AraEval: An Arabic Multi-Task Evaluation Suite for Large Language Models

Alhanoof Althnian[1,3], Norah Alzahrani[1], Shaykhah Z. Alsubaie[1], Eman Albilali[1,3]
Ahmed Abdelali[1], Nouf M. Alotaibi[1], M Saiful Bari[2], Yazeed Alnumay[2]
Abdulhamed alothaimen[1], Maryam Saif[2], Shahad D. Alzaidi[2], Faisal Mirza[2]
Yousef Almushayqih[2], Mohammed Al Saleem[2], Ghadah Alabduljabbar[1]
Abdulmohsen Al-Thubaity[1], Areeb Alowisheq[1], Nora Al-Twairesh[1,3]
[1] HUMAIN, [2]Saudi Data and AI Authority (SDAIA), [3]King Saud University
Riyadh, Saudi Arabia
Correspondence: aalthnian@humain.com

Abstract

The rapid advancements of Large Language models (LLMs) necessitate robust benchmarks. In this paper, we present AraEval, a pioneering and comprehensive evaluation suite specifically developed to assess the advanced knowledge, reasoning, truthfulness, and instruction-following capabilities of foundation models in the Arabic context. AraEval includes a diverse set of evaluation tasks that test various dimensions of knowledge and reasoning, with a total of 24,378 samples. These tasks cover areas such as linguistic understanding, factual recall, logical inference, commonsense reasoning, mathematical problem-solving, and domain-specific expertise, ensuring that the evaluation goes beyond basic language comprehension. It covers multiple domains of knowledge, such as science, history, religion, and literature, ensuring that the LLMs are tested on a broad spectrum of topics relevant to Arabic-speaking contexts. AraEval is designed to facilitate comparisons across different foundation models, enabling LLM developers and users to benchmark performance effectively. In addition, it provides diagnostic insights to identify specific areas where models excel or struggle, guiding further development. AraEval datasets can be found at https://huggingface.co/collections/humain-ai/araeval-datasets-687760e04b12a7afb429a4a0.

1 Introduction

With the unprecedented scaling of large language models (LLMs) (OpenAI, 2022; Google, 2024; Anthropic, 2022; Dubey et al., 2024; Mistral, 2024; Team et al., 2024; Liu et al., 2024; Team, 2024), algorithmic intelligence has reached new frontiers (Guo et al., 2025; Jaech et al., 2024) across numerous domains, demonstrating remarkable abilities in tasks ranging from creative writing (Gómez-Rodríguez, 2023), program synthesis (Jimenez et al., 2023; Khan et al., 2024), instruction following (Zhou et al., 2023), knowledge extraction (Hendrycks et al., 2021; Wang et al., 2024b) to rich scientific reasoning (Mialon et al., 2023; Rein et al., 2023). The field has witnessed breakthroughs, with models matching or surpassing expert human performance (Glazer et al., 2024) - from solving olympiad-level problems (AlphaCode Team, 2023; Chervonyi et al., 2025) to generating research-level insights (Google, 2025; OpenAI, 2025) - catalyzing massive industry investments [1] and research efforts (Workshop et al., 2022; Lovenia et al., 2024; LAION-AI, 2025; Lozhkov et al., 2024). As model capabilities rapidly expand and emerge on a different scale (Wei et al., 2022; Srivastava et al., 2022), systematic evaluation (Laskar et al., 2023; Phan et al., 2025) serves as a vital proxy for decision making across the ecosystem, enabling key stakeholders - from developers and regulators to investors, researchers, and industry practitioners - to make informed strategic choices (Handa et al., 2025) about model development, deployment, and adoption (Latent Space, 2024).

Despite progress, the evaluation landscape remains significantly skewed towards English and other high-resource languages (Joshi et al., 2020), creating a significant gap in our understanding of LLM capabilities in different linguistic and cultural contexts. In addition to that Yong et al. (2023) showed that safety or instruction following don't generalize with low-resource languages. This disparity is particularly pronounced for Arabic, the fifth most spoken language worldwide with more than 400 million speakers (Eberhard et al., 2020) and rich dialectal variations spanning more than 20 countries. Although recent years have seen the emergence of Arabic-specific language models (Bari et al., 2025; Abbas et al., 2025; Sengupta et al., 2023b; Huang et al., 2023) and the increasing integration of Arabic in multilingual models (Team,

[1] https://openai.com/index/announcing-the-stargate-project/

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33037–33061
November 4-9, 2025 ©2025 Association for Computational Linguistics

2024; Mistral, 2024; Jaech et al., 2024), comprehensive evaluation frameworks for assessing their capabilities remain limited.

Existing Arabic evaluation efforts have primarily focused on translating english benchmarks (Huang et al., 2023; OpenAI, 2025; Sengupta et al., 2023b) or targeted towards only knowledge base questions (Koto et al., 2024; Almazrouei et al., 2023), lacking the systematic multi-task assessment necessary for understanding model performance across diverse linguistic phenomena and real-world applications. Notable initiatives like ArabicMMLU (Koto et al., 2024), Exams (Hardalov et al., 2020), ACVA (Huang et al., 2023), Belebele (Bandarkar et al., 2023), and AraDiCE (Mousi et al., 2024), along with various leaderboard efforts (El Filali et al., 2024), have established foundational work in Arabic language evaluation. Recent work by Bari et al. (2025) and Abbas et al. (2025) have attempted to address these limitations through human evaluation, but this approach faces inherent challenges of *scalability* and *consistency*, being vulnerable to variations in setup, prompt design, individual assessor biases, and temporal factors.

In this work, we introduce AraEval, a comprehensive Arabic multi-task evaluation suite designed to rigorously assess large language models (LLMs). AraEval introduces a collection of **novel**, carefully designed **holistic** Arabic language benchmarking evaluation datasets that address these critical limitations. AraEval serves as a native Arabic benchmark, ensuring cultural, linguistic, and normative alignment with Arabic-speaking communities. Our contributions include:

1. AraEval includes **24,378 novel** samples across knowledge, reasoning, truthfulness,and instruction-following (Table 1).
2. AraEval facilitates detailed diagnostic assessments of model performance, enabling the identification of specific strengths and weaknesses in reasoning, instruction-following, and knowledge retention. (Figures 1, 3, 4 and 7 and tables 9 to 12)
3. AraEval includes higher Arabic token coverage than ArabicMMLU and OpenAI's Arabic-translated MMMLU (Figure 5 and table 18).
4. AraEval supports both log-probability-based and API-based evaluation schemes, facilitating seamless assessment of both open and close-source models.

2 AraEval Evaluation Suite

We contribute seven datasets of Arabic benchmarks, which vary in capabilities as shown in Table 1.

Task	Type	Dataset	Test Split	Dev Split
Knowledge	MCQ	AraPro	5001	110
Knowledge	MCQ	IEN MCQ	9990	190
Knowledge	Boolean	IEN TF	5823	190
Reasoning	MCQ	AraMath	605	5
Reasoning	MCQ	ETEC	1887	5
Instruction following	Generation	AraIFEval	536	-
Truthfulness	MCQ	AraTruthfulQA	536	5
Total			24,378	

Table 1: AraEval tasks splits statistics.

2.1 Design Principles

To establish a comprehensive Arabic benchmark for evaluating LLMs across diverse tasks, we developed our datasets based on the following principles:

Human-curated or human-validated: Every dataset of AraEval is meticulously created by experts or rigorously validated by humans to ensure the highest standards of quality and relevance. This guarantees that the questions, answers, and annotations are both accurate and meaningful, reflecting real-world scenarios and challenges. The validation criteria were task-specific, and human validators received specialized training on the respective tasks before beginning the validation process. The validaiton process was conducted by three humans where majority agreement was taken as the final verdict. Guidelines for human annotators to create/validate the datasets can be found in Section G.

Granularity for fine-grained evaluation: Our datasets are designed with a high level of granularity, enabling detailed evaluation and nuanced insights into model performance. Fine-grained labels allow for the analysis of specific areas of strength and weakness, making the datasets particularly useful for diagnostic and comparative studies.

Cultural and normative alignment: All datasets are thoughtfully aligned with Arabic culture, values, and norms. This ensures the content is appropriate, contextually relevant, and reflective of the diverse realities of Arabic-speaking communities, allowing for more authentic and reliable evaluations.

2.2 Datasets Overview

2.2.1 AraPro

This dataset comprises 5,001 multiple-choice questions (MCQs) carefully crafted by university professors across 19 distinct knowledge domains.

These experts were selected and instructed to create MCQs that reflect the competencies expected of professionals in their respective fields. Therefore, the questions evaluate LLMs in achieving professional-level competency within these domains. To ensure cultural and geographic neutrality, we did not impose any restrictions tying the content to Saudi Arabia or any specific country, as shown in the guideline in Table 14. A detailed breakdown of the knowledge domains and the corresponding number of questions is provided in Table 12, while we show subject categories distribution in Figure 7.

2.2.2 IEN

The global pandemic of COVID-19 has challenged the world and inevitably the education sector. In Saudi Arabia, the Ministry of Education responded by launching the IEN[2] platform as part of its broader e-learning and distance education strategy.

The IEN platform includes a vast repository of more than 1.5 million questions and answers, meticulously classified into varying levels of difficulty. This extensive database not only supports differentiated learning, but also enables customized assessments that address the unique needs and abilities of students at every stage of their educational journey.

A representative subset that covers all grades, subjects and levels of difficulty was randomly selected from the IEN platform as shown in Table 1, the selection contains 5,823 samples as true/false questions and 9,990 MCQs. Figure 1 shows the detailed distributions of the questions and subjects per grade level. Table 10 and Table 11 provide more granular details about the dataset.

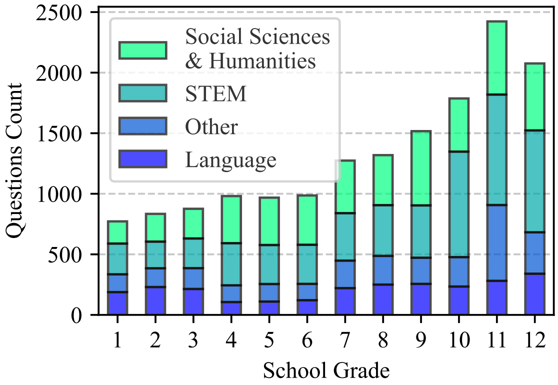


Figure 1: Course and grade level coverage for TF and MCQ IEN datasets combined.

2.2.3 AraMath

AraMath consists of 605 MCQs derived from ArMath (Alghamdi et al., 2022), which includes mathematical word problems, and the solution is an equation that solves the problem. We reformulated the dataset and converted it to a multiple-choice problem (MCQ). The correct answer is extracted from the equation by parsing the formulas, and three random distractors were generated to complete the set of options.

Human annotators meticulously reviewed and validated the dataset to ensure the accuracy of the equations in representing the mathematical word problems. They also assessed choice distinctiveness, verifying that all answer choices were unique and free of duplicates, and answer correctness, ensuring that the labeled answer corresponded to the correct choice.

