DIALECTALARABICMMLU: Benchmarking Dialectal Capabilities in Arabic and Multilingual Language Models

Malik H. Altakrori[1] Nizar Habash[2,3] Abed Alhakim Freihat[3]
Younes Samih[1] Kirill Chirkunov[3] Muhammed AbuOdeh[3]
Radu Florian[1] Teresa Lynn[3] Preslav Nakov[3] Alham Fikri Aji[3]
[1]IBM Research AI [2]New York University Abu Dhabi
[3]Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi,UAE
malik.altakrori@ibm.com, nizar.habash@nyu.edu, alham.fikri@mbzuai.ac.ae

Abstract

We present DIALECTALARABICMMLU, a new benchmark for evaluating the performance of large language models (LLMs) across Arabic dialects. While recently developed Arabic and multilingual benchmarks have advanced LLM evaluation for Modern Standard Arabic (MSA), dialectal varieties remain underrepresented despite their prevalence in everyday communication. DIALECTALARABICMMLU extends the MMLU-Redux framework through manual translation and adaptation of 3K multiple-choice question–answer pairs into five major dialects (Syrian, Egyptian, Emirati, Saudi, and Moroccan), yielding a total of 15K QA pairs across 32 academic and professional domains (22K QA pairs when also including English and MSA). The benchmark enables systematic assessment of LLM reasoning and comprehension beyond MSA, supporting both task-based and linguistic analysis. We evaluate 19 open-weight Arabic and multilingual LLMs (1B–13B parameters) and report substantial performance variation across dialects, revealing persistent gaps in dialectal generalization. DIALECTALARABICMMLU provides the first unified, human-curated resource for measuring dialectal understanding in Arabic, thus promoting more inclusive evaluation and future model development.

Keywords: Arabic dialects, benchmark, question answering, large language models, evaluation, MMLU.

1. Introduction

The rise of large language models (LLMs) has been enabled by the development of robust evaluation benchmarks capable of assessing not only their overall performance on Natural Language Processing (NLP) tasks, but also their linguistic adaptability across languages and varieties. Many recent efforts have therefore focused on multilingual benchmarks that evaluate model capabilities beyond English. In question-answering (QA), benchmarks are often first developed in English and later extended to other languages through translation-based adaptation, as in MLQA (Lewis et al., 2020) and TyDi QA (Clark et al., 2020). Such approaches have proven effective for broadening cross-lingual evaluation coverage.

In the case of Arabic, evaluation has largely centered on Modern Standard Arabic (MSA), the standardized written variety used in formal communication. This focus overlooks the diglossic nature of Arabic, where MSA coexists with diverse regional dialects that differ substantially in morphology, syntax, lexicon, and usage (Ferguson, 1959). While recent Arabic benchmarks have advanced coverage and modeling for MSA, they provide limited insight into LLM behavior on *Arabic dialects*, which dominate everyday communication, social media, and spoken interaction. As a result, current evaluation practices offer only a partial picture of real-world Arabic language understanding.

Here, we bridge this gap by introducing DIALECTALARABICMMLU,[1] a new benchmark specifically designed to evaluate LLM capabilities across five major Arabic dialects: Syrian, Egyptian, Emirati, Saudi, and Moroccan. Building upon the English-based MMLU-Redux (Hendrycks et al., 2020), we adopt the established paradigm of adapting high-quality English QA benchmarks, but extend it to dialectal Arabic through fully manual translation.

By situating dialects, not MSA, as the primary target of evaluation, DIALECTALARABICMMLU establishes a framework for quantifying dialectal understanding, reasoning, and general knowledge in Arabic. The benchmark enables controlled comparison across parallel dialectal variants while preserving semantic equivalence. Our contributions can be summarized as follows:

- We introduce DIALECTALARABICMMLU, the first large-scale, human-curated benchmark specifically designed to evaluate the reasoning and comprehension capabilities of LLMs across five major Arabic dialects.
- We develop a comprehensive dataset of more than 3K QA pairs per dialect (plus MSA and English), a total of 15K QA pairs, spanning 32 academic and professional domains, all produced and validated by native speakers to ensure linguistic fidelity and naturalness.

[1]Available on HuggingFace: Dialectal-Arabic-MMLU

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3199–3219
11–16 May 2026.

- We evaluate 19 open-weight Arabic and multilingual LLMs (ranging from 1B to 13B parameters) under three experimental settings (default, oracle, and dialect identification) to systematically assess the impact of dialectal variation on model performance.
- We conduct a detailed analysis of model behavior across dialects, revealing substantial performance disparities and highlighting the need for dialect-aware evaluation and training strategies for Arabic-enabled LLMs.

2. Related Work

Several benchmarks have been introduced to enable LLM evaluation for Arabic. ArabicMMLU (Koto et al., 2024) created an MMLU-like framework for MSA, drawing on real school and professional exams from around the Arab world. LAraBench (Abdelali et al., 2024) collected a comprehensive suite of 61 datasets spanning 33 diverse tasks across text and speech, establishing a multi-domain, multi-task evaluation platform for Arabic. More recently, BALSAM (Al-Matham et al., 2025) emerged as a community-driven, unified benchmark encompassing 78 NLP tasks from 14 broad categories with over 52K examples, with curated data covering diverse domains and various Arabic dialects. However, none of the above work had dialectal Arabic evaluation of LLMs as its main focus.

Complementing these efforts, 3LM (Boussaha et al., 2025) focused on bridging Arabic, STEM, and code, introducing a set of rigorous benchmarks to evaluate Arabic-enabled LLMs on scientific reasoning and programming tasks. It extended prior efforts by targeting domains that require structured problem-solving, such as mathematics, physics, and computer science, where Arabic benchmarks have been notably lacking. These initiatives, together with Arabic-centric models such as Jais (Sengupta et al., 2023), ALLaM (Bari et al., 2025), and Fanar (Abbas et al., 2025) significantly advanced evaluation for MSA. Yet, they remain centered on Modern Standard Arabic and offer only minimal insight into dialectal Arabic performance.

Extensive research, benchmarks, and shared tasks have targeted dialectal Arabic *identification* and *translation* (Bouamor et al., 2019; Abdul-Mageed et al., 2021, 2022, 2023, 2024), which was complemented by several datasets and tools (Zaidan and Callison-Burch, 2014; Bouamor et al., 2014; Salama et al., 2014; Alsarsour et al., 2018; Abu Kwaik et al., 2018; Salameh et al., 2018; Bouamor et al., 2018; Abdelali et al., 2021; Baimukan et al., 2022). Unlike this work, which has focused on just two tasks, our focus here is on general dialectal understanding.

Moreover, new benchmarks have begun to address dialectal and cultural dimensions in Arabic, albeit only partially. PALM (Alwajih et al., 2025) introduced a year-long, human-curated dataset of over 10K instruction–response pairs covering all 22 Arab countries in both MSA and dialects, across 20 culturally salient topics. While PALM effectively exposed critical gaps in model performance on culturally grounded and dialectal instructions, its design follows an instruction-tuning paradigm rather than a multitask reasoning framework.

AraDiCE (Mousi et al., 2025) offers the most comprehensive attempt so far at LLM evaluation for dialectal Arabic, contributing 45K post-edited dialectal examples across Egyptian, Syrian, and Gulf varieties of Arabic and introducing a fine-grained cultural question–answering component. While being a significant step toward dialectal evaluation by extending MMLU-style tasks to Arabic dialectal varieties, it is largely derived from machine-translation followed by post-editing. In contrast, our benchmark is entirely human-translated by native speakers of the target Arabic dialects and double-checked by other native speakers, ensuring linguistic naturalness, idiomatic precision, and cultural authenticity.

Belebele (Bandarkar et al., 2024) extends coverage to several Arabic dialects in a controlled reading comprehension setting, but remains limited in task scope and does not assess multi-domain reasoning or knowledge in Arabic dialects.

JEEM (Kadaoui et al., 2025) expands benchmarking into a multimodal setting by evaluating image captioning and visual question–answering across Jordanian, Emirati, Egyptian, and Moroccan Arabic dialects, revealing that current Arabic vision–language models (VLMs), including GPT-4V (OpenAI et al., 2024), struggle with dialect-specific visual understanding and show uneven competence across dialects.

Alyah (Alkaabi et al., 2026) is an Emirati-dialect benchmark that focuses on evaluation for Gulf and Emirati Arabic, offering 1,173 culturally grounded examples sourced from native speakers and covering idioms, greetings, figurative expressions, etiquette norms, and other dialect-specific linguistic features. However, it is limited to a single dialect, and is smaller in size compared to our dataset.

In contrast to the above work, we introduce a large-scale, parallel benchmark explicitly designed for understanding multiple dialectal Arabic. Unlike prior benchmarks that position dialects as secondary to MSA or cover a single dialect, our benchmark features balanced coverage across major regional varieties and relies exclusively on human-curated translation and dialectal adaptation, enabling principled evaluation of reasoning and comprehension across dialects.

