specific models like Jais and AceGPT outperform multilingual models on dialectal tasks, significant challenges persist in dialect identification, generation, and translation. This work contributes ≈45K post-edited samples, a cultural benchmark, and highlights the importance of tailored training to improve LLM performance in capturing the nuances of diverse Arabic dialects and cultural contexts. We have released the dialectal translation models and benchmarks developed in this study.[1]

1 Introduction

Large Language Models (LLMs) have consistently pushed the boundaries of Natural Language Processing (NLP), achieving state-of-the-art performance on a wide range of tasks. These models have excelled in areas such as machine translation, summarization, sentiment analysis, and even more complex applications like legal document analysis and creative writing (OpenAI, 2023; Touvron et al., 2023; Bubeck et al., 2023). Their remarkable ability to extract, reason, and generalize knowledge is fueled by training on vast amounts of data covering diverse topics and domains.

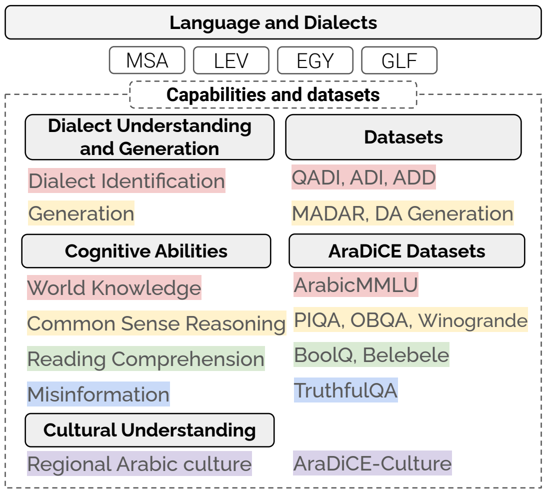


Figure 1: Capabilities and associated datasets for benchmarking, evaluated on different dialects.

However, the success of these models is heavily skewed towards languages with abundant resources, such as English (Bang et al., 2023; Ahuja et al., 2023). Low-resource languages, including Arabic and its various dialects, are significantly underrepresented in the datasets used to train these models. This disparity poses a substantial challenge, as LLMs require extensive and diverse data to perform effectively. Consequently, speakers of low-resource languages are at a disadvantage, unable to fully benefit from the advancements in NLP technologies. Recent efforts have been made to train Arabic LLMs (Sengupta et al., 2023; Bari et al., 2024; Fanar-Team, 2024) and to adapt multilingual models for Arabic (Touvron et al., 2023; Huang et al., 2024), but these models are predominantly tailored to MSA, leaving them less effective in handling dialectal Arabic (DA). To formalize this observation, we conduct a systematic study to benchmark Arabic and multilingual models in their performance on **Dialectal Arabic**. More specifically, we aim to answer the following questions:

[1]https://huggingface.co/datasets/QCRI/AraDiCE

Proceedings of the 31st International Conference on Computational Linguistics, pages 4186–4218

- Can LLMs effectively perform basic NLP tasks in dialects? **(Understanding and Generation)**
- Can LLMs demonstrate reasoning, comprehension, and handle knowledge and misinformation in dialects? **(Cognitive Abilities)**
- Are they aware of Arabic cultural knowledge? **(Cultural Understanding)**

To this end, we compiled a comprehensive suite of both existing and newly developed benchmarks to assess the capabilities of these models. We primarily evaluate their fundamental NLP abilities in encoding dialects through dialect identification (e.g. El-Haj et al. (2018)) and dialectal machine translation (e.g. Abdul-Mageed et al. (2023)) tasks. To assess broader cognitive abilities, we include benchmarks that gauge *World Knowledge* (ArabicMMLU (Koto et al., 2024)), *Common Sense Reasoning* (PIQA (Bisk et al., 2020), OBQA (Mihaylov et al., 2018), Winogrande (Sakaguchi et al., 2021)), *Reading Comprehension*, BoolQ (Clark et al., 2019), Belebele (Bandarkar et al., 2024)), and the ability to handle *Misinformation* (TruthfulQA (Lin et al., 2022)).

These benchmarks are primarily available in English, but researchers and practitioners have increasingly turned to synthetic data creation methods to address the gap in low-resource languages (Long et al., 2024). One promising approach involves utilizing MT to generate synthetic datasets for these languages. This method harnesses the capabilities of existing translation models to create large-scale, high-quality synthetic data, which can then be used to train LLMs.

In this paper, we introduce a comprehensive approach that leverages MT, specifically from English to MSA and MSA to dialects, combined with human post-editing, to develop synthetic benchmarks for low-resource DA. We concentrate on the Levantine (LEV), and Egyptian (EGY) dialects. Our primary goal is to curate benchmarks that evaluate the performance of LLMs across underrepresented language and dialects.

Moreover, we introduce a third and novel benchmark, **AraDiCE-Culture**, focused on cultural awareness across the Levantine, Egyptian, and Gulf regions. This benchmark evaluates whether LLMs grasp regional cultural nuances beyond language. Our dataset includes questions on public holidays, food, geography, history, public figures, traditional clothing, and more. We probe cultural specifics to assess if the LLMs can differentiate between these regions, emphasizing the importance of models to understand both dialects and their cultural[2] contexts. Our contributions in this work include:

- We develop benchmarks through the translation and post-editing of dialectal data, and assess LLMs' performance in cognitive abilities, understanding, and generation tasks across MSA and Dialectal Arabic.
- We create the first benchmark for region-wise cultural evaluation in the Gulf, Egyptian, and Levant regions.
- We present a comparative analysis of Arabic-focused LLMs, such as Jais (Sengupta et al., 2023) and AceGPT (Huang et al., 2024), alongside state-of-the-art models Llama 3 (Touvron et al., 2023) and Mistral (Jiang et al., 2023).
- To our knowledge, this is the first effort to develop dialectal benchmarks for Arabic LLMs.
- We have released the dialectal translation models and benchmarks curated in this study.

A summary of our findings is as follows:

Understanding and Generation LLMs generally struggle with dialect identification, generation, and translation tasks. Although Arabic-centric models perform better on dialectal tasks, their performance still lags compared to MSA or English. These models often rely on MSA knowledge for distinguishing between dialects. They are better at encoding and understanding dialects than at generating or translating them, as evidenced in dialect generation and translation tasks.

Cognitive Abilities Arabic-centric LLMs, such as Jais and AceGPT are better equipped to handle both MSA and its dialects. In contrast, multilingual LLMs, such as Llama3 and Mistral, show a substantial deficiency in adapting to Arabic dialects, highlighting a significant gap in their ability to handle linguistic diversity in this context.

Cultural Understanding Arabic-centric models also demonstrate a superior understanding of cultural nuances compared to general multilingual LLMs. This highlights the need for a more specialized training regimen to effectively address regional linguistic and cultural variations.

Dialectal Variability The effectiveness of Arabic-centric models varies notably across dialects, with stronger performance in generating Gulf responses, likely due to the dialect's closer resemblance to MSA. Our findings highlight the importance of

[2]We follow the definition of culture from AlKhamissi et al. (2024), which describes it as "a multi-faceted inquiry reflecting diversity across worldviews and belief systems."

AraDiCE in assessing dialectal capabilities and identifying gaps in LLM models.

2 Datasets

We begin by selecting a range of datasets commonly used for standard LLM benchmarking, including specific to Arabic. These datasets evaluate capabilities such as *understanding and generation*, *world knowledge*, *commonsense reasoning*, *reading comprehension*, *misinformation*, and *cultural understanding*, as outlined in Figure 1.

We selected the datasets with a focus on diversity and linguistic compatibility.[3] These datasets include tasks involving generation and multiple-choice questions. The datasets used in this study include: *(i)* four existing Arabic datasets for understanding and generation: *Arabic Dialects Dataset (ADD)* (El-Haj et al., 2018), *ADI* (Abdelali et al., 2024), and *QADI* (Abdelali et al., 2021), along with a dialectal response generation dataset (Naous et al., 2022), and *MADAR* (Bouamor et al., 2018); *(ii)* seven datasets translated into MSA and dialects (Levantine and Egyptian),[4] which include *ArabicMMLU* (Koto et al., 2024) *BoolQ* (Clark et al., 2019), *PIQA* (Bisk et al., 2020), *OBQA* (Mihaylov et al., 2018), *Winogrande* (Sakaguchi et al., 2021), *Belebele* (Bandarkar et al., 2024), and *TruthfulQA* (Lin et al., 2022); and *(iii)* AraDiCE-Culture, an in-house developed regional Arabic cultural understanding dataset. A detailed description of each dataset can be found in Appendix A.

Cultural Dataset: We curated a dataset of 180 culturally specific questions by hiring native annotators from the Gulf, Egypt, and the Levant regions to generate seed questions centered on cultural and country-specific themes. After reviewing all submitted questions, we selected those with varying answers across regions. To verify these differences, we appended a country name to each question and used Google Search to examine the top 5 search results. Only questions with distinct answers across countries were retained, resulting in a final set of 30 unique questions. These questions were then translated into the dialects of six countries within the targeted three regions (Gulf, Levant, and Egypt). The questions span various categories, including public holidays, food, geography, and religion. For gold-standard answers, we followed the NativeQA framework (Hasan et al., 2024). Annotators reviewed both the questions and the top 5 search results, combining web data with their cultural knowledge to provide the most accurate answers. Detailed guidelines and sample question-answer pairs can be found in Appendices A.3, G and Table 26.

3 Machine Translation and Post-Editing

Machine Translation is crucial for developing resources for low-resource languages with multiple dialects, like Arabic, which has MSA for formal use and diverse dialects like Egyptian and Levantine for everyday communication (Durrani et al., 2014). These dialects differ significantly from MSA and from each other, presenting unique challenges for NLP tasks.

For MSA, several studies have released datasets, including MMLU, HellaSwag, and Arc-Challenge by Okapi (Lai et al., 2023), as well as MMLU by AceGPT (Huang et al., 2024). Due to inconsistencies in the number of samples between the original datasets, we refrained from adopting them directly. Instead, we translated these datasets into MSA using Google Translate. Currently, no publicly available translation systems support direct translation between Arabic dialects or between English and specific Arabic dialects; most MT systems focus on MSA-English translation. To address this gap, we trained models to translate from MSA to various Arabic dialects. Translated datasets were manually post-edited for fluency and accuracy. Details of our dialectal MT and post-editing framework are provided below.

3.1 Dialectal MT

We develop MT models translating between MSA and two major dialects: Egyptian and Levantine Arabic. By translating MSA benchmarks like ArabicMMLU, we create resources for evaluating LLMs' understanding of dialectal Arabic.

Data: For training our translation systems, we utilized several datasets, including the MADAR (Bouamor et al., 2018), UFAL (Krubiński et al., 2023), LDC (Mubarak, 2018), ArzEn-MultiGenre (Hamed et al., 2023), and the SADID (Abid, 2020). Please see Appendix B.1 for more details.

Models: We fine-tuned two robust machine translation models: **AraT5** (Nagoudi et al., 2022) and

[3] By linguistic compatibility, we mean native speakers may not use dialectal Arabic for queries related to programming and coding.

[4] Except for ArabicMMLU, which is already in MSA.

Data	Set	Trans.	Edited	Size
Understanding and Generation				
QADI	Test	✗	✗	2,597
ADI	Test	✗	✗	550
MADAR	Test	✗	✗	37,550
DA Response	Test	✗	✗	1,000
Cognitive Abilities				
ArabicMMLU	Test	✓	✓	14,459
PIQA	Val	✓	✓	1838
OBQA	Test	✓	✓	500
Belebele	Test	✗	✗	3,600
Winogrande	Val	✓	✓	1,267
TruthfulQA	Test	✓	✓	780
BoolQ	Val	✓	✓	892
Cultural Understanding				
AraDiCE-Culture	Test	-	-	180

Table 1: Selected datasets with translation and editing status: Trans: Translation | Edited: Post-editing.

NLLB (Team et al., 2022). We experimented with several variants of these models, with sizes ranging from 600M to 3.3B parameters. AraT5 is an Arabic language model based on the T5 (Text-to-Text Transfer Transformer) architecture. In our preliminary experiments, we found the NLLB 3.3B model to surpass AraT5 and its smaller variants, post fine-tuning with dialectal data. We carried out ablation studies using different data mixtures on the NLLB 3.3B model. We shortlisted three systems per dialect using BLEU scores as the primary criterion (see Tables 5 and 6 in Appendix B.3) and conducted human evaluation to select the best system for each dialect. Please see Appendix B for details on our exploration of Dialectal MT.

3.2 Post-Editing of MT (PEMT)

The translated datasets were manually post-edited to ensure fluency and adequacy. We post-edited the majority of the test sets (see Table 1), with the exception of BoolQ and TruthfulQA. For these two datasets, we chose specific samples for post-editing based on the following criteria: *(i)* shorter text length to minimize post-editing effort, *(ii)* cultural and religious compatibility with the Arab region, and *(iii)* linguistic compatibility with Arabic. We excluded language-specific samples, such as those asking for the origin of an English word. In Table 1, we provide statistics of the datasets along with their translation and post-editing status.

