

ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic

Fajri Koto¹ Haonan Li¹ Sara Shatnawi¹ Jad Doughman¹
Abdelrahman Boda Sadallah¹ Aisha Alraeesi¹ Khalid Almubarak²
Zaid Alyafeai³ Neha Sengupta⁴ Shady Shehata¹ Nizar Habash^{1,5}
Preslav Nakov¹ Timothy Baldwin^{1,6}

¹Department of Natural Language Processing, MBZUAI

²Prince Sattam bin Abdulaziz University ³King Fahd University of Petroleum and Minerals

⁴Core42 ⁵New York University Abu Dhabi ⁶The University of Melbourne

{fajri.koto,haonan.li,sara.shatnawi,jad.doughman,abdelrahman.sadallah,aisha.alraeesi}@mbzuai.ac.ae

Abstract

The focus of language model evaluation has transitioned towards reasoning and knowledge-intensive tasks, driven by advancements in pre-training large models. While state-of-the-art models are partially trained on large Arabic texts, evaluating their performance in Arabic remains challenging due to the limited availability of relevant datasets. To bridge this gap, we present ArabicMMLU, the first multi-task language understanding benchmark for the Arabic language, sourced from school exams across diverse educational levels in different countries spanning North Africa, the Levant, and the Gulf regions. Our data comprises 40 tasks and 14,575 multiple-choice questions in Modern Standard Arabic (MSA) and is carefully constructed by collaborating with native speakers in the region. Our comprehensive evaluations of 35 models reveal substantial room for improvement, particularly among the best open-source models. Notably, BLOOMZ, mT0, LLaMA2, and Falcon struggle to achieve a score of 50%, while even the top-performing Arabic-centric model only achieves a score of 62.3%.¹

1 Introduction

Although large language models (LLMs) such as GPT-3.5 (Ouyang et al., 2022), BLOOMZ (Muenighoff et al., 2022), and Jais (Sengupta et al., 2023) have been pretrained with substantial coverage of Modern Standard Arabic (MSA), their reasoning and knowledge assessments are primarily conducted using datasets translated from English to Arabic (Sengupta et al., 2023; Huang et al., 2023), which means there is limited capacity to evaluate

¹Data and code can be accessed at: <https://github.com/mbzuai-nlp/ArabicMMLU>

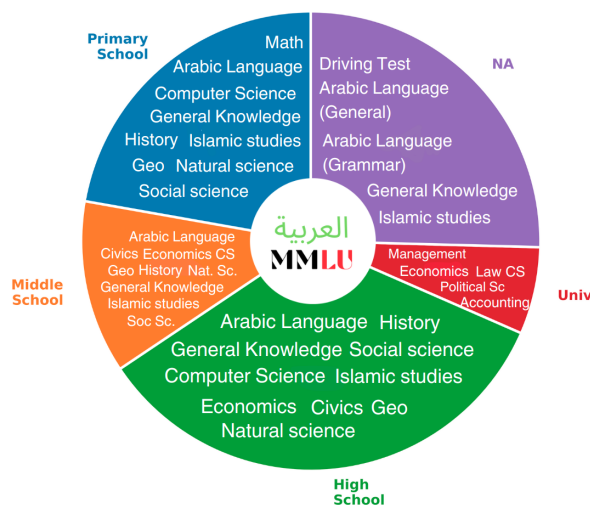


Figure 1: Distribution of educational levels and corresponding subjects in ArabicMMLU. “NA” denotes other levels.

content specific to Arabic. This reliance on translation systems not only demonstrates an Anglocentric approach (Ramesh et al., 2023; Talat et al., 2022) but also potentially introduces errors and biases. Given that Arabic is one of the most widely-spoken languages in the world, with a speaker population of over 400 million people (Shoufan and Alameri, 2015; Diab et al., 2017), it is critically important that datasets be constructed for the language that are regionally- and culturally-localized.

The evaluation of language models has increasingly shifted from linguistically-centric tasks, such as part-of-speech (POS) tagging and named entity recognition (NER), towards reasoning and knowledge evaluation. This shift is evidenced in evaluations of models like GPT-4 (OpenAI, 2023), LLaMA2 (Touvron et al., 2023), and LLM360 (Liu et al., 2023) on various commonsense reasoning datasets (Zellers et al., 2019; Huang et al., 2019;

Koto et al., 2022, 2024), mathematical problems (Amini et al., 2019; Cobbe et al., 2021), coding challenges (Chen et al., 2021; Austin et al., 2021; Yu et al., 2024), and school exams (Hendrycks et al., 2021; Li et al., 2023; Koto et al., 2023). One notable dataset for knowledge evaluation is MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2021), which comprises multiple-choice questions across various subjects based on the US education system. In recent Arabic-centric LLMs like Jais (Sengupta et al., 2023) and AceGPT (Huang et al., 2023), knowledge evaluation was carried out using MMLU translated from English to Arabic.

To comprehensively evaluate the reasoning and knowledge capabilities of Arabic LLMs in local Arabic-speaking contexts, we introduce ArabicMMLU, styled around MMLU and sourced from school exams across Arabic-speaking countries spanning North Africa, the Levant, and the Gulf regions. ArabicMMLU was constructed through collaboration with native Arabic speakers from Jordan, Egypt, UAE, Lebanon, and Saudi Arabia (KSA), ensuring rich local context, particularly in the subject areas of history, geography, law, civics education, and driving tests. Figure 1 summarizes the distribution of education levels and corresponding subjects in ArabicMMLU. The proportion of primary school, middle school, high school, and university level questions in ArabicMMLU are 22.2%, 12.2%, 34%, and 6.1%, respectively, with the remaining questions categorized as “NA”.

Our contributions can be summarized as follows:

- We introduce the first Arabic MMLU-style dataset in Modern Standard Arabic (MSA), featuring 40 tasks covering various subject areas and educational levels across eight countries. Over 50% of the questions in our dataset are tailored to Arabic-specific contexts.
- We evaluate 22 open-source multilingual models, 11 open-source Arabic-centric models, and 2 closed-source models. GPT-4 achieves the best performance, while the open-source models struggle to achieve scores above 60%.
- We conduct a thorough analysis of the top-performing open-source models across various dimensions, encompassing: (1) individual subject areas, education levels, countries, and Arabic-specific topics; (2) few-shot inference performance; (3) model confidence; and (4) the influence of negation.

2 Related Work

2.1 Language Models in Arabic

Early Arabic pretrained language models typically had less than 2 billion parameters and were primarily monolingual. These models can be classified into three main categories: encoder-only, decoder-only, and encoder–decoder models. The encoder-only models, such as AraBERT (Antoun et al., 2020), CAMELBERT (Inoue et al., 2021), AraELECTRA (Antoun et al., 2021a), and ARBERT & MARBERT (Abdul-Mageed et al., 2021), are mainly from the BERT family. AraGPT2 (Antoun et al., 2021b), on the other hand, is a decoder-only model available in different sizes ranging from 135M to 1.4B parameters. Examples of encoder–decoder models include AraT5 (Nagoudi et al., 2022) and AraBART (Kamal Eddine et al., 2022).

Jais (Sengupta et al., 2023) and AceGPT (Huang et al., 2023) are two recent Arabic-centric decoder-only models with parameter sizes of up to 30B and 13B, respectively. Jais is pretrained on a corpus of 72 billion Arabic tokens, while AceGPT builds upon LLaMA2 and is enhanced with reinforcement learning from AI feedback (Lee et al., 2023) to localize the model to Arabic values and culture. Both models are bilingual (English and Arabic), and were fine-tuned on various instruction datasets.