2.2.4 ETEC

The Education & Training Evaluation Commission (ETEC)[3]serves as an independent regulatory body responsible for evaluating, measuring, and accrediting qualifications in education and training in both the public and private sectors in Saudi Arabia. Its role includes ensuring and enhancing the quality and efficiency of educational and training institutions, programs, and their outcomes. The commission offers more than 42 types of qualification tests spanning all educational levels from K12 to professional levels. A subset of 1887 MCQs were chosen from different types of tests that include: a) Qudurati: A series of tests offered to students from 3rd grade elementary school to 10th grade to assess their level of general aptitude in comprehension, analysis, reasoning, and application, focusing on their readiness for learning. b) Professional Educational Occupation License Test: A standardized assessment tool to measure applicants' competency in general and specialized educational standards for on-the-job teachers.

2.2.5 AraIFEval

AraIFEval is an Arabic instruction-following (IF) evaluation benchmark designed to automatically assess language models' compliance with specified instructions through verifiable methods. The dataset consists of 535 instances, each containing two to four verifiable instructions that can be validated using deterministic programming approaches.

[2] https://ien.edu.sa/

[3] https://etec.gov.sa/en/

An example of the AraIFEval dataset with verifiable instructions is shown in Section E.3.

We created a collection of 23 Arabic verifiable instructions, inspired by Zhou et al. (2023). To construct the dataset, we randomly selected open questions from our data to serve as seed prompts. We generated IF prompts by randomly combining two to four instructions for each prompt, carefully ensuring logical consistency and avoiding contradictions between instructions. The dataset was then reviewed by humans for quality assurance. The Arabic verifiable instructions are presented in Section F, while the dataset distribution is detailed in Figure 8. To enable automatic response verification, we implemented regex-based category phrase checking. We followed Zhou et al. (2023)'s evaluation approach to assess instruction-following capabilities following strict and loose criteria. Similar to Fourrier et al. (2024), we only report strict accuracy in this work.

2.2.6 AraTruthfulQA

Inspired by TruthfulQA Lin et al. (2021), this benchmark evaluates the truthfulness of LLM responses to questions designed to elicit common misconceptions. The benchmark targets questions that some individuals may answer incorrectly due to false beliefs or misinformation. It comprises questions spanning diverse categories with a particular emphasis on prevalent misconceptions in the Arab world. To ensure cultural relevance, we reviewed TruthfulQA dataset and selected 287 questions that align with Arabic cultural norms and beliefs, ensuring they are culturally appropriate and broadly acceptable across the Arabic-speaking world. To avoid regional or cultural skew, we instructed the human annotators to use Modern Standard Arabic (MSA) for translation (see Table 17). Additionally, we crafted 249 culturally relevant questions of similar complexity and depth, specifically addressing common misconceptions in the Arab world, further enhancing the benchmark's comprehensiveness.

3 Experiments

3.1 Setup

In this paper, we integrate the `AraEval` benchmark with the LM Evaluation Harness framework (Gao et al., 2024). We evaluate models in zero-shot and few-shot settings, using test and dev sets; except for AraIFEval, where only zero-shot results are reported. To mitigate token bias (Alzahrani et al., 2024), we ensured a balanced distribution of correct answer positions across four-choice MCQ datasets such as AraMath, ETEC, and AraPro (see Figure 9). For the few-shot experiments in IEN-MCQs, IEN-TF, and AraPro, we selected examples from the same domain as the target question to reduce the impact of out-of-domain samples. A complete list of models evaluated, including both open-source and closed-source models, is provided in Appendix C.

3.1.1 Open Models Setup

For open-source models, where weights are accessible, we evaluated performance using log-probability–based scoring on the MCQ datasets and reported normalized accuracy. We used labels (A, B, C, D, etc.) to compute log probabilities, with the exception of AraTruthfulQA, where we calculated the log-probability of the choice label followed by the context of the choices. AraIFEval was implemented as a generation task in the LM Evaluation Harness, and we report both prompt-level and instruction-level strict accuracies. Additional details on metrics are provided in Section I.

3.1.2 Closed Models Setup

To evaluate the closed-source models for the `AraEval` suite, we implemented a generation-based evaluation using the LM Evaluation Harness framework. Since closed models can only be accessed through APIs and do not provide token-level probabilities (logprobs), we adapted all benchmark tasks in `AraEval` to a generation-based format to suit such models. We set the generation temperature to 0.0 to ensure consistency and determinism in the model responses. For the multiple-choice tasks, such as the IEN datasets MCQ and TF, ETEC, AraMath, AraPro, and AraTruthfulQA, we applied filters that extract the model's selected answer from its generated response. Such filters ensure that the extracted response corresponds exactly to one of the provided answer choices. After processing the model outputs, accuracy was calculated by comparing the extracted responses to the gold-standard labels using an exact match criterion.

3.2 Baselines

We evaluate a range of Arabic and state-of-the-art multilingual models to assess the utility of our evaluation suite. To this end, we design a series of experiments that: (1) compare model performance across various tasks, analyzing fine-grained results across different domains, (2) examine knowledge retention across different model sizes within the same family, and (3) compare base and instruct (chat) models to assess their relative strengths. Our evaluation covers models shown in Table 8, considering variants with 7B, 13B, 30B, and 70+B parameters to study scaling trends and performance variations.

3.3 Results

Zero-shot results for instruct models are shown in Table 2, while zero-shot for base models and five-shot results for base and instruct models are presented in Table 3, Table 4, and Table 5, respectively, in Appendix A. All results in this section for open-source models are based on log-probability evaluation, except for AraIFEval. We report normalized accuracy for the tasks, and similar to Fourrier et al. (2024), we report strict prompt-level and instruction-level accuracy Zhou et al. (2023) for AraIFEval (see Section I for more details). We also evaluate open-source models using generation-based evaluation and provide a comparative analysis with log-probability evaluation in Section B.

The results reveal notable performance variations across models, model sizes, and shot settings. GPT4o, Claude, and Gemini demonstrate the highest performance across most tasks, consistently outperforming other models. Qwen 32B and 72B models and ALLaM 34B follow closely, showing robust performance across multiple tasks, especially in IEN MCQs and IEN TF. Llama 70B performs well but lags behind top-tier models, particularly in reasoning and advanced knowledge tasks including ETEC, AraPro, and AraMath, where its scores remain in the high 60s to low 70s. Among the Arabic models, these tasks remain challenging to Jais-family models where they underperform, while the AceGPT 32B model demonstrates improved performance; however, it falls short of achieving 70% accuracy.

The impact of model scaling varies across different types of tasks. For example, AraMath shows the most significant improvements with scaling, where Qwen 7B achieves an accuracy of 71.24% that increases to 92.07% with Qwen 32B. Similarly, Llama 3.3 70B achieves 69.92% compared to 32.73% with Llama 3.1 8B. Conversely, AraTruthfulQA do not exhibit the same level of improvement. For example, the Qwen models—7B, 14B, and 72B—achieve comparable accuracy rates of 52.8%, 58.4%, and 57.84%, respectively, while the Qwen 32B model outperforms them slightly with a higher accuracy of 61.19%.

The results highlight distinct patterns in task difficulty levels. Certain tasks, such as IEN MCQ and IEN TF, demonstrate consistently high accuracy across multiple models, suggesting a lower level of difficulty. This outcome is expected, as these tasks primarily consist of questions covering K01–K12 school subjects, which involve fundamental concepts and factual recall, making them easier for language models to handle. However, These two datasets are designed with multiple difficulty levels, as shown in Table 10 and Table 11, enabling the creation of more challenging subsets if needed. Other advanced knowledge and reasoning tasks, such as ETEC, AraPro, and AraMath, show a wider variance in scores, highlighting higher difficulty level. For ETEC, performance varies significantly across models, with Claude Sonnet 3.5 (86.06%) and Gemini Pro 1.5 (83.31%) achieving high scores, but Llama 8B is struggling at 45.89%. Similar trends are seen in AraMath and AraPro, where high variance is observed across models, with GPT4o achieving 81.16% and 80.86%, respectively, and Llama 8B scoring 32.73% and 52.51%, respectively. AraIFEval exhibit consistently low performance across all model families, indicating inherent difficulty. Even the strongest models achieve relatively low scores, compared to other tasks, with Claude sonnet 3.5 at 53.73%.

Most models benefit from few-shot prompting, but the degree of improvement varies. For instance, Qwen models show substantial improvements, particularly Qwen 7B, which gains over 10% in IEN MCQ, while Jais-family models struggle with few-shot prompting, with Jais-13B experiencing a performance drop in ETEC from 48.97% to 26.66%. Instruct models consistently outperform base models, particularly in AraMath, AraIFEval, and AraTruthfulQA. For example, Qwen 72B-Instruct scores 87.51% on AraIFEval, while its base counterpart achieves only 50.31%, highlighting the impact of instruction tuning on instruction following. Similarly, in AraTruthfulQA, ALLaM 34B Instruct scores 81.53%, whereas its base version achieves 64.18%, in five-shot setting showing that fine-tuning improves truthfulness and misinformation resistance. However, for simpler knowledge-based tasks like IEN MCQ, the gap is smaller. In some cases, base models outperform their instruct counterparts, as seen in IEN MCQ, where Qwen

72B Base scores 90.77%, surpassing the 86.54% of its instruct version. Few-shot prompting benefits base models more than instruct models, as seen in the AraMath task, where Qwen 72B improves from 88.60% (0-shot) to 95.87% (5-shot). Overall, instruction tuning significantly enhances reasoning, alignment, and reliability, while larger base models still perform well in factual retrieval.

4 Analysis

4.1 Cross-Models Analysis

AraEval aggregates 7 datasets into a single score representing general Arabic capabilities. Inspired by Fourrier et al. (2024), we take the average normalized score across benchmarks, which is defined as:

$$\text{Norm. Score} = 100 \cdot \frac{\text{Raw Score} - \text{Baseline}}{100 - \text{Baseline}} \quad (1)$$

This transformation assigns a normalized score of 0% for the random baseline and 100% for a perfect score, with the rest linearly interpolated. In effect, this unifies score variances across benchmarks; it increases the contribution of benchmarks with high random baselines, such as true/false benchmarks, such that their scores span $[0, 100]$ instead of $[50, 100]$. The final score is the mean of the 7 normalized benchmark scores. Five-shot evaluation is used whenever applicable to decouple formatting from base model evaluation.

Figure 2 illustrates the relationship between model size and AraEval accuracy for several prominent model families, including Qwen 2.5, Llama 3, Jais Family, AceGPT v2, Fanar, ALLaM, and ALLaM Adapted. Across all model families, there is a consistent trend of increasing accuracy as model size scales from 7B to 70B parameters. This suggests that larger models are better equipped to capture the complexities of the Arabic language, benefiting from richer parameterization. While all models demonstrated performance gains with increased size, ALLaM Base exhibited the most significant improvements, particularly in the small-to-mid size range (7B–30B), indicating the effectiveness of its architecture and training data for Arabic-specific tasks. The sensitivity of AraEval to variations in model scale—from 7B to 70B parameters—further highlights the benchmark's robustness. It effectively captures nuanced performance differences, making it particularly well-suited for fine-grained comparisons across diverse model configurations.