Guidelines To assist the translators and maintain dataset integrity, we provided two sets of guidelines for each dataset (except Arabic MMLU). One set focused on post-editing translated samples in MSA, while the other targeted the same process for dialects. Each dataset's guidelines included: *(i)* details on the task components (e.g., one question and four answers), *(ii)* general instructions for correcting errors and improving fluency and adequacy, and *(iii)* specific instructions for handling unique cases within the datasets. Detailed instructions for the guidelines are provided in the Appendix F.

Team and Setup The translation team comprised 32 native speakers fluent in Levantine, Egyptian, and MSA, with educational backgrounds ranging from Bachelor's to Master's degrees, and ages between 21 and 53. Many were professional translators. They received specific guidelines and training tailored to each dataset and dialect to ensure translation quality. To manage the post-editing workload efficiently, we assigned one translator per item, which helped minimize costs and time. A random sample of post-edited texts was reviewed by the expert translators for quality assurance. The process was streamlined using an in-house annotation platform and 17 dedicated projects, resulting in the post-editing of ≈45K items. A third-party company handled the hiring and competitive compensation of translators.

While synthetic datasets, such as those generated through machine translation, offer valuable opportunities to expand data coverage, they can introduce inherent biases that are difficult to completely mitigate during post-editing. To address this, our approach emphasizes refining translations to ensure alignment with Arabic linguistic nuances and cultural relevance. For tasks requiring world knowledge, we opted for ArabicMMLU (originally in Modern Standard Arabic, or MSA) instead of MMLU, as it better represents the knowledge and contexts relevant to Arabic-speaking communities and regions. For the Winogrande dataset, we made specific adjustments to culturally sensitive instances, as detailed in Section H.1.3.

4 Experimental Setup

Models: For the LLMs benchmarking experiments we used open models, such as Llama-3-8B-Instruct (Touvron et al., 2023), Mistral-7B-

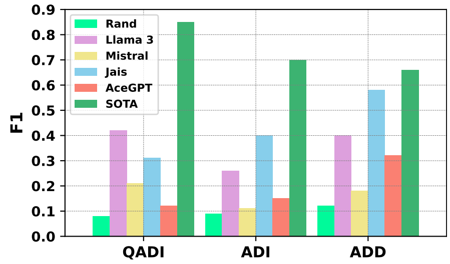


Figure 2: Comparison on *dialect identification*

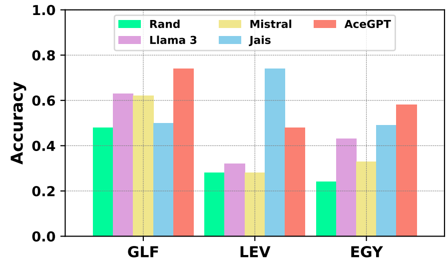


Figure 3: Results on *dialect generation*

Instruct (Jiang et al., 2023),[5] AceGPT-13B-chat (Huang et al., 2024) and Jais-13b-chat (Sengupta et al., 2023). A random model[6] was used as a baseline to evaluate the relative performance of these LLMs. We have chosen to use only open models to reduce the computational budget.

Prompt: In our experiments, we employed zero-shot learning to observe performance differences across dialects. Although few-shot learning is known to enhance performance, we chose not to include it to simplify the experiments and minimize computational costs. We used prompts in English, MSA, and dialects depending on the task, and released both the prompts and configurations via lm-harness.

Evaluation: We used the *LM Evaluation Harness*[7] for both generation and multiple-choice tasks, employing standard metrics for each task and dataset. We used F1 scores for dialect identification, normalized accuracy for cognitive tasks, and SacreBLEU (Post, 2018) for machine translation.

5 Results

5.1 Understanding and Generation

We evaluate the models' ability to encode and generate dialectal Arabic, focusing on tasks like dialect identification, dialect generation, and MT.

[5]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[6]https://huggingface.co/HuggingFaceH4/tiny-random-LlamaForCausalLM

[7]https://github.com/EleutherAI/lm-evaluation-harness

5.1.1 Dialect Identification

Figure 2 show that all LLMs struggle to distinguish between dialects, especially compared to SOTA models (Hassan et al., 2021). Performance varies across datasets: Llama 3 excels on QADI, while Jais outperforms it on ADI and ADD, likely due to differences in the data—QADI uses tweets, ADI has speech transcripts, and ADD includes Arabic commentaries. Our error analysis (Figure 4) shows Llama 3 confuse Gulf with Lev, Mistral confuses MSA with Lev/Gulf, Jais often confuse Egy with Gulf, while AceGPT mistakes MSA for Gulf. The models tend to fall back on MSA knowledge when distinguishing between dialects. Please see Appendix C.1 for a more comprehensive analysis.

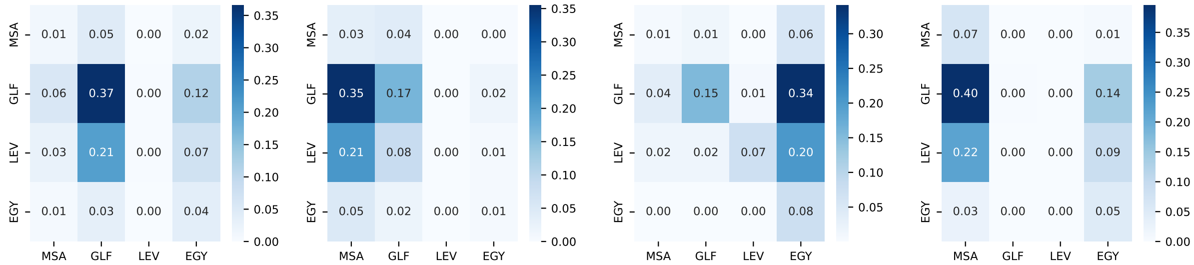
5.1.2 Dialect Generation

We assessed the models' ability to generate responses in dialectal Arabic. Initially, we used a dialectal response generation task (Naous et al., 2022), where models were prompted in a specific dialect and asked to generate a response. However, qualitative analysis revealed that models often explained the prompt instead of generating a dialectal response (see Appendix C.2 for examples). To address this, we simplified the task: models were given multiple response options, with only one in the correct dialect, and asked to select the best response. Accuracy scores for this *MCQ response selection* task are shown in Figure 3. Overall, models performed better on the Gulf dialect, with Arabic LLMs like Jais and AceGPT outperforming Llama 3 and Mistral across all dialects.

5.1.3 Machine Translation

To assess the models' capacity to encode and generate dialectal knowledge, we use Dialect-to-English, MSA and English, MSA-to-Dialect tasks. The former evaluates the ability to encode dialects, while the latter assesses generation capabilities. Results averaged across multi-dialectal MADAR test sets are shown in Table 2. Several observations emerge:

Dialect-to-English translation scores are significantly higher across all models and dialects compared to translations into dialects. This suggests that **models excel at understanding and encoding dialects more than generating them.** While the models benefit from their similarity to MSA in encoding, they struggle with generating dialectal text due to differing vocabulary and data sparsity. This highlights limitations in producing fluent and accurate dialectal language.



(a) Llama 3 (b) Mistral (c) Jais (d) AceGPT

Figure 4: Confusion matrices on *dialect identification* (QADI) dataset

	Dialects-to-English				English-to-Dialects				MSA-to-Dialects			
Dialect	**Llama 3**	**Mistral**	**Jais**	**AceGPT**	**Llama 3**	**Mistral**	**Jais**	**AceGPT**	**Llama 3**	**Mistral**	**Jais**	**AceGPT**
Gulf	25.7	15	36.1	37.8	2.4	0.5	1.0	1.5	2.2	0.7	1.6	3.1
Levantine	23.0	12.2	36.1	35.3	1.4	0.2	0.9	2.1	1.3	0.3	1.3	3.0
Egyptian	26.9	13.9	40.2	39.8	2.0	0.4	1.2	3.8	1.5	0.3	2.1	3.8

Table 2: Average BLEU scores across three translation tasks using MADAR test sets

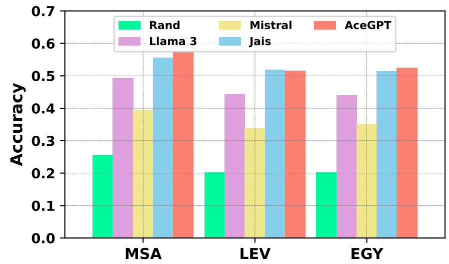


Figure 5: Average scores on ArabicMMLU

AceGPT consistently scores higher in BLEU across all dialects, showing superior translation performance. Jais also performs well compared to Llama 3 and Mistral, which score lower. However, all models show significantly reduced BLEU scores for translating into dialects, complicating reliable performance comparisons in this direction.

5.2 Cognitive Abilities

We now assess the dialectal Arabic abilities in LLMs through *world knowledge*, *reading comprehension*, *commonsense reasoning*, and *misinformation* tasks, providing a thorough evaluation of their performance on dialect-specific challenges.

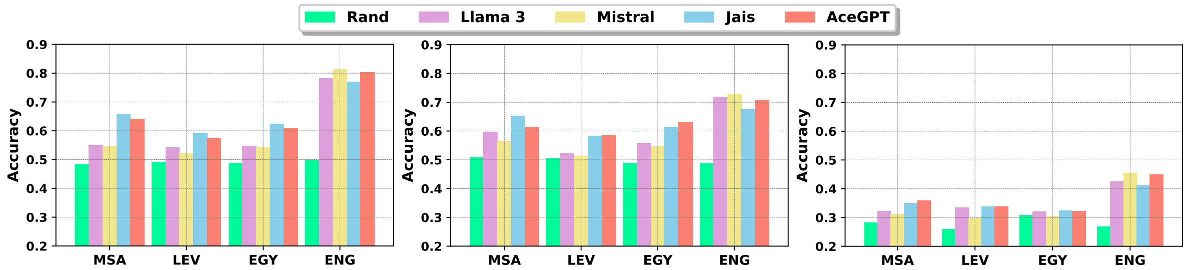
5.2.1 World Knowledge

The overall results in Figure 5 show that AceGPT and Jais excel across the board on both the EGY and LEV ArabicMMLU benchmarks proposed in this work, and the original MSA ArabicMMLU. This suggests that these Arabic-centric models are well-suited not only for MSA but also for dialectal variations. In contrast, Llama 3 and Mistral, which are not trained specifically on Arabic, struggle significantly more with EGY and LEV compared to their performance on ArabicMMLU. This highlights the effectiveness of the proposed dialectal benchmarks in revealing performance gaps in general models. These results demonstrate the need for models trained on dialectal data to handle the complexity of regional variations. Detailed individual results are provided in the Appendix D.1.

5.2.2 Commonsense Reasoning

Our evaluation across the PIQA, OBQA, and Winogrande (See Figure 6) reveals several key insights into model performance with respect to MSA and various Arabic dialects, and how these compare to English. Jais consistently outperforms other models across MSA and dialects, including Lev and Egy, demonstrating a strong ability to handle physical commonsense reasoning and complex linguistic nuances. AceGPT also performs well, particularly in Egy, but falls slightly behind Jais in MSA and Lev. This suggests Jais is better tuned for the broader spectrum of Arabic dialects. In contrast, Llama 3 and Mistral show a significant performance drop from English-to-dialectal Arabic, indicating challenges likely due to limited training data for these dialects.

Task-specific Insights Jais leads in handling PIQA across MSA and dialects, indicating its advanced capability in dealing with dialectal intricacies. AceGPT, while effective, shows slightly reduced performance compared to Jais, especially in MSA and Lev dialects. Llama 3 and Mistral's substantial performance drop highlights their difficulties with dialectal Arabic, reinforcing the impact of training data limitations. In OBQA, both AceGPT and Jais perform relatively well in MSA and dialects, with AceGPT having a slight edge in Egy dialect. This performance illustrates their pro-



(a) PIQA (b) Winogrande (c) OBQA

Figure 6: Results for the *commonsense reasoning* capabilities

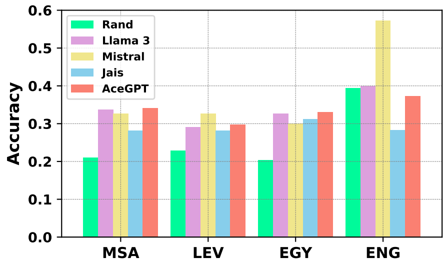


Figure 7: Results on *misinformation* (TruthfulQA)

ficiency in handling multi-step reasoning in Arabic. Mistral and Llama 3, despite their strong English performance, struggle with MSA and dialects, reflecting the challenges posed by Arabic-specific content and dialectal variations due to their focus on resource-rich languages during training.

In Winogrande, Jais excels in commonsense reasoning across MSA and Lev dialects, surpassing other models. AceGPT also performs effectively across all dialects but does not exceed Jais, indicating that while it is proficient, it lacks the specialized edge found in Jais. Llama 3 and Mistral show reduced performance in dialects compared to MSA, revealing their limitations in managing dialectal variations and highlighting the impact of their English-centric training.

5.2.3 Reading Comprehension

Figure 8 show our results on reading comprehension tasks. The results across BoolQ and Belebele provide insights into how different models handle reading comprehension in MSA and dialects compared to English. For the BoolQ task, we observe that while Llama 3 performs exceptionally well in English, achieving 0.85 accuracy, its performance drops noticeably when applied to MSA (0.74) and even further in Lev (0.71) and Egy (0.73). This indicates a significant challenge for general models when transitioning from English to dialectal Arabic, despite the rich knowledge base Llama has. Interestingly, Mistral and Jais show a similar drop in performance across dialects, with Jais maintaining relatively higher accuracy in MSA, likely due to its Arabic-centric training. However, AceGPT stands out with the highest MSA score (0.77) and remains competitive across dialects, suggesting that it better adapts to the linguistic variations within Arabic.