Arabic is also present in multilingual models. This includes earlier models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), and more recent LLMs such as BLOOMZ (Muennighoff et al., 2022), mT0 (Muennighoff et al., 2022), Falcon (Penedo et al., 2023), GPT-3.5 (Ouyang et al., 2022), and GPT-4 (OpenAI, 2023). In the original papers, only GPT-4 was evaluated in Arabic in terms of its reasoning and knowledge capabilities, using the English–Arabic translated MMLU dataset, reporting an accuracy of 80%.

2.2 Arabic Benchmarks for Evaluating Language Models

Arabic is included in various multilingual benchmarks for natural language understanding and generation, such as XGLUE (Liang et al., 2020), XTREME (Hu et al., 2020), XTREME-R (Ruder et al., 2021) and GEM (Gehrmann et al., 2021). In recent years, several Arabic-centric benchmarks have been released, such as Dolphin (Nagoudi et al., 2023), OCRA (Elmadany et al., 2023), and LARA-Bench (Abdelali et al., 2024). Many tasks in these benchmarks involve classification, such as natural lan-

History subject, High school exam in Jordan	
وصل الأمير عبدالله بن الحسين إلى مدينة معان في	Prince Abdullah bin Al Hussein arrived in the city of Ma'an in
أ- 21 تشرين الثاني عام 1920 م ب- 21 تشرين الأول عام 1921 م ج- 22 تشرين الثاني عام 1920 م د- 21 تشرين الأول عام 1920 م	A. November 21, 1920 AD B. October 21, 1921 AD C. November 22, 1920 AD D. October 21, 1920 AD
History subject, High school exam in Egypt	
انتهت حملة فريرز 1807 وخرجت من مصر عن طريق	Frazer's 1807 expedition ended and he left Egypt via...
أ- المفاوضات ب- المساعي ج- الحميدة الوساطة د- السلم	A. Negotiations B. Endeavours C. Benign mediation D. Ladder
Driving test exam in the UAE	
ما هي السرعة القصوى القانونية على الطرق العامرة في أبو ظبي؟	What is the legal maximum speed on busy roads in Abu Dhabi?
أ- 70 كم/الساعة ب- 100 كم/الساعة ج- 90 كم/الساعة د- 80 كم/الساعة	A. 70 km/h B. 100 km/h C. 90 km/h D. 80 km/h

Figure 2: Examples of two history questions and one driving test question from Jordan, Egypt, and UAE, respectively. **Left** is the original text and **right** is the English translation for illustrative purposes. The bold options are the correct answer keys.

guage inference (Conneau et al., 2018), POS tagging (Darwish et al., 2017), named entity recognition (Pan et al., 2017), and summarization (Ladhak et al., 2020). There are three notable question answering datasets: TyDiQA (Clark et al., 2020), Arabic-SQuAD (Mozannar et al., 2019), and MLQA (Lewis et al., 2020). These datasets primarily focus on reading comprehension and question answering, unlike the MMLU dataset (Hendrycks et al., 2021) which evaluates reasoning and knowledge in real-world settings, in the form of multiple-choice questions. Related, EXAMs (Hardalov et al., 2020) is a dataset based on multilingual school exams, which contains a subset of about 500 Arabic questions.

3 ArabicMMLU

In the Middle East, the education system mostly follows the K12 system, consisting of six years of primary school, three years of middle school, and three years of high school.^{2,3} Many education systems in countries within the region, such as Egypt

²<https://www.pwc.com/m1/en/industries/education/publications/understanding-middle-east-education.pdf>

³With the exception of the UAE, which follows a 4-4-4 structure for primary, middle, and high schools.

Social science subject, Middle school exam in Palestine	
تسهم تنمية قيم المواطنة بشكل كبير في	The development of citizenship values contributes significantly to
أ- وحدة واستقرار المجتمع ب- تعزيز الولاء والانتماء ج- تحقيق العدالة د- جميع ما سبق	A. Unity and stability of society B. Enhancing loyalty and belonging C. Achieving justice D. All of the above
Math subject, Primary school exam in Jordan	
عدد المئات والعشرات في الرقم 700 هو	The number of hundreds and tens in the number 700 is
أ- 700 = 7 مئات = 7 عشرات ب- 700 = 700 hundreds = 7 tens ج- 700 = 70 مئات = 7 عشرات د- 700 = 7 مئات = 7 عشرات	A. 700 = 7 hundreds = 70 tens B. 700 = 700 hundreds = 7 tens C. 700 = 70 hundreds = 7 tens D. 700 = 7 hundreds = 7 tens
Bar (legal) exam in Morocco	
يمكن أن تقام الدعوى المدنية التابعة ضد	Ancillary civil action may be brought against
أ- الفاعلين الأصليين فقط ب- الفاعلين الأصليين والمشاركين والمساهمين فقط ج- الفاعلين الأصليين والمساهمين والمشاركين وضد ورثتهم أو الأشخاص المسؤولين مدنيا عنهم د- جميع الاجوبة خاطئة	A. Original actors only B. The original actors, participants and contributors only C. The original perpetrators, shareholders, and participants, and against their heirs or persons who are civilly responsible D. All answers are wrong

Figure 3: Examples of social science, math, and bar exam questions from Palestine, Jordan, and Morocco, respectively. **Left** is the original text and **right** is the English translation for illustrative purposes. The bold options are the correct answer keys.

and KSA, prioritize Islamic studies alongside subjects like mathematics, natural science, social science, and geography.⁴ In public schools, Arabic is commonly used for teaching and assessment, while in international schools, English is the predominant language of instruction for most subjects, following either the UK or USA curriculum. When designing ArabicMMLU, we excluded questions in English and only included questions in Arabic.

ArabicMMLU is an Arabic multiple-choice question-answering dataset comprising 40 tasks spanning a wide range of subjects and education levels. The questions are sourced from eight different countries in North Africa (Morocco and Egypt), the Levant (Jordan, Palestine, and Lebanon), and the Gulf (UAE, Kuwait, and KSA). Each question has 2–5 candidate answers, with one correct answer. Table 1 provides details of the subjects in ArabicMMLU. The subjects are drawn from different education levels (primary school, middle school,

⁴https://www.tabahfoundation.org/wp-content/uploads/2018/12/TabahFuturesInitiative-Islamic-Education_En.pdf

Group	Subjects
STEM	Natural Science (P, M), Math (P), Physics (H), Biology (H), Computer Science (P, M, H, U)
Social science	Social Science (P, M), Civics education (M, H), Geography (P, M, H), Economics (M, H, U), Political Science (U)
Humanities	Islamic studies (P, M, H, U, NA), History (P, M, H), Accounting (U), Law (U), Philosophy (H)
Language	Arabic Language (P, M, H), Arabic Language - General (NA), Arabic Language - Grammar (NA)
Other	Management (U), General Knowledge (P, M, NA), Driving Test (NA)

Table 1: Subject areas in ArabicMMLU. “P”, “M”, “H”, “U”, and “NA” indicate that questions in the subject are available in primary school, middle school, high school, university and professional, and others, respectively.

high school, university, and professional) and are categorized into STEM, social science, humanities, language, and other fields.

Figures 2 and 3 showcase various examples of ArabicMMLU questions, with some focusing on history, driving tests, social science, and bar exams, all of which are pertinent to Arabic-specific norms and cultures. Notably, Arabic multiple-choice questions sometimes use Arabic-script characters (أ, ب, ج, د, ه) rather than Latin-script characters (e.g. A, B, C, D, E). This differs from many other languages, where the answer options are strictly in Latin script (even if the local writing script is not Latin, as with Mandarin Chinese). In prior work (Hendrycks et al., 2021; Koto et al., 2023; Li et al., 2023), answering these multiple-choice questions has relied on the probability of the alphabetic options. We experiment with both Arabic and Latin script outputs in Section 4.