Although performance generally improved with size, diminishing returns became apparent beyond the 30B parameter mark for Qwen2.5 and for ALLaM instruct scaling from 7B to 30B. For these models, the accuracy gains were marginal compared to the more substantial improvements observed when scaling from 7B to 30B in ALLaM base. This suggests potential saturation points where further parameter increases yield limited benefits. This ability to detect performance plateaus is critical for guiding model scaling decisions and optimizing resource allocation.

Instruct models consistently outperform their Base counterparts across all size categories, underscoring the benchmark's ability to reflect improvements from fine-tuning strategies aimed at aligning models with user instructions.

4.2 Fine-Grained Analysis

While average evaluation metrics provide a general overview of LLMs performance, fine-grained assessments offer deeper insights into specific capabilities and areas needing improvement. This detailed evaluation is crucial for understanding the strengths and weaknesses of LLMs in various contexts. Several approaches were proposed to reveal the fine-graind capabilities of models. FAC^{2}E (Wang et al., 2024a) proposed a framework for better understanding LLM capabilities by dissociating Language and Congitive capabilities allowing for a more detailed analysis of LLM performance. Similarly, the "FLASK" (Ye et al., 2024) evaluation protocol decomposes overall scoring into specific skill sets for each instruction, providing a fine-grained evaluation that enhances interpretability and reliability. To this extent, AraEval benchmark offers a deeper insight into the capabilities of LLMs by pinpointing model scoring not only at an overall view but more deeper such as grade, subject, and difficulty level, See Figure 3, 4 and 6. The variations in the figures indicate that the models performances varies and provide insightful remarks about how each model performs when compared to others, and at the same time will identify the gap or the deficiencies the model might suffer from. In Figure 3(b), it is noticeable that “Above average” questions have more variance between the models compared to “Average” or “Below average” questions. Table 6 reports the performance of all instruct models on the IEN datasets, broken down by difficulty level.

Further, subjects like “Language” in Figure 3(c),

Model	IEN		AraPro	AraMath	ETEC	AraTruthfulQA	AraIFEval	
	MCQ	TF					Prompt	Instruction
ALLaM 7B-Instruct	93.07	83.17	73.57	67.44	69.26	65.30	61.94	83.75
Llama-3.1-8B-Instruct	59.24	72.83	52.51	32.73	45.89	54.29	53.36	79.32
Qwen2.5-7B-Instruct	66.20	78.79	64.63	71.24	64.23	52.8	28.17	65.19
Fanar-1-9B-Instruct	79.51	79.70	66.63	59.67	62.53	79.48	44.59	77.13
ALLaM Adapted 13B-Instruct	**93.39**	84.23	74.69	78.68	73.98	67.16	59.33	83.14
Jais-family-13B-chat	63.01	69.05	57.53	42.64	48.97	56.53	17.16	54.27
Qwen2.5-14B-Instruct	80.32	77.25	69.11	80.17	72.66	58.4	68.66	86.76
ALLaM 34B-Instruct	93.21	87.19	79.52	60.50	74.46	78.36	67.16	86.76
AceGPT-v2-32B-chat	81.61	80.92	67.19	64.13	65.08	65.11	25.75	63.41
Jais-family-30B-16k-chat	74.70	68.62	62.79	50.74	53.37	63.99	16.60	54.95
Jais-family-30B-8k-chat	72.78	70.62	61.27	42.64	53.63	62.69	16.79	54.68
Qwen2.5-32B-Instruct	84.71	81.97	71.81	92.07	78.91	61.19	56.90	82.87
ALLaM Adapted 70B-Instruct	92.43	85.88	75.82	73.22	76.26	81.72	65.49	85.39
Jais-adapted-70B-chat	74.41	76.85	64.59	50.74	56.76	71.46	27.05	65.05
Llama-3.3-70B-Instruct	79.68	78.50	70.49	69.92	69.00	67.16	70.90	88.60
Qwen2.5-72B-Instruct	86.90	87.12	74.69	89.26	78.96	57.84	67.72	87.51
GPT-4o	92.07	**89.87**	80.86	81.16	79.23	87.69	70.90	88.12
Gemini pro 1.5	89.33	85.73	76.22	**96.36**	83.31	88.43	**74.81**	**90.17**
Claude Sonnet 3.5	92.45	89.64	**81.46**	88.6	**86.06**	**90.67**	53.73	80.14
Random baseline	30.77	50	25	25	25	23.46	0	0

Table 2: Zero-shot results of instruct models on all `AraEval` benchmarks.

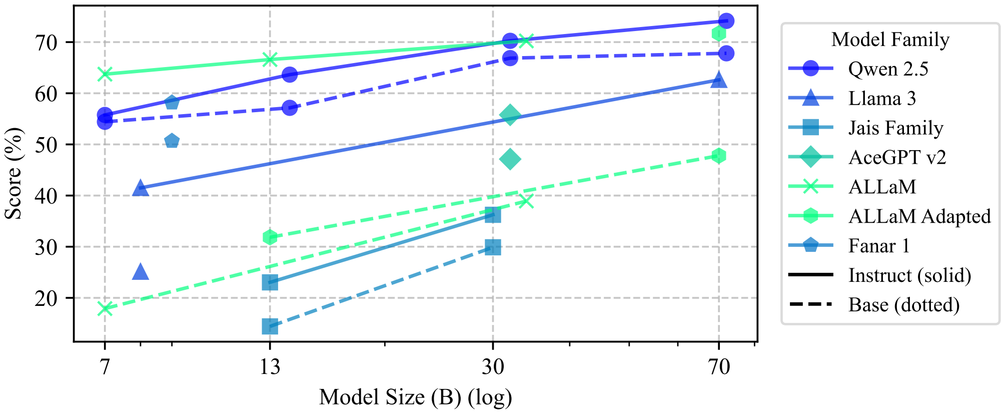


Figure 2: LLMs performance on `AraEval` for various model sizes. Instruct models are in solid lines, while Base models are in dashed lines.

and "Humanities" in Figure 4 show similar trends where the performance of the models varies widely. Such nuances and observations are useful and insightful and reflect the utility of a high-quality benchmark.

4.3 Vocabulary Coverage Analysis

A robust evaluation of large language models in Arabic requires not only challenging tasks, but also a comprehensive vocabulary coverage. In this work, we assess the vocabulary coverage of several models across the Arabic datasets within our proposed benchmark `AraEval`, and compare it against two widely used benchmarks in the community, including Arabic MMLU (Koto et al., 2024) and OpenAI MMMLU (translated to Arabic) (OpenAI, 2024).

As shown in Figure 5, the vocabulary coverage values are averaged across all models. `AraEval` achieves 74.05% coverage of Arabic tokens, closely aligning with OpenAI Arabic MMMLU (74.17%), while surpassing Arabic MMLU (66.38%). This coverage ensures that `AraEval` incorporates a diverse range of Arabic tokens, including domain-specific tokens from science, history, and literature.

This rich token representation makes `AraEval` a more faithful and challenging benchmark for evaluating LLM performance in Arabic. A detailed

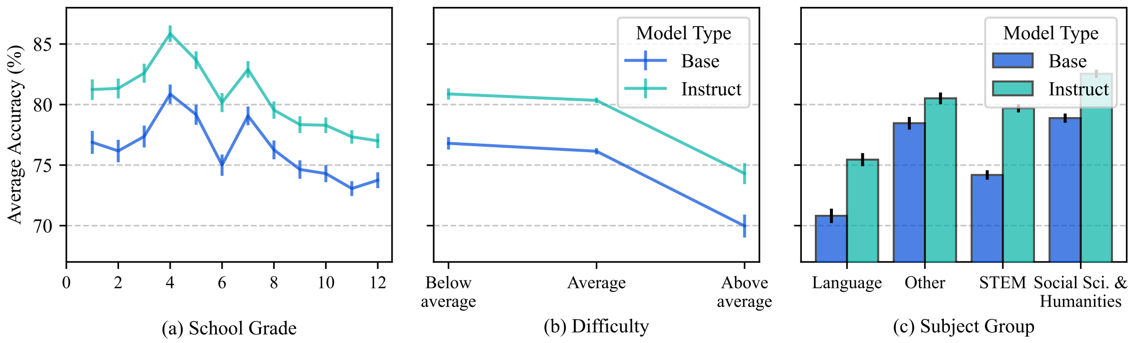


Figure 3: Average accuracies on all evaluated models for various IEN MCQ subsets. Error bars represent 95% confidence intervals of the average accuracy across all models.

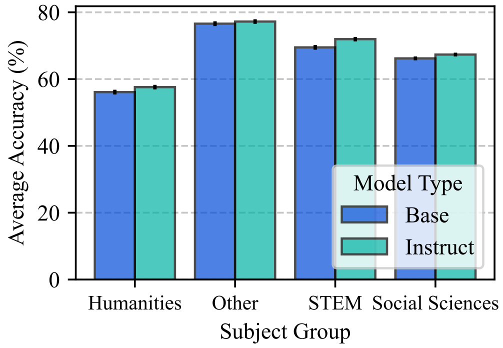


Figure 4: Average accuracies on all evaluated models for various AraPro subsets. Error bars represent 95% confidence intervals for the average across models.

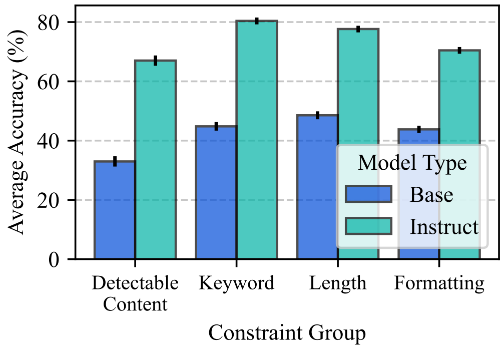


Figure 6: Average accuracies on all evaluated models for various AraIFEval constraint subsets. Error bars represent 95% confidence intervals across models.

breakdown of the vocabulary coverage is provided in Table 18.

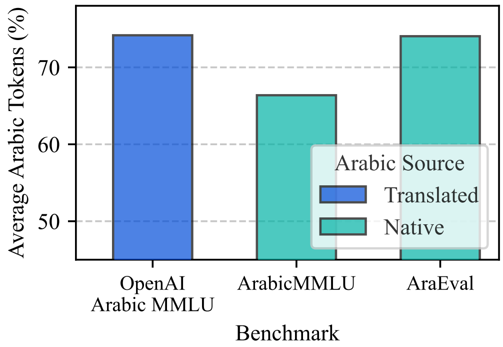


Figure 5: Average Arabic vocabulary coverage across various tokenizers. Details are presented in Table 18. AraEval covers a large portion of Arabic vocabulary without using translated data.