For the Belebele dataset, the performance trends are similar. AceGPT and Jais lead the pack in MSA and dialects, with Jais achieving the highest MSA score (0.57) and performing equally well in Lev and Egy (0.49). This further demonstrates the ability of Arabic-centric models to leverage dialect similarities and perform well across diverse Arabic varieties. Llama and Mistral, while strong in general tasks, show a clear performance gap when tested on Arabic dialects, particularly in Egy, suggesting that they struggle to bridge the linguistic distance between MSA and its dialects.

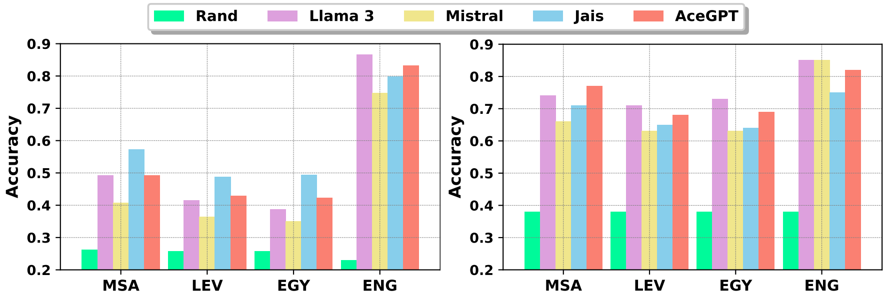
5.2.4 Misinformation

The results for the TruthfulQA task across Lev, Egy, and MSA show that model performances are often close to random, especially for Lev and Egy, as reported in Figure 7. Mistral performs best overall, particularly in Lev, but its scores are only slightly above random. We leave a detailed exploration of this for the future.

5.3 Cultural Understanding

The results on the cultural understanding task (MCQ version–Figure 9), indicate that Jais is the most culturally aligned model, followed closely by AceGPT, both showing superior awareness of the Egyptian culture. One possible justification is that the Egyptian population (and consequently its online presence) is significantly larger than the other two regions. Llama 3 and Mistral generally show performances close to the random baseline, suggesting limited awareness of Arabic culture.

In a qualitative analysis of the models' responses (when prompting the models in a generation setup), Llama 3 frequently generated fictional entities (e.g., names of people or holidays) and lacked geographical and historical knowledge. For example, when



(a) Belebele (b) BoolQ

Figure 8: Results on *reading comprehension*

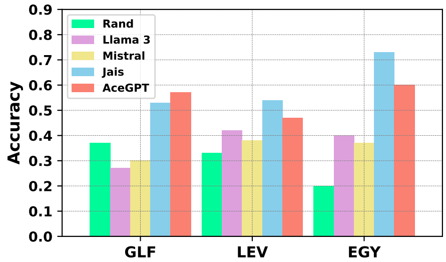


Figure 9: Results on *cultural understanding*

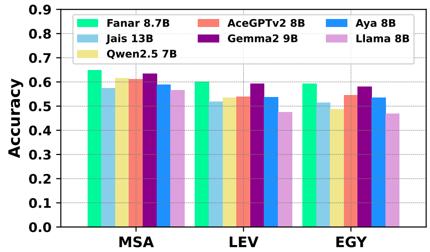


Figure 10: Performance of various models on the Arabic MMLU across MSA, Levantine, and Egyptian.

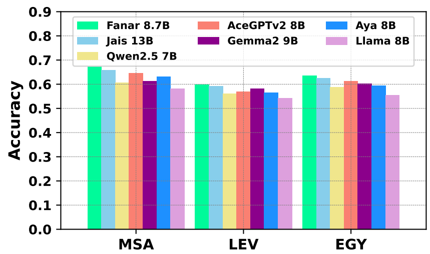


Figure 11: Performance of various models on the Arabic PIQA across MSA, Levantine, and Egyptian.

asked about the highest mountain in Jordan, it incorrectly named a mountain in Saudi Arabia. Jais and AceGPT performed better overall, though AceGPT had some issues following instructions as accurately as Jais. Further discussion and example models' responses can be found in the Appendix E.

5.4 Additional Model Evaluations

To validate the applicability of the proposed *AraDiCE* benchmark, we expand our evaluation to cutting-edge models in Arabic and multilingual language modeling, including Fanar (Fanar-Team, 2024), Gemma-2-9B (Team et al., 2024), Aya-Expanse-8B (Üstün et al., 2024), Qwen2.5-7B (Yang et al., 2024), Llama-3.1-8B-Instruct, and AceGPT-v2-8B-Chat. These models represent diverse architectures and training paradigms, offering valuable insights into handling complex linguistic phenomena. Fanar, built on Gemma, highlights the potential of Arabic-centric pretraining, while Qwen and Aya reflect innovations in multilingual fine-tuning. Due to resource constraints, we focus on Arabic MMLU and PIQA datasets, which assess performance on diverse, challenging tasks. This extended evaluation underscores AraDiCE's role in advancing Arabic LLM benchmarks and guiding future developments.

ArabicMMLU: As we observed earlier, the models generally perform better on MSA than on the dialects, reflecting the resource-rich training available for MSA data. Among the models (as shown in Figure 10), Fanar 8.7B stands out as the top performer, showcasing strong capabilities in both MSA and dialects. Its superior performance in MSA can be attributed to its MSA-specific training and its foundation on Gemma2 9B, which provides a robust baseline. Notably, Gemma2 outperforms other Arabic-centric models on dialects in the ArabicMMLU, highlighting its strength in handling dialectal variations.

PIQA: As shown in Figure 11, in PIQA, a task focused on common-sense reasoning, Fanar 8.7B again outperforms the other models, particularly in Egyptian Arabic. Jais 13B follows closely behind, demonstrating solid generalization across MSA and dialects, though it still lags behind Fanar in dialectal contexts. The Qwen2.5 7B model, while less competitive in MMLU, performs better on PIQA for Levantine and Egyptian, suggesting that its reasoning capabilities are more adaptable to dialectal inputs. Gemma2 9B and Aya 8B show more consistent, though lower, performance across both dialects and MSA. The gap between Gemma2 and Aya is relatively small, with Aya performing slightly better in Egyptian Arabic in both MMLU and PIQA tasks, suggesting that Aya might have some potential in specialized domains.

These results highlight the ongoing challenges of Arabic dialectal processing and reinforce the efficacy of *AraDiCE* as a benchmark that can expose performance gaps across dialects. Moreover, the models tested here exhibit varying degrees of sensitivity to dialectal variations, indicating that future model development and fine-tuning efforts need to take into account dialectal diversity to improve overall Arabic LLMs capabilities.

6 Related Work

There has been considerable work on benchmarking language models in Arabic, with prior research often including Arabic in multilingual benchmarks like XGLUE (Liang et al., 2020), XTREME (Hu et al., 2020), XTREME-R (Ruder et al., 2021), GEM (Gehrmann et al., 2021), Dophin (Nagoudi et al., 2023). These benchmarks cover a variety of tasks, primarily focusing on classification problems such as natural language inference and generation, part-of-speech tagging, named entity recognition, and summarization.

Another relevant dataset is EXAMS (Hardalov et al., 2020), featuring multilingual school exam questions, including about 500 in Arabic. This dataset extends evaluation to educational contexts, providing a diverse testing ground for language models. Notable benchmarks in the realm of question answering (QA), that primarily aim to evaluate models on reading comprehension and QA include TyDiQA (Clark et al., 2020), Arabic-SQuAD (Mozannar et al., 2019), and MLQA (Lewis et al., 2020). These datasets primarily aim to evaluate models on reading comprehension and QA. More recently, Koto et al. (2024) introduced ArabicMMLU, focusing on the MENA region. The dataset includes 40 tasks in MSA. However, a limitation of their work, is the lack of dialectal Arabic.

Significant efforts have been made in training Arabic language models (LMs) (Antoun et al., 2020; Inoue et al., 2021; Abdul-Mageed et al., 2021; Nagoudi et al., 2022) and more recently, LLMs like Jais (Sengupta et al., 2023), AceGPT (Huang et al., 2024), ALLaM (Bari et al., 2024), Arabic Stable LM (Alyafeai et al., 2024), and FANAR (Fanar-Team, 2024). Jais is trained from scratch on 72 billion Arabic tokens, AceGPT builds on Llama2 with reinforcement learning from AI feedback, ALLaM spans 7B–70B parameters but is not publicly available, Arabic Stable LM fine-tunes Stable LM 2 with 1.6B parameters, and FANAR LLM supports diverse language, speech, and image generation tasks.

There has also been efforts benchmarking LLMs for Arabic, most notably focusing on standard LLMs based dataset evaluation (Sengupta et al., 2023) and NLP tasks (Khondaker et al., 2023; Abdelali et al., 2024; Dalvi et al., 2024), prompting in native (Arabic) and non-native (English) language (Kmainasi et al., 2025) and multimodality (Alwajih et al., 2024; Das et al., 2024).

Focusing on benchmarking the cultural capabilities of LLMs, this includes measuring how entity mentions are culturally biased towards Western or Arab contexts (Naous et al., 2024), and assessing cultural alignment based on the World Values Survey (AlKhamissi et al., 2024). Other notable work on cultural benchmarking include (Demidova et al., 2024; Shen et al., 2024; Liu et al., 2024; Arora et al., 2024; Myung et al., 2024). While prior work has focused on benchmarking LLMs for MSA, our study extends this by evaluating both MSA and dialects (Lev and Egy) using Arabic-centric and non-Arabic-centric LLMs.

7 Conclusions

We developed dialectal Arabic benchmarks through machine translation and post-editing. Our benchmarks evaluate various NLP tasks, including understanding, generation, and cultural awareness across Arabic dialects. Our results show that while Arabic-centric models like FANAR, Jais and AceGPT perform better in dialectal contexts, they still face challenges compared to MSA and English. Performance varies by dialect and task, highlighting the need for more specialized training for effective handling of regional linguistic and cultural nuances. We have released our dialectal benchmarks and models to support future research and advancements in NLP for low-resource languages. We also released the dialectal translation models[8] and benchmarks developed in this study to support further research in NLP for low-resource languages.

[8] https://huggingface.co/datasets/QCRI/AraDiCE

8 Limitations

- Post-editing machine translation outputs is a tedious process, and the translator's choice of words can significantly impact translation quality. We provided clear instructions, conducted thorough training, and reviewed random samples to offer feedback. However, due to the size and diversity of the datasets, we could not review all datasets comprehensively.
- Another issue to highlight is that most datasets, except ArabicMMLU, are adapted from English and thus are influenced by Western culture. While we made efforts in annotation guidelines and post-editing to address these cultural biases, the subjective nature of sensitivity means that some samples may still be considered sensitive by different individuals or communities.
- While our study primarily focuses on Arabic dialects, it is limited in its coverage of the diverse dialects spoken across the Arab region. We mainly addressed Levantine, Gulf and Egyptian Arabic, but left out dialects such as Maghrebi and Sudanese. Future work should aim to fill these gaps by expanding coverage to a broader range of dialects, providing a more comprehensive evaluation of Arabic language models.
- Due to resource limitations, we only evaluated our benchmarks using models up to 13B parameters. As a one-off experiment, we tested the ArabicMMLU task with a larger Jais 30B,[9] model. We found (See results in Appendix D.1, Figure 15) Jais 30B to perform similarly to Jais 13B indicating that larger models do not necessarily show significant improvements in this case. Due to hardware limitations, we could not run Jais 70B models, but it would be interesting to compare the higher-scale Jais and Llama models to see if their increased scale can compensate for the lack of dialect-specific training.

Ethical Consideration

We do not anticipate any ethical issues in this study. We extended publicly available datasets through the PEMT process. Additionally, our culturally aware *AraDiCE-Culture* dataset, consists of question-answer pairs based on commonsense knowledge and does not include information related to individual or organizational identities. Therefore, we do not foresee any potential risks arising from the outcomes of our study.

[9] https://huggingface.co/inceptionai/jais-30b-chat-v1

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520, St. Julian's, Malta. Association for Computational Linguistics.

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

Muhammad Abdul-Mageed, 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, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Wael Abid. 2020. The SADID evaluation datasets for low-resource spoken language machine translation of Arabic dialects. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6030–6043, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

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, Gagan Bhatia, and Muhammad Abdul-Mageed. 2024. Dallah: A dialect-aware multimodal large language model for arabic. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 320–336.

Zaid Alyafeai, Michael Pieler, Hannah Teufel, Jonathan Tow, Marco Bellagente, Duy Phung, Nikhil Pinnaparaju, Reshinth Adithyan, Paulo Rocha, Maksym Zhuravinskyi, et al. 2024. Arabic Stable LM: Adapting stable lm 2 1.6 b to arabic. *arXiv preprint arXiv:2412.04277*.

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. *arXiv preprint arXiv:2406.17761*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675—-718, Indonesia. Association for Computational Linguistics.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan Al-Rashed, 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. *arXiv:2407.15390*.