3.1 Data Construction

The data construction process involved a total of 10 Arabic native speakers from different countries: 6 internal workers (1 Jordanian, 1 Egyptian, 1 Lebanese, 1 from UAE, and 2 from KSA) and 4 external workers (3 Jordanian and 1 Egyptian).

During the first stage of data collection, the internal workers were tasked with collecting relevant

sources for data collection. These sources were URLs containing the questions, which needed to be publicly available.

In the second stage, all workers were asked to manually scrape the data within a 2-month period. The task was to collect metadata, including the source (URL of the source document), country, subject, level, question, multiple-choice options, and the correct answer key. Each external worker was assigned to gather 2,000 questions, while internal workers were tasked with gathering 1,000–2,000 questions each. Our internal workers are Master’s students and Research Assistants in Computer Science, while the external workers hold Bachelor’s degrees. We ensured competitive compensation for the workers, exceeding the monthly average wage in each respective country.

During manual data scraping, workers were instructed to include only questions accompanied by an answer key, and to discard questions containing multi-modal information (e.g., images, videos, or tables). If a question had additional contextual information (e.g., a passage referenced by several questions), the context was required to be included with each corresponding question.

3.2 Quality Control

While our workers are native speakers of Modern Standard Arabic with at least Bachelor’s degrees, we maintain the quality of our dataset construction through meticulous steps. Firstly, we conducted a 1-hour workshop before the data collection stage to clarify the process. Secondly, we automatically filtered out repetitive questions and those without an answer key, reducing the initial set of over 15,000 questions to 14,575 unique questions. Finally, we assessed the accuracy of our data collection by having two native Arabic speakers annotate 100 randomly sampled questions. They were provided with all metadata, including the answer key, and tasked with verifying the correctness of each sample using any available resources (e.g., search engines). We found that 96% of the questions and answer keys match on average, while the remaining could have incorrect answer keys. This 96% score is considered to represent the maximum score meaningfully achievable for ArabicMMLU.

3.3 Data Statistics

Table 2 presents detailed statistics of ArabicMMLU, categorized by education level and subject area. The distribution of questions across education lev-

Group	# Questions	# Chars	
		Question	Answer
Primary	3239	43.6	30.4
Middle	1775	58.3	54.6
High	4963	76.7	66.0
Univ	892	69.1	97.3
NA	3706	311.4	52.7
STEM	3220	60.0	49.4
Social Science	3540	62.2	57.5
Humanities	3655	57.1	60.2
Language	1661	647.3	45.1
Other	2499	57.7	59.1

Table 2: Average question and answer length (in characters) for each education group and subject area.

Country	STEM	Social	Hum.	Lang.	Other	TOTAL
Jordan	1086	2163	1579	362	863	6053
Egypt	1012	581	335	324	254	2506
Palestine	860	585	600	2	-	2047
Morocco	-	-	317	-	-	317
Lebanon	-	-	-	-	239	239
UAE	-	-	56	-	128	184
Kuwait	-	-	-	-	111	111
KSA	67	-	-	-	37	104
Other	195	211	768	973	867	3014
TOTAL	3220	3540	3655	1661	2499	14575

Table 3: The distribution of ArabicMMLU sources by country, categorized according to subject areas. “Social”, “Hum.”, and “Lang.” denote social science, humanities, and Arabic language, respectively.

els varies, with primary school having the largest number, around 4.9K, followed by high school with 3.2K. Questions and candidate options are generally longer at the high school and university levels. Additionally, we observe that questions in the “NA” (other) category are four times longer (in characters) than those in school exams. This is expected since this category includes subjects like Arabic language (General) and Arabic language (Grammar), where questions typically involve lengthy paragraphs as context. For a detailed breakdown of questions for each subject in each education level, please refer to the Appendix (Table 7).

For subject areas, they are reasonably evenly distributed, particularly for STEM, social science, and the humanities, each consisting of roughly 3.2K to 3.5K questions. There are only minor differences in question length between these three subject areas. However, for the language category, the average question length (in characters) is 10 times longer than other categories.

<p>[COUNTRY] في [LEVEL] [SUBJECT] هذا سؤال اختر الإجابة الصحيحة! سؤال: [QUESTION] [OPTION] إجابة:</p>
<p>-----</p> <p>This is a [SUBJECT] question for [LEVEL] in [COUNTRY]. Select the correct answer! Question: [INPUT] [OPTION] Answer:</p>

Figure 4: Prompt templates in Arabic and English.

Table 3 further shows the distribution of questions across the eight countries from which questions were collected, with Jordan, Egypt, and Palestine being the top three sources. Various subjects within the social sciences, humanities, and other categories (such as driving tests) often include Arabic-specific content, representing 57.7% of the dataset. While STEM questions are more aligned with the English MMLU, it is worth noting that differences in curriculum between North Africa, the Levant, the Gulf regions, and the USA may influence variations in assessment question design.

4 Experiments

4.1 Set-Up

Our experiments focus on zero-shot and few-shot settings across 35 models. This includes 22 open-source multilingual models (XGLM (Lin et al., 2022), BLOOMZ (Muennighoff et al., 2022), mT0 (Muennighoff et al., 2022), Falcon (Penedo et al., 2023), and LLaMA2 (Touvron et al., 2023), across various sizes), 11 open-source Arabic-centric models (AraT5 (Nagoudi et al., 2022), AraGPT2 (Antoun et al., 2021b), AceGPT (Huang et al., 2023) and Jais (Sengupta et al., 2023), also across various sizes), and 2 closed-source models (GPT-3.5: gpt-3.5-turbo (Ouyang et al., 2022) and GPT-4: gpt-4-0613 (OpenAI, 2023)).

We initially conducted experiments with four settings: (1) Arabic prompt and Arabic alphabetic output, (2) Arabic prompt and English (i.e. Latin script) alphabetic output, (3) English prompt and Arabic alphabetic output, and (4) English prompt and English alphabetic output. Figure 4 illustrates the Arabic and English prompts. The placeholders [SUBJECT], [LEVEL], and [COUNTRY] are replaced with the corresponding Arabic and English words, while the placeholders [INPUT] and [OPTION] are in Arabic. The choice of the al-