5 Conclusion

In this paper, we introduced AraEval, a comprehensive benchmark designed to evaluate different advanced capabilities of foundation models within the Arabic context. Our evaluation highlights the robustness and diversity of the datasets within AraEval, offering key insights into their effectiveness in distinguishing model capabilities. Tasks like AraMath, AraPro, ETEC, and AraIFEval prove highly challenging, effectively differentiating models, making them strong indicators of true model competency. AraTruthfulQA effectively measures a model's susceptibility to misinformation, revealing clear differences in truthfulness across models. Conversely, IEN MCQ and IEN TF capture less advanced knowledge that some base models can handle. These findings emphasize the value of AraEval as a benchmarking tool for Arabic LLMs. We release the main results using log-probability scoring due to its efficiency and replicability, while also providing generation-based evaluation results as they better reflect end-user expectations in real-world applications. By releasing AraEval, we aim to support further research into advanced Arabic prompting strategies and provide a strong foundation for future evaluations, paving the way for more targeted advancements in Arabic NLP.

6 Limitations

Despite AraEval's contribution to addressing the gap in comprehensive assessment datasets, several limitations warrant consideration. First, the dataset's reliance on multiple-choice questions (MCQ) and true/false formats inherently constrains the evaluation of language models' capabilities. These structured response formats may not adequately assess deeper levels of comprehension or the ability to generate creative solutions that more closely align with real-world applications.

Second, some AraEval's datasets, mainly IEN and ETEC, focus on the Saudi curriculum which may introduce potential cultural bias. This geographical and cultural specificity may limit the generalizability of the dataset to educational contexts in other regions and cultures, potentially overlooking important cultural nuances and educational approaches from diverse Arab educational systems.

Third, the current benchmark's scope is limited to text-based assessments, excluding evaluation capabilities for multi-modal models. This limitation becomes particularly significant as artificial intelligence increasingly requires the ability to process and synthesize information across various modalities, including visual, auditory, and textual data. However, some recent work has been conducted to address this issue (Das et al., 2024; Ghaboura et al., 2024).

Fourth, our dataset curation process emphasized Arabic cultural alignment. However, Arabic is a pluricentric language that spans many regions and subcultures. We attempted to collect and filter data in such a manner that conforms to the majority of Arabic communities. However, we acknowledge that the annotators and datasets are sourced predominately from Saudi Arabia, which could induce Saudi biases.

These limitations suggest opportunities for future work to develop more comprehensive evaluation frameworks that incorporate open-ended responses, diverse cultural perspectives, and multimodal assessment capabilities.

7 Ethical Considerations

All authors of this work acknowledge and adhere to the ACL Code of Ethics, upholding its principles throughout the research process. All domain experts and annotators involved in the creation and review of the datasets are official employees, who are fairly compensated based on mutually agreed-upon wage standards and working hours. These employment agreements fully comply with local labor regulations. Furthermore, we prioritize clear communication about how data and annotations are utilized, obtaining informed consent from domain experts and annotators before incorporating their contributions into our research. We are also dedicated to safeguarding their privacy throughout the annotation and data creation process, fostering an ethical and respectful research environment.

Acknowledgments

We would like to thank the Ministry of Education for their support with the IEN dataset and the Education & Training Evaluation Commission (ETEC) for their support with the ETEC dataset. We are also grateful to Nora Alturayeif, Amjad Felemban, Eyad Alqayisi, Atheer Almusallam, Amjad Alsaeed, Amal Almohammadi, Badr Almalki, Alhanouf Alotaibi, Khulud Al Awadh, Maha Awaji, Maryam Alshuwayman, Lama Rajeh, and Yaser Alharbi for their valuable contributions to the development of this work.

References

Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, et al. 2025. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

Saja AL-Tawalbeh and Mohammad Al-Smadi. 2020. A benchmark arabic dataset for commonsense explanation. *ArXiv*, abs/2012.10251.

Reem Alghamdi, Zhenwen Liang, and Xiangliang Zhang. 2022. Armath: a dataset for solving arabic math word problems. In *Proceedings of the Language Resources and Evaluation Conference*, pages 351–362, Marseille, France. European Language Resources Association.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, Julien Launay, and Badreddine Noune. 2023. AlGhafa evaluation benchmark for Arabic language models. In *Proceedings of ArabicNLP 2023*, pages 244–275, Singapore (Hybrid). Association for Computational Linguistics.

AlphaCode Team. 2023. Alphacode 2 technical report. *Google DeepMind*.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets:

Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

Anthropic. 2022. The claude 3 model family: Opus, sonnet, haiku.

AI Anthropic. 2024. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 3:6.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

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 *The Thirteenth International Conference on Learning Representations*.

Yuri Chervonyi, Trieu H Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, Junehyuk Jung, Vikas Verma, Quoc V Le, and Thang Luong. 2025. Gold-medalist performance in solving olympiad geometry with alphageometry2. *arXiv preprint arXiv:2502.03544*.

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. *Preprint*, arXiv:2110.14168.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2020. *Ethnologue: Languages of the World*, 23 edition. SIL International, Dallas.

Ali El Filali, Hamza Alobeidli, Clémentine Fourrier, Basma El Amel Boussaha, Ruxandra Cojocaru, Nathan Habib, and Hakim Hacid. 2024. Open arabic llm leaderboard. https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard.

Aaron Grattafiori et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

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. A framework for few-shot language model evaluation.

Team Gemini. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Sara Ghaboura, Ahmed Heakl, Omkar Thawakar, Ali Alharthi, Ines Riahi, Abduljalil Saif, Jorma Laaksonen, Fahad S Khan, Salman Khan, and Rao M Anwer. 2024. Camel-bench: A comprehensive arabic lmm benchmark. *arXiv preprint arXiv:2410.18976*.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*.

Paul Gómez-Rodríguez, Carlos Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Google. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Google. 2025. Google gemini: Deep research. https://blog.google/products/gemini/google-gemini-deep-research/. Accessed: 2025-02-11.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. 2025. Which economic tasks are performed with ai? evidence from millions of claude conversations. *Anthropic*.

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. *Preprint*, arXiv:2009.03300.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, et al. 2023. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*.

Aaron Hurst and et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Inception. 2024. Jais family model card. *Hugging Face*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Mohammad Abdullah Matin Khan, M Saiful Bari, Do Long, Weishi Wang, Md Rizwan Parvez, and Shafiq Joty. 2024. XCodeEval: An execution-based large scale multilingual multitask benchmark for code understanding, generation, translation and retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6766–6805, Bangkok, Thailand. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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

LAION-AI. 2025. Open-assistant: A chat-based assistant for task understanding and dynamic interaction. https://github.com/LAION-AI/Open-Assistant. Accessed: 2025-02-12.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Latent Space. 2024. In the arena: How lmsys changed llm benchmarking forever. https://www.latent.space/p/lmarena. Blog post (accessed: 11 February 2025).

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li, Benyou Wang, and Jinchao Xu. 2024. Alignment at pre-training! towards native alignment for arabic LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Ngee Chia Tai, Ayu

Purwarianti, Sebastian Ruder, William Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng-Xin Yong, and Samuel Cahyawijaya. 2024. Seacrowd: A multilingual multimodal data hub and benchmark suite for southeast asian languages. *arXiv preprint arXiv: 2406.10118*.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2024. Starcoder 2 and the stack v2: The next generation. *Preprint*, arXiv:2402.19173.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*.

Mistral. 2024. Au large.

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

Ahmad Mustapha, Hadi Al-Khansa, Hadi Al-Mubasher, Aya Mourad, Ranam Hamoud, Hasan El-Husseini, Marwah Al-Sakkaf, and Mariette Awad. 2024. Arastem: A native arabic multiple choice question benchmark for evaluating llms knowledge in stem subjects. *Preprint*, arXiv:2501.00559.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2024. Multilingual massive multitask language understanding (mmmlu).

OpenAI. 2025. Introducing deep research. https://openai.com/index/introducing-deep-research/. Accessed: 2025-02-11.

OpenAI. 2025. simple-evals. https://github.com/openai/simple-evals. Accessed: 2025-02-11.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, et al. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Zhaozhi Qian, Faroq Altam, Muhammad Alqurishi, and Riad Souissi. 2024. Cameleval: Advancing culturally aligned arabic language models and benchmarks. *Preprint*, arXiv:2409.12623.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14182–14214, Bangkok, Thailand. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023a. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023b. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh*

International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *Preprint*, arXiv:2402.11548.

Aarohi Srivastava and et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.*, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Reka Team, Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, et al. 2024. Reka core, flash, and edge: A series of powerful multimodal language models. *arXiv preprint arXiv:2404.12387*.

Xiaoqiang Wang, Lingfei Wu, Tengfei Ma, and Bang Liu. 2024a. FAC^{2}E: Better understanding large language model capabilities by dissociating language and cognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13228–13243, Miami, Florida, USA. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Zeyu Wang. 2024. CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. Quick and (not so) dirty: Unsupervised selection of justification sentences for multi-hop question answering. 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 2578–2589, Hong Kong, China. Association for Computational Linguistics.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. Flask: Fine-grained language model evaluation based on alignment skill sets. *Preprint*, arXiv:2307.10928.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schuetze. 2024. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055, Miami, Florida, USA. Association for Computational Linguistics.

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.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. Available: https://github.com/google-research/google-research/tree/master/instruction_following_eval.

Jianqing Zhu, Huang Huang, Zhihang Lin, Juhao Liang, Zhengyang Tang, Khalid Almubarak, Mosen Alharthi, Bang An, Juncai He, Xiangbo Wu, Fei Yu, Junying Chen, Zhuoheng Ma, Yuhao Du, Yan Hu, He Zhang, Emad A. Alghamdi, Lian Zhang, Ruoyu Sun, Haizhou Li, Benyou Wang, and Jinchao Xu. 2024. Second language (arabic) acquisition of llms via progressive vocabulary expansion. *arXiv preprint arXiv:2412.12310*.

A Additional Results

In addition to zero-shot results for instruct models in Table 2 in the main paper, we also show the five-shot instruct models results in Table 5. Also, the base models results in zero-shot and five-shot settings are presented in Tables 3 and 4, respectively. Further, the results of all instruct models across difficulty level for the IEN datasets are presented in Table 6.

B Generation-based Evaluation

We present a comparative analysis of open-source models performance on the AraEval datasets using a generation-based zero-shot setting (shown in Table 7) and contrasting the results with those obtained under zero-shot log-probability scoring in Table 2.

The ALLaM model family exhibits strong robustness and stability across both log-probability and generation-based evaluation paradigms. In contrast, the Qwen models show marked improvement under generation-based evaluation—particularly on the AraTruthful dataset—highlighting their strength in open-ended generation tasks. Conversely, the Jais-family models consistently underperform in the generation setting, suggesting potential limitations in alignment, reasoning capabilities, or instruction following.

At the dataset level, IEN MCQ and AraPro show strong agreement between log-probability and generation-based evaluations, with most models retaining similar rankings, suggesting these datasets are less sensitive to prompting variations. AraMath and ETEC display moderate shifts in model performance rankings, indicating some influence of evaluation paradigm. In contrast, AraTruthfulQA shows the most pronounced divergence, where models like Qwen see notable gains under generation, reflecting its sensitivity to open-ended reasoning and alignment capabilities.