Dialect	Question	Ch. 1	Ch. 2	Ch. 3	Ch. 4
MSA	هل تتغير السمات مع التقدم في العمر؟	لا	تتغير قليلاً	فقط لدى النساء	تتغير كثيرًا
EGY	هل الصفات الشخصية بتتغيّر لما الواحد بيكبر في السن؟	لا	بتتغير شوية	بس عند الستات	بتتغير كتير
KSA	تتغير السمات مع التقدم في العمر؟	لا	تتغير شوية	بس عند الحريم	تتغير كثير
MAG	واش كيتبدلو السمات (ديال الشخصية) مع التقدم فالعمر؟	لا	كتبدّل شوية	غير عند العيالات	كتتبدّل بزاف
SYR	بتتغير ميزات البني آدم بس يكبر؟	لأ	بتتغير شوي	بس عند النسوان	بتتغير كتير
UAE	هل تتغير الصفات الشخص مع كبر العمر؟	لا	تتغير شوي	فقط عند النساء	تتغير بشكل كبير
ENG	Do traits change with age?	No	They change a little	Only for women	They change quite a lot

Table 1: An example from the **Human Aging** domain, showing the variation in the questions and the choices across dialects/languages. (QID 20, the correct answer is choice 2: "*They change a little.*")

3. Data Collection and Quality Assurance

DIALECTALARABICMMLU, is based on the translation of over 3,135 English (**ENG**) multiple-choice question–answering (MCQA) pairs into five dialects representing the geographical spread of the Arab World: Egypt (**EGY**), Morocco (**MAG**), Saudi Arabia (**KSA**), Syria (**SYR**), and the United Arab Emirates (**UAE**). We also include **MSA** and English: a total of 21,945 MCQA pairs (see Table 1 for an example).

Dataset Our dataset is based on MMLU-Redux-v2 (Gema et al., 2025),[2] a high-quality re-annotated subset of MMLU. We selected 32 domains from it and translated the corresponding QA-pairs to the above five dialects. Below are the fields and the domains in each field:

- **Humanities**: High School US History, High School World History, International Law, Moral Scenarios, Philosophy, Prehistory, Professional Law, World Religions;
- **Stem**: Abstract Algebra, Anatomy, Astronomy, College Computer Science, Conceptual Physics, Elementary Mathematics, High School Chemistry;
- **Social Sciences**: High School Geography, High School Macroeconomics, High School Psychology, Professional Psychology, Public Relations, Security Studies, Sociology, US Foreign Policy;
- **Other**: Business Ethics, Clinical Knowledge, College Medicine, Global Facts, Human Aging, Management, Marketing, Nutrition, Virology.

The translation was carried out in two main phases: manual translation and quality assurance.

Manual Translation The translations were outsourced to a language service provider (LSP) and carried out manually by teams of native or near-native speakers of each dialect. To ensure consistency, the process was guided by a detailed translation guidelines document (see Appendix A).

[2]Available from HuggingFace: mmlu-redux-2.0

The guidelines emphasized three principles:

- **Correctness**: accurately reflect the source meaning;
- **Naturalness**: sound authentic in the target dialect, even if close to MSA;
- **Simplicity**: use concise, conversational language.

Additional instructions included avoiding over-dialectalization, using MSA terms where contextually natural, and respecting natural spelling variation in the dialectal orthography.

Each dialect team had members with three roles:

- **Translator (Native/Near-native)**: produced the initial translation;
- **Reviewer (Native)**: reviewed and accepted/rejected annotations with justification;
- **Adjudicator**: resolved the disagreements and ensured the final quality.

Before starting, the translators attended two training sessions focusing on the workflow and the guidelines, to clarify the expected outcome and to align across the teams.

Quality Assurance To assess the translation quality, we carried out an in-house validation step. For each dialect, we sampled 32 QA pairs from eight domains, and native speakers independently scored the translations on a Likert scale of 1–5:

1. The translation is incorrect.
2. The translation is partially correct.
   (e.g., contains some inaccuracies, contains MSA where dialectal terms exist, etc.).
3. The translation is acceptable
   (e.g., contains the original meaning but could be improved in terms of formulation, fluency).
4. The translation is good
   (but I would translate it differently).
5. The translation is very good.

This evaluation step showed that only about 5% of the translations contained some inaccuracies. Almost all dialectical translations had average assessments between good (4) and very good (5) levels, except for the UAE dialect, which was only marginally below good (at 3.94).

Dial.	SYR	UAE	KSA	MSA	MAG	EGY
SYR	1.0	.45	.55	.51	.34	.52
UAE		1.0	.59	.61	.37	.51
KSA			1.0	.70	.39	.61
MSA				1.0	.41	.59
MAG					1.0	.37
EGY						1.0

Table 2: Binary Jaccard word-level similarity between the dialectal Arabic pairs.

We further performed linguistic analysis, which revealed that some of the UAE translations had a bias towards the use of Saudi dialectal words instead of more common terms used in the region. For Syrian, some translators had a tendency to translate the concept rather than the original text. Additionally, there were issues with the translation of pronouns. This evaluation effectively resulted in a second round of translation-review revision for the UAE and Syrian QA pairs, improving the quality of the final dataset.

Similarity between Dialects To further understand the nature of the data, we examined the lexical variations between the dialects using Jaccard similarity. We took the frequency counts for all words in both the question and the choice columns and passed them through a preprocessing step using the CAMeL Tools (Obeid et al., 2020) for Unicode character normalization, dediacritization, and whitespace tokenization. Table 2 shows the Jaccard similarity between the various dialect pairs.

When comparing our results to those of Salameh et al. (2018), we observe several differences. While we found MAG to be the most distant dialect from MSA, they reported it as relatively closer, and whereas they placed KSA much further from MSA, we found it to be the closest one. These discrepancies may stem from differences in data: their study used travel expressions, while our MMLU questions are more technical.

In both cases, EGY is closer to MSA than SYR. We provide the weighted Jaccard similarity measure (Table 12), and the Manhattan (Table 13) and Euclidean (Table 14) distances in Appendix C.

Dataset Statistics Table 3 shows some dataset statistics: number of dialects, domains, questions, translations, including both average and total counts. Table 4 further zooms into the average length of the questions and the answers across the investigated dialects/languages in our dataset: we can see that the lengths are similar across the Arabic varieties, with English being slightly longer. We provide detailed question counts for each dialect–topic combination in Table 15 in Appendix D.

	Characteristic	Value
1.	# of dialects	5, plus ENG & MSA
2.	# of domains	32 domains (in 4 fields)
3.	# of Qs/domain	$\simeq$ 98.0 Qs (68–100)
4.	# of Qs/dialect	3,135 Qs
5.	# of Qs translated	15,675 Qs (5 * 3,135)
6.	# of Qs in total	21,945 Qs (7 * 3,135)

Table 3: Dataset statistics.

Dialect	Questions		Choices	
	Chars	Words	Chars	Words
EGY	171.5	29.5	130.4	22.1
KSA	170.3	29.8	129.9	21.8
MAG	179.8	30.0	137.3	22.5
SYR	162.7	27.8	121.0	20.4
UAE	163.1	27.7	125.5	20.9
MSA	177.5	30.3	132.3	22.1
ENG	212.5	35.9	157.1	24.7

Table 4: Average question and choice length (characters and words) across the dialects.

4. Experimental Setup

In this section, we present our experimental setup, the LLMs we experiment with, and the evaluation setup.

4.1. Evaluated Arabic-Enabled LLMs

We evaluated language models of 1B to 13B parameters. These models are considered small- to medium-size compared to frontier models such as the 120B version of OpenAI's GPT-OSS.

We based our LLM selection on (Ouda, 2025), which provides a comprehensive survey of Arabic language models across different sizes, including both open-weight and proprietary ones. For our experiments, we restrict ourselves to open-weight LLMs only.[3]

We made sure that we included the three main Arabic-enabled LLMs that were developed in the Arab region: ALLaM (Bari et al., 2025), Fanar (Abbas et al., 2025), and Jais (Sengupta et al., 2023), in addition to recent, multilingual models of comparable sizes such as Google's Gemma-3 (Team, 2025) and Cohere Labs' Command R7B (Alnumay et al., 2025). Note that these models do not distinguish between MSA and dialects; rather, all the dialects and MSA are considered as just Arabic. A list of these models, their sizes, and whether they support Arabic, English, or both, is provided in Table 5.

[3]All selected models are publicly available on HuggingFace.

	Family	Model	Size	Ar	En
1	inceptionai	jais-13b-chat	13.0	•	•
2	google	gemma-3-12b-it	12.2	•	•
3	MBZUAI-Paris	Nile-Chat-12B	11.8	•	
4	silma-ai	SILMA-9B-Instruct	9.2	•	•
5	QCRI	Fanar-1-9B-Instruct	8.8	•	•
6	CohereLabs	command-r7b-arabic	8.0	•	•
7	CohereLabs	aya-expanse-8b	8.0	•	•
8	tiiuae	Falcon-H1-7B-Instruct	7.6	•	•
9	mistralai	Mistral-7B-Instruct	7.2	•	•
10	ALLaM-AI	ALLaM-7B-Instruct	7.0	•	•
11	Navid-AI	Yehia-7B	7.0	•	•
12	inceptionai	jais-6p7b-chat	6.8	•	•
13	google	gemma-3-4b-it	4.3	•	•
14	Qwen	Qwen3-4B-Instruct*	4.0	•	•
15	MBZUAI-Paris	Nile-Chat-4B	3.9	•	
16	UBC-NLP	NileChat-3B	3.1	•	•
17	tiiuae	Falcon-H1-3B-Instruct	3.1	•	•
18	inceptionai	jais-2p7b-chat	2.7	•	•
19	stabilityai	ar-stablelm-2-chat	1.6	•	

Table 5: The evaluated language models. (* Based on Qwen2 language support)

4.2. QA Evaluation Setup

For our experiments, we adopted the LM-Eval-Harness framework (Gao et al., 2024), which is a community-supported tool that contains a suite of evaluation tasks to measure the performance of LLMs. We developed custom evaluation modules based on the original MMLU configuration and extended it for our scenarios:

- **Default Setting** This setting preserves the original MMLU prompt without any dialectal cues. It varies, however, based on the evaluated domain, e.g., for the *Abstract Algebra* domain, the prompt will be *"The following are multiple choice questions (with answers) about abstract algebra."* followed by the multiple choices, and concluded with *"Answer:"*
- **Oracle Setting** This setting introduces explicit dialectal conditioning by specifying the dialect of the question as part of the prompt. As a result, the prompt is modified based on the dialects as well. For the same domain, *Abstract Algebra*, the first part of the prompt in the oracle setting will be *"The following are multiple choice questions (with answers) about abstract algebra in the Egyptian dialect."*
- **Dialect Identification** The Dialect Identification setting asks the model to infer the dialect of the input: each of our five dialects or MSA. Here, the topic is irrelevant and, as a result, we have a fixed prompt: *"The following are multiple-choice questions (with answers) for Arabic dialect identification."*

All tasks follow a multiple-choice format to align with MMLU. We use log-likelihood evaluation, appending each answer option to the prompt and selecting the one with the highest log-likelihood as the model's prediction. Then, the accuracy is determined by exact match with the gold answer. Each experiment is repeated five times, and we report the average accuracy across runs and across the 32 topics per dialect/language (unless stated otherwise). This setup ensures transparency, comparability, and controlled analysis of model sensitivity to dialectal variation.

5. Experimental Results and Analysis

In this section, we discuss the experiments and the analysis, organized around three key questions:

5.1. MSA vs. DA in QA Performance

How do LLMs perform on Question–Answering tasks in MSA compared to dialectal Arabic? Table 6 shows the accuracy for various LLMs when evaluated on the default DIALECTALARABICMMLU setting for QA. For each model–dialect pair, we report the average accuracy over the 32 domains, with the experiment repeated five times to ensure stability and robustness.

We can see that our newly developed dialectal Arabic evaluation dataset is effective for testing the dialectal capabilities of LLMs by highlighting the performance gap for English vs. MSA and dialects, which is easy to see given the parallel nature of the questions and the answers in the dataset.

We further see that the performance consistently declines across all dialects compared to MSA and English, and this trend holds consistently across all Arabic-enabled LLMs we evaluated.

Finally, although comparing and ranking individual models is tempting, we deliberately refrain from doing so. Instead, we focus on the *average performance across all models*.[4] We argue that this offers a more holistic perspective on the current state of the art and may yield deeper insights than analyzing individual models in isolation. One observation supporting this view is that some larger models, both Arabic-centric and multilingual, perform worse than smaller ones. Understanding this would require investigating each model's training process, including the base model (if any) and the datasets. Given the limited availability of such information, a fine-grained comparative analysis is impractical, which motivates our emphasis on studying the aggregate trends instead.

[4] Jais-2 was released after we performed the experiments and the detailed analysis. We included it in Table 6 for comparison, but we exclude it from the average and we do not include it in any other tables or figures.

Model	Size↓	EGY	KSA	MAG	SYR	UAE	DA Avg	MSA	ENG
jais-13b-chat	13.0	48.0	48.2	44.7	45.4	48.5	47.0	52.0	55.3
gemma-3-12b-it	12.2	61.5	58.5	54.3	57.4	61.7	58.7	62.6	73.7
Nile-Chat-12B	11.8	**61.9**	**60.6**	**55.8**	**58.9**	**62.5**	**59.9**	**63.8**	72.8
SILMA-9B-Instruct	9.2	55.3	54.0	48.7	52.0	55.3	53.1	57.6	72.4
Fanar-1-9B-Instruct	8.8	58.5	56.6	53.6	54.6	58.0	56.2	61.3	70.4
command-r7b-arabic	8.0	53.5	52.8	50.2	52.2	55.0	52.7	57.7	67.4
aya-expanse-8b	8.0	51.8	50.1	47.2	49.2	52.1	50.1	54.0	63.3
Falcon-H1-7B-Instruct	7.6	59.1	58.1	52.6	55.8	60.2	57.2	62.4	**76.5**
Mistral-7B-Instruct	7.2	33.9	35.2	34.0	33.3	36.2	34.5	38.0	62.8
ALLaM-7B-Instruct	7.0	56.6	56.2	53.4	55.3	58.2	56.0	60.3	66.7
Yehia-7B	7.0	53.7	53.6	50.5	52.9	55.3	53.2	58.5	62.5
jais-6p7b-chat	6.8	42.8	44.6	40.2	41.5	45.3	42.9	48.2	52.9
gemma-3-4b-it	4.3	41.0	40.5	36.4	38.0	43.3	39.8	44.1	54.6
Qwen3-4B-Instruct	4.0	28.7	27.1	26.7	26.6	28.8	27.6	31.2	65.5
Nile-Chat-4B	3.9	48.3	47.0	42.4	45.3	47.8	46.2	49.5	59.3
Falcon-H1-3B-Instruct	3.1	46.1	44.7	41.7	43.1	46.2	44.3	48.4	67.9
NileChat-3B	3.1	54.3	51.8	52.8	51.0	53.7	52.7	55.6	64.3
jais-2p7b-chat	2.7	38.2	40.4	34.4	37.7	41.5	38.4	43.4	47.1
ar-stablelm-2-chat	1.6	36.4	36.5	35.5	36.1	36.4	36.2	37.3	38.3
Average	6.8	48.9	48.2	45.0	46.6	49.8	47.7	51.9	62.8
Jais-2-8B-Chat	8.1	56.9	55.7	53.0	54.1	56.8	55.3	59.0	65.3
Jais-2-70B-Chat	72.4	68.0	68.0	65.0	66.0	68.8	67.2	70.7	76.2

Table 6: Accuracy scores for the default DIALECTALARABICMMLU setting. (Average of 5 runs for the 32 different topics. Random chance = $\frac{1}{4}$. Size↓: Sorted in descending order. **Bold**: Maximum per column.)

5.2. DA Identification vs. QA Performance

To what extent does a model's proficiency in recognizing dialectal Arabic correlate with its Question–Answering performance for the same dialect? We evaluate the performance of the LLMs as per the setup described in Section 4.2. To establish a baseline, we use the CAMeL Tools **D**ialect **ID**entification (DID) tool (Obeid et al., 2020), which classifies Arabic texts into 26 dialects: MSA or one of 25 cities in 15 Arab countries; the tool can return a city, a country, or a region. In our experiments, we used CAMeL Tools DID with two configurations: $\text{DID}_{country}$ and $\text{DID}_{aligned}$. For $\text{DID}_{country}$, we used the tool out-of-the-box, where the only post-processing we did was to remap the labels, e.g., *Syria* is mapped to *SYR*. For $\text{DID}_{aligned}$, we aligned the CAMeL Tools country labels to our ones. This alignment is based on apriori geographical and dialectal groupings: **EGY** (Egypt, Sudan), **KSA** (Saudi Arabia, Yemen, Baghdad/Iraq), **MAG** (Morocco, Algeria, Tunisia, Libya), **SYR** (Syria, Jordan, Lebanon, Palestine, Mosul/Iraq), and **UAE** (Qatar, Oman, Basra/Iraq).

Table 7 shows the accuracy for predicting the dialect of the question. Several observations can be made from these results. First, there is a huge difference between the average performance on MSA and the Arabic dialects, and many models perform worse than random.

Moreover, CAMeL Tools, an off-the-shelf tool with minimal alignment, achieved the best identification accuracy for three of five dialects, as well as on both the dialectal and overall averages.

Finally, we emphasize the high risk resulting from combining MSA and Arabic dialects as one language. As demonstrated in Table 7, command-r7b-arabic which scored among the highest MSA and total average scores, performs extremely poorly on UAE and KSA dialects.

To answer the question at the beginning of this section, we conducted a Pearson correlation analysis on the average dialectal performance, and the score of each dialect separately[5]. We observed a moderate positive correlation between the MCQ and the dialect ID tasks as indicated by a Pearson correlation $r = 0.431$, which, however, is not statistically significant ($p = 0.07$). Similarly, the correlation was not statistically significant for EGY, KSA, SYR and UAE with $p = 0.18$, $p = 0.79$, $p = 0.12$, and $p = 0.07$, respectively. The difference was statistically significant only for MAG with $p = 0.04$, for a moderate positive correlation of $r = 0.483$.

We investigated this behavior further using the Oracle setting explained in Section 4.2, where we infused the prompt with extra information about the dialect. Based on the results above, our intuition is that a model that cannot identify the dialect will not benefit from being told what that dialect is.

[5]See Fig. 4 in Appendix G.