Model (#parameters)	STEM	Social Science	Humanities	Arabic Language	Other	Average
Random	29.5	28.9	28.6	25.8	32.3	29.0
XGLM (1.7B)	30.0	30.5	31.2	28.1	34.6	31.0
XGLM (2.9B)	30.0	30.7	31.4	28.1	35.3	31.2
XGLM (4.5B)	27.6	29.1	28.5	26.7	34.7	29.3
XGLM (7.5B)	27.8	29.3	29.5	27.9	33.0	29.5
BLOOMZ (560M)	32.7	30.4	31.9	26.5	36.6	31.9
BLOOMZ (1.1B)	30.4	26.5	30.1	25.1	28.1	28.4
BLOOMZ (1.7B)	35.3	39.0	37.4	37.2	39.6	37.7
BLOOMZ (3B)	40.4	44.5	43.8	40.9	48.5	43.7
BLOOMZ (7B)	43.2	48.0	49.1	49.9	49.9	47.8
mT0 _{small} (300M)	31.1	30.5	29.4	29.4	33.2	30.7
mT0 _{base} (580M)	30.2	30.9	31.5	28.2	34.4	31.2
mT0 _{large} (1.2B)	31.1	31.7	31.6	29.7	35.7	32.0
mT0 _{xl} (3.7B)	38.7	42.3	40.1	43.9	43.5	41.4
mT0 _{xxl} (13B)	42.7	45.4	43.4	46.0	46.0	44.5
LLaMA2 (7B)	33.7	32.8	33.5	28.4	36.7	33.4
LLaMA2-chat (7B)	34.5	32.9	31.5	30.9	37.0	33.4
LLaMA2 (13B)	32.9	35.0	37.8	35.8	39.3	36.1
LLaMA2-chat (13B)	36.2	34.8	34.2	35.3	40.7	36.0
Falcon (7B)	29.8	29.9	31.5	29.0	35.1	31.1
Falcon-instruct (7B)	28.4	29.5	27.3	21.3	29.1	27.7
Falcon (40B)	34.9	33.8	36.2	30.1	37.4	34.8
Falcon-instruct (40B)	33.8	30.9	33.9	28.9	36.2	33.0
AraT5 (220M)	29.9	30.3	33.0	28.4	32.0	31.0
AraT5v2 (220M)	31.4	30.7	32.8	27.4	34.7	31.7
AraGPT2 (1.7B)	33.0	31.5	35.8	29.8	37.4	33.7
AceGPT (7B)	35.4	35.9	36.2	31.1	41.7	36.3
AceGPT-chat (7B)	41.2	45.3	47.8	41.5	51.5	45.6
AceGPT (13B)	42.7	45.5	48.3	42.4	50.7	46.1
AceGPT-chat (13B)	47.3	52.8	53.9	50.5	58.5	52.6
Jais (13B)	30.3	31.4	33.6	28.1	36.3	32.2
Jais-chat (13B)	50.5	56.1	59.3	39.9	61.7	54.8
Jais (30B)	39.5	45.6	50.5	34.6	49.1	44.8
Jais-chat (30B)	56.2	60.5	65.5	62.0	68.2	62.3
GPT-3.5 (175B)	53.8	57.0	57.5	57.6	63.8	57.7
GPT-4 (NA)	70.2	70.4	73.2	72.8	76.9	72.5

Table 4: Zero-shot LLM performance (% accuracy), combined across subject groups. “Average” means the average across all questions in ArabicMMLU.

phabetic output (English vs. Arabic) is adjusted in [OPTION]. See Appendix B (Figure 10) for examples of the full input in both English and Arabic.

Following previous studies (Koto et al., 2023; Li et al., 2023), for open-source models, we determine the answer based on the highest probability among all possible options. In the case of English alphabetic output, we measure the probability of the first generated token being A, B, C, D, or E. For Arabic, we measure the probability of the first generated token being أ, ب, ج, د, or ه. For closed-source models, we determine the answer based on the first token generated in the text using a regular expression. If there is no match, we assign a random answer.

4.2 Results

To evaluate the influence of prompt language, we initially benchmarked the open-source models using all four prompt settings (Section 4.1), as depicted in Figure 5. We observe that the optimal configuration across all models is to use an English prompt and English alphabetic output. Predictably, the Arabic-specific LLMs — Jais-chat (30B) and AceGPT-chat (13B) — demonstrate the greatest robustness when employing Arabic alphabetic output. Please refer to Appendix for complete results of all prompt settings across the open-source models. For the remaining experiments, we will report based on the setting of English prompt and English alphabetic output.

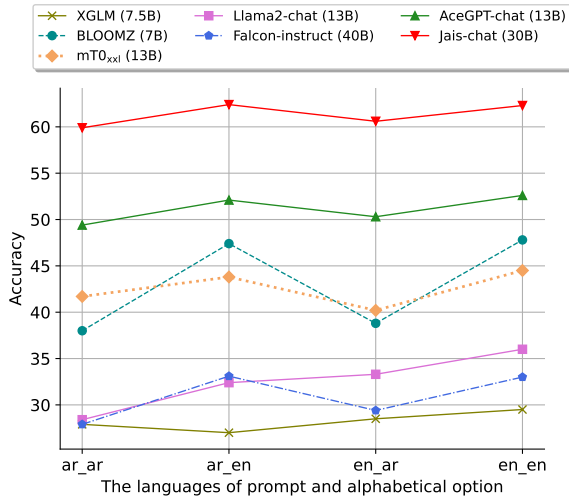


Figure 5: LLM performance with different prompt settings. ar_en means that the prompt template is in Arabic and the alphabetic option is in English (the Latin script).

Results across all models Table 4 shows the full results of all models, grouped by subject area. As expected, the Arabic-centric model Jais-chat (30B) emerges as the top-performing open-source model, boasting an average score of 62.3%, surpassing GPT-3.5 by 4.6 points. Compared to AceGPT-chat (13B), both Jais-chat models (13B and 30B) exhibit substantially higher accuracy in areas including STEM, Social Science, Humanities, and Others. For multilingual models such as BLOOMZ (7B) and mT0 (13B), their performance lags behind Jais, with a disparity of more than 14 points. XGLM, LLaMA2, and Falcon perform at a level close to random, suggesting their limited proficiency in Arabic. GPT-4 achieves the highest accuracy, with a score of 72.5%, surpassing Jais-chat (30B) by 10 points. It is noteworthy that in the GPT-4 technical report (OpenAI, 2023), the accuracy of the English-Arabic translated MMLU dataset is reported as 80%, which is 8 points higher than our ArabicMMLU results. One possible explanation for this difference is that our ArabicMMLU presents a greater challenge due to its inclusion of a higher proportion of Arabic-specific content.

Furthermore, we notice a trend of increasing accuracy with larger models, with the exception of XGLM. For example, BLOOMZ (7B) achieves an accuracy 15.9 points higher than BLOOMZ (560M), while mT0 (13B) shows a 13.8-point increase compared to mT0 (300M). This trend is also evident in AceGPT and Jais, although it is less pronounced in LLaMA2 and Falcon, which are

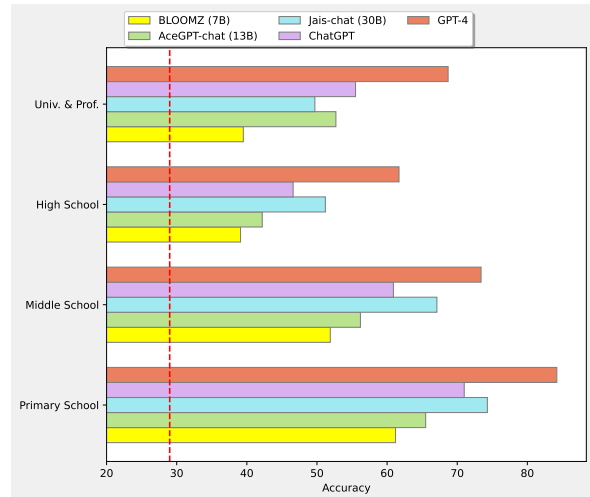


Figure 6: LLM performance across different education levels.

Country	# Q.	BLOOMZ	AceGPT	Jais
UAE	128	29.7	46.9	48.4
Egypt	830	45.0	48.0	55.6
Lebanon	239	55.6	62.8	69.5
Jordan	2532	45.6	51.5	59.8
Kuwait	111	44.1	53.2	58.6
KSA	37	32.4	54.1	56.8
Palestine	152	42.1	52.6	63.8
Morocco	314	25.9	52.7	33.1

Table 5: Average performance on subjects with Arabic-specific context, grouped by countries. Here we use BLOOMZ (7B), AceGPT-chat (13B), and Jais-chat (30B).

English-centric models.