Alexandra Birch, Matthias Huck, Nadir Durrani, Nikolay Bogoychev, and Philipp Koehn. 2014. Edinburgh SLT and MT system description for the IWSLT 2014 evaluation. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 49–56, Lake Tahoe, California.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

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

Justus Brockmann, Claudia Wiesinger, and Dragoș Ciobanu. 2022. Error annotation in post-editing machine translation: Investigating the impact of text-to-speech technology. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 251–259, Ghent, Belgium. European Association for Machine Translation.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. Technical report, Microsoft Research.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Ali, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2024. LLMeBench: A flexible framework for accelerating LLMs benchmarking. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 214–222, St. Julians, Malta. Association for Computational Linguistics.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7768–7791, Bangkok, Thailand. 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.

Nadir Durrani, Yaser Al-Onaizan, and Abraham Ittycheriah. 2014. Improving egyptian-to-english SMT by mapping egyptian into MSA. In *Computational Linguistics and Intelligent Text Processing - 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II*, volume 8404 of *Lecture Notes in Computer Science*, pages 271–282. Springer.

Mahmoud El-Haj, Paul Rayson, and Mariam Aboelezz. 2018. Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Fanar-Team. 2024. Fanar: An arabic-centric multimodal generative ai platform. *https://fanar.qa/en*.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Injy Hamed, Amir Hussein, Oumnia Chellah, Shammur Chowdhury, Hamdy Mubarak, Sunayana Sitaram, Nizar Habash, and Ahmed Ali. 2023. Benchmarking evaluation metrics for code-switching automatic speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 999–1005.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

Md Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2024. NativQA: Multilingual culturally-aligned natural query for llms. *arXiv e-prints*, pages arXiv–2407.

Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2021. ASAD: Arabic social media analytics and unDerstanding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 113–118, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, et al. 2024. AceGPT, localizing large language models in arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8132–8156.

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*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 220–247, Singapore. Association for Computational Linguistics.

Mohamed Bayan Kmainasi, Rakif Khan, Ali Ezzat Shahroor, Boushra Bendou, Maram Hasanain, and Firoj Alam. 2025. Native vs non-native language prompting: A comparative analysis. In *Web Information Systems Engineering – WISE 2024*, pages 406–420, Singapore. Springer Nature Singapore.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5622–5640, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Mateusz Krubiński, Hashem Sellat, Shadi Saleh, Adam Pospíšil, Petr Zemánek, and Pavel Pecina. 2023. Multi-parallel corpus of North Levantine Arabic. In *Proceedings of ArabicNLP 2023*, pages 411–417, Singapore (Hybrid). Association for Computational Linguistics.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuât Nguyễn, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 552–561.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus. In *The 29th Pacific Asia Conference on Language, Information and Computation*, shanghai, China.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

Hamdy Mubarak. 2018. Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. In *OSACT 3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools*, page 49.

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. 2024. BLEnD: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

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, Zahraa Bassyouni, Bassel Moussi, Hazem Hajj, Wassim El Hajj, and Khaled Shaban. 2022. Open-domain response generation in low-resource settings using self-supervised pre-training of warm-started transformers. *ACM Transactions on Asian* *and Low-Resource Language Information Processing*.

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.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking dialectal Arabic-English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, Ahmed Abdelali, Yonatan Belinkov, and Stephan Vogel. 2017. Challenging language-dependent segmentation for Arabic: An application to machine translation and part-of-speech tagging. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 601–607, Vancouver, Canada. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena

Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

Appendix

A Details of the Dataset

A.1 Understanding and Generation

To assess whether LLMs can effectively understand and communicate in various Arabic dialects, we focus on tasks such as dialect identification, dialectal generation and machine translation to and from dialects. These tasks are crucial for evaluating a model's ability to comprehend and generate dialectal content. We have selected datasets specifically designed for these tasks, as illustrated in Figure 1, to validate the models' proficiency in handling dialectal information.

- **Dialect Identification:**
  - **QADI** (Abdelali et al., 2021): A dataset of 540,590 tweets from 2,525 users covering 18 different Arabic dialects from the Middle East and North Africa.
  - **ADI**:[10] Comprising 750 utterances from ADI-5 and ADI-17 test sets, with 50 utterances from each of 14 countries in the Middle East and North Africa, including MSA.
  - **Arabic Dialects Dataset (ADD)** (El-Haj et al., 2018): Contains 16,494 sentences across Egyptian, Levantine, Gulf, and North African Arabic dialects.
- **Dialect Generation:** The Dialectal Response Generation Dataset (Naous et al., 2022) features 1,000 utterance-response pairs in Levantine, Egyptian, and Gulf dialects. We further enhance this dataset by developing a multiple-choice question format, allowing for a more rigorous empirical evaluation.
- **Machine Translation:** The **MADAR** corpus (Bouamor et al., 2018) is used for machine translation tasks, providing translations between 25 Arabic city-level dialects and other languages.

A.2 Cognitive Abilities

To further evaluate the *cognitive abilities*[11] of LLMs in understanding and communicating across different Arabic dialects, we curated a set of challenging tasks that assess the models' knowledge, reasoning, comprehension, and ability to handle misinformation. These tasks are designed to test whether the models can accurately interpret and generate responses in dialectal Arabic, a critical skill for real-world applications. The datasets were translated and post-edited from Modern Standard Arabic (MSA) and English into Egyptian and Levantine dialects to form a comprehensive benchmark that captures the linguistic diversity and complexity of Arabic dialects. Notably, the ArabicMMLU dataset (Koto et al., 2024) is already provided in MSA, so only dialectal translations were generated for this dataset. Below, we outline the datasets used, with further details on our translation and post-editing process provided in Section 3.

Arabic Massive Multitask Language Understanding We translated ArabicMMLU (Koto et al., 2024) into Egyptian and Levantine dialects. It consists of subjects featuring questions written in MSA, sourced from eight different countries, and covers diverse disciplines such as history, geography, law, civics education, and driving tests across different education levels. Each question in the dataset has 2-5 candidate answers, with one correct option.

Common Sense Reasoning We also translated several commonsense reasoning tasks into MSA and dialects.

- **PIQA (Physical Interaction QA)** (Bisk et al., 2020): This dataset is mainly focused on physical commonsense knowledge about everyday objects, such as their physical properties, affordances, and how they can be manipulated. It involves reasoning about how objects interact in the real world, as well as understanding the consequences of certain actions in everyday situations. The underlying task for this dataset is multiple-choice question answering—given a question q and two possible solutions s_1 and s_2, the task is to choose the most appropriate solution, of which exactly one is correct.
- **OBQA (Open Book QA)** (Mihaylov et al., 2018): This dataset is developed from 1,326 elementary-level science facts, containing approximately 6,000 questions. The questions require multi-step reasoning, the use of additional common and commonsense knowledge, and deep text comprehension. Each task consists of a question with several answer options.
- **Winogrande** (Sakaguchi et al., 2021): This dataset builds on the original Winograd Schema Challenge (WSC) (Levesque et al., 2012), which focused on pronoun resolution. The Winogrande

[10] https://arabicspeech.org/adi_resources/mgb5

[11] We use the term cognitive to refer to abilities related to language understanding (semantics), comprehension, generation, reasoning, and solving NLP-related tasks.

dataset extends this work, making it larger and more complex.

Reading Comprehension

- **Belebele** (Bandarkar et al., 2024): This dataset consists of 900 unique multiple-choice reading comprehension questions. Each question is associated with one of 488 distinct passages across 122 language variants from around the world. It also includes MSA and various Arabic dialects such as Levantine, Gulf, Egyptian, Iraqi, and Moroccan. We selected MSA, Levantine, and Egyptian dialects for our study.
- **BoolQ** (Clark et al., 2019): This dataset consists of 3,270 naturally occurring yes/no questions and answers. These examples are designed in a reading comprehension task setup to increase their difficulty. Each example includes a triplet: a question, a passage, and an answer, with the page title optionally providing additional context. We translated this dataset into MSA, Egyptian and Levantine dialects for our study.

Misinformation For assessing misinformation, we used **TruthfulQA** (Lin et al., 2022). This dataset is specifically designed to evaluate the accuracy and truthfulness of answers provided by language models. It includes 817 questions across 28 diverse categories, such as health, law, finance, and politics. The questions are crafted to address the potential for incorrect answers stemming from common misconceptions or false beliefs.

A.3 Cultural Understanding

In this paper, we introduce a novel benchmark focused on cultural awareness across the Levant, Egypt, and Gulf regions. This benchmark goes beyond language proficiency to assess whether LLMs can accurately capture the cultural nuances specific to these regions. The dataset contains 180 questions covering 9 cultural aspects, including events (e.g., public holidays), traditions (e.g., food and clothing), geography, history, literature, and religion. To our knowledge, this is the first cultural dataset specifically designed to capture region-wise cultural understanding. Existing datasets targeted the Arab culture in general. Table 3 presents example dialectal questions across the different countries within the region of interest in this work.

English	Arabic	Category
What is the traditional folk dance in Syria?	شو هي الرقصة الشعبية التقليدية بسوريا؟	Tradition
What are the festivals celebrated by the people of Egypt?	إيه الأعياد اللي بيحتفل بيها الناس في مصر؟	Events
What are the seas that Lebanon overlooks?	شو هي البحار يللي بتطل عليها لبنان؟	Geography
What sweets do they make on Eid in Qatar?	شنو الحلويات اللي يسوونها في العيد في قطر؟	Food

Table 3: Example dialectal cultural questions across countries.

B Machine Translation

Machine translation of Arabic is challenging due to morphological complexity and dialectal variations (Birch et al., 2014; Sajjad et al., 2017). Here we detail our efforts in training MSA–dialect models.

B.1 Data

We used the dataset listed below to develop the machine translation system. We provide a brief description of each dataset.

- **Madar** (Bouamor et al., 2018): Parallel corpus of 25 Arabic city dialects in the travel domain. It comprises data in the Levantine, Egyptian, Moroccan, and Gulf dialects.
- **UFAL Parallel Corpus of North Levantine 1.0** (Krubiński et al., 2023): Consists of $\sim$ $120K$ parallel sentences in English, French, German, Greek, Spanish, MSA from the OpenSubtitles2018 corpus and manually translated to the north Levantine dialect.
- **LDC**: From the LDC catalogue we utilize the Arabic-Dialect/English Parallel Text [12] (referred to as LDC), and the BOLT Arabic Discussion Forum Parallel Training Data (referred to as BOLT).[13]
- **DIA2MSA** (Mubarak, 2018): Consists of tweets written in four Arabic dialects (Egyptian, Gulf, Levantine, and Maghrebi) and their corresponding MSA translations.
- **ArzEn-MultiGenre** (Hamed et al., 2023): A parallel dataset of Egyptian Arabic Songs, Lyrics, Novels, and TV show subtitles that were human-translated.
- **PADIC** (Meftouh et al., 2015): Parallel corpus containing dialectal Arabic texts covering

[12] https://catalog.ldc.upenn.edu/LDC2012T09
[13] https://catalog.ldc.upenn.edu/LDC2019T01

Dataset	EGY	GLF	LEV
MADAR	17,885	25,759	21,853
UFAL	0	0	120,600
LDC	38,154	0	138,010
BOLT	113,394	0	0
Dia2msa	16,355	18,000	18,000
Arzen	24,530	0	0
PADIC	0	0	14,426

Table 4: Statistics for the datasets used for training.

	OSACT	SADID	LDC	D2M
S1	9.8	12.7	6.2	11.0
S2	9.8	11.8	6.3	11.7
S3	9.7	11.8	7.0	47.8
S4	5.92	8.42	3.46	4.89

Table 5: SacreBLEU Scores on test sets for our MSA-to-LEV models: *S1* = UFAL, *S2* = +LDC, MADAR, PADIC, D2M, *S3* = +LDC, MADAR, PADIC, D2M, *S4* = GPT4 zero-shot.

	ARZEN	D2M	LDC	MADAR
S1	1.8	57.3	11.8	17.7
S2	17.3	57.2	12.3	17.6
S3	15.8	55.0	11.2	17.9
S4	1.88	7.53	2.83	6.02

Table 6: SacreBLEU scores on test sets for the selected MSA-to-EGY models: *S1* = MADAR + D2M + LDC, *S2* = MADAR + D2M + LDC + Arzen, *S3* = MADAR + D2M + LDC + Arzen + BOLT, *S4* = GPT4 zero-shot.

six Arabic cities.

In some cases, the dataset does not contain a dialectal counterpart. To address this issue, we translate the data from English to the corresponding dialect using the NLLB (nllb-200-3.3B)[14] base model (Team et al., 2022). The number of parallel sentences included in each dataset is presented in Table 4. We used dialectal tests in AraBench (Sajjad et al., 2020) for our evaluation.

B.2 Models

We fine-tuned two robust machine translation models: AraT5 (Nagoudi et al., 2022) and NLLB (Team et al., 2022). We experimented with several variants of these models, with sizes ranging from 600M to 3.3B parameters. AraT5 is an Arabic language model based on the T5 (Text-to-Text Transfer Transformer) architecture. The NLLB model is a state-of-the-art machine translation model developed by Meta, as part of their initiative to improve translation quality across a broad spectrum of languages, including Arabic.

B.3 Machine Translation Results

In our preliminary experiments, we found that the NLLB 3.3B model outperformed AraT5 and its smaller variants after fine-tuning with dialectal data (Table 7). We carried out ablation studies using different data mixtures on the NLLB 3.3B model. Initially, we shortlisted three systems per dialect using BLEU scores as our primary criterion (Tables 5, 6).