These findings emphasize the importance of evaluation choice when assessing Arabic language models. While log-probability scoring provides efficiency and replicability, generation-based evaluation better captures end-user expectations in real-world applications.

C Evaluated Models

Table 8 outlines the LLMs used in our evaluation with additional details.

Size	Model	Access
7B	Qwen 2.5 (Qwen et al., 2025)	weights
8B	Llama 3.1 (et al., 2024)	weights
9B	Fanar I (Abbas et al., 2025)	weights
7B	ALLaM (Bari et al., 2025)	weights
14B	Qwen 2.5	weights
13B	Jais family 13b chat (Sengupta et al., 2023a; Inception, 2024)	weights
13B	ALLaM Adapted	weights
32B	Qwen 2.5	weights
30B	Jais family 30b 8k-chat	weights
30B	Jais family 30b 16k-chat	weights
32B	AceGPT (Zhu et al., 2024; Liang et al., 2024)	weights
34B	ALLaM	weights
72B	Qwen 2.5	weights
70B	Llama 3.3	weights
70B	Jais-adapted 70b-chat	weights
70B	ALLaM Adapted	weights
—	GPT4o (Hurst and et al., 2024)	API
—	Gemini pro 1.5 (Gemini, 2024)	API
—	Claude 3.5 Sonnet (Anthropic, 2024)	API

Table 8: Instruct models considered

D Related Work

Evaluating LLMs requires comprehensive benchmark datasets that assess knowledge, reasoning, and language understanding. These datasets can be categorized into general-purpose and domain-specific types, ensuring models are both broadly competent and specialized.

D.1 General-Purpose Datasets

General-purpose datasets evaluate a model's versatility across tasks like question-answering, translation, and commonsense reasoning. The Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2021) measures general knowledge across 57 subjects, with adaptations for languages such as Korean (KMMLU) (Son et al., 2024), Turkish (TurkishMMLU) (Yüksel et al., 2024), and Chinese (CMMLU) (Li et al., 2024). OpenAI has also translated MMLU into 14 languages, including Arabic (OpenAI, 2024).

HellaSwag (Zellers et al., 2019) evaluates commonsense reasoning through multiple-choice questions, with multilingual extensions like XCOPA (Ponti et al., 2020) and mCSQA (Sakai et al., 2024). Grade School Math 8K (GSM8K) (Cobbe et al., 2021) focuses on quantitative reasoning, extended to ten languages via MGSM (Shi et al., 2023). Finally, BigBench (Srivastava and et al., 2023) offers over 200 diverse tasks to test LLM capabilities across various domains.

Model	IEN		AraPro	AraMath	ETEC	AraTruthfulQA	AraIFEval	
	MCQ	TF					Prompt	Instruction
ALLaM 7B Base	58.79	57.15	49.41	20.33	39.27	44.78	3.73	29.56
Llama-3.1-8B	64.36	54.63	51.07	26.61	42.82	54.29	7.28	41.50
Qwen2.5-7B	76.95	78.00	61.75	67.93	59.94	71.08	6.72	44.57
Fanar-1-9B	80.16	74.84	64.89	49.75	58.24	58.21	15.30	53.52
ALLaM Adapted 13B Base	63.36	67.18	54.85	23.14	40.75	50	6.53	38.50
Jais-family-13B	37.99	54.82	31.15	31.90	28.46	50	6.90	40.75
Qwen2.5-14B	83.55	68.38	68.45	79.17	70.06	66.98	10.82	47.78
ALLaM 34B Base	83.42	55.90	72.71	48.10	62.43	53.54	**17.16**	**55.15**
AceGPT-v2-32B	78.40	66.96	65.85	54.71	58.88	63.81	8.02	45.26
Jais-family-30B-16k	67.00	55.71	54.29	28.10	42.24	48.88	11.01	45.12
Jais-family-30B-8k	58.65	62.22	55.21	26.12	42.82	48.13	11.57	48.74
Qwen2.5-32B	84.99	**82.38**	71.43	81.82	76.15	73.13	11.75	46.35
ALLaM Adapted 70B Base	75.76	76.13	64.19	35.54	55.11	59.33	3.17	24.30
Jais-adapted-70B	70.27	61.48	61.79	37.69	44.78	61.19	9.89	43.21
Qwen2.5-72B	**88.83**	80.73	**73.89**	**88.60**	**78.48**	**78.73**	14.93	50.31
Random baseline	30.77	50	25	25	25	23.46	0	0

Table 3: Zero-shot results of base models on all AraEval benchmarks.

Model	IEN		AraPro	AraMath	ETEC	AraTruthfulQA
	MCQ	TF				
ALLaM 7B Base	63.63	64.93	55.77	18.02	43.46	43.28
Llama-3.1-8B	71.05	63.85	59.29	39.67	48.01	51.49
Qwen2.5-7B	81.52	79.65	66.55	75.70	65.50	75.75
Fanar-1-9B	82.57	79.87	67.59	61.16	63.06	73.13
ALLaM Adapted 13B Base	72.43	71.29	62.93	23.47	51.19	59.70
Jais-family-13B	32.11	59.54	40.35	26.45	33.28	42.35
Qwen2.5-14B	86.56	83.75	72.53	92.56	76.21	83.96
ALLaM 34B Base	86.22	81.93	77.16	51.74	65.45	64.18
AceGPT-v2-32B	82.95	80.82	70.11	66.45	66.30	72.95
Jais-family-30B-16k	72.78	70.60	65.09	35.87	51.99	53.36
Jais-family-30B-8k	71.36	69.23	63.05	32.23	51.35	52.24
Qwen2.5-32B	87.84	**86.16**	74.99	94.05	80.29	82.28
ALLaM Adapted 70B Base	83.01	77.69	72.45	48.26	63.38	79.48
Jais-adapted-70B	78.07	75.17	66.97	51.24	52.46	77.24
Qwen2.5-72B	**90.77**	85.80	**77.86**	**95.87**	**82.88**	**84.33**
Random baseline	30.77	50	25	25	25	23.46

Table 4: Five-shot results of base models on all AraEval benchmarks

D.2 Domain-Specific Datasets

Domain-specific datasets evaluate LLMs in specialized fields. ARC-Challenge (Yadav et al., 2019) tests science reasoning, with Arabic versions like Okapi ARC-Challenge (Lai et al., 2023) and Al-Ghafa Evaluation Benchmark (Almazrouei et al., 2023). Minerva Math (Lewkowycz et al., 2022) assesses mathematical reasoning, while CausalBench (Wang, 2024) evaluates causal inference across textual, mathematical, and coding domains. Multi-MedQA (Singhal et al., 2023) combines six medical datasets to evaluate clinical knowledge, making it essential for healthcare-related tasks.

D.3 Arabic Datasets

Few datasets have been explicitly developed to evaluate LLMs in Arabic, but recent efforts have made significant progress. One notable example is ArabicMMLU (Koto et al., 2024), a comprehensive multiple-choice question benchmark designed to assess reasoning and knowledge capabilities of LLMs in Modern Standard Arabic. Developed with input from native speakers across North Africa, the Levant, and the Gulf, it includes 14,575 questions spanning 40 diverse tasks. These tasks cover subjects such as STEM, social sciences, humanities, and the Arabic language, sourced from educational ma-

Model	IEN		AraPro	AraMath	ETEC	AraTruthfulQA
	MCQ	TF				
ALLaM 7B-Instruct	92.28	83.96	72.13	71.57	67.25	67.72
Llama-3.1-8B-Instruct	64.96	60.78	57.45	35.70	47.59	58.58
Qwen2.5-7B-Instruct	78.10	78.34	65.97	71.74	65.24	69.96
Fanar-1-9B-Instruct	81.87	80.23	68.33	54.71	65.29	81.53
ALLaM Adapted 13B-Instruct	92.64	83.69	74.93	75.04	73.77	70.34
Jais-family-13B-chat	53.65	59.75	32.99	26.61	26.66	48.69
Qwen2.5-14B-Instruct	80.73	80.10	71.31	82.81	73.82	70.34
ALLaM 34B-Instruct	**92.84**	87.50	80.70	62.81	74.09	81.53
AceGPT-v2-32B-chat	82.93	72.78	68.23	64.46	66.19	67.54
Jais-family-30B-16k-chat	71.11	63.63	62.57	41.49	48.75	61.75
Jais-family-30B-8k-chat	67.31	72.59	60.61	33.39	45.20	59.7
Qwen2.5-32B-Instruct	84.45	82.45	73.45	91.90	78.11	76.12
ALLaM Adapted 70B-Instruct	92.18	85.59	76.74	74.88	75.73	84.14
Jais-adapted-70B-chat	77.33	76.66	68.23	45.62	57.82	77.43
Llama-3.3-70B-Instruct	80.90	79.79	72.53	70.91	68.20	70.71
Qwen2.5-72B-Instruct	86.54	86.79	75.66	92.89	79.33	71.27
GPT-4o	91.43	89.63	81.46	83.47	79.92	90.11
Gemini pro 1.5	85.67	87.21	78.28	**94.88**	84.42	84.14
Claude Sonnet 3.5	92.64	**90.74**	**83.96**	79.83	**86.96**	**93.47**
Random baseline	30.77	50	25	25	25	23.46

Table 5: Five-shot results of instruct models on all AraEval benchmarks.

Model	IEN MCQ			IEN TF		
	Below Avg.	Avg.	Above Avg.	Below Avg.	Avg.	Above Avg.
ALLaM-7B-Instruct	92.97	93.47	88.79	82.87	83.53	79.10
Llama-3.1-8B-Instruct	60.91	59.39	52.84	74.26	72.61	69.03
Qwen2.5-7B-Instruct	67.23	66.72	57.30	80.51	78.28	77.99
Fanar-1-9B-Instruct	80.43	79.89	72.50	80.88	79.50	76.87
ALLaM Adapted 13B-Instruct	**94.33**	93.47	89.86	85.29	84.27	78.36
Jais-family-13B-chat	64.34	63.33	55.61	68.97	68.92	71.64
Qwen2.5-14B-Instruct	80.59	80.81	73.73	79.49	76.62	75.75
ALLaM-34B-Instruct	93.84	**93.48**	88.33	87.79	87.39	80.97
AceGPT-v2-32B-Chat	83.59	81.81	73.73	83.38	80.50	75.00
Jais-family-30B-16k-chat	76.50	74.99	66.36	68.01	68.82	68.66
Jais-family-30B-8k-chat	74.43	72.83	67.59	70.00	70.70	72.39
Qwen2.5-32B-Instruct	85.93	84.84	79.72	84.26	81.48	77.99
ALLaM Adapted 70B-Instruct	93.84	92.45	88.33	87.06	85.86	80.22
Jais-adapted-70B-chat	75.57	74.56	69.43	77.28	76.92	73.51
Llama-3.3-70B-Instruct	80.15	79.91	75.73	79.93	77.90	80.60
Qwen2.5-72B-Instruct	87.90	87.17	80.95	88.16	87.13	81.72
GPT-4o	92.53	92.22	89.09	**90.66**	88.80	**83.21**
Gemini pro 1.5	90.46	89.35	85.87	86.76	85.20	82.46
Claude Sonnet 3.5	92.69	92.60	**90.02**	**90.66**	**89.51**	81.72

Table 6: Zero-shot results of instruct models on IEN MCQ and IEN TF datasets across difficulty levels.

terials in various Arabic-speaking countries. The dataset reflects a range of educational levels.