Results across education levels Figure 6 depicts the average scores of the top-performing models (BLOOMZ, AceGPT-chat, Jais-chat, GPT-3.5, and GPT-4) across different education levels. We observe that ArabicMMLU questions are more challenging at the high school level compared to the primary and middle school levels. Specifically, for high school questions, GPT-4 achieves a score of only 61.7%, while Jais-chat scores 51.2%. Interestingly, we notice that the model accuracy at the university level is higher than for high school. This could be attributed to the relatively small portion (i.e., 6%) of university-level questions in ArabicMMLU, which potentially skews the results.

Results by country We present the performance of open-source models on selected subjects that potentially contain Arabic-specific contexts. These subjects include history, geography, civics, political

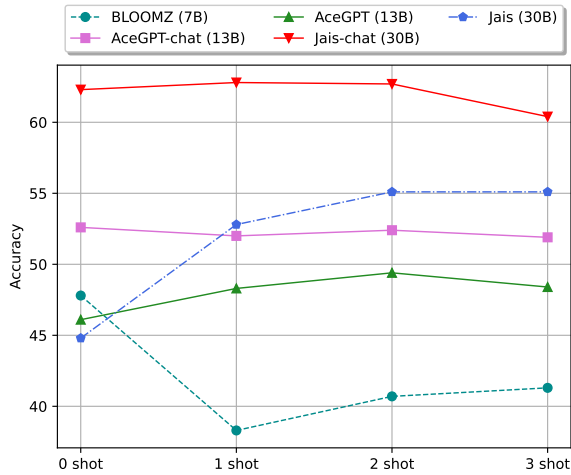


Figure 7: Few-shot performance (% accuracy) of LLMs averaged across all questions.

science, law, and driving tests, grouped by country in Table 5. We observe that BLOOMZ performs less well on questions sourced from the UAE and Morocco compared to other countries, while Jais performs best overall except in questions sourced from Morocco.

4.3 Analysis

We focus our more detailed analysis in this section solely on the best open-source models, namely BLOOMZ, AceGPT, and Jais, providing researchers and the community with insights to better understand these models and opportunities for future improvements.

Few-shot performance While all the results in Section 4.2 were based on zero-shot learning, we observe in Figure 7 that when we move to few-shot learning, results for base models improve but those for instruction-tuned models deteriorate. Specifically, AceGPT and Jais show an improvement of 2–10 points when using few-shot learning, but the results for BLOOMZ and Jais-chat drop. These findings are consistent with prior research over IndoMMLU (Koto et al., 2023) and CMMLU (Li et al., 2023).

Model confidence We analyze whether BLOOMZ, AceGPT, and Jais are well-calibrated in answering ArabicMMLU questions by comparing the probability of the correct answers with the actual accuracy for each task (i.e., subject and level combination). The answer probability is obtained through softmax normalization across the available candidate answers. In Figure 8, we observe that

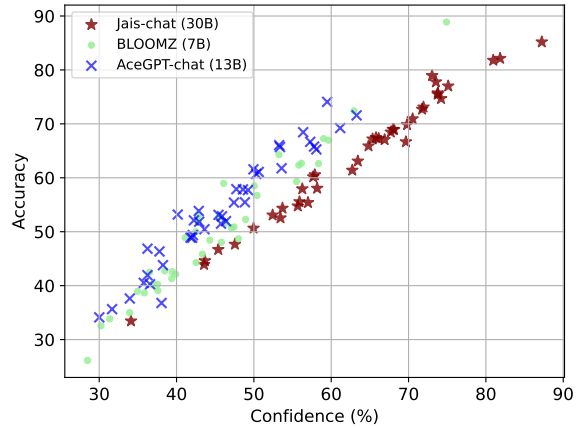


Figure 8: Zero-shot calibration of BLOOMZ, AceGPT-chat, and Jais-chat across 40 tasks. Confidence (%) denotes the average probability scores in percentage.

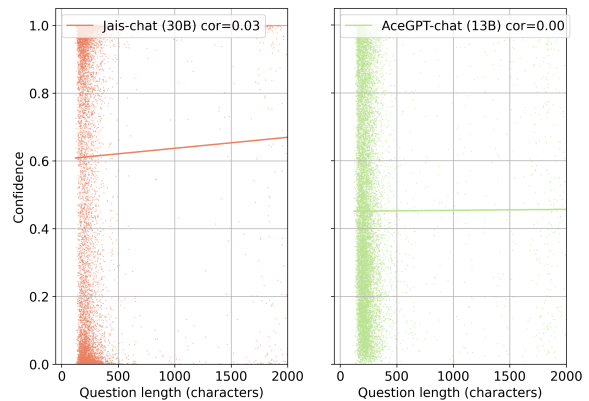


Figure 9: Correlation between model confidence and question length.

the three open-source models are well calibrated with correlation scores $r > 0.9$.

Additionally, we investigate the correlation between model confidence and question length in Figure 9. We find no correlation between the length of the questions and the model confidence for either Jais or AceGPT.

Impact of negation Despite negation being an absolutely foundational linguistic phenomenon, LLMs have been shown to be worryingly insensitive to its effects in English (Kassner and Schütze, 2020; Hosseini et al., 2021; Truong et al., 2023). We thus perform an analysis of LLM performance over questions in ArabicMMLU with and without negation to determine whether this observation ports across to Arabic. We utilize specific negation phrases to identify questions containing negations in Arabic. These include: لا (no), ليس (is not), ليست (is not), لم (did not), من غير (without),

Model	W/o negation	W/ negation
<i>Geography (8.0%)</i>		
BLOOMZ (7B)	42.8	42.2
AceGPT-chat (13B)	48.7	53.2
Jais-chat (30B)	56.9	48.6
<i>Biology (6.7%)</i>		
BLOOMZ (7B)	35.2	31.6
AceGPT-chat (13B)	37.7	35.8
Jais-chat (30B)	47.0	43.2
<i>Economics (13.3%)</i>		
BLOOMZ (7B)	52.5	37.7
AceGPT-chat (13B)	50.1	51.9
Jais-chat (30B)	60.4	54.5

Table 6: Model accuracy in answering questions with and without negation in Geography, Biology, and Economics. The number following the subject name indicates the proportion of negated questions.

باستثناء (excluding), and دون (without). To prevent ambiguity, the term ما is omitted, as it can also mean “what”. After applying this filtering, we obtain 816 questions. We randomly inspected 100 random samples and found the detection accuracy for negation to exceed 95%.

Table 6 presents the accuracy of the LLMs in answering questions with and without negation in the top three subjects containing negation (Geography, Biology, and Economics). Overall, negated questions generally exhibit slightly lower accuracy, particularly in Biology and Economics. However, for Geography, the models actually achieve higher accuracy.

4.4 Discussion

Our experiments show that open-source LLMs perform poorly on ArabicMMLU questions, particularly multilingual models. Furthermore, the Arabic-centric LLMs still struggle to capture Arabic cultural knowledge across all education levels. This highlights a significant need for improvement in Arabic LLMs. In contrast, GPT-4 demonstrates remarkable performance across all tasks, surpassing all other models. However, it remains unclear whether the success of GPT-4 results from scaling up the dataset and model size or simply from memorization (given that all questions were taken from public sources).

5 Conclusion and Future Work

We introduce ArabicMMLU, the first large-scale multi-task language understanding dataset designed to evaluate real-world knowledge in Arabic. Through experiments with over 14K multiple-choice questions spanning various subjects and education levels, we observed that Arabic-centric LLMs outperform multilingual LLMs, albeit with lower accuracy than GPT-4. We envision ArabicMMLU as a valuable resource for tracking the real-world knowledge and reasoning capabilities of future Arabic LLMs. For future work, ArabicMMLU can be extended to include short-answer or essay questions, different modalities (i.e., images, audio, video), larger region coverage, and more questions in professional settings. This will enhance the evaluation to better reflect real-world scenarios.