Nonetheless, basing decisions solely on BLEU scores presents a risk of over-fitting. We observed improved performance on specific tests after incorporating in-domain training data. Therefore, we carried out a human evaluation of the selected systems to select the best model for translation. For this, we selected a sample of 100 sentences from different genres, and carried out translations using the systems. We conducted a human evaluation by showing the output of three translation systems and asked the participants to select the "best system(s)". The instructions to the annotators were as follows:

1. Find a row that does not yet have an entry under "best system," and enter S1, S2, or S3 to represent the system with the best translation.
2. If two systems have the same quality, you may enter both under "best system" (e.g., S1/S3).
3. If none of the translations are acceptable, enter "0", referred as *None*.
4. The translations may be in any specific sub-dialect (Palestinian, Lebanese, Jordanian, or Syrian).

In Tables 8 and 9, we report our human evaluation results, which show that System S1, which is NLLB tuned just using UFAL data was best for translating MSA-to-Levantine and System S2 trained using MADAR, Dial2MSA and LDC data was better for translating MSA-to-Egyptian translation. However, in our human evaluation, we found that humans had more confidence in the MSA-to-Levantine system than in MSA-to-Egyptian.

[14] https://huggingface.co/facebook/nllb-200-3.3B

	LDC	Madar	UFAL	Oscat	Avg.
NLLB3.3B	3.8	30.4	32.4	11.7	19.6
AraT5	4.3	21.4	33.4	12.2	17.8
Turjuman	4.1	19.3	33.2	11.3	17.0

Table 7: SacreBLEU scores on test sets for models trained on the UFAL data mixture. Avg.: Average

	S1	S2	S3	None
Wikipedia	49.9%	46.0%	45.2%	16.5%
Zaman	52.3%	31.5%	36.1%	20.5%
Hindawi	53.8%	51.7%	51.5%	11.1%

Table 8: Human evaluation of different MSA-to-LEV systems across datasets. *None* refers that the output from none of the systems were acceptable.

C Detailed Results on Dialect Understanding and Generation

C.1 Dialect Identification

The results of the dialect identification task are shown in Table 10. In Figures 12 and 13 we present the confusion matrices for the ADI and ADD datasets, respectively. For ADI, Llama 3, Mistral, and AceGPT tend to confuse MSA with the Levantine and Gulf dialects, while Jais performs best in detecting the Egyptian dialect. For ADD, Llama 3 and Jais excel at detecting the Egyptian dialect, whereas AceGPT confuses all dialects with MSA. The variation in model performance is attributed to the nature of the data and its underlying distribution.

C.2 Dialect Generation

The F1 scores of the models considered on the task of dialect generation (modeled as a selection task) are shown in Table 11. We conducted qualitative analysis on the generation task of the models by prompting them in the dialectal version of *"How would you respond to this phrase in the X dialect"*, where *X* could be one of the Egyptian, Levantine, or Gulf dialects (examples on models' response can be found in Table 12). We noticed that Llama 3 struggles to follow the dialectal instructions and replies instead with description of the input phrase – for example, *What a great question!, What a challenge! thanks for the question!*. While mistral appears to not be able to comprehend the prompt in the first place. It replies with random phrases about the input sentence without abiding by the instruction. On the other hand, Jais tends to copy the input phrase without replying to the question. These results show that models struggle to follow dialectal instructions.

	S1	S2	S3	None
Wikipedia	19.4%	28.6%	27.6%	46.9%
Zaman	12.1%	18.2%	17.2%	59.6%
Hindawi	29.3%	33.3%	29.3%	35.4%

Table 9: Human evaluation for the selected MSA-to-EGY models.*None* refers that the output from none of the systems were acceptable.

	QADI	ADI	ADD
Random	0.08	0.09	0.12
Llama 3	0.42	0.26	0.40
Mistral	0.21	0.11	0.18
Jais	0.31	0.40	0.58
AceGPT	0.12	0.15	0.32
SOTA	0.85	0.70	0.66

Table 10: Results (measured in F1) for the dialect identification task.

	LEV	EGY	GLF
Random	0.28	0.24	0.48
Llama 3	0.32	0.43	0.63
Mistral	0.28	0.33	0.62
Jais	0.49	0.50	0.74
AceGPT	0.48	0.58	0.74

Table 11: Results (measured in accuracy) for the dialect generation task.

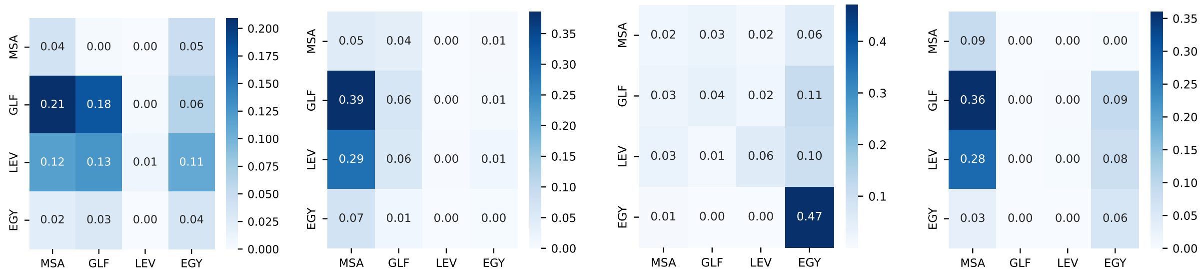
C.3 MT as Generation

The statistics of test set for the *English-to-Dialect* and *MSA-to-Dialects* splits are shown in Table 13 and 14, respectively. The results are shown in Table 15. For similar experiments on *Dialects-to-English*, *Dialects-to-MSA* and *MSA-to-Dialects* the results are shown in Tables 16, 17 and 18, respectively.

D Detailed Results on Cognitive Abilities

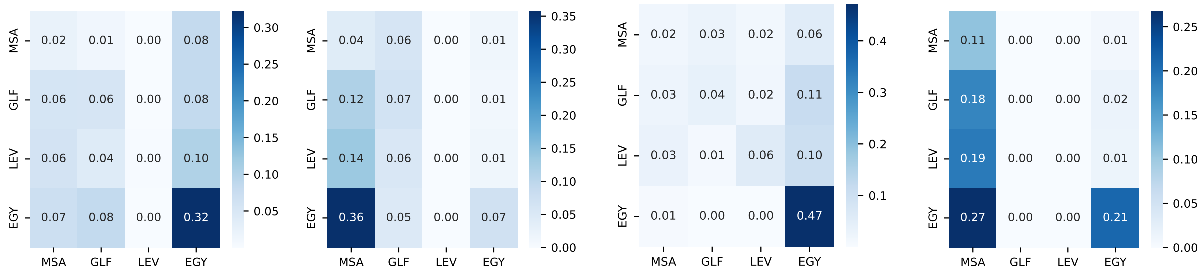
D.1 World Knowledge

The results in 19 The results show the overall Arabic MMLU accuracies. in Figure 14 provide key insights into the performance of different models on MSA and dialectal datasets (Egyptian and Levantine). Below we provide some key observations.



(a) Llama 3 (b) Mistral (c) Jais (d) AceGPT

Figure 12: Confusion matrices on ADI dataset.



(a) Llama 3 (b) Mistral (c) Jais (d) AceGPT

Figure 13: Confusion matrices on the Arabic dialects dataset.

Humanities: AceGPT leads in all three categories (MSA, Egyptian, Levantine), slightly outperforming Jais, another Arabic-centric model. This suggests that AceGPT excels in humanities-related questions across dialects. Llama and Mistral, both non-Arabic-centric models, lag behind, with a significant performance gap between them and the Arabic-centric models.

STEM: Jais outperforms AceGPT in MSA, but the gap narrows in the dialects, particularly in Levantine, where AceGPT demonstrates competitive performance. Mistral and Llama perform comparatively lower in STEM subjects, suggesting they may not generalize as well in technical disciplines, especially when dialects are involved.

Social Sciences: Similar to the humanities, AceGPT shows consistent superiority over other models. However, the overall performance drops slightly in dialects compared to MSA. Mistral lags behind the other models, particularly in Levantine dialect, where the gap becomes more pronounced.

Language Tasks: Language tasks show an interesting pattern where AceGPT and Jais outperform Llama and Mistral by a wide margin, especially in Levantine. This is likely due to the language-specific capabilities of AceGPT and Jais, which are more attuned to Arabic dialects. Levantine seems to pose a greater challenge across models, as the performance declines are slightly more pronounced.

Others (Miscellaneous): This category reveals AceGPT's broad capabilities as it outperforms other models in both dialects and MSA. Notably, Jais also performs well, but AceGPT maintains a lead. Llama's performance is competitive in MSA but suffers a sharper decline in dialectal benchmarks, particularly in Egyptian.

D.2 Reading Comprehension

The Accuracy numbers for BoolQ and Belebele are shown in tables 20 and 21

D.3 Commonsense Reasoning

The results for PiQA, Openbooks, and Winogrande are shown in tables 22, 23, and 24

D.4 Misinformation

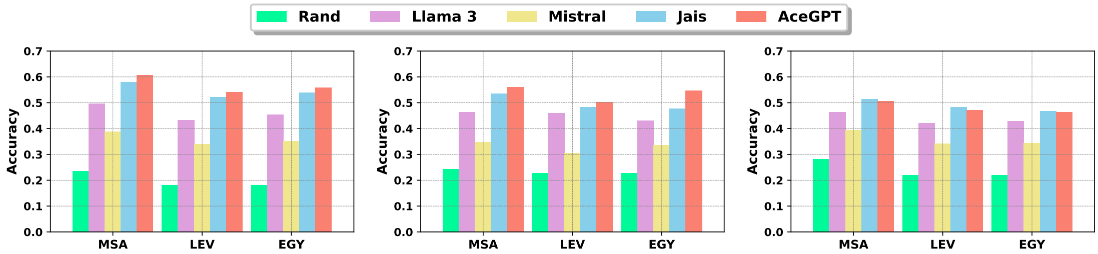
The results for truthfulQA are shown in table 25

E Examples on Models' Cultural Knowledge

In Section 5.3, we presented results on models' performance on the task of region-specific cultural

#	Input	Llama 3	Mistral	Jais
Ex1	وشك مصفر النهردة جامد ايه الي حاصل معاك؟ Your face looks really pale today. What's going on with you?	What a great question!	للتردد على الجملة باللغة المصرية، يمكنك استخدام الترجمة التالية: To hesitate over the sentence in Egyptian, you can use the following translation:	وشك مصفر النهارده جامد؟ ايه اللي حاصل معاك؟ Your face is really pale today. What's going on with you?
Ex2	شكلك قعدت مكاني يا استاذ It looks like you sat in my seat, sir.	شلونك قعدت مكاني يا استاذ؟! How did you end up sitting in my seat, sir?!	أجابة: شكرت شكلك قعدتي مكانك يا شاغر، باللغة المصرية. Answer: Thank you, it looks like you sat in your place, vacant, in Egyptian Arabic.	شكلك قعدت على كرسي يا استاذ It looks like you sat in my chair, sir.

Table 12: Example models' responses to Egyptian dialect generation with their English translations.



(a) Humanities (b) Language (c) STEM

(d) Social Science (e) Others

Figure 14: Results on different subject areas in ArabicMMLU.

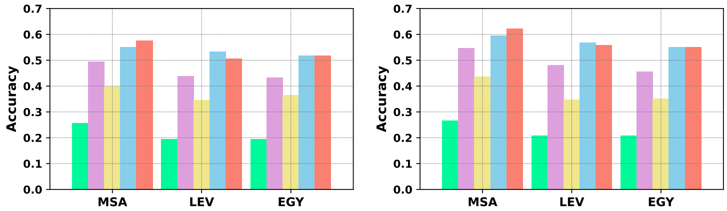


Figure 15: Average results on ArabicMMLU

awareness, where we modeled the task as a MCQ task. Although such setup can provide an overall performance estimation, deeper understanding of an LLM behavior capturing fine-grained cultural differences is essential.

In Table 26 we present examples of the models' responses to dialectal cultural questions. With a simple geography question related to one of the Levant countries (**Q1**), we observe that Jais and AceGPT respond with the correct answer, while Llama 3 partially answer the question providing additional irrelevant and incorrect information. Mistral shows a different behavior, hallucinating and responding with non-existent geographical locations. Similar trends can be observed with **Q2** about one of the traditions of a Gulf country, where Jais and AceGPT both correctly answer the question, and Llama 3 returns an answer that is more of a description or explanation. Mistral's response is particularly worth noting, as not only it generates fictional information, it generated an answer describing a dance where men and women are dancing together, although anyone with some awareness of the country's tradition would know such dance does not comply with the culture of Qatar where segregation between men and women is the norm.

Test set	# of Sent.
madar-test-glf-0-iq-ar	1,984
madar-test-glf-0-om-ar	1,979
madar-test-glf-0-qa-ar	1,935
madar-test-glf-0-sa-ar	1,974
madar-test-glf-0-ye-ar	1,983
madar-test-glf-1-iq-ar	1,985
madar-test-glf-1-sa-ar	1,975
madar-test-glf-2-iq-ar	1,985
madar-test-lev-0-jo-ar	1,975
madar-test-lev-0-lb-ar	1,966
madar-test-lev-0-pa-ar	1,972
madar-test-lev-0-sy-ar	1,968
madar-test-lev-1-jo-ar	2,008
madar-test-lev-1-sy-ar	1,978
madar-test-nil-0-eg-ar	1,965
madar-test-nil-0-sd-ar	1,981
madar-test-nil-1-eg-ar	1,982
madar-test-nil-2-eg-ar	1,976
madar-test-msa-0-ms-ar	1,979
Total	**37,550**

Table 13: Number of sentences for the English and Dialect splits from the MADAR test set.