Another important contribution is AraSTEM (Mustapha et al., 2024), which focuses on STEM subjects like mathematics, physics, chemistry, biology, computer science, and medicine. This dataset comprises multiple-choice questions sourced from elementary, secondary, and higher education levels, ensuring broad coverage of difficulty and topics. It was carefully compiled from multiple internet sources to ensure diversity and comprehensiveness.

Efforts to adapt existing English evaluation datasets for Arabic include the AlGhafa Arabic LLM Benchmark (Almazrouei et al., 2023). This benchmark consists of 11 datasets translated or modified from English benchmarks, verified by

Model	IEN		AraPro	AraMath	ETEC	AraTruthfulQA
	MCQ	TF				
ALLaM 7B-Instruct	92.96	82.91	73.61	67.93	69.21	65.3
Llama-3.1-8B-Instruct	71.62	72.78	58.97	35.54	52.68	66.23
Qwen2.5-7B-Instruct	78.57	78.67	64.99	71.4	64.65	82.84
Fanar-1-9B-Instruct	75.27	80.2	66.71	56.86	61.05	77.61
ALLaM Adapted 13B-Instruct	**93.47**	83.63	74.83	79.01	74.14	66.98
Jais-family-13b-chat	55.52	55.42	57.71	44.13	45.95	55.41
Qwen2.5-14B-Instruct	82.06	78.04	69.37	79.17	72.71	84.51
ALLaM 34B-Instruct	93.24	86.18	79.56	60.5	74.72	77.8
AceGPT-v2-32B-Chat	82.66	62.56	68.25	69.75	67.62	82.46
Jais-family-30b-16k-chat	59.26	39.91	54.75	46.78	29.15	67.54
Jais-family-30b-8k-chat	55.66	41.23	50.09	34.38	23.26	62.5
Qwen2.5-32B-Instruct	84.27	82.74	71.89	91.24	78.48	83.58
ALLaM Adapted 70B-Instruct	92.51	85.63	75.82	73.22	76.47	82.28
Jais-adapted-70b-chat	59.62	48.12	61.31	44.96	25.65	75.56
Llama-3.3-70B-Instruct	82.74	80.18	72.89	71.07	71.65	77.8
Qwen2.5-72B-Instruct	88.29	86.97	74.45	89.09	79.12	86.94
GPT-4o	92.07	**89.87**	80.86	81.16	79.23	87.69
Gemini pro 1.5	89.33	85.73	76.22	**96.36**	83.31	88.43
Claude Sonnet 3.5	92.45	89.64	**81.46**	88.6	**86.06**	**90.67**
Random baseline	30.77	50	25	25	25	23.46

Table 7: Zero-shot results of instruct models on AraEval benchmarks using generation-based setting.

native Arabic speakers. Similarly, the Benchmark Arabic Dataset for Commonsense Explanation (AL-Tawalbeh and Al-Smadi, 2020) translates the original English ComVE task into Arabic. It contains 12,000 instances, each presenting an Arabic sentence that defies commonsense, accompanied by three explanatory options. The task is to identify the best explanation for why the sentence is nonsensical.

Qian et al. (2024) introduced CamelEval, a suite of three test sets designed to evaluate general instruction following, factuality, and cultural alignment in Arabic. Each test set includes 805 carefully curated cases reflecting the nuances of the Arabic language and culture.

While these datasets significantly advance the evaluation of Arabic LLMs, they also exhibit certain limitations. For instance, ArabicMMLU and AraSTEM may not fully capture the diversity of educational systems, cultural nuances, and historical contexts across Arabic-speaking countries. Despite sourcing questions from multiple regions, ArabicMMLU might struggle to encompass the full spectrum of curricula and perspectives in the Arab world. Similarly, AraSTEM, while focusing on STEM subjects, may not adequately represent the varied educational strategies and cultural contexts found in different Arabic-speaking nations.

Additionally, translating English datasets into Arabic, such as in the case of AlGhafa and the Benchmark Arabic Dataset for Commonsense Explanation, presents challenges. Translations may fail to preserve cultural nuances and contextual meanings inherent in the original language, leading to potential misinterpretations. Furthermore, these datasets may not align well with the educational curricula and cultural contexts of Arabic-speaking countries, where educational systems and cultural norms vary significantly. This misalignment can result in evaluations that do not accurately reflect the capabilities of Arabic-centric LLMs in real-world applications.

E AraEval Datasets

In this section, we detail each dataset used in AraEval, including fine-grained analyses, task statistics, and example samples.

E.1 Domain and Subject Distribution

Table 10 and Table 11 show distribution for both IEN MCQ and IEN TF, respectively, in terms of study stage, difficulty level, and subjects.

AraPro subjects distribution is presented in Table 12 and category distribution in Figure 7. For AraIFEval, we show the distribution of constraint groups in Figure 8, while Table 9 shows the distribution of instructions, where each sample comprises multiple instructions.

Category	Count	Percent (%)
number words at least	265	18.09
number paragraphs	225	15.36
response language	139	9.49
title	135	9.22
keyword frequency	135	9.22
number words at most	87	5.94
include keywords	63	4.30
forbidden words	60	4.10
number bullets	48	3.28
letter frequency	46	3.14
postscript	34	2.32
first word in i-th paragraph	33	2.25
check end	27	1.84
number sentences at least	25	1.71
minimum number highlighted section	22	1.50
json format	21	1.43
multiple sections	20	1.37
quotation	20	1.37
number placeholder	14	0.96
repeat prompt	13	0.89
two responses	12	0.82
number sentences at most	11	0.75
no commas	10	0.68
Total	**1465**	**–**

Table 9: Category distribution and percentage of AraIFEval dataset.

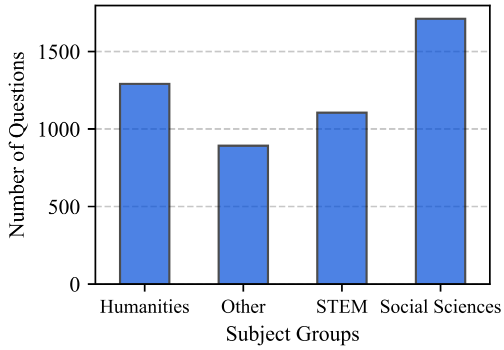


Figure 7: Subject distribution of AraPro.

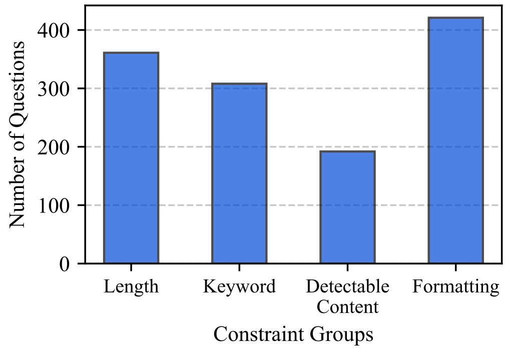


Figure 8: Constraint distribution of AraIFEval.

Category	#Subject/Specialty	#Questions
In terms of study stages		
Secondary education	17	3747
Primary education	10	3739
Intermediate education	11	2504
In terms of difficulty level		
Below average	17	1834
Average	17	7505
Above Average	17	651
In terms of Levels		
K01	8	551
K02	8	583
K03	8	595
K04	9	680
K05	9	660
K06	9	670
K07	10	769
K08	10	892
K09	11	906
K10	13	1057
K11	13	1293
K12	13	1240
Breakdown by Subject/Specialty		
Social Studies and National Ed	–	844
Biology	–	178
Research and Information Sour	–	92
Family and Health Education	–	854
Physical Education	–	517
Art Education	–	829
Computer Science	–	1003
Mathematics	–	799
Science	–	944
Administrative Sciences	–	284
Islamic Studies	–	1209
Behavioral Sciences	–	267
Physics	–	239
Chemistry	–	220
English Language	–	637
Arabic Language	–	980
Environmental Science	–	93
Total	**17**	**9990**

Table 10: Statistics of IEN MCQs.

E.2 MCQ Datasets Distribution

Figure Figure 9 shows the options distribution in `AraEval` datasets.

E.3 Dataset Examples

Figure 10 illustrates the construction of verifiable instructions in AraIFEval: the upper part shows the original (normal) instruction, while the bottom part shows the instruction after adding verifiable prompts.

F AraIFEval Prompts

Table 13 shows the instructions categories prompts in AraIFEval.

G Dataset Curation and Validation

The guidelines for domain experts on creating AraPro can be found in Table 14, while the validation

Category	#Subject/Specialty	#Questions
In terms of study stages		
Secondary education	17	2539
Primary education	10	1678
Intermediate education	11	1606
In terms of difficulty level		
Below average	17	1360
Average	17	4195
Above average	17	268
In terms of levels		
K01	8	221
K02	8	251
K03	8	281
K04	9	301
K05	9	308
K06	9	316
K07	10	505
K08	11	490
K09	11	611
K10	13	730
K11	13	973
K12	13	836
Breakdown by Subject/Specialty		
Social Studies and Nation	–	482
Biology	–	159
Research and Information	–	99
Family and Health Educat.	–	453
Physical Education	–	421
Art Education	–	380
Computer Science	–	598
Mathematics	–	507
Science	–	421
Administrative Sciences	–	161
Islamic Studies	–	558
Behavioral Sciences	–	233
Physics	–	133
Chemistry	–	197
English Language	–	394
Arabic Language	–	530
Environmental Science	–	97
Total	**17**	**5823**

Table 11: Statistics of IEN TF.

Subject	#Question
Breakdown by Subject/Specialty	
Sociology	403
Biology	212
Management	197
Arabic Literature	558
Economics	397
History	297
Computing	199
Religion	299
Sports	396
Mathematics	200
Politics	414
Physics	97
Chemistry	200
Arabic Linguistics	434
Finance	100
Human Resources	200
Engineering	98
Psychology	200
Earth Sciences	100
Total	**5001**

Table 12: Statistics of AraPro.