Model	Size↓	EGY	KSA	MAG	SYR	UAE	MSA	DA Avg	Avg All
jais-13b-chat	**13.0**	40.0	24.1	10.1	16.8	13.4	26.7	20.9	21.8
gemma-3-12b-it	12.2	36.0	9.5	64.0	**71.0**	4.8	83.9	37.1	44.9
Nile-Chat-12B	11.8	30.5	0.6	22.6	19.9	0.5	82.7	14.8	26.1
SILMA-9B-Instruct	9.2	54.8	13.1	59.4	29.6	10.2	58.5	33.4	37.6
Fanar-1-9B-Instruct	8.8	**84.2**	9.3	45.5	8.6	0.9	58.8	29.7	34.5
command-r7b-arabic	8.0	23.4	6.4	28.8	17.5	0.5	87.2	15.3	27.3
aya-expanse-8b	8.0	38.2	3.9	17.2	3.7	1.4	59.0	12.9	20.6
Falcon-H1-7B-Instruct	7.6	63.5	2.4	24.3	6.4	0.4	74.8	19.4	28.6
Mistral-7B-Instruct	7.2	37.2	0.9	1.4	7.0	0.3	60.5	9.4	17.9
ALLaM-7B-Instruct	7.0	42.4	23.0	55.5	22.0	4.0	**95.4**	29.4	40.4
Yehia-7B	7.0	15.6	11.9	29.3	5.4	3.8	94.9	13.2	26.8
jais-6p7b-chat	6.8	19.5	3.4	5.6	6.8	7.2	68.6	8.5	18.5
gemma-3-4b-it	4.3	30.5	19.8	37.6	21.0	11.5	12.2	24.1	22.1
Qwen3-4B-Instruct	4.0	47.9	10.8	13.8	12.0	10.4	21.1	19.0	19.3
Nile-Chat-4B	3.9	48.4	4.8	11.4	12.9	13.2	11.5	18.1	17.0
Falcon-H1-3B-Instruct	3.1	49.0	3.0	3.9	4.6	2.0	60.0	12.5	20.4
NileChat-3B	3.1	24.8	2.3	7.7	11.8	0.2	80.5	9.4	21.2
jais-2p7b-chat	2.7	11.1	2.0	9.3	5.7	19.6	60.4	9.5	18.0
ar-stablelm-2-chat	1.6	11.0	4.8	31.5	10.0	5.3	46.5	12.5	18.2
Average	6.8	37.3	8.2	25.2	15.4	5.8	60.2	18.4	25.3
CAMeL Tools-DID$_{country}$	–	53.9	10.0	70.2	23.7	0.0*	73.9	31.6	37.4
CAMeL Tools-DID$_{aligned}$	–	57.4	**31.0**	**79.4**	64.4	**29.6**	73.9	**52.4**	**56.3**

Table 7: Recall scores for the Dialect Identification setting. (The average of 5 runs for the 32 different topics. The random chance is $\frac{1}{6} \simeq 16.7$. * No labels for UAE cities.)

Model	EGY	KSA	MAG	SYR	UAE	MSA	ENG
jais-13b-chat	-1.0	-1.1	-0.6	-0.8	-1.6	0.0	0.0
gemma-3-12b-it	-9.0	-5.0	-11.4	-16.3	-7.8	0.3	0.1
Nile-Chat-12B	-1.6	-3.1	-1.7	-2.3	-3.0	0.1	0.0
SILMA-9B-Instruct	-2.1	-1.3	-1.1	-0.5	-0.9	0.0	0.0
Fanar-1-9B-Instruct	-2.7	-1.1	-2.6	-1.0	-1.5	-0.1	-0.1
command-r7b-arabic	-0.7	-0.5	-1.4	-0.3	-1.1	0.0	0.1
aya-expanse-8b	-1.7	-0.8	-0.6	-0.8	-2.2	-0.1	0.0
Falcon-H1-7B-Instruct	-3.8	-2.7	-1.3	-3.0	-2.8	-0.2	-0.1
Mistral-7B-Instruct	-1.5	-1.4	-0.4	0.4	-1.3	0.1	-0.1
ALLaM-7B-Instruct	-2.1	-2.1	-3.0	-2.9	-1.8	0.1	0.2
Yehia-7B	-1.1	-0.5	-2.0	-0.6	-1.4	0.0	0.0
jais-6p7b-chat	-0.3	-1.9	-1.1	-1.2	-3.7	0.0	0.0
gemma-3-4b-it	-7.0	-9.4	-2.6	-8.4	-9.7	0.0	0.4
Qwen3-4B-Instruct	3.9	1.6	0.3	2.2	2.1	-0.2	0.0
Nile-Chat-4B	-2.5	-1.4	-2.1	-1.4	-2.2	0.0	0.2
Falcon-H1-3B-Instruct	-2.8	-2.2	-2.1	-1.5	-2.3	-0.1	0.1
NileChat-3B	-1.7	-0.3	-2.1	-1.1	-1.6	0.0	0.1
jais-2p7b-chat	0.2	-0.2	0.0	0.1	-2.0	0.0	0.0
ar-stablelm-2-chat	-0.2	0.3	-0.5	-1.2	0.4	0.0	0.0
Average	-2.0	-1.7	-1.9	-2.1	-2.3	-0.0	0.0

Table 8: Difference between the accuracy scores for Oracle − Default.

The results are shown in Table 8, where we can see the difference in accuracy between the Oracle setting, where we inject the dialect ID in the prompt and the default setting, where the prompt contains no explicit information about the dialect.[6]

We did not expect the added dialect label to improve performance, as it provides little useful signal to the LLMs. Indeed, the oracle setting led to a statistically significant drop across all dialects. While the link between dialect identification and QA remains unclear, explicitly priming the model with the dialect label consistently degraded performance.

[6]The exact results are in Table 16 in Appendix E.

Model	MSA	MADLAD	Google	ENG
jais-13b-chat	52.0	47.6	50.5	55.3
gemma-3-12b-it	62.6	60.8	66.8	73.7
Nile-Chat-12B	63.8	60.3	65.5	72.8
SILMA-9B-Instruct	57.6	60.0	64.6	72.4
Fanar-1-9B-Instruct	61.3	58.3	63.1	70.4
command-r7b-arabic	57.7	57.0	61.8	67.4
aya-expanse-8b	54.0	54.5	57.2	63.3
Falcon-H1-7B-Instruct	62.4	63.5	69.1	76.5
Mistral-7B-Instruct	38.0	52.0	56.3	62.8
ALLaM-7B-Instruct	60.3	55.6	61.1	66.7
Yehia-7B	58.5	52.4	56.5	62.5
jais-6p7b-chat	48.2	46.2	48.3	52.9
gemma-3-4b-it	44.1	48.5	50.4	54.6
Qwen3-4B-Instruct	31.2	54.3	59.4	65.5
Nile-Chat-4B	49.5	49.1	54.2	59.3
Falcon-H1-3B-Instruct	48.4	55.5	61.5	67.9
NileChat-3B	55.6	54.2	59.0	64.3
jais-2p7b-chat	43.4	41.7	44.4	47.1
ar-stablelm-2-chat	37.3	37.0	37.4	38.3
Average	**51.9**	**53.1**	**57.2**	**62.8**

Table 9: Experiments with MSA and English input, as well as for two machine-translations of the MSA input to English: MADLAD and Google.

5.3. Improving Dialectal QA through MT

Can machine translation mitigate data scarcity in dialectal QA? In this experiment, we investigate whether translating the dialectal questions to English (or to MSA) can help language models perform better on the QA task.

Choosing a translation model. To perform the translation, we experimented with two machine translation tools: (*i*) Google Translate (paid) API, which is a commercial translating tool, and (*ii*) Google's MADLAD–400 (Kudugunta et al., 2023), which is a 7B/10B parameters, open-weight translation model.

Translating the MSA question to English We translated the MSA input into English using the two tools above and evaluated the same set of Arabic-enabled LLMs. Table 9 reports the performance on the original **MSA** input, on the **MADLAD**-400 and **Google** translations, and on the original **ENG** questions. We can see that translating MSA into English improves the average performance compared to using the original MSA input for both translation models, although the results still fall short of those obtained when using the original English input. A paired t-test indicates no statistically significant difference between the original MSA and the MADLAD-translated English (T-stat=0.78, P-value=0.45). In contrast, the following four comparisons are statistically significant:

- MSA vs. Google (T-stat=3.30, P-value$\leq$.000)
- MADLAD vs. Google (T-stat=-12.12, P-value$\leq$.000)
- MADLAD vs. ENG (T-stat=-14.51, P-value$\leq$.000)
- Google vs. ENG (T-stat=-15.15, P-value$\leq$.000).

Based on these results, we decided to use the Google API translations when doing the evaluation. We provide the MADLAD results in Appendix E (see Table 17) for reproducibility given that the technical details of the current Google API translation model is not public and it is not clear if/when this particular model would be replaced.

Translating the dialectal Arabic input to English & to MSA Table 10 shows the effect of translating the dialectal Arabic input to English and to MSA (using English as a pivot). The main observation is that translating to English, on average, yields performance gains when compared to using the original dialectal Arabic input (statistically significant: (T-stat=-3.54, P-value=0.002). Most of this gain is driven by two multilingual models, namely *Mistral-7B-Instruct* and *Qwen3-4B-Instruct* with an increase of 17.9 and 27.4 points absolute, respectively.

In contrast, when translating to MSA, nearly all the performance gains vanish to the point where the average difference between using the original and the translated questions drops from 5.6 to 0.3 points absolute, resulting in a statistically insignificant difference (T-stat=-1.22, P-value=0.24) in performance between using the original dialectal Arabic questions and translating them to MSA. One potential explanation is that translation errors that occur when translating to English cause more errors when translating to MSA. This behavior is consistent across all Arabic dialects as can be inferred from the average scores and their standard deviation values.

6. Conclusion and Future Work

We introduced DIALECTALARABICMMLU, a new benchmark for evaluating large language models (LLMs) across major Arabic dialects. Our work addresses a persistent gap in current Arabic NLP evaluation, which has largely focused on Modern Standard Arabic (MSA) while neglecting the linguistic diversity of real-world Arabic usage.