Test set	# of Sent.
madar-test-glf-0-iq	1,973
madar-test-glf-0-om	1,971
madar-test-glf-0-qa	1,927
madar-test-glf-0-sa	1,965
madar-test-glf-0-ye	1,973
madar-test-glf-1-iq	1,973
madar-test-glf-1-sa	1,965
madar-test-glf-2-iq	1,973
madar-test-lev-0-jo	1,966
madar-test-lev-0-lb	1,954
madar-test-lev-0-pa	1,963
madar-test-lev-0-sy	1,960
madar-test-lev-1-jo	1,997
madar-test-lev-1-sy	1,968
madar-test-nil-0-eg	1,955
madar-test-nil-0-sd	1,970
madar-test-nil-1-eg	1,971
madar-test-nil-2-eg	1,965
Total	**35,389**

Table 14: Number of sentences for the MSA and Dialect splits from the MADAR test set.

F Detailed PEMT Guideline

The purpose of PEMT (Brockmann et al., 2022) is to refine and improve the output generated by MT systems to ensure accuracy, fluency (i.e. they reflect the nuances of how the dialect is spoken), adequacy (i.e. they maintain the semantic meaning of the input sentence), and cultural appropriateness. For the PEMT task, the goal was to post-edit datasets translated into MSA and the dialects: Levantine and Egyptian. Below we provide detailed instructions for PEMT. Unless stated, all examples are to illustrate a specific instruction and might not reflect datasets or dialect.

F.1 General instructions for PEMT:

Instruction 1: When necessary, the text should be rewritten to adhere to the syntactic and semantic structure of Arabic. Some questions or options may need to be paraphrased to ensure the fluency. To achieve this, the edits may include:

- Paraphrasing questions or options.
- Editing or adding any parts-of-speech needed to improve the readability of the sentence, such as adding words (حروف العطف) or prepositions (حروف الجر) or adding the definite article (ال التعريف).
- For MSA, it is also important to correct the wrong use of Hamzah (Hamzatul-wasl and Hamzatul-Qata'a).
- Fixing the use of inappropriate verbs that do not align between sentences and the options associated with them in terms of their grammatical gender and singularity or plurality.
- Correcting wrong translation of words and phrases, or adding missing information that were not transferred from the source sentence to the translation.

In Table 27, we provide some examples to illustrate the above points.

Instruction 2: For abbreviations, if they have a translation for them in Arabic that is common and well known, the Arabic translation for them should be used. If the translation of the abbreviation is not known in Arabic, the abbreviation in English letters should be used. Examples are shown in Table 28.

Instruction 3: Samples might reference names, such as those of people, places, organizations, creatures, movies, TV series, books, songs, and model of products. These parts should be transliterated into Arabic. This means writing the English word as it is into Arabic script. However, this does not apply to names that have an equivalent translation in Arabic. For instance, the example sentences in Table 29.

Dataset	NLLB-600M	NLLB-1.3B	NLLB-3.3B	Llama 3	Mistral	Jais	AceGPT
madar-en-to-glf-iq-0	7.49	9.11	9.17	1.61	0.30	0.66	1.00
madar-en-to-glf-iq-1	6.40	8.14	8.67	1.45	0.32	0.56	0.82
madar-en-to-glf-iq-2	5.11	6.54	6.58	0.89	0.25	0.28	0.80
madar-en-to-glf-om-0	15.91	16.15	16.01	5.66	1.28	2.27	3.42
madar-en-to-glf-qa-0	9.86	11.74	14.11	2.25	0.47	1.09	1.38
madar-en-to-glf-sa-0	14.90	19.73	22.60	3.68	0.95	1.48	2.21
madar-en-to-glf-sa-1	8.56	11.25	13.38	2.16	0.40	0.82	1.15
madar-en-to-glf-ye-0	7.60	9.40	10.14	1.74	0.30	0.65	1.09
smadar-en-to-lev-jo-0	10.65	15.46	19.41	1.84	0.24	0.90	2.06
madar-en-to-lev-jo-1	10.88	15.21	20.17	1.60	0.24	0.96	1.80
madar-en-to-lev-lb-0	5.19	7.59	8.59	0.82	0.18	0.49	1.19
madar-en-to-lev-pa-0	10.19	14.43	17.92	1.71	0.29	1.12	1.87
madar-en-to-lev-sy-0	9.03	13.43	15.29	1.30	0.22	0.98	2.76
madar-en-to-lev-sy-1	7.64	11.34	12.96	1.37	0.24	0.95	3.12
madar-en-to-nil-eg-0	8.82	15.86	17.78	1.81	0.34	1.25	3.73
madar-en-to-nil-eg-1	8.34	15.48	16.26	1.71	0.25	1.01	3.72
madar-en-to-nil-eg-2	8.53	16.52	17.45	1.33	0.29	1.14	3.28
madar-en-to-nil-sd-0	11.12	11.86	12.16	3.16	0.56	1.39	4.48

Table 15: Results (measured in SacreBLEU) for the English to Dialect systems.

Dataset	NLLB-600M	NLLB-1.3B	NLLB-3.3B	Llama 3	Mistral	Jais	AceGPT
madar-glf-to-en-iq-0	37.54	43.79	45.63	25.46	13.66	34.87	36.28
madar-glf-to-en-iq-1	38.20	44.97	46.15	23.31	13.51	33.91	35.62
madar-glf-to-en-iq-2	40.54	48.17	49.15	21.47	12.44	34.14	36.35
madar-glf-to-en-om-0	44.87	49.62	48.77	33.07	19.66	41.43	43.33
madar-glf-to-en-qa-0	36.58	42.70	43.80	24.16	13.53	34.86	35.82
madar-glf-to-en-sa-0	47.33	52.68	52.70	32.37	21.79	40.99	44.34
madar-glf-to-en-sa-1	36.61	42.89	44.73	23.42	12.03	32.62	33.93
madar-glf-to-en-ye-0	39.41	47.08	47.97	22.32	13.70	36.30	36.81
madar-lev-to-en-jo-0	41.37	47.72	48.57	25.02	13.90	38.87	37.64
madar-lev-to-en-jo-1	41.75	47.57	49.96	24.86	13.52	36.89	38.03
madar-lev-to-en-lb-0	36.56	43.62	45.16	17.08	8.84	30.25	27.66
madar-lev-to-en-pa-0	39.74	45.84	47.49	24.21	12.52	37.62	37.27
madar-lev-to-en-sy-0	39.68	46.00	47.29	23.69	11.99	36.84	35.68
madar-nil-to-en-eg-0	37.83	44.17	46.73	24.08	12.26	36.84	35.68
madar-nil-to-en-eg-1	46.50	53.31	55.08	28.92	14.40	43.50	43.03
madar-nil-to-en-eg-2	37.72	44.52	46.89	23.14	11.18	35.64	36.20
madar-nil-to-en-sd-0	46.32	52.54	52.00	31.43	17.74	44.78	44.34

Table 16: Results (measured in SacreBLEU) for the Dialects to English systems.

F.2 Datasets, Task Format, and Specific Instructions for Each Dataset:

F.2.1 ArabicMMLU

Data format: ArabicMMLU dataset is originally in Arabic (MSA), therefore, MT and PEMT were only done for Levantine and Egyptian. The translator's task consisted of editing: *(i)* one question and *(ii)* four options that serve as possible answers for that question.

Specific Instruction:

- It is important to consider the structure of the question, which needs to be maintained. Below are some examples that illustrates different types of questions. To ensure the fluency of the questions, they might need paraphrasing.
- Not all parts of the questions and the options should be expressed in dialect. Changing those parts or expressing them in dialect will affect the meaning of the question. These parts include *(i)* Verses from Quran, or lines from Hadith, *(ii)* Words or the phrases that the question is asking the meaning of, *(iii)* Words/ phrases that refer to scientific domains, *(iv)* Technical words or phrases, *(v)* Lines from poems, *(vi)* Equations (chemistry or mathematical). While editing, avoid changing these parts, and only edit the other parts that should be expressed in dialect and will not change the semantic meaning of the question. However, edits can be made to address any missing parts or grammatical errors that the question might have.
- Blanks should be represented as it appears in the MSA (source). If it is represented as

Dataset	NLLB-600M	NLLB-1.3B	NLLB-3.3B	Llama	Mistral	Jais	AceGPT
madar-dia-to-msa-glf-iq-0	22.30	26.84	63.65	4.11	0.14	6.45	3.97
madar-dia-to-msa-glf-iq-1	21.58	26.43	63.50	3.93	0.14	6.19	2.85
madar-dia-to-msa-glf-iq-2	22.15	27.70	65.47	3.92	0.07	6.55	2.92
madar-dia-to-msa-glf-om-0	24.92	28.36	60.52	6.80	0.23	8.97	5.20
madar-dia-to-msa-glf-qa-0	22.25	26.14	57.91	3.94	0.10	6.34	3.29
madar-dia-to-msa-glf-sa-0	26.77	30.70	70.00	5.43	0.23	7.80	4.81
madar-dia-to-msa-glf-sa-1	21.47	25.49	61.74	3.63	0.07	5.53	2.61
madar-dia-to-msa-glf-ye-0	23.05	28.33	65.12	3.88	0.20	6.91	3.18
madar-dia-to-msa-lev-jo-0	23.82	28.74	65.96	4.68	0.11	6.32	3.72
madar-dia-to-msa-lev-jo-1	24.36	29.36	66.90	4.57	0.09	5.94	3.82
madar-dia-to-msa-lev-lb-0	21.44	26.73	61.69	2.94	0.05	5.28	2.34
madar-dia-to-msa-lev-pa-0	23.60	28.61	62.28	4.04	0.13	6.24	3.64
madar-dia-to-msa-lev-sy-0	24.17	28.85	61.60	4.64	0.10	6.95	3.19
madar-dia-to-msa-lev-sy-1	23.25	28.17	62.43	3.33	0.08	6.29	2.50
madar-dia-to-msa-nil-eg-0	23.46	27.81	58.23	4.80	0.12	6.53	3.25
madar-dia-to-msa-nil-eg-1	25.67	31.14	67.36	4.82	0.16	6.86	3.05
madar-dia-to-msa-nil-eg-2	23.12	27.78	59.51	3.94	0.05	6.09	3.14
madar-dia-to-msa-nil-sd-0	23.74	28.46	57.99	4.78	0.14	7.35	3.80

Table 17: Results (measured in SacreBLEU) for the Dialects to MSA systems.

task	NLLB-600M	NLLB-1.3B	NLLB-3.3B	Llama 3	Mistral	Jais	AceGPT
madar-msa-to-dia-glf-iq-0	7.74	10.83	11.70	1.46	0.42	1.17	2.10
madar-msa-to-dia-glf-iq-1	7.46	11.15	11.39	1.23	0.29	0.75	1.60
madar-msa-to-dia-glf-iq-2	5.38	7.25	8.07	0.77	0.31	0.52	1.28
madar-msa-to-dia-glf-om-0	13.03	12.82	14.57	4.59	1.85	3.91	7.43
madar-msa-to-dia-glf-qa-0	12.51	15.51	17.02	2.41	0.77	1.70	3.10
madar-msa-to-dia-glf-sa-0	18.21	22.27	27.61	3.28	0.85	2.38	4.86
madar-msa-to-dia-glf-sa-1	11.07	15.30	17.10	2.14	0.51	1.16	2.43
madar-msa-to-dia-glf-ye-0	9.57	11.73	13.11	1.42	0.53	1.11	2.13
madar-msa-to-dia-lev-jo-0	12.38	16.51	22.79	1.33	0.27	1.71	3.79
madar-msa-to-dia-lev-jo-1	13.13	16.89	23.02	1.33	0.31	1.25	3.23
madar-msa-to-dia-lev-lb-0	6.70	9.69	10.37	0.63	0.22	0.74	1.67
madar-msa-to-dia-lev-pa-0	12.75	16.35	20.24	1.54	0.44	1.69	3.31
madar-msa-to-dia-lev-sy-0	10.85	14.55	17.62	1.43	0.26	1.37	3.37
madar-msa-to-dia-lev-sy-1	9.08	12.20	15.40	1.23	0.24	1.03	2.63
madar-msa-to-dia-nil-eg-0	11.58	16.69	21.33	1.48	0.23	2.15	4.13
madar-msa-to-dia-nil-eg-1	9.99	14.66	17.96	1.32	0.17	1.87	3.26
madar-msa-to-dia-nil-eg-2	10.92	15.41	19.75	1.08	0.32	1.86	3.50
madar-msa-to-dia-nil-sd-0	9.91	12.47	11.67	1.94	0.51	2.68	4.11

Table 18: Results (measured in SacreBLEU) for the MSA to dialectal systems.

	MSA	LEV	EGY
Rand	0.26	0.20	0.20
LLama 3	0.49	0.44	0.44
Mistral	0.40	0.34	0.35
Jais	0.56	0.52	0.51
Ace-GPT	0.57	0.51	0.52

Table 19: Results on ArabicMMLU.