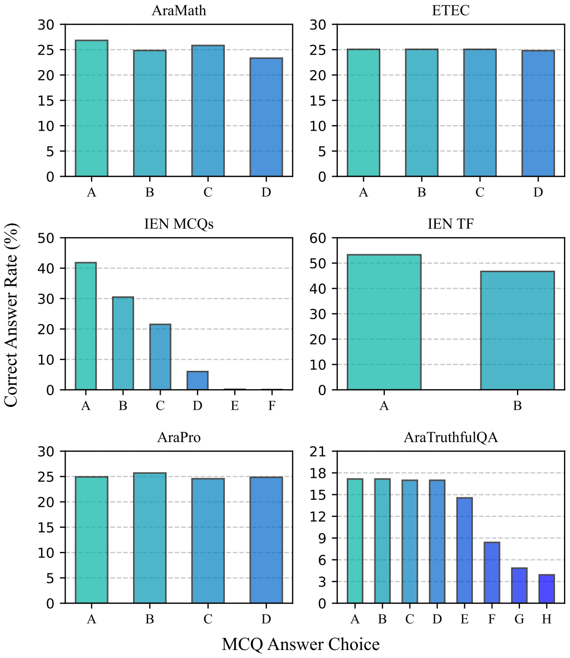


Figure 9: Distribution percentage of the correct answer in each MCQ dataset of `AraEval`.

guidelines for AraMath are presented in Table 15. The guideline for validation of AraIFEval is detailed in Table 16, and the guidelines for AraTruthfulQA are provided in Table 17.

H Tokenizer Vocabulary Coverage

Table 18 shows the models' vocabulary coverage across the Arabic datasets within `AraEval` compared to MMLU and OpenAI MMLU benchmarks.

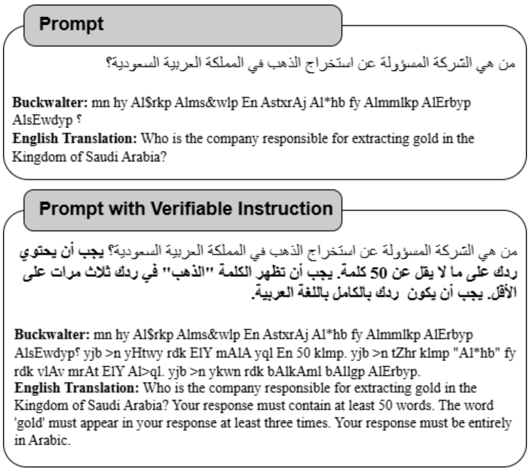
I Evaluation Metrics

I.1 Normalized Accuracy

For all `AraEval` benchmarks, except AraIFEval, we used normalized accuracy as a metric, which simply selects the pre-defined answer completion to each question that maximizes the sum of log likelihood, normalized by the answers token lengths:

$$\frac{1}{m}\sum_{i=1}^{m}\log P(a_i|q_1,...,q_n,a_1,...,a_{i-1}) \quad (2)$$

Where the tested prompt is $[q_1,...,q_n,a_1,...,a_m]$, where q_i represents the i^{th} question token (with n total) and a_i represents the i^{th} answer token (with m total). This is identical to choosing the prompt



Prompt

من هي الشركة المسؤولة عن استخراج الذهب في المملكة العربية السعودية؟

Buckwalter: mn hy Al$rkp Alms&wlp En AstxrAj Al*hb fy Almmlkp AlErbyp AlsEwdyp ؟
English Translation: Who is the company responsible for extracting gold in the Kingdom of Saudi Arabia?

Prompt with Verifiable Instruction

من هي الشركة المسؤولة عن استخراج الذهب في المملكة العربية السعودية؟ **يجب أن يحتوي ردك على ما لا يقل عن 50 كلمة. يجب أن تظهر الكلمة "الذهب" في ردك ثلاث مرات على الأقل. يجب أن يكون ردك بالكامل باللغة العربية.**

Buckwalter: mn hy Al$rkp Alms&wlp En AstxrAj Al*hb fy Almmlkp AlErbyp AlsEwdyp؟ yjb >n yHtwy rdk ElY mAlA yql En 50 klmp. yjb >n tZhr klmp "Al*hb" fy rdk vlAv mrAt ElY Al>ql. yjb >n ykwn rdk bAlkAml bAllgp AlErbyp.
English Translation: Who is the company responsible for extracting gold in the Kingdom of Saudi Arabia? Your response must contain at least 50 words. The word 'gold' must appear in your response at least three times. Your response must be entirely in Arabic.

Figure 10: Example of verifiable instruction created of an existing instruction in Arabic.

which the maximum geometric mean probability over the answer's tokens, since Equation (2) can be rearranged to:

$$log\left[\left(\prod_{i=1}^{m} P(a_i|q_1, ..., q_n, a_1, ..., a_{i-1})\right)^{\frac{1}{m}}\right] \quad (3)$$

and $\log(\cdot)$ is a monotonic function, which is maximized by maximizing its argument.

If the question's prompt asks the model answer an MCQ by outputing the current answer key (A, B, C, or D), then the answer is a single token (the answer key). In such case, we simply pick the token out of the 4 with the highest probability. However, AraTruthfulQA has 4 pre-defined answers that are not presented as multiple choice answers, but we evaluate the models likelihood to generate these answers and select the most probable one as the model's answer.

I.2 AraIFEval Metrics

Inspired by (Zhou et al., 2023), AraIFEval consists of four metrics, loose prompt-level, loose instruction-level, strict prompt-level, and strict instruction-level accuracies. Which are defined as follows:

1. Strict: The instruction is followed without post-processing.
2. Loose: The instruction is followed using any combination of:
   - Removing the first paragraph.
   - Removing the last paragraph.
   - Removing markdown artifacts.
3. Prompt-level: The proportion of prompts for which the model follows *all* the instructions.
4. Instruction-level: The proportion of instructions that the model follows.

I.3 Mitigating Risks of Data Contamination

Most of the AraEval datasets were carefully designed to minimize contamination risks, with components that are either fully original or constructed through controlled human annotation and validation. For instance, AraPro was authored entirely by university professors across diverse knowledge domains, ensuring complete originality. AraIFEval was built from scratch using manually designed multi-instruction prompts and validated through human and deterministic checks. AraTruthfulQA combines 287 culturally filtered, human-translated items from TruthfulQA with 249 newly authored, culturally grounded questions, thereby extending beyond the original benchmark. Finally, the IEN datasets were provided directly by the Ministry of Education; since the IEN platform is closed and inaccessible to the public, the risk of contamination is further minimized.

J GPU Time

GPU time for running evaluation on `AraEval` datasets is reported in Table 19.

Dataset	7B	13B	30B	70B
AraPro (0 shot)	447.65	969.77	4326.20	9770.33
AraPro (5 shot)	328.78	576.82	1434.85	2459.53
IEN MCQ (0 shot)	420.02	463.81	1268.43	2129.42
IEN MCQ (5 shot)	552.10	867.71	2875.39	4196.97
IEN TF (0 shot)	269.64	357.27	1232.53	1686.52
IEN TF (5 shot)	321.30	514.43	1344.34	2677.28
AraMath (0 shot)	44.55	62.17	1676.55	3623.28
AraMath (5 shot)	61.19	94.08	253.83	396.62
ETEC (0 shot)	153.76	172.00	351.70	550.40
ETEC (5 shot)	226.07	367.63	1031.75	1685.91
AraIFEval (0 shot)	7051.31	6954.25	29382.06	29724.12
AraTruthfulQA (0 shot)	514.21	844.75	4443.01	9924.95
AraTruthfulQA (5 shot)	250.30	494.59	1226.33	2111.18

Table 19: GPU time for different model sizes. The reported time is in seconds and is the average across all models of the corresponding size.

Instruction Category	Prompt
include_keywords	قم بتضمين الكلمات المفتاحية في ردك qm btDmyn AlklmAt AlmfAtAHyp (keyword1), (keyword2) fy rdk. Include the keywords (keyword1) and (keyword2) in your response
keyword_frequency	يجب أن تظهر الكلمة (word) في ردك (N) مرة yjb >n tZhr Alklmp (word) fy rdk (N) mrp The word (word) must appear in your response (N) times
forbidden_words	لا تقم بتضمين الكلمات المحظورة lA tqm btDmyn AlklmAt AlmHZwrp Do not include the (forbidden word)
letter_frequency	يجب أن يظهر الحرف (letter) في ردك (N) مرة yjb >n yZhr AlHrf (letter) fy rdk (N) mrp. The letter (letter) must appear (N) times in your response
response_language	يجب أن يكون ردك بالكامل باللغة (language) ولا يُسمح بأي لغة أخرى yjb >n ykwn rdk bAlkAml bAllgp (language) wlA ysmH blgp >xrY Your response must be entirely in (language), and no other language is allowed
number_paragraphs	يجب أن يحتوي ردك على عدد معين من الفقرات yjb >n yHtwy rdk ElY Edd mEyn mn AlfqrAt Your response must contain (N) paragraphs
number_words_at_least	أجب بما لا يقل عن (N)كلمة >jb bmA lA yql En (N) klmp Answer with at least (N) words
number_words_at_most	أجب بما لا يزيد عن (N)كلمة >jb bmA lAyzyd En (N) klmp Answer with (N) words at most
number_sentences_at_least	أجب بما لا يقل عن (N) جملة >jb bmA lA yql En (N) jmlp Answer with at least (N) sentences
number_sentences_at_most	أجب بما لا يزيد عن (N) جملة >jb bmA lA yzyd En (N) jmlp Answer with (N) sentences at most
first_word_in_i-th_paragraph	يجب أن تحوي الإجابة على عدد معين من الفقرات وتبدأ إحدى الفقرات بكلمة محددة yjb >n tHwy Al<jAbp ElY Edd mEyn mn AlfqrAt wtbd> <HdY AlfqrAt bklmp mHddp) The answer must contain a specific number of paragraphs, with one of the paragraphs starting with a specific word
postscript	يرجي إضافة ملاحظة توضيحية في نهاية ردك تبدأ ب (postscript marker) yrjy <DAfp mlAHZp twDyHyp fy nhAyp rdk tbd> b (postscript marker) Please add a clarifying note at the end of your response, starting with (postscript marker)
number_placeholder	يجب أن يحوي ردك على عدد من مواضع الترميز تمثل بأقواس مربعة yjb >n yHwy rdk Ely Edd mn mwADE Altrmyz tmvl b>qwAs mrbEp Your response must contain at least (N) placeholders, represented using square brackets
number_bullets	يجب أن يحتوي ردك على عدد معين من النقاط yjb >n yHtwy rdk ElY Edd mEyn mn AlnqAT Your response must contain a specific number of points.
title	يجب أن يحتوي ردك على عنوان بين أقواس مزدوجة yjb >n yHtwy rdk ElY EnwAn byn >qwAs mzdwjp Your response must include a title enclosed in double angle brackets
minimum_number_high-lighted_section	قم بتسليط الضوء على عدد م من الأقسام على الأقل qm btslyT AlDw' ElY Edd m mn Al>qsAm ElY Al>ql Highlight at least Highlight at least sections.
multiple_sections	يجب أن يحتوي ردك على عدد م من الأقسام. ضع علامة على بداية كل قسم yjb >n yHtwy rdk ElY Edd m mn Al>qsAm . DE ElAmp ElY bdAyp kl qsm Your response must contain N sections. Place a section separator at the beginning of each section
json_format	يجب أن يكون الرد بالكامل بتنسيق JSON yjb >n ykwn Alrd bAlkAml btnsyq JSON Your response must be entirely formatted in JSON
repeat_prompt	كرر المدخل دون تغيير ثم قدم إجابتك krr Almdxl dwn tgyyr vm qdm <jAbtk Repeat the input without modification then respond to the prompt
two_responses	قدم إجابتين مختلفتين. الردود فقط يجب فصلها بـ ٦ رموز نجوم qdm <jAbtyn mxtlftyn. Alrdwd fqT yjb fSlhA b 6 rmwz njwm Provide two different answers. The responses should only be separated by six asterisk symbols
end_checker	انه ردك بالعبارة المحددة Anh rdk bAlEbArp AlmHddp End your response with specific phrase
quotation	يجب أن يكون ردك بالكامل بين علامات اقتباس مزدوجة yjb >n ykwn rdk byn ElAmAt AqtbAs mzdwjp Your response should be between double quotation mark
no-comma	امتنع عن استخدام فواصل في ردك AmtnE En AstxdAm fwASl fy rdk Don’t use comma in your response

Table 13: Instructions categories prompts. We used buckwalter transliteration to transliterate Arabic instructions.