Limitations

Although we believe our benchmark will significantly contribute to the advancement of Arabic LLMs, it is important to acknowledge limitations that need to be addressed in future work. We outline these limitations as follows:

Limited diversity ArabicMMLU does not represent all Arabic countries equally. For example, we have collected over 6K multiple-choice questions from Jordan, while other countries are represented with only 100 questions or, in some cases, not at all. This is largely due to the availability of publicly-accessible exams in each country; some countries have digitized their exams, but not others. Additionally, our search for relevant Arabic content across the internet was not exhaustive.

Dialectical Arabic is excluded The dataset primarily focuses on Modern Standard Arabic (MSA). However, multilingual and Arabic LLMs are often exposed to both MSA and dialectal Arabic.

Text-based questions only ArabicMMLU is focused solely on text-based assessment, and the exploration of multimodal questions is left for future work.

Ethical Considerations

It is important to emphasize that our experimental results do not provide conclusive answers regarding the performance of LLMs in Arabic. This issue becomes even more vexing when discussing the GPT-4 results, which outperformed all models, due

to a lack of sufficient information about its training regimen. As such, we cannot assert that the model’s pretraining data is free from contamination.

Acknowledgements

We extend our gratitude to all collaborators from Jordan, Egypt, Lebanon, UAE, and Saudi Arabia who participated in the data collection process. We also acknowledge the contributions of Samta Kamboj, Sarah Al Barri, and Onkar Pandit from Core42, who assisted in collecting the Arabic Language question dataset.

References

- Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Yousseif Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. [LARA-Bench: Benchmarking Arabic AI with large language models](#). arXiv preprint arXiv:2305.14982.
- 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.
- Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. [MathQA: Towards interpretable math word problem solving with operation-based formalisms](#). 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 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. [AraBERT: Transformer-based model for Arabic language understanding](#). In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2021a. [AraELECTRA: Pre-training text discriminators for Arabic language understanding](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2021b. [AraGPT2: Pre-trained transformer for Arabic language generation](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- 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.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, and Mohamed Eldesouki. 2017. [Arabic POS tagging: Don’t abandon feature engineering just yet](#). In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 130–137, Valencia, Spain. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.
- Mona Diab, Nizar Habash, and Imed Zitouni. 2017. [NLP for Arabic and related languages](#). *Traitement Automatique des Langues*, 58(3):9–13.

- AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. [ORCA: A challenging benchmark for arabic language understanding](#). *arXiv preprint arXiv:2212.10758*.
- 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.
- 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.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R. De- von Hjelm, Alessandro Sordani, and Aaron C. Courville. 2021. [Understanding by understanding not: Modeling negation in language models](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1301–1312. Association for Computational Linguistics.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization](#). In *Proceedings of ICML 2020*.
- Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, et al. 2023. [AceGPT, localizing large language models in Arabic](#). *arXiv preprint arXiv:2309.12053*.
- Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. [Cosmos QA: Machine reading comprehension with contextual commonsense reasoning](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.
- Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. [The interplay of variant, size, and task type in Arabic pre-trained language models](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. [ArABART: a pretrained Arabic sequence-to-sequence model for abstractive summarization](#). In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 31–42, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Nora Kassner and Hinrich Schütze. 2020. [Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7811–7818. Association for Computational Linguistics.
- Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. [Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.
- Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. [Cloze evaluation for deeper understanding of commonsense stories in Indonesian](#). In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 8–16, Dublin, Ireland. Association for Computational Linguistics.
- Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. [IndoCulture: Exploring geographically-influenced cultural commonsense reasoning across eleven Indonesian provinces](#). *arXiv preprint arXiv:2404.01854*.
- Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. [WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and

- Sushant Prakash. 2023. [RLAIF: Scaling reinforcement learning from human feedback with AI feedback](#). arXiv preprint arXiv:2309.00267.
- 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.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. CMMLU: Measuring massive multitask language understanding in Chinese. *arXiv preprint arXiv:2306.09212*.
- 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.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. [Few-shot learning with multilingual generative language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Timothy Baldwin, and Eric P. Xing. 2023. [LLM360: Towards fully transparent open-source LLMs](#). *ArXiv*, abs/2312.06550.
- 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.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.
- 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.
- OpenAI. 2023. GPT-4 technical report. *ArXiv*, abs/2303.08774.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. [Cross-lingual name tagging and linking for 282 languages](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. [Fairness in language models beyond English: Gaps and challenges](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. 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.
- 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*.
- Abdulhadi Shoufan and Sumaya Alameri. 2015. [Natural language processing for dialectal Arabic: A survey](#).

- In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China. Association for Computational Linguistics.
- Zeeraq Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. [You reap what you sow: On the challenges of bias evaluation under multilingual settings](#). In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruiti Bhosale, et al. 2023. Llama2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. [Language models are not naysayers: an analysis of language models on negation benchmarks](#). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Hao Yu, Bo Shen, Dezhi Ran, Jiaxin Zhang, Qi Zhang, Yuchi Ma, Guangtai Liang, Ying Li, Qianxiang Wang, and Tao Xie. 2024. CoderEval: A benchmark of pragmatic code generation with generative pre-trained models. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, pages 1–12.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Data Statistics

Table 7 presents the distribution of ArabicMMLU data categorized by subject across different education levels.

Subject	#question
Primary school	
Arabic Language	255
Computer Science	193
General Knowledge	165
Geography	60
History	105
Islamic Studies	1002
Math	412
Natural Science	339
Social Science	708
Middle school	
Arabic Language	30
Civics	239
Computer Science	30
Economics	90
General Knowledge	175
Geography	275
History	206
Islamic Studies	241
Natural Science	245
Social Science	244
High school	
Arabic Language	393
Biology	1412
Civics	90
Computer Science	264
Economics	363
Geography	1041
History	763
Islamic Studies	337
Philosophy	42
Physics	258
University and Professional	
Accounting	77
Computer Science	67
Economics	140
Management	78
Political Science	213
Law	317
Other / NA	
Arabic Language (General)	615
Arabic Language (Grammar)	368
Driving Test	1214
General Knowledge	867
Islamic Studies	642
Total	14575

Table 7: The distribution of ArabicMMLU for each subject in different education levels.

B Examples

Figure 10 illustrates a complete example of prompts used in this study. This example features a Natural Science question with prompts provided in both Arabic and English.

هذا سؤال في العلوم الطبيعية للمدرسة الابتدائية في الأردن. اختر الإجابة الصحيحة

سؤال: واحد مما يلي عضوا ليس من أعضاء الحس:

أ- الأنف
ب- العين
ج- الدماغ
د- الأذن

إجابة:

This is a Natural Science question for primary school in Jordan. Select the correct answer!

Question: واحد مما يلي عضوا ليس من أعضاء الحس

A. الأنف
B. العين
C. الدماغ
D. الأذن

Answer:

Figure 10: Example of prompt input in Arabic and English.

C Detailed Experiment Results

Table 8 presents the detailed zero-shot results across subjects and education levels, while Table 9, Table 10, Table 11 display the results with different prompts and alphabetic outputs (complementing the main result at Table 8).