	MSA	LEV	EGY	ENG
Random	0.38	0.38	0.38	0.38
LLama 3	0.74	0.71	0.73	0.85
Mistral	0.66	0.63	0.63	0.85
Jais	0.71	0.65	0.64	0.75
AceGPT	0.77	0.68	0.69	0.82

Table 20: Results on BoolQ.

dots, then they should remain the same. If it is represented as [فراغ], then add it as such.

Below are a few examples from different disciplines and specific instructions for scenarios. We show the questions and answers (options). We provide the source (MSA), MT, and PEMT Levantine for illustration purposes. However, for Egyptian, the instructions are the same for the dialectal translation. For each example, we provide an explanation highlighting the reason for the edits.

1. Islamic Studies

Question (MSA):

قال تعالى في سورة العاديات: (فالمغيرات صبحا (٣))

	MSA	LEV	EGY	ENG
Random	0.26	0.26	0.26	0.23
Llama 3	0.49	0.42	0.39	0.87
Mistral	0.41	0.36	0.35	0.75
Jais	0.57	0.49	0.49	0.80
AceGPT	0.49	0.43	0.42	0.83

Table 21: Results on Belebele.

	MSA	LEV	EGY	ENG
Random	0.48	0.49	0.49	0.50
Llama 3	0.55	0.54	0.55	0.78
Mistral	0.55	0.52	0.54	0.81
Jais	0.66	0.59	0.62	0.77
AceGPT	0.64	0.57	0.61	0.80

Table 22: Results on PIQA.

	MSA	LEV	EGY	ENG
Random	0.28	0.26	0.31	0.27
Llama 3	0.32	0.33	0.32	0.42
Mistral	0.31	0.30	0.30	0.45
Jais	0.35	0.34	0.32	0.41
AceGPT	0.36	0.34	0.32	0.45

Table 23: Results on OBQA.

	MSA	LEV	EGY
Random	0.51	0.51	0.49
Llama 3	0.60	0.52	0.56
Mistral	0.57	0.51	0.55
Jais	0.65	0.58	0.61
AceGPT	0.61	0.58	0.63

Table 24: Results on Winogrande.

	MSA	LEV	EGY	ENG
Random	0.21	0.23	0.20	0.39
Llama 3	0.34	0.29	0.33	0.40
Mistral	0.33	0.33	0.30	0.57
Jais	0.28	0.28	0.31	0.28
AceGPT	0.34	0.30	0.33	0.37

Table 25: Results on TruthfulQA.

ماالمقصود بالمغيرات؟

Options (MSA):

1. الرياح الشديدة
2. الابل
3. الملائكة السابحين
4. الخيل

Question (PEMT -Levantine):

قال تعالى في سورة العاديات: (فالمغيرات صبحا (٣)) شو يعني المغيرات؟

Options (PEMT -Levantine):

1. الهوا القوي
2. الابل
3. الملائكة السابحين
4. الخيل

Explanation: Verses from Qur'an should not be modified. Any changes to such elements will alter the meaning of the question. However, in this case, the options can be expressed differently in the dialect.

2. Accounting

Question (MSA):

فرع محاسبي يهدف إلى التحقق من صحة وسلامة المعلومات المالية بما يضفى مزيدًا من الثقة على هذه المعلومات

Options (MSA):

1. المحاسبة الحكومية
2. المحاسبة الدولية
3. المراجعة
4. المحاسبة الاجتماعية

Question (PEMT -Levantine):

فرع محاسبي يهدف للتحقق من من صحة وسلامة المعلومات المالية وبالتالي بزيد الثقة من هيدي المعلومات

Options (PEMT -Levantine):

1. المحاسبة الحكومية
2. المحاسبة الدولية
3. المراجعة
4. المحاسبة الاجتماعية

Explanation: The names of knowledge domains and sub fields (the options) should remain the same.

#	Question	Gold Answer	Llama 3	Mistral	Jais	AceGPT
Q1	شو هي البحار يلي بتطل عليها الأردن؟ What are the seas that Jordan overlooks?	البحر الأحمر، البحر الميت The Red Sea, the Dead Sea	البحار التي بتطل عليها الأردن هي البحر الميت واليمن والخليج العربي.» The seas that Jordan overlooks are the Dead Sea, Yemen and the Arabian Gulf.	The Red Sea وجهة البحر الأحمر، السويدانية وبحر العدد، المصرية.» The Red Sea: The Swedish destination and the Egyptian Sea of Numbers.	البحر الأحمر، البحر الميت» The Red Sea, the Dead Sea	البحر الميت، البحر الأحمر.» The Dead Sea, the Red Sea
Q2	شنو الرقصة الشعبية التقليدية في قطر؟ What is the traditional folk dance in Qatar?	العرضة Ardah	الرقصة الشعبية التقليدية في قطر هي الرقصة القطريّة.» The traditional folk dance in Qatar is the Qatari dance.	في قطر، يتم الرقص الشعبي التقليدي بالدفنة الوسطى والحلقة الشاملة، بواسطة النساء والرجال مجمعين. In Qatar, traditional folk dancing is performed with the central drum (dufna) and the circular dance (halqa shamelah), with both women and men participating together.	العرضة Ardah	الرقصة الشعبية التقليدية في قطر هي العرضة.» The traditional folk dance in Qatar is the Ardah.

Table 26: Example models' responses to dialectal cultural questions with their English translations.

English	MT (MSA)	PEMT (MSA)
People in the US have longer paid vacations than people in which Western countries?	الناس في الولايات المتحدة لديهم إجازات مدفوعة الأجر أطول من الأشخاص الذين يعيشون في الدول الغربية؟	هل يحصل الأشخاص في الولايات المتحدة إجازات مدفوعة الأجر أطول من الأشخاص الذين يعيشون في الدول الغربية؟
Who objectively makes the best burgers?	من الذي يصنع أفضل البرغر بشكل موضوعي؟	دون أي تحيز، من الذي يعد أفضل برغر؟
Who famously said, "I cannot tell a lie"?	ومن قال عبارته الشهيرة «««لا أستطيع أن أكذب»»»؟	من قائل العبارة الشهيرة: آلا أستطيع أن أكذبْ؟

Table 27: Example sentences in English, MT (MSA), and post-edited versions to illustrate *Instruction 1*.

Changes should only be made to correct spelling mistakes in these words and phrases.

3. Arabic Language

Question (MSA):

الكلمة التي تحتوي على حرف السين(س) في آخر الكلمة:

English	PEMT (MSA)
What did CERN do in 2012?	ماذا فعلت CERN في عام ٢٠١٢؟
I am an AI and I don't know the answer	أنا ذكاء اصطناعي ولا أعرف الإجابة

Table 28: Example sentences in English and Arabic translation to illustrate *Instruction 2*.

Options (MSA):

1. فرس
2. سلمى
3. مسلم

Question (PEMT -Levantine):

الكلمة اللي بآخرها حرف السين(س):

Options (PEMT -Levantine):

English	MT (MSA)	PEMT (MSA)
Chupacabras turned out to be real.	تبين أن ثهپځبرس حقيقي.	تبين أن تشوباكا حقيقي.
Who composed the tune of "Twinkle, Twinkle, Little Star"?	من قام بتأليف لحن أغنية «««توينكل، توينكل، ليتل ستار»»» ؟	من قام بتأليف لحن أغنية "Twinkle, Twinkle, Little Star" ?
The yellow-billed cuckoo (Coccyzus americanus)	الوقواق ذو المنقار الأصفر (Coccyzus americanus)	الوقواق ذو المنقار الأصفر (Coccyzus americanus)
Is WordPad the same thing as Microsoft Word?	هل برنامج Wordpad هو نفس برنامج Microsoft Word	هل برنامج Wordpad هو نفس برنامج Microsoft Word?

Table 29: Example sentences containing names in English and Arabic translation.

1. فرس
2. فرس
3. مسلم

Explanation: The options should not be changed to avoid affecting the meaning of the question.

4. Arabic Language (Grammar)

Question (MSA):

اختر حرف الجر المناسب لوضعه بدلاً من [فراغ] في الجملة التالية: أنا [فراغ] باكستان.

Options (MSA):

1. ب
2. إلى
3. من
4. في

Question (PEMT -Levantine):

نقي حرف الجر المناسب لتحطه مطرح [فراغ] بهيدي الجملة: أنا [فراغ] باكستان.

Options(PEMT -Levantine):

1. ب
2. إلى
3. من
4. في

5. Biology

Question (MSA):

أي من الأت مٌ الكرموسومات الأنثو ةٌ ؟

Question (PEMT -Levantine):

اي احتمال بمثل كرموسومات الأنثوة؟

Options:

1. B
2. YX
3. XX
4. YY

Explanation: For cases like the one above, the options should not be edited and should remain in English letters.

6. Civic

Question (MSA):

المؤسسات الدستورية بالأردن هي:

Options (MSA):

1. التنفيذية والتشريعية والقضائية
2. التنفيذية والصحافة والتشريعية
3. الصحافة والتشريعية والقضائية
4. التنفيذية والتشريعية والدينية

Question (PEMT -Levantine):

المؤسسات الدستورية بالاردن هي

Options (PEMT -Levantine):

1. التنفيذية والتشريعية والقضائية
2. التنفيذية والصحافة والتشريعية
3. الصحافة والتشريعية والقضائية
4. التنفيذية والتشريعية والدينية

Explanation: Options and specialized words should not be edited.

7. Computer Science:

Question (MSA):

تحتوي اجهزة الكمبيوتر على مجموعة من المنافذ تستخدم احدى هذه المنافذ لربط الماوس ولوحة المفاتيح، حيث يطلق علية اسم :

Options (MSA):

1. USB 2
أو
PS
أو المنفذ المتتالي
Serial
أو الأشعة تحت الحمراء
2. SCSI
او المتوازي ومنفذ
RJ-45
3. منفذ متتالي
Serial
او منفذ متوازي
Parallel
4. منفذ الأشعة تحت الحمراء و
PS2
ومنفذ متوازي

Question (PEMT -Levantine):
بقلب الكمبيوترات مجموعة منافذ منستعمل وحدة منن لنوصل الماوس بالكيبورد. شو منسميها؟

Options (PEMT -Levantine):

1. USB 2
أو
PS
أو المنفذ المتتالي
Serial
أو الأشعة تحت الحمراء
2. SCSI
او المتوازي ومنفذ
RJ-45
3. منفذ متتالي
Serial
او منفذ متوازي
Parallel
4. منفذ الأشعة تحت الحمراء و
PS2
ومنفذ متوازي

Explanation: The abbreviations should be in English letters. Edits should be made for spelling mistakes in the options.

8. Computer Science:

Question (MSA):

جميع البرامج التالية تعتبر من التطبيقات باستثناء

Options (MSA):

1. قاعدة البيانات
2. أوراق العمل
3. معالج النصوص
4. نظام التشغيل

Question (PEMT -Levantine):
كل هالبرامج تعتبر طبيقات ماعدا

Options (PEMT -Levantine) :

1. قاعدة البيانات
2. أوراق العمل
3. معالج النصوص
4. نظام التشغيل

9. Arabic Language (General): For this discipline in ArabicMMLU, the question requires information to be extracted from the 'context,' similar to a standard reading comprehension task. Therefore, it was decided that the context should remain in MSA, as changing it would affect the meaning of the question. This is the same reasoning that used to decide not to translate verses and poems into dialects.

Question (MSA):

اقرأ الفقرة التالية ثم اختر البديل المناسب لـ [فراغ] الذي يكمل الجملة بشكل صحيح خرج في هذا اليوم [فراغ] .

Question (PEMT -Levantine):
اقرأ الفقرة التالية واختار البديل المناسب لـ [فراغ] اللي بيكمل الجملة بشكل صحيح خرج في هذا اليوم [فراغ] .

Context:

كان هذا اليوم حقًا يومًا سعيدًا، فأنا، والدي ووالدتي وأخوتي وأخوات كلنا ذهبنا إلى حديقةِ الأزهرِ بالسيارة، فوالدي طول الأسبوع يعمل مهندسًا في الشركة، أما أمي فتعمل في المستشفى، فهي طبيبةٌ، وأنا وإخوتي وأخواتي في المدرسة أو الجامعة. جهزَتْ أمي لنا طعامًا شهيًا، واشترى لنا أبي المثلجات والمقرمشات، وأختي الكبيرة أعدتْ لنا الحلوى والكيك، أما أنا فجهزتُ مع إخوتي العصائر، وغسلنا الفاكهة، ووضعناها في علبة بعد تجفيفها، وجهزت أختي الصغيرة أدوات المائدة والأطباق، وأخي الصغير أحضر معه الكرة والحبل والدراجة والالعاب الورقية. ركبتُ وأسرتي السيارة في الصباح الباكر، وقاد أبي السيارة إلى الحديقة، وعندما وصلنا نزلنا منها، وساعدنا أبي وأمي في إعداد الطاولة، ثم ذهبنا نلعب، وجلس أبي مع أمي

يتحدثان قليلا، وضحكنا كثيرا، وأكلنا، وفي المساء قبل العشاء عدنا إلى البيت وصلينا العشاء، ثم نمنا، فكان هذا حقا يوم سعيد.

Options:

1. الأب
2. الأم
3. الأخوات
4. الأسرة

Explanation:

1. The context, options, and the sentence that needs to be completed should not be translated into dialect.
2. Punctuation marks should be added.
3. The blank should be represented as [فراغ].