DIALECTALARABICMMLU extends the MMLU-Redux framework through high-quality, human-curated translations of more than 3K question–answer pairs into five dialects (Syrian, Egyptian, Emirati, Saudi, and Moroccan) resulting in a corpus of over 15K (21K when including MSA and English) multiple-choice QA instances spanning 32 academic and professional domains.

Model	Dialectal Q&As Translated to English						Dialectal Q&As Translated to MSA*					
	EGY	KSA	MAG	SYR	UAE	Avg	EGY	KSA	MAG	SYR	UAE	Avg
jais-13b-chat	-0.1	0.3	0.8	1.4	0.9	0.6	1.4	0.8	1.4	2.4	-0.2	1.1
gemma-3-12b-it	2.2	3.3	2.8	3.3	1.7	2.6	-2.4	0.4	-0.5	-0.5	-2.4	-1.1
Nile-Chat-12B	1.2	1.2	1.2	1.5	0.0	1.0	-2.4	-1.8	-0.4	-1.0	-2.9	-1.7
SILMA-9B-Instruct	6.7	6.0	7.1	6.7	6.8	6.6	-1.6	-1.0	2.0	0.4	-0.5	-0.2
Fanar-1-9B-Instruct	1.4	1.4	0.7	3.0	2.0	1.8	-1.3	0.2	-1.0	1.0	-0.3	-0.2
command-r7b-arabic	5.1	4.3	3.5	3.1	3.8	4.0	0.4	1.0	0.3	1.5	-0.8	0.5
aya-expanse-8b	4.5	4.3	4.2	4.2	4.0	4.2	-1.2	0.6	2.5	0.8	-0.4	0.5
Falcon-H1-7B-Instruct	6.7	5.7	6.2	7.2	5.1	6.1	-2.0	-0.8	0.5	0.0	-2.2	-0.9
Mistral-7B-Instruct	19.8	18.0	15.6	18.2	17.7	17.9	1.8	-0.2	0.8	1.8	-0.2	0.8
ALLaM-7B-Instruct	1.5	1.3	-0.4	2.8	1.5	1.3	-0.3	-0.4	-0.6	-0.3	-0.1	-0.4
Yehia-7B	1.0	-0.4	-1.3	-0.2	0.7	0.0	0.5	-0.1	1.3	-0.4	-0.6	0.1
jais-6p7b-chat	3.1	1.3	2.2	2.9	1.2	2.1	3.2	0.7	3.9	2.5	0.1	2.1
gemma-3-4b-it	7.2	8.2	9.0	9.6	5.5	8.0	1.2	1.8	3.4	2.9	-0.8	1.8
Qwen3-4B-Instruct	27.8	28.4	24.4	28.4	28.1	27.4	1.0	2.3	1.8	2.2	1.1	1.7
Nile-Chat-4B	2.6	4.5	4.9	6.1	5.4	4.7	-1.4	0.0	1.5	0.5	0.0	0.1
Falcon-H1-3B-Instruct	12.8	12.9	11.4	13.8	12.5	12.7	-0.4	1.3	1.6	1.7	0.6	1.0
NileChat-3B	2.3	2.8	-2.0	3.9	3.4	2.1	-1.9	-0.3	-3.7	0.4	-1.2	-1.3
jais-2p7b-chat	4.4	1.9	5.8	4.4	1.3	3.6	2.9	1.3	5.2	1.4	0.1	2.2
ar-stablelm-2-chat	-0.5	1.1	-1.5	1.3	1.1	0.3	0.4	0.2	0.0	-0.8	0.3	0.0
Average (SD)	5.8 (7.2)	5.6 (7.2)	5.0 (6.6)	6.4 (7.0)	5.4 (7.0)	5.6 (6.9)	-0.1 (1.7)	0.3 (1.0)	1.1 (2.0)	0.9 (1.2)	-0.6 (1.0)	0.3 (1.1)

Table 10: Change in accuracy when using the translated inputs instead of the original inputs. The translation is performed using Google's translation API (* The dialectal Arabic questions were translated to English first and then to MSA).

Through comprehensive experiments with nineteen open-weight Arabic and multilingual LLMs, we demonstrated that model performance drops substantially across dialects compared to MSA and English. We further showed that explicit dialect conditioning does not consistently improve the performance and that a model's ability to identify a dialect only moderately correlates with its ability to reason in that dialect. These findings underscore the need for dedicated resources and training strategies that explicitly target dialectal Arabic.

In future work, we aim to expand coverage to additional Arabic dialects and domains, including low-resource varieties and specialized professional contexts. Second, we want to add auxiliary tasks that probe lexical, syntactic, and pragmatic understanding in dialects. Finally, we envision the benchmark serving as a foundation for fine-tuning and adaptation, encouraging the development of LLMs that can reason and communicate effectively across the full spectrum of Arabic varieties.

7. Acknowledgments

This work was conducted as part of the **IBM–MBZUAI AI Center of Excellence**.

The authors gratefully acknowledge the contributions of Alya Almsouti, Hamad Alshehhi, Mohammad Anwar, Samar Mohamed Magdy, and Tareq Almsouti for their assistance with the data annotation process.

Ethics and Broader Impact

We followed ethical research and data management practices at all stages of data collection, translation, and validation. All question–answer pairs originate from publicly available and educational sources contained in MMLU-Redux, and carry no personal or sensitive information. All dialectal translations were produced by qualified speakers through a paid language service provider under informed consent, ensuring fair compensation and professional oversight. No personally identifiable or user-generated content was collected.

As dialectal Arabic is inherently diverse, we recognize the potential for bias arising from regional, social, or stylistic variation in translation. To minimize this, all data underwent multi-stage review by annotators from different dialectal backgrounds, with guidelines emphasizing neutrality, inclusivity, and linguistic authenticity. Nevertheless, residual biases reflecting the translators' linguistic preferences or educational backgrounds may persist.

The benchmark is intended exclusively for research and educational purposes. By providing an open, transparent, and reproducible evaluation framework, we aim to promote progress in Arabic NLP and raise awareness of dialectal variation as a key dimension of Arabic LLMs. We encourage responsible use of our dataset, with careful consideration of the potential downstream impact of Arabic LLM evaluation and deployment.

Limitations

While DIALECTALARABICMMLU represents an important step toward evaluating LLMs across Arabic dialects, several limitations should be acknowledged. First, despite our focus on five major dialects (Syrian, Egyptian, Emirati, Saudi, and Moroccan), the benchmark does not yet cover the full spectrum of dialectal variation across the Arabic-speaking world. Numerous regional, urban–rural, and sociolectal sub-varieties exist within each dialect group, reflecting differences in geography, age, education, and social context, which our dataset does not explicitly represent.

Second, dialectal Arabic lacks standardized orthography, which introduces inherent variability in spelling and transcription. Although all items were manually curated by native speakers and validated for linguistic fidelity, residual inconsistencies may still affect model evaluation. Similarly, human translation and adjudication introduce subjective judgment, which, while mitigated through multi-stage review, cannot be entirely eliminated.

Third, our experiments are limited to open-weight models of moderate size (between 1B and 13B parameters). Results for larger proprietary models, which are often stronger on multilingual tasks, remain to be explored. Finally, as our benchmark is derived from question–answering tasks, it captures only a subset of dialectal capabilities; future work should complement it with generative, conversational, and multimodal evaluations.

These limitations provide avenues for future refinement and broader representational coverage.

8. Bibliographical References

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

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, 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.

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

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

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. Shami: A corpus of Levantine Arabic dialects. In

Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Rawan Al-Matham, Kareem Darwish, Raghad Al-Rasheed, Waad Alshammari, Muneera Alhoshan, Amal Almazrua, Asma Al Wazrah, Mais Alheraki, Firoj Alam, Preslav Nakov, Norah Alzahrani, Eman alBilali, Nizar Habash, Abdelrahman El-Sheikh, Muhammad Elmallah, Haonan Li, Hamdy Mubarak, Mohamed Anwar, Zaid Alyafeai, Ahmed Abdelali, Nora Altwairesh, Maram Hasanain, Abdulmohsen Al Thubaity, Shady Shehata, Bashar Alhafni, Injy Hamed, Go Inoue, Khalid Elmadani, Ossama Obeid, Fatima Haouari, Tamer Elsayed, Emad Alghamdi, Khalid Almubarak, Saied Alshahrani, Ola Aljarrah, Safa Alajlan, Areej Alshaqarawi, Maryam Alshihri, Sultana Alghurabi, Atikah Alzeghayer, Afrah Altamimi, Abdullah Alfaifi, and Abdulrahman AlOsaimy. 2025. BALSAM: A platform for benchmarking Arabic large language models. *ArXiv preprint*, arXiv:2507.22603.

Omar Alkaabi, Ahmed Alzubaidi, Hamza Alobeidli, Shaikha Alsuwaidi, Mohammed Alyafeai, Leen AlQadi, Basma El Amel Boussaha, and Hakim Hacid. 2026. Alyah: An Emirati dialect benchmark for evaluating Arabic large language models. https://huggingface.co/blog/tiiuae/emirati-benchmarks.

Yazeed Alnumay, Alexandre Barbet, Anna Bialas, William Darling, Shaan Desai, Joan Devassy, Kyle Duffy, Stephanie Howe, Olivia Lasche, Justin Lee, Anirudh Shrinivason, and Jennifer Tracey. 2025. Command R7B Arabic: A small, enterprise focused, multilingual, and culturally aware Arabic LLM. *ArXiv preprint*, arXiv:2503.14603.

Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. 2018. DART: A large dataset of dialectal Arabic tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, Ahmed Oumar El-Shangiti, Aisha Alraeesi, Mohammed Anwar AL-Ghrawi, Abdulrahman S. Al-Batati, Elgizouli Mohamed, Noha Taha Elgindi, Muhammed Saeed, Houdaifa Atou, Issam Ait Yahia, Abdelhak Bouayad, Mohammed Machrouh, Amal Makouar, Dania Alkawi, Mukhtar Mohamed, Safaa Taher Abdelfadil, Amine Ziad Ounnoughene, Anfel Rouabhia, Rwaa Assi, Ahmed Sorkatti, Mohamedou Cheikh Tourad, Anis Koubaa, Ismail Berrada, Mustafa Jarrar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, ACL 2025, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical aggregation of dialectal data for Arabic dialect identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4586–4596, Marseille, France.

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)*, page 749–775.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2025. ALLam: Large language models for Arabic and English. In *Proceedings of the Thirteenth International Conference on Learning Representations*, ICLR '2025, Singapore.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Ke-

mal 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.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy.

Basma El Amel Boussaha, Leen AlQadi, Mugariya Farooq, Shaikha Alsuwaidi, Giulia Campesan, Ahmed Alzubaidi, Mohammed Alyafeai, and Hakim Hacid. 2025. 3LM: Bridging Arabic, STEM, and code through benchmarking. *ArXiv preprint*, arXiv:2507.15850.

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.

Charles A Ferguson. 1959. Diglossia. *word*, 15(2):325–340.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2025. Are we done with MMLU? *ArXiv preprint*, arXiv:2406.04127.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *ArXiv preprint*, arXiv:2009.03300.

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

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.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. *ArXiv preprint*, arXiv:2309.04662.

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.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source Python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey

Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, ukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, ukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 technical report. *ArXiv preprint*, arXiv:2303.08774.

Karim Ouda. 2025. Arabic LLM models. https://huggingface.co/blog/silma-ai/arabic-llm-models-list. Hugging Face Blog.

Ahmed Salama, Houda Bouamor, Behrang Mohit, and Kemal Oflazer. 2014. YouDACC: the Youtube dialectal Arabic comment corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1246–1251, Reykjavik, Iceland.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *ArXiv preprint*, arXiv:2308.16149.

Gemma Team. 2025. Gemma 3. *ArXiv preprint*, arXiv:2503.19786.

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

A. Translation Guidelines

Below are the guidelines provided to the translation team. Two online training sessions were held prior to commencing the task, to ensure that that the guidelines were clear and that all translators understood what was expected from them. The training session focused on Egyptian Arabic as an example as it was the commonly understood dialect amongst the speakers of the various dialects.

- The translation requests are presented in an Excel sheet:
  Column A - English source text
  Column B - MSA translation (this column will be hidden)
  Column C - Dialect translation
  Column D - Reject/Accept (leave blank if translation request is accepted)
  Column E - Justification (If Column D is "Reject")

- The translation task will be completed sheet by sheet by two translators. One who translates and one who reviews the translation.

- If there is a problem with the source text, the reviewer can refuse to do the translation by providing a justification for his/her rejection in Column E.

- The rejected lines are reviewed by the team leader who decides what to do with the rejected translations.

- In addition, the team leader reviews the accepted translations randomly to check the quality of the review process.

Please pay attention to the following instructions:

1. Only work on a sheet if you are confident in your translation capabilities for the specific domain of those questions.

2. Do not translate in cases where you doubt your understanding of the question/answers. In such cases, simply move to the next line to avoid wasting time.

3. The English questions/answers should be used as the main reference for translation. However, it is acceptable to use the MSA translation (unhide Column B) as an additional reference when needed.

4. Support tools: It is acceptable to use dictionaries or term bases to find the most suitable term for a word in a sentence. However, automatic translation is strictly forbidden as these translated texts will be part of a scientific study that cannot be influenced by translation technology.

5. When translating, be simple, concise, and concrete.

6. The final translation should read like a question-answer conversation with a friend. For example:
   En: *Can you help me find my bag? Of course.*
   MSA: هل يمكنك مساعدتي في ايجاد امتعتي؟ بالطبع
   Egy DA:ساعدني القى شنطتي ممكن تساعدني القى شنطتي؟ اكيد

7. MSA terms may be used, depending on the context. However, do not try to translate domain-specific words into a dialect if the context does not allow for it.

8. In some cases, the resulting translation may seem very similar to MSA (especially in short sentences). This is ok if it seems like a more natural translation. Do not try to invent terms to make your translation "appear dialectal."

9. Before delivering the translation to the reviewer, carry out your own review first, and make changes when necessary. This will make the entire translation process more efficient for everyone.

10. Use the spelling and writing style you normally use for your dialect. Since there is no single standardized spelling for dialect words, feel free to write them in the way you normally would. For example, in MSA, you can indicate the future tense by using سـ at the beginning of a verb or by adding سوف (e.g., سأتصل بك or سوف أتصل بك for "*I will call you*"). However, in the Egyptian dialect, the future tense is usually marked with هـ at the beginning of the verb, like هكلمك ("*I will call you*"). Some speakers, however, prefer to use حـ instead of هـ, so they might write حكلمك instead of هكلمك.

B. Annotators' Demographics

The annotators' demographics are given in Table 11.

ID	Dataset	Native	Residence	Age	Gender	Degree	Task	Background
A0	MAG	MAG	MAG	30s	F	BA	A	RA, LA
A1	MAG	MAG	MAG	30s	F	BA	T, V	LA
A2	MAG	MAG	MAG	40s	M	PhD	T	LA, RA, CT
A3	MAG	MAG	MAG	30s	M	MBA	T, V	CT, PT
A4a	KSA	SYR	SYR	40s	M	BA	T, V	LA, RA, CT
A5	KSA	SYR	SYR	40s	F	BA	T, V	LA
A6	KSA	SYR	SYR	50s	F	BA	A	LA, PT
A7	EGY	EGY	EGY	30s	F	BA	T, V	LA, PT
A8	EGY	SYR	SYR	40s	F	PhD	T, V	CT
A9	EGY	EGY	EGY	30s	F	BA	A	LA, RA, CT
A4b	SYR	SYR	SYR	40s	M	BA	A	LA, RA, CT
A10	SYR	SYR	SYR	20s	F	BA	T, V	CT, RA
A11	SYR	SYR	SYR	30s	F	BA	T, V	CT, PT
A12	SYR	SYR	SYR	40s	F	BA	T	LA, RA, PT
A13	UAE	PAL	UAE	20s	M	BA	T, V	LA
A14	UAE	PAL	UAE	40s	M	BA	A	LA
A15	UAE	PAL	UAE	40s	F	BA	T, V	LA

Table 11: Annotator demographics and roles for the dialect annotation task. There are three roles: Approve (A), Translate (T), and Validate (V). The annotator background experience includes: Certified Teacher (CT), Private Tutor (PT), Linguistic Annotator (LA), and Research Assistant (RA). All annotators are native speakers of Arabic, and they worked on their own native dialect or on dialects of a country where they had resided for 13–24 years (if they worked on a different dialect).

C. Additional Similarity Scores

We provide additional similarity scores, namely, the weighted Jaccard similarity measure (Table 12), and the Manhattan (Table 13) and Euclidean (Table 14) distances in support of Sec. 3: similarity between dialects.

Dialect	SYR	UAE	KSA	MSA	MAG	EGY
SYR	1.000	0.554	0.601	0.533	0.442	0.593
UAE		1.000	0.665	0.612	0.471	0.626
KSA			1.000	0.681	0.469	0.647
MSA				1.000	0.455	0.601
MAG					1.000	0.464
EGY						1.000

Table 12: Jaccard weighted similarity.

Dialect	SYR	UAE	KSA	MSA	MAG	EGY
SYR	0.000	0.713	0.630	0.776	0.988	0.649
UAE		0.000	0.518	0.625	0.936	0.594
KSA			0.000	0.500	0.956	0.562
MSA				0.000	1.000	0.661
MAG					0.000	0.972
EGY						0.000

Table 13: Manhattan (L1) distance.

Dialect	SYR	UAE	KSA	MSA	MAG	EGY
SYR	0.000	0.715	0.648	0.756	0.758	0.745
UAE		0.000	0.506	0.653	0.843	0.448
KSA			0.000	0.492	0.864	0.541
MSA				0.000	1.000	0.648
MAG					0.000	0.874
EGY						0.000

Table 14: Euclidean (L2) distance.

D. Dialects and Language Distribution per Domain

Table 15 gives a detailed statistics of the number of questions for each domain–dialect pair. While the number of questions per topic may vary, the total row shows an equal number of question per dialect/language.