Section	Guidelines
Objective	The goal of these MCQs is to evaluate Large Language Models (LLMs) in achieving professional-level competency in your field of expertise. Each question should reflect real-world knowledge, critical thinking, and problem-solving skills relevant to industry standards. The data you create will only be used for research purposes.
Question Structure	Each MCQ should consist of: • A clear and concise question that assesses knowledge, application, or analysis. • Four answer choices (A, B, C, D), with only one correct answer.
Guidelines for Crafting Questions	• Ensure relevance to key competencies in the profession. • Avoid ambiguity, excessive complexity, or unnecessary jargon. • Use practical scenarios, case studies, or problem-solving situations where possible. • Maintain a mix of basic, intermediate, and advanced questions. • Avoid testing trivial facts; focus on meaningful concepts.
Answer Choices	• One clear correct answer that is indisputably accurate. • Three plausible distractors that are incorrect but not obviously wrong.
Example Question Format	**Question:** What is the primary purpose of risk assessment in cybersecurity? • A) To eliminate all potential threats • B) To identify, analyze, and mitigate security risks • C) To ensure compliance with industry regulations only • D) To monitor network traffic for suspicious activity **Correct Answer:** B) To identify, analyze, and mitigate security risks **Domain:** Computing
Submission Format	• Provide questions in a structured format (Question, Options, Correct Answer, Domain). • Ensure accuracy and relevance. • Submit questions in a spreadsheet as instructed.
Review Process	All questions will be reviewed for accuracy, clarity, and alignment with professional competencies before finalization.

Table 14: Guidelines for Creating AraPro Dataset.

Section	Guidelines
Objective	The purpose of this validation process is to ensure the accuracy, consistency, and quality of a dataset containing mathematical word problems. Annotators are responsible for verifying the correctness of equations, answer choices, and labels to maintain data integrity. This dataset is used to evaluate mathematical reasoning capability of Large Language Models (LLMs). The data will be used for research purposes only.
Dataset Components	Each data entry consists of: - **Mathematical Word Problem**: A problem statement requiring mathematical reasoning. - **Equation**: The corresponding mathematical equation representing the problem. - **Answer Choices (A, B, C, D)**: Four distinct answer options. - **Correct Answer**: The solution to the problem. - **Answer Label**: The letter (A, B, C, or D) corresponding to the correct choice.
Validation Criteria	**1. Accuracy of Equations** - Verify that the equation correctly represents the given word problem. - Ensure the mathematical formulation aligns with the intended logic. - Check for errors in mathematical symbols, operations, and missing components. **2. Choice Distinctiveness** - Confirm that all four answer choices are unique and do not repeat. - Ensure that distractor options are plausible but incorrect. - Avoid choices that are too similar (e.g., minor rounding differences). **3. Answer Correctness** - Solve the problem independently and compare it with the provided correct answer. - Cross-check that the correct answer matches the labeled answer choice. - If errors are found, provide corrected answers and labels. **4. Presence of Correct Answer** - Ensure that the correct answer is one of the four given choices. - If the correct answer is missing from the options, flag the entry for correction. **5. Formatting and Consistency** - Ensure uniform formatting across all dataset entries. - Verify that symbols, units, and mathematical notation follow standard conventions. **6. Logical Soundness** - Assess whether the problem makes sense mathematically and linguistically. - Check for unintended biases or misleading wording.
Annotation Process	1. Read the problem statement carefully and understand its context. 2. Examine the provided equation and ensure it correctly models the problem. 3. Verify that the correct answer is calculated accurately. 4. Confirm that all answer choices are unique and logically reasonable. 5. Check that the correct answer exists within the four given choices. 6. Cross-check the labeled answer against the correct answer. 7. If discrepancies are found, document corrections and flag the entry for review.
Error Reporting & Corrections	Annotators should log any errors found, specifying: - **Entry ID**: The unique identifier of the dataset entry. - **Issue Type**: (Equation Error, Answer Mismatch, Duplicate Choices, Missing Correct Answer, Formatting Issue, etc.). - **Correction**: The revised equation, answer choice, or label. - **Comments**: Additional notes explaining the error.
Final Review & Approval	- After validation, a second-level review may be conducted to ensure error-free dataset entries. - Approved entries will be included in the final dataset, while flagged entries undergo correction and re-evaluation.

Table 15: Guidelines for Human Annotators to validate AraMath Dataset.

Section	Guidelines
Objective	The purpose of this task is to ensure that each instance in this data accurately represents its instructed prompt and instruction categories. Annotators review the dataset for logical consistency, completeness, and correctness. This dataset is used to evaluate instruction following capability of Large Language Models (LLMs). The data will be used for research purposes only.
Dataset Components	Each data entry consists of: - **Instructed Prompt**: A textual prompt containing verifiable instructions. - **Instruction Categories**: A set of verifiable instructions used in the prompt.
Validation Criteria	**1. Contradiction Check** - Ensure that no contradictory instructions exist within the instructed prompt. - Flag instances where conflicting instructions lead to logical inconsistencies. **2. Instruction Completeness** - Verify that all instruction categories in the instruction set are explicitly mentioned in the instructed prompt. - If an instruction is missing, annotate it as an omission. **3. Prompt Coverage** - Ensure that all instructions present in the instructed prompt are correctly identified in the instruction set. - If additional, unlisted instructions are found, flag them for review. **4. Logical Coherence** - Assess whether the prompt flows naturally and follows a coherent structure. - Check for redundant, unclear, or ambiguous wording. **5. Formatting and Standardization** - Verify that instruction labels and categories follow the predefined taxonomy. - Ensure proper punctuation, spelling, and grammar for clarity.
Annotation Process	1. Read the instructed prompt carefully to understand its structure and intent. 2. Compare the instruction categories with the prompt to check for completeness. 3. Identify and flag any contradictory instructions within the prompt. 4. Verify that no instruction is missing from the instruction set. 5. Ensure that no extra, unlisted instructions are present in the prompt. 6. Check for formatting, clarity, and coherence issues. 7. Document errors and suggest corrections where necessary.
Error Reporting & Corrections	Annotators should log errors using the following details: - **Entry ID**: Unique identifier of the dataset instance. - **Issue Type**: (Contradiction, Missing Instruction, Extra Instruction, Formatting Issue, etc.). - **Correction**: Suggested revision for the prompt or instruction set. - **Comments**: Additional explanation of the issue.
Final Review & Approval	- A second-level review may be conducted to ensure high-quality validation. - Approved entries are included in the final dataset, while flagged entries undergo correction and re-evaluation.

Table 16: Guidelines for Validation of AraIFEval Dataset.

Section	Guidelines
Objective	The purpose of this review process is to evaluate multiple-choice questions (MCQs) for alignment with Arabic cultural norms and beliefs, ensuring that the content is appropriate, respectful, and contextually relevant. Additionally, validated MCQs should be translated into Arabic while maintaining their original meaning and intent. This dataset is used to evaluate truthfulness of Large Language Models (LLMs). The data will be used for research purposes only.
Dataset Components	Each MCQ consists of: - **Question**: The main stem of the MCQ. - **Four Answer Choices**: Options (A, B, C, D), with only one correct answer. - **Correct Answer Label**: The letter corresponding to the correct answer.
Validation Criteria	**1. Cultural Alignment** - Ensure that the question and answer choices do not conflict with Arabic cultural and social values. - Avoid topics that may be considered sensitive or inappropriate in an Arabic cultural context. - Verify that examples, names, and scenarios used in the MCQ are relevant and culturally recognizable. **4. Translation Guidelines** - Translate only the MCQs that align with Arabic cultural norms. - Maintain the original intent and meaning of the question while using culturally appropriate phrasing. - Adapt idiomatic expressions or region-specific references to ensure clarity for Arabic speakers. - Use Modern Standard Arabic (MSA) for translation, avoiding dialect-specific terms.
Annotation Process	1. Read the MCQ carefully and assess its cultural appropriateness. 2. If the MCQ is **not aligned**, flag it and provide a justification. 3. If the MCQ is **aligned**, proceed with translation while preserving accuracy and clarity. 4. Ensure that all answer choices remain meaningful and distinguishable after translation. 5. Verify that the correct answer remains unchanged in meaning. 6. Document any modifications made during translation for transparency.

Table 17: Guidelines for Reviewing and Translating TruthfulQA dataset.

Benchmark	ALLaM-7B	ALLaM-34B	ALLaM-Adapted	Jais-Family	Jais-Adapted	Qwen-2.5*	Llama-3**
AraIFEval	7.80	7.54	9.72	6.64	8.98	37.29	35.79
ETEC	32.37	33.34	38.10	28.39	35.53	67.22	58.74
IEN MCQs	53.64	56.15	60.22	48.33	56.82	77.34	63.36
IEN TF	36.24	36.84	42.24	32.21	39.26	71.20	59.70
AraPro	44.18	46.73	50.81	39.87	48.39	73.53	61.82
AraTruthfulQA	17.92	17.60	21.67	15.46	20.01	53.56	49.54
AraMath	5.63	5.19	7.26	5.61	6.41	26.35	38.68
AraEval	**72.02**	**75.37**	**77.67**	68.26	75.96	**82.66**	66.38
OpenAI Arabic MMMLU	71.33	74.69	75.89	**73.08**	**79.54**	80.20	64.45
Arabic MMLU	61.60	63.02	68.17	57.04	65.60	79.95	**69.25**
Vocabulary Token Statistics							
Arabic tokens	29,552	36,028	37,195	43,857	32,046	3,990	3,769
Arabic and math tokens	29,643	36,065	37,236	44,947	32,137	4,311	4,995

*Tokenizer identical to AceGPT-V2 8B/70B's.

**Tokenizer identical to AceGPT-V2 32B's.

Table 18: Vocabulary coverage across Arabic benchmarks and model tokenizers.