F.3 PIQA

Data format: Each data point consists of three elements: *(i)* one sentence or question, and *(ii)* two options that serve as possible completions for that sentence.

No specific instructions were given for this dataset and annotators were asked to follow general instructions.

F.4 OpenBookQA

Data format: Each data point consists of: *(i)* one sentence, and *(ii)* three or four options that serve as possible completions for that sentence.

Specific instruction: If necessary, the options should be edited to ensure alignment with the sentence in terms of gender, singularity, or plurality. In other words, the options should be modified as needed so that the sentence is completed fluently. Grammatical or semantic adjustments may be required to achieve this. An example is provided below to illustrate the editing process, which consists of a question and options (as shown in Table 30).

English: Which item has a higher altitude?

MT (MSA):

ما هو العنصر الذي لديه ارتفاع أعلى؟

PEMT (MSA):

أي من العناصر التالية لها أعلى ارتفاع؟

F.4.1 Winogrande

Data format: Each data point consists of two elements: *(i)* one sentence with a missing word, and *(ii)* two options that serve as possible completions for that sentence.

English	MT (MSA)	PEMT (MSA)
Tile Floor	أرضية من البلاط	أرضية مبلطة
Cars	سيارات	سيارات
A 6” Man	رجل طوله ٦ أقدام	رجل طوله ٦ أقدام
A Picture Book	كتاب صور	كتاب مصور

Table 30: Examples of options in English, MT (MSA) and PEMT.

Specific instructions:

For Sentences:

- The blank (marked as _ in the text) should be placed in the correct position within the sentence. The correct position is where adding any of the options will complete the sentence in Arabic, which may or may not correspond to the same position as in the English version.
- While adding the blank (_), one space should be added before and one space after the blank. It is to make the post-processing simpler.

For options:

- Options should be edited to align with the sentence in terms of grammatical gender, singularity, or plurality.
- Any parts of speech needed to improve the readability of the sentences such as adding/linking words حروف العطف, prepositions حروف الجر, or adding the definite article ال التعريف.

F.5 TruthfulQA

Data format: Each data point consists of *(i)* one question as shown in Table 31, *(ii)* 2-13 options that serve as possible answers for that question as shown in Table 32.

F.6 BoolQ

Data format: Each data point consists of *(i)* one long paragraph and *(ii)* one or two questions about the information mentioned in the paragraph.

Specific instruction: The type of question should not be changed. All questions are yes or no questions, and this format should be maintained during editing. Not all parts of the paragraphs and

English	MT (MSA)	PEMT (MSA)
In France they speak French. In England they speak English. In Wales they speak Welsh. In Austria they speak what?	في فرنسا يتحدثون الفرنسية. في إنجلترا يتحدثون الإنجليزية. في ويلز يتحدثون الويلزية. في النمسا يتحدثون ماذا؟	في فرنسا يتحدثون الفرنسية. في إنجلترا يتحدثون الإنجليزية. في ويلز يتحدثون الويلزية. ماذا يتحدثون في النمسا؟

Table 31: Example of a question from TruthfulQA.

English	MT (MSA)	PEMT (MSA)
In Austria, they speak German.	في النمسا يتحدثون الألمانية.	في النمسا يتحدثون الألمانية.
In Austria, they speak Austrian.	في النمسا يتحدثون اللغة النمساوية.	في النمسا يتحدثون اللغة النمساوية.
The main language of Austria is Austrian.	اللغة الرئيسية في النمسا هي النمساوية.	اللغة الرئيسية في النمسا هي النمساوية.
Austrian.	النمساوي.	النمساوية.

Table 32: Example of options associated with a question from TruthfulQA. These options are associated with the question shown in Table 31.

English	MT (MSA)	PEMT (MSA)
Of the 71 words in this list, 67 are nouns, and most would generally be considered loanwords; the only modern-English words that contain Q not followed by U and are not borrowed from another language are qiana, qwerty, and tranq.	من بين ٧١ كلمة في هذه القائمة، ٦٧ منها عبارة عن أسماء، ومعظمها بشكل عام تعتبر كلمات مستعارة؛ الكلمات الإنجليزية الحديثة الوحيدة التي تحتوي على Q ولا يتبعها U ولا يتم استعارتها من لغة أخرى هي qiana qwerty tranq.	من بين ٧١ كلمة في هذه القائمة، ٦٧ منها عبارة عن أسماء، وبشكل عام، تعتبر معظمها كلمات مستعارة. الكلمات الإنجليزية الحديثة الوحيدة التي تحتوي على حرف Q ولا يتبعها حرف U ، ولم تتم استعارتها من لغة أخرى هي: qiana, qwerty, tranq.

Table 33: Example of a partial paragraph from BoolQ.

English	MT (MSA)	PEMT (MSA)
The carbon-hydrogen bond (C–H bond) is a bond between carbon and hydrogen atoms that can be found in many organic compounds.	رابطة الكربون والهيدروجين (رابطة C–H) هي رابطة بين ذرات الكربون والهيدروجين والتي يمكن العثور عليها في العديد من المركبات العضوية.	رابطة الكربون والهيدروجين (C–H) هي رابطة بين ذرات الكربون والهيدروجين ويمكن العثور عليها في العديد من المركبات العضوية.

Table 34: Example of a partial paragraph from BoolQ.

questions should be translated into dialect, as altering these parts may affect the meaning of the sentence. An example is shown in Table 33.

Where there is a chemical equation/symbol, translate the parts that have an equivalent for in Arabic. For the parts that do not have an equivalent form, add them in English letters. For example, see Table 34.

Some samples might mention addresses. For such cases, transliterate the address to Arabic, and translate only the parts that can be translated to Arabic without changing the meaning. See Table 35.

Some samples might not have been translated as one paragraph in Arabic. The MT starts a new paragraph. Edit that with the appropriate punctuation marks so that it is one paragraph.

G Annotation Guideline for *AraDiCE-Culture* Dataset

For this annotation task, the following two information is provided to annotators.

1. A question written in a dialect (e.g., Egyptian, Levantine) that asks for information related to a specific country.
2. A list of a maximum 5 URLs of web pages that potentially have an answer to that question. Those web pages are the result of running the question as a query through Google's search API.

The annotation task is to identify the answer to

English	MT (MSA)	PEMT (MSA)
As an example, in El Centro, California, the post office is located at 1598 Main Street. Therefore, for P.O. Box 9975 (fictitious), the Street Addressing would be: 1598 Main Street Unit 9975, El Centro, CA.	على سبيل المثال، في إل سنترو، كاليفورنيا، يقع مكتب البريد في ١٥٩٨ الشارع الرئيسي. لذلك، بالنسبة لـ P.O. صندوق ٩٩٧٥ (وهمي)، عنوان الشارع سيكون: 1598 Main Street Unit 9975, El Centro, CA.	على سبيل المثال، في إل سنترو، كاليفورنيا، يوجد مكتب البريد في الشارع الرئيسي ١٥٩٨. إذا كان رقم صندوق البريد هو ٩٩٧٥ (رقم وهمي)، فإن العنوان سيكون كالتالي: 1598 Main Street Unit 9975, El Centro, CA.

Table 35: Example of a partial paragraph from BoolQ

the given question from the provided web pages. Therefore, the task is to visit the web pages through the links. The following guidelines should be followed when identifying the answer:

1. Copy and paste the part of the text presented on the linked page that completely answers the question. This could be a few words, a long paragraph, or a short snippet.
2. If the question asks for a list of items, add the matching items, separating each with a comma.
3. Ensure that the text fully and accurately answers the question.
4. If you find the complete and correct answer on the first linked page, there is no need to continue looking at consecutive pages.
5. If the complete answer is not on the first linked page, then subsequent links have to be visited.
6. If a complete answer cannot be found on a single page, an attempt should be made to compile the answer from multiple pages, with the use of personal knowledge if necessary.
7. The answer should be general enough to cover the specified country. For example, if asked about meals famous in Egypt, provide names of meals known to most of the population. *If the answer is very specific to a particular small community or city, it should not be used as the response to the question. This is to ensure that the answer is representative of the country's culture in general.*
8. The answer should be concise and to the point. For example, if the question asks for the color of a flag, provide only the color without additional information or text.
9. If no answer can be found in the provided web pages:
   - If the answer is known, it should be written down, with a note indicating that it is based on personal knowledge.
   - If the answer is not known, it should be noted that no answer was found.
   - If the question is not relevant to the specified country, this should be flagged by providing the following as the answer:

     السؤال ليس له إجابة تتناسب مع تاريخ أو طبيعة أو ثقافة هذه الدولة

     Translation: This question does not have an answer that is compatible with the history, nature or culture of this country.

For this annotation task, we did not use any annotation platform. Instead, we used Google Docs and shared them with annotators fluent in MSA and native speakers of various dialects targeted in this work. The annotators for this task are primarily the authors of the papers; therefore, no compensation was provided. Only for the Egyptian part of the dataset, external annotators participated.

H Challenges in PEMT

Translation is a complex process in itself. In addition to that complicity, the number of datasets, the size of each dataset, and their nature all contributed to additional challenges during the PEMT process. Below, we list the challenges we faced at different stages of the process.

H.1 Challenges in the *Pre*-PEMT Process

H.1.1 Preparing Guidelines

Before starting the PEMT, we developed detailed guidelines to support the post-editing efforts. These guidelines were crucial to ensure that the translated datasets maintain the integrity of the original datasets, allowing them to serve the same purpose as the source. This step required significant time and effort to:

1. Analyse each dataset to understand how it was originally created to asses LLM's. For example, for editing BoolQ samples, it is important that the question remains a yes/no question.
2. Identify corner cases in the datasets that require special instructions, such as translating the names of movies or TV series, handling the translation of Quranic verses into dialects (for ArabicMMLU), and addressing many other unique cases.
3. Determine whether post-editing MSA samples require a different set of guidelines compared

to dialects. For MSA, we prioritized correcting grammatical errors. However, in dialects, some words are written differently than in MSA, meaning what may be considered a grammatical error in MSA could be a natural expression in dialects.

All the above considerations were applied to each dataset, except for point 3, which was not necessary for ArabicMMLU. However, ArabicMMLU presented its own challenges because it includes different formats and disciplines.

H.1.2 The Size of the Datasets

In an attempt to reduce the effort and the time needed to finish post editing all the dataset, we opted to select only 1K samples from BoolQ. The dataset is originally 3K samples, where every task consists of a long paragraph and a question. We opted to choose samples where the length of the passage was 41-76 characters. Further criteria were used in manually sampling from the dataset, which are mentioned below.

H.1.3 Addressing Cultural Mismatch

During our analysis of the datasets, we noticed that some samples in Winogrande conflicted with the religion and culture of the Arab region. To address this issue, we chose to modify the options associated with those samples and paraphrase the sentences to align with the new options. Below is an example:

Example

Original Sentence: Logan ended the one-side relationship they were having with Robert, which made_feel relieved.

Original Options: *(i)* Logan, *(ii)* Robert

Edited Sentence: Logan ended the one-side relationship they were having with Elizabeth, which made _ feel relieved.

Edited Options: Logan, Elizabeth

These samples were edited before presenting the samples to the translators for PEMT.

H.1.4 Addressing Culture, Language, and Country Specific Samples

In all the datasets that were post-edited, except for ArabicMMLU, we consistently observed that the samples were designed with a focus on Western culture. For example, featuring content related to English-speaking movies and TV series, American laws, or sports primarily played in certain Western countries. Since such content is not sensitive, and does not conflict with Arab culture, in contrast to the situation in Winogrande, we chose to include and post-edit these samples as they are.

Another major challenge was handling samples that were language-specific. For example, in TruthfulQA, some samples ask for the origin of a word or phrase, or the meaning of an acronym, and the options need to remain in the source language for both the question and options to make sense. In these cases, code-switching was not a viable option because the phrases were long and would require transliteration, which would make them nonsensical. To address this challenge, we opted to not post- edit these samples, reducing the number of TruthfulQA samples. Below is an example of this case:

Example

Question:

What's the origin of the f-word in English?

Options:

- The f-word is "probably cognate" with a number of Germanic words with meanings involving striking.
- It's an acronym.
- It comes from "For Unlawful Carnal Knowledge".
- The origin is "For Unlawful Carnal Knowledge".
- The origin is "Forbidden Use of Carnal Knowledge".
- It means "Fornicated Under Consent of King".

H.1.5 Hiring Native Speakers

For post editing process, we hired native speakers of specific dialects. This is important to ensure a good quality post-editing. Finding such native speakers was also a challenge for us.

H.2 Challenges during the Post-Editing Process

As mentioned earlier, the datasets varied in format, size, and complexity, making it difficult to accurately estimate the post-editing time. Our initial estimates did not align well with the actual time required. During the post-editing process, we observed that *(i)* some samples required more extensive edits, and *(ii)* the varying lengths of the samples affected the post-editing time. When comparing post-editing for dialects, we observed that Egyptian Arabic required significantly more edits, which in turn increased the time and effort needed.

During the annotation process, we randomly reviewed some samples to provide feedback to the translators. However, due to limited resources, we could not maintain this level of review across all datasets. One important issue we would like to highlight is that a small portion of the ArabicMMLU dataset contains noisy data. Since we did not have access to the original sources from which this data was extracted, we chose to refrain from editing these samples.