Domain	Egyptian	Saudi	Marrocan	Syrian	Emirati	MSA	English
abstract_algebra	100	100	100	100	100	100	100
anatomy	100	100	100	100	100	100	100
astronomy	97	97	97	97	97	97	97
business_ethics	99	99	99	99	99	99	99
clinical_knowledge	100	100	100	100	100	100	100
college_computer_science	100	100	100	100	100	100	100
college_medicine	99	99	99	99	99	99	99
conceptual_physics	100	100	100	100	100	100	100
elementary_mathematics	100	100	100	100	100	100	100
global_facts	96	96	96	96	96	96	96
high_school_chemistry	99	99	99	99	99	99	99
high_school_geography	100	100	100	100	100	100	100
high_school_macroeconomics	90	90	90	90	90	90	90
high_school_psychology	100	100	100	100	100	100	100
high_school_us_history	100	100	100	100	100	100	100
high_school_world_history	100	100	100	100	100	100	100
human_aging	99	99	99	99	99	99	99
international_law	100	100	100	100	100	100	100
management	100	100	100	100	100	100	100
marketing	68	68	68	68	68	68	68
moral_scenarios	98	98	98	98	98	98	98
nutrition	94	94	94	94	94	94	94
philosophy	100	100	100	100	100	100	100
prehistory	100	100	100	100	100	100	100
professional_law	99	99	99	99	99	99	99
professional_psychology	100	100	100	100	100	100	100
public_relations	98	98	98	98	98	98	98
security_studies	100	100	100	100	100	100	100
sociology	100	100	100	100	100	100	100
us_foreign_policy	99	99	99	99	99	99	99
virology	100	100	100	100	100	100	100
world_religions	100	100	100	100	100	100	100
TOTAL	3,135	3,135	3,135	3,135	3,135	3,135	3,135

Table 15: Number of questions per domain and dialect.

E. Additional Results

In this section, we provide additional results in support of Sec. 5. In Table 16, we show the accuracy scores for the Oracle DIALECTALARABICMMLU setting, to compliment the results in Sec. 5.2. In Table 17, we show the change in accuracy when the input is translated using the MADLAD translation model instead of the Google Translate API (Table 10).

Model	Size↓	EGY	KSA	MAG	SYR	UAE	DA Avg	MSA	ENG
jais-13b-chat	**13.0**	47.0	47.1	44.1	44.6	46.9	45.9	52.0	55.3
gemma-3-12b-it	12.2	52.5	53.5	42.9	41.1	53.9	48.8	62.9	73.8
Nile-Chat-12B	11.8	**60.3**	**57.5**	**54.1**	**56.6**	**59.5**	**57.6**	**63.9**	72.8
SILMA-9B-Instruct	9.2	53.2	52.7	47.6	51.5	54.4	51.9	57.6	72.4
Fanar-1-9B-Instruct	8.8	55.8	55.5	51.0	53.6	56.5	54.5	61.2	70.3
command-r7b-arabic	8.0	52.8	52.3	48.8	51.9	53.9	51.9	57.7	67.5
aya-expanse-8b	8.0	50.1	49.3	46.6	48.4	49.9	48.9	53.9	63.3
Falcon-H1-7B-Instruct	7.6	55.3	55.4	51.3	52.8	57.4	54.4	62.2	**76.4**
Mistral-7B-Instruct	7.2	32.4	33.8	33.6	33.7	34.9	33.7	38.1	62.7
ALLaM-7B-Instruct	7.0	54.5	54.1	50.4	52.4	56.4	53.6	60.4	66.9
Yehia-7B	7.0	52.6	53.1	48.5	52.3	53.9	52.1	58.5	62.5
jais-6p7b-chat	6.8	42.5	42.7	39.1	40.3	41.6	41.2	48.2	52.9
gemma-3-4b-it	4.3	34.0	31.1	33.8	29.6	33.6	32.4	44.1	55.0
Qwen3-4B-Instruct	4.0	32.6	28.7	27.0	28.8	30.9	29.6	31.0	65.5
Nile-Chat-4B	3.9	45.8	45.6	40.3	43.9	45.6	44.2	49.5	59.5
Falcon-H1-3B-Instruct	3.1	43.3	42.5	39.6	41.6	43.9	42.2	48.3	68.0
NileChat-3B	3.1	52.6	51.5	50.7	49.9	52.1	51.4	55.6	64.4
jais-2p7b-chat	2.7	38.4	40.2	34.4	37.8	39.5	38.1	43.4	47.1
ar-stablelm-2-chat	1.6	36.2	36.8	35.0	34.9	36.8	35.9	37.3	38.3
Average	6.8	46.9	46.5	43.1	44.5	47.5	45.7	51.9	62.9

Table 16: Accuracy scores for the Oracle DIALECTALARABICMMLU setting. (Average of 5 runs for the 32 different topics. Random chance = $\frac{1}{4}$. Size↓: Sorted in descending order. **Bold**: Maximum per column.)

	Dialectal Q&As translated to English						Dialectal Q&As translated to MSA*					
Model	**EGY**	**KSA**	**MAG**	**SYR**	**UAE**	**DA Avg**	**EGY**	**KSA**	**MAG**	**SYR**	**UAE**	**DA Avg**
jais-13b-chat	-2.6	-4.1	-6.5	-3.2	-3.4	-4.0	-2.5	-4.7	-6.1	-3.8	-2.6	-4.0
gemma-3-12b-it	-6.2	-5.3	-8.5	-5.9	-5.7	-6.3	-6.8	-6.9	-10.2	-6.6	-8.1	-7.7
Nile-Chat-12B	-5.9	-5.9	-9.2	-7.4	-6.0	-6.8	-6.9	-7.9	-10.7	-8.7	-7.7	-8.4
SILMA-9B-Instruct	-1.0	-0.7	-4.2	-1.1	0.0	-1.4	-5.0	-4.9	-6.1	-5.2	-3.5	-5.0
Fanar-1-9B-Instruct	-5.5	-4.8	-10.1	-5.7	-5.0	-6.2	-6.3	-6.3	-10.5	-6.8	-5.7	-7.1
command-r7b-arabic	-1.6	-4.1	-7.4	-5.5	-2.5	-4.2	-4.0	-4.0	-8.3	-5.2	-4.3	-5.1
aya-expanse-8b	-2.0	-1.7	-6.6	-2.0	-0.8	-2.6	-3.7	-3.4	-7.8	-4.0	-3.3	-4.5
Falcon-H1-7B-Instruct	-0.6	-2.7	-5.7	-1.9	-1.9	-2.6	-6.4	-7.5	-9.0	-7.4	-7.4	-7.6
Mistral-7B-Instruct	14.7	11.4	6.7	12.4	11.8	11.4	0.7	-1.4	-2.5	0.5	-0.9	-0.7
ALLaM-7B-Instruct	-4.2	-6.3	-9.9	-6.0	-5.9	-6.5	-5.1	-5.6	-9.8	-6.5	-4.9	-6.5
Yehia-7B	-5.8	-5.8	-10.1	-6.6	-5.3	-6.7	-2.7	-4.6	-7.8	-7.1	-5.2	-5.5
jais-6p7b-chat	-0.6	-2.5	-3.5	-1.2	-1.8	-1.9	-1.2	-3.8	-4.3	-1.6	-2.6	-2.7
gemma-3-4b-it	3.2	3.8	2.7	5.3	3.0	3.7	-0.6	0.1	-0.7	0.3	-1.0	-0.4
Qwen3-4B-Instruct	21.0	21.1	12.7	19.5	20.5	19.0	-0.5	1.1	0.0	0.4	0.1	0.2
Nile-Chat-4B	-3.3	-2.5	-3.6	-1.8	-1.1	-2.5	-5.1	-6.1	-5.9	-4.2	-4.5	-5.2
Falcon-H1-3B-Instruct	4.4	4.9	1.1	3.3	5.4	3.9	-4.0	-2.8	-5.4	-3.5	-4.5	-4.0
NileChat-3B	-4.1	-2.5	-11.6	-4.4	-2.7	-5.1	-6.0	-5.3	-11.8	-6.1	-6.1	-7.0
jais-2p7b-chat	0.9	-1.1	-0.4	0.8	-0.3	0.0	0.0	-2.7	-1.3	-1.2	-1.7	-1.3
ar-stablelm-2-chat	-2.6	-1.7	-2.1	-0.5	-0.7	-1.6	-2.2	-1.9	-4.4	-1.5	-2.2	-2.5
Average (SD)	-0.1 (7.0)	-0.6 (6.8)	-4.0 (6.3)	-0.6 (6.8)	-0.1 (6.6)	-1.1 (6.7)	-3.6 (2.5)	-4.1 (2.5)	-6.5 (3.6)	-4.1(2.9)	-4.0 (2.4)	-4.5 (2.6)

Table 17: Change in accuracy when using the translated instead of the original inputs. The translation is performed using MADLAD (* The dialectal Arabic questions were translated to English first and then to MSA).

F. Supporting Figures

In this section, we provide a set of supporting figures to show the general performance trend of all the evaluated models across the various domains.

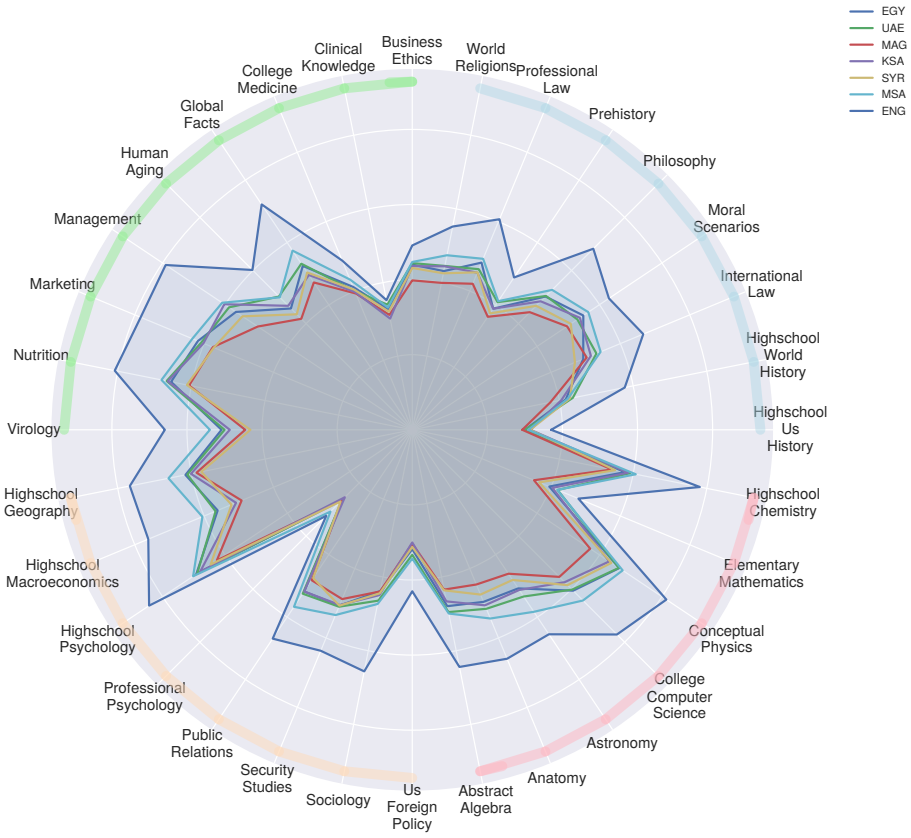


Figure 1: Average DIALECTALARABICMMLU accuracy per domain (Fields: Humanities, Stem, Social Sciences, and Other).