Subject	BLOOMZ	AceGPT-chat	Jais-chat	GPT-3.5	GPT-4
Primary School					
Arabic Language	64.3	60.7	63.1	65.1	80.6
Computer Science	62.6	65.3	68.9	66.8	80.5
General Knowledge	62.3	66.0	74.7	75.9	77.2
Geography	50.9	57.9	61.4	66.7	82.5
History	48.0	52.0	75.5	56.9	71.6
Islamic Studies	67.0	71.6	81.8	74.0	89.8
Math	41.3	48.9	57.9	58.2	76.0
Natural Science	67.3	68.5	82.1	80.4	88.7
Social Science	62.7	69.2	75.7	74.3	84.7
Middle School					
Arabic Language	51.9	51.9	77.8	55.6	85.2
Civics	40.3	40.3	60.2	45.3	62.7
Computer Science	88.9	74.1	85.2	81.5	96.3
Economics	72.4	66.7	77.0	77.0	81.6
General Knowledge	59.0	65.3	70.5	67.6	78.6
Geography	50.7	57.7	67.3	62.5	75.4
History	52.7	61.1	68.5	62.6	71.9
Islamic Studies	56.7	55.5	73.1	62.6	73.9
Natural Science	51.7	61.6	69.8	70.2	87.2
Social Science	42.7	49.4	54.4	49.8	57.7
High School					
Arabic Language	33.8	35.6	44.6	36.7	44.6
Biology	35.0	37.6	46.7	42.4	59.6
Civics	39.1	36.8	47.7	39.1	44.8
Computer Science	42.1	52.1	55.6	57.9	74.7
Economics	45.8	48.9	58.1	56.7	71.1
Geography	40.2	46.3	53.1	49.0	66.1
History	38.9	40.5	50.6	42.7	54.1
Islamic Studies	52.8	51.3	66.9	62.4	76.7
Philosophy	59.0	53.8	66.7	59.0	74.4
Physics	32.5	34.1	43.9	42.0	61.6
University and Professional					
Accounting	50.0	55.4	55.4	59.5	73.0
Computer Science	48.4	53.1	67.2	57.8	78.1
Economics	48.9	43.8	52.6	52.6	62.8
Management	48.7	65.8	78.9	64.5	80.3
Political Science	44.3	52.9	54.8	51.4	66.7
Law	25.9	52.7	33.1	55.8	66.9
Other / NA					
Arabic Language (General)	58.5	57.8	72.7	66.7	84.5
Arabic Language (Grammar)	42.5	46.8	60.5	59.7	77.3
Driving Test	52.3	61.8	65.9	68.3	79.5
General Knowledge	42.5	50.4	68.9	54.5	72.5
Islamic Studies	38.7	41.9	67.4	44.0	71.8

Table 8: Zero-shot LLM performance (% accuracy) with **English prompt and English alphabetic output**, for each subject and education level. The models are BLOOMZ (7B), AceGPT-chat (13B), Jais-chat (30B), GPT-3.5 (175B), and GPT-4.

Model (#parameters)	STEM	Social Science	Humanities	Arabic Language	Other	Average
Random	29.5	28.9	28.6	25.8	32.3	29.0
XGLM (1.7B)	28.8	28.0	25.4	25.1	28.7	27.3
XGLM (2.9B)	28.8	26.9	26.8	26.5	32.3	28.2
XGLM (4.5B)	30.5	27.0	27.1	26.6	32.0	28.6
XGLM (7.5B)	30.3	25.7	25.8	25.0	32.6	27.9
BLOOMZ (560M)	29.3	26.3	27.1	23.7	27.2	27.0
BLOOMZ (1.1B)	31.3	28.1	31.0	28.3	29.0	29.7
BLOOMZ (1.7B)	32.5	34.9	35.2	30.4	35.1	34.0
BLOOMZ (3B)	38.3	42.6	40.0	36.2	39.5	39.7
BLOOMZ (7B)	37.7	40.5	34.8	38.2	39.6	38.0
mT0 _{small} (300M)	29.1	28.7	26.0	22.5	27.3	27.2
mT0 _{base} (580M)	30.2	30.5	33.1	24.8	34.3	31.1
mT0 _{large} (1.2B)	29.4	28.8	23.9	22.7	27.2	26.7
mT0 _{xl} (3.7B)	39.0	40.2	39.5	41.3	43.7	40.5
mT0 _{xxl} (13B)	40.3	43.5	41.3	38.6	43.3	41.7
LLaMA2 (7B)	31.7	31.3	33.2	27.2	32.2	31.6
LLaMA2-chat (7B)	31.5	31.4	30.9	26.6	30.9	30.7
LLaMA2 (13B)	31.8	31.7	32.5	29.3	38.4	32.8
LLaMA2-chat (13B)	30.8	30.2	25.4	24.7	29.7	28.4
Falcon (7B)	29.3	27.6	26.8	23.8	28.1	27.4
Falcon-instruct (7B)	28.9	28.7	26.5	22.3	27.5	27.3
Falcon (40B)	30.1	30.3	31.1	24.8	31.5	30.0
Falcon-instruct (40B)	29.3	29.0	27.6	22.9	28.2	27.9
AraT5 (220M)	28.2	25.7	23.5	24.2	26.7	25.7
AraT5v2 (220M)	31.2	31.1	33.0	27.7	34.5	31.8
AraGPT2 (1.7B)	29.9	30.5	31.6	28.1	35.1	31.2
AceGPT (7B)	31.8	28.2	29.9	28.0	31.9	30.0
AceGPT-chat (7B)	42.9	47.7	50.5	42.6	52.5	47.6
AceGPT (13B)	38.4	42.0	42.1	36.8	41.8	40.6
AceGPT-chat (13B)	44.3	50.9	49.0	50.8	53.8	49.4
Jais (13B)	31.6	34.4	35.9	29.7	38.5	34.3
Jais-chat (13B)	51.6	55.1	57.3	41.1	59.3	54.0
Jais (30B)	33.2	35.1	34.4	27.7	39.4	34.4
Jais-chat (30B)	53.3	57.9	62.9	60.0	66.8	59.9

Table 9: Zero-shot LLM performance (% accuracy) with **Arabic prompt and Arabic alphabetic output**, combined across subject groups. “Average” means the average across all questions in ArabicMMLU.