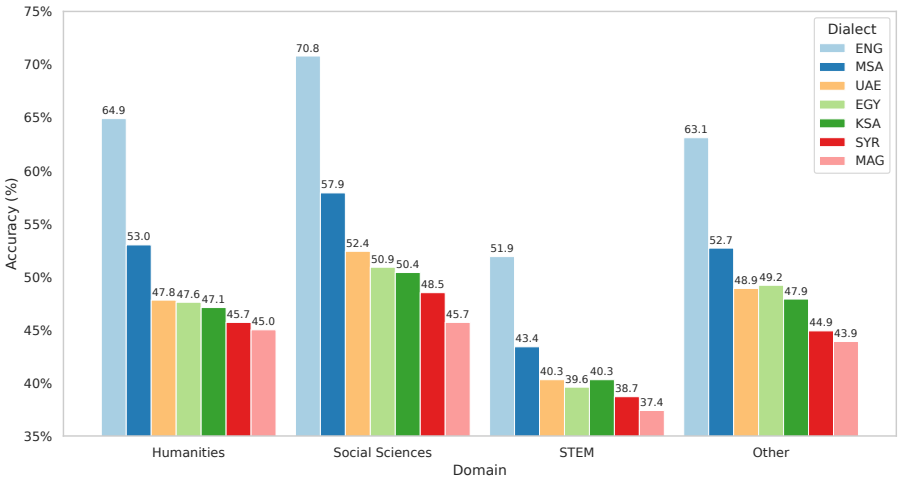
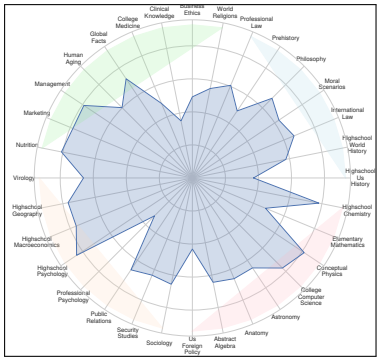
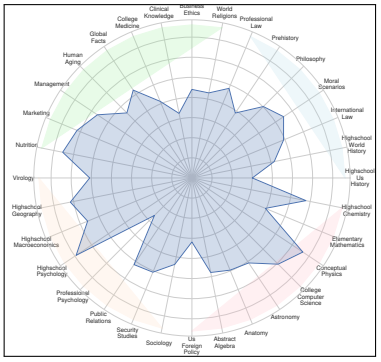


Figure 2: Average DIALECTALARABICMMLU accuracy grouped by field.

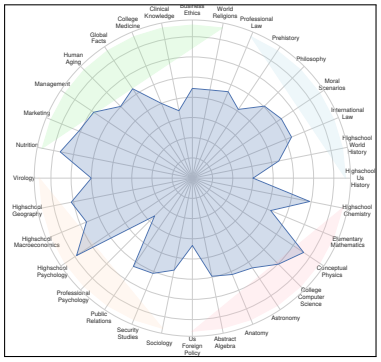


(a) English

(b) MSA

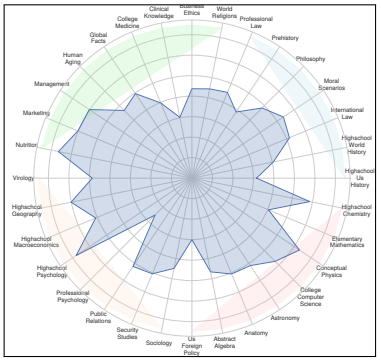


(c) EGY

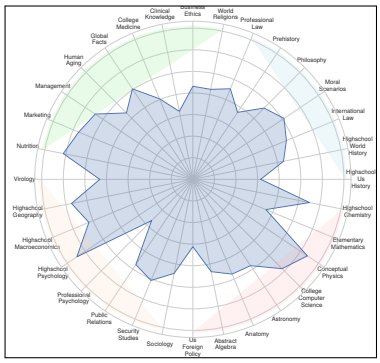


(d) UAE

(e) MAG



(f) KSA



(g) SYR

Figure 3: Average DIALECTALARABICMMLU accuracy over domains for each dialect.

G. The Correlation between MCQ and Dialect Identification Accuracy

Fig. 4 shows the correlation between each model's performance on the dialect identification task vs. the MCQ task (right) as well as the average performance over all the models (left), for each dialect.

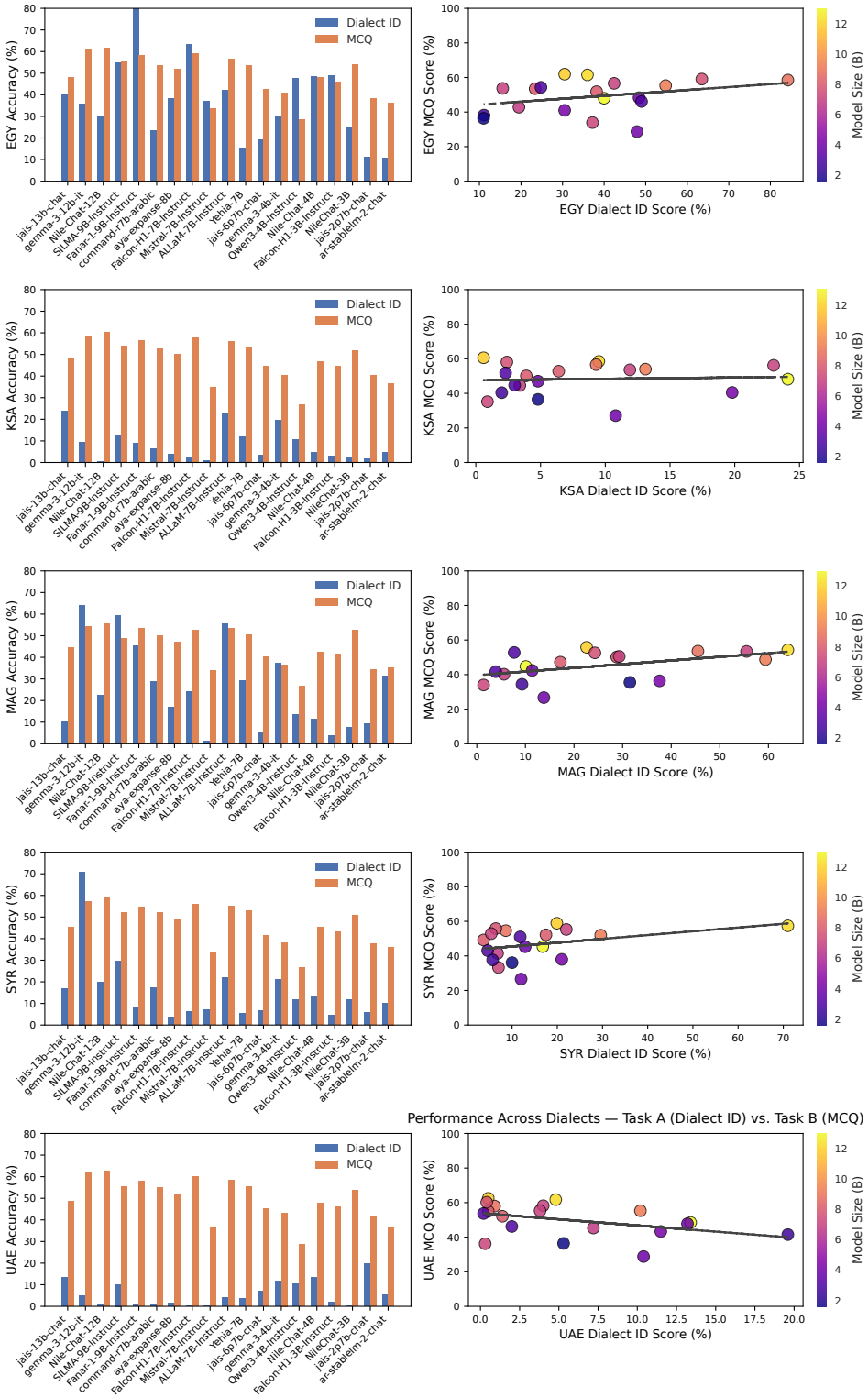


Figure 4: The correlation between MCQ accuracy and dialect identification accuracy per model (left) and on average (right).