Model (#parameters)	STEM	Social Science	Humanities	Arabic Language	Other	Average
Random	29.5	28.9	28.6	25.8	32.3	29.0
XGLM (1.7B)	29.9	30.7	30.8	27.7	34.8	30.9
XGLM (2.9B)	29.4	30.7	31.2	27.9	34.4	30.9
XGLM (4.5B)	28.8	29.8	30.5	27.4	31.5	29.8
XGLM (7.5B)	27.7	27.5	24.8	26.5	29.3	27.0
BLOOMZ (560M)	31.2	30.9	33.1	28.1	35.7	32.0
BLOOMZ (1.1B)	30.3	26.7	31.1	25.5	27.5	28.6
BLOOMZ (1.7B)	36.3	38.8	38.7	38.0	39.1	38.2
BLOOMZ (3B)	40.5	45.5	44.3	43.9	47.8	44.3
BLOOMZ (7B)	43.3	47.4	47.5	50.9	50.4	47.4
mT0 _{small} (300M)	30.7	30.7	31.4	28.0	34.5	31.2
mT0 _{base} (580M)	30.6	31.0	31.6	29.3	35.7	31.7
mT0 _{large} (1.2B)	30.0	30.0	29.9	29.2	34.6	30.7
mT0 _{xl} (3.7B)	39.5	43.9	40.9	43.4	43.9	42.1
mT0 _{xxl} (13B)	41.2	45.2	43.3	46.7	43.8	43.8
LLaMA2 (7B)	32.2	29.0	31.4	27.2	30.3	30.3
LLaMA2-chat (7B)	31.7	30.7	29.5	30.3	31.4	30.7
LLaMA2 (13B)	33.2	34.1	35.3	31.7	39.5	34.9
LLaMA2-chat (13B)	33.3	30.8	30.7	31.5	36.2	32.4
Falcon (7B)	29.8	30.6	31.4	28.2	35.1	31.1
Falcon-instruct (7B)	29.9	30.7	31.5	28.0	35.1	31.2
Falcon (40B)	34.8	31.8	34.3	29.9	38.6	34.1
Falcon-instruct (40B)	33.3	29.3	33.3	30.9	39.3	33.1
AraT5 (220M)	29.9	30.3	33.0	28.5	32.0	31.0
AraT5v2 (220M)	29.1	29.8	31.1	28.3	33.6	30.5
AraGPT2 (1.7B)	33.0	31.5	35.8	29.6	37.4	33.7
AceGPT (7B)	33.6	32.3	35.2	27.6	38.9	33.9
AceGPT-chat (7B)	42.4	47.2	49.8	41.4	51.3	46.9
AceGPT (13B)	43.2	46.6	47.5	42.4	50.0	46.2
AceGPT-chat (13B)	46.7	53.2	52.8	50.7	57.3	52.1
Jais (13B)	32.5	35.1	34.3	28.2	37.4	33.9
Jais-chat (13B)	52.4	56.6	60.0	42.5	60.4	55.6
Jais (30B)	39.6	45.1	49.0	32.9	49.1	44.2
Jais-chat (30B)	55.7	59.7	67.5	61.4	68.3	62.4

Table 10: Zero-shot LLM performance (% accuracy) with **Arabic prompt and English alphabetic output**, combined across subject groups. “Average” means the average across all questions in ArabicMMLU.

Model (#parameters)	STEM	Social Science	Humanities	Arabic Language	Other	Average
Random	29.5	28.9	28.6	25.8	32.3	29.0
XGLM (1.7B)	30.0	29.9	26.7	27.2	29.6	28.7
XGLM (2.9B)	29.1	27.2	29.5	27.8	31.0	28.9
XGLM (4.5B)	29.8	26.8	26.9	27.6	31.8	28.4
XGLM (7.5B)	30.4	26.3	26.7	27.8	32.4	28.5
BLOOMZ (560M)	29.5	25.9	26.3	23.9	27.1	26.8
BLOOMZ (1.1B)	31.3	29.3	30.4	28.3	29.3	29.9
BLOOMZ (1.7B)	32.0	33.5	33.6	30.0	34.3	32.9
BLOOMZ (3B)	39.3	42.0	41.8	35.2	40.9	40.4
BLOOMZ (7B)	37.6	41.3	36.2	38.3	40.8	38.8
mT0 _{small} (300M)	29.1	28.4	27.0	22.7	27.5	27.4
mT0 _{base} (580M)	29.5	30.3	33.3	25.3	32.6	30.7
mT0 _{large} (1.2B)	28.6	28.3	24.6	22.7	27.3	26.6
mT0 _{xl} (3.7B)	36.8	38.9	37.7	39.8	43.2	39.0
mT0 _{xxl} (13B)	39.1	41.9	40.0	36.7	42.1	40.2
LLaMA2 (7B)	33.0	31.2	35.5	29.5	34.4	33.0
LLaMA2-chat (7B)	34.5	33.1	31.3	27.7	34.9	32.6
LLaMA2 (13B)	33.5	30.9	31.7	30.6	35.0	32.3
LLaMA2-chat (13B)	34.8	33.6	31.7	28.7	36.6	33.3
Falcon (7B)	29.9	30.3	34.4	27.7	32.5	31.3
Falcon-instruct (7B)	28.5	28.4	28.9	23.0	27.8	27.8
Falcon (40B)	32.4	31.6	34.7	26.9	33.3	32.3
Falcon-instruct (40B)	30.3	31.2	29.5	23.5	29.4	29.4
AraT5 (220M)	28.1	25.7	23.4	24.8	26.7	25.7
AraT5v2 (220M)	31.3	30.0	32.9	27.1	32.9	31.2
AraGPT2 (1.7B)	29.9	30.5	31.6	28.1	35.1	31.2
AceGPT (7B)	28.6	26.5	25.7	26.1	27.7	26.9
AceGPT-chat (7B)	43.0	46.5	49.4	42.8	52.2	47.0
AceGPT (13B)	37.6	38.9	40.1	34.3	43.6	39.2
AceGPT-chat (13B)	46.4	50.9	50.1	50.2	54.7	50.3
Jais (13B)	30.5	32.0	34.5	28.7	36.3	32.7
Jais-chat (13B)	49.2	53.4	55.8	38.5	59.5	52.4
Jais (30B)	39.1	43.0	47.5	32.9	49.1	43.2
Jais-chat (30B)	54.7	58.8	63.3	59.7	67.4	60.6

Table 11: Zero-shot LLM performance (% accuracy) with **English prompt and Arabic alphabetic output**, combined across subject groups. “Average” means the average across all questions in ArabicMMLU.

D Model Artifacts

Table 12 lists the sources of pre-trained models used in this study. All models are sourced from Huggingface (Wolf et al., 2020).

Models (#parameters)	Source
XGLM (564M)	facebook/xglm-564M
XGLM (1.7B)	facebook/xglm-1.7B
XGLM (2.9B)	facebook/xglm-2.9B
XGLM (4.5B)	facebook/xglm-4.5B
XGLM (7.5B)	facebook/xglm-7.5B
BLOOMZ (560M)	bigscience/bloomz-560m
BLOOMZ (1.1B)	bigscience/bloomz-1b1
BLOOMZ (1.7B)	bigscience/bloomz-1b7
BLOOMZ (3B)	bigscience/bloomz-3b
BLOOMZ (7.1B)	bigscience/bloomz-7b1
mT0 _{small} (300M)	bigscience/mt0-small
mT0 _{base} (580M)	bigscience/mt0-base
mT0 _{large} (1.2B)	bigscience/mt0-large
mT0 _{xl} (3.7B)	bigscience/mt0-xl
mT0 _{xxl} (13B)	bigscience/mt0-xxl
LLaMA2 (7B)	meta-llama/Llama-2-7b
LLaMA2-chat (7B)	meta-llama/Llama-2-7b-chat
LLaMA2 (13B)	meta-llama/Llama-2-13b
LLaMA2-chat (13B)	meta-llama/Llama-2-13b-chat
Falcon (7B)	tiiuae/falcon-7b
Falcon-instruct (7B)	tiiuae/falcon-7b-instruct
Falcon (40B)	tiiuae/falcon-40b
Falcon-instruct (40B)	tiiuae/falcon-40b-instruct
AraT5 (220M)	UBC-NLP/AraT5-base
AraT5v2 (220M)	UBC-NLP/AraT5v2-base-1024
AraGPT2 (1.7BB)	aubmindlab/aragpt2-mega
AceGPT (7B)	FreedomIntelligence/AceGPT-7B
AceGPT-chat (7B)	FreedomIntelligence/AceGPT-7B-chat
AceGPT (13B)	FreedomIntelligence/AceGPT-13B
AceGPT-chat (13B)	FreedomIntelligence/AceGPT-13B-chat
Jais (13B)	core42/jais-13b
Jais-chat (13B)	core42/jais-13b-chat
Jais (30B)	core42/jais-30b-v3
Jais-chat (30B)	core42/jais-30b-chat-v3

Table 12: With the exception of GPT-3.5 and GPT-4, all the models used in this study were sourced from Huggingface (Wolf et al., 2020).