From Theory to Practice: A New Paradigm for Arabic Language Model Evaluation

Anonymous ACL submission

Abstract

This paper addresses critical gaps in Arabic language model evaluation by establishing comprehensive theoretical guidelines and introducing a novel evaluation framework. We first analyze existing Arabic evaluation datasets, identifying significant issues in linguistic accuracy, cultural alignment, and methodological rigor. To address these limitations, we present the Arabic Depth Mini Dataset (ADMD), a carefully curated collection of 490 questions spanning ten major domains. Using ADMD, we evaluate five leading language models: GPT-4, Claude 3.5 Sonnet, Gemini Flash 1.5, CommandR 100B, and Qwen-Max. Our results reveal significant variations in model performance across different domains, with particular challenges in areas requiring deep cultural understanding and specialized knowledge. Claude 3.5 Sonnet demonstrated the highest overall accuracy at 30%, showing notable strengths in mathematical, Arabic and islamic domains. This work provides both theoretical foundations and practical insights for improving Arabic language model evaluation, emphasizing the importance of cultural competence alongside technical capabilities.

1 Introduction

The evaluation of Arabic large-language models (LLMs) presents unique challenges that extend beyond conventional metrics of linguistic accuracy. As these models become increasingly prevalent in various applications, the need for comprehensive and culturally aware evaluation frameworks has become critical. Recent developments in Arabic LLM evaluation have produced several datasets, including GPTArEval (Khondaker et al., 2023), Ghafa (Almazrouei et al., 2023), and ArabicMMLU from openAI (OpenAI, 2024), each attempting to address different aspects of model assessment. However, these efforts often fail to provide a comprehensive evaluation that includes both technical proficiency and cultural understanding.

Current evaluation approaches frequently rely on translated content (Romanou et al., 2024) or simplified metrics that fail to capture the nuances of Arabic language and culture (OpenAI, 2024). This limitation is particularly evident in specialized domains such as Islamic studies, classical literature, and technical fields where cultural context and domain expertise are crucial. Furthermore, existing datasets often exhibit inconsistencies in linguistic standards and cultural representation, potentially resulting in misleading assessments of model capabilities.

Our work addresses these challenges through three main contributions. First, we establish theoretical guidelines for Arabic evaluation datasets that encompass linguistic standards, cultural alignment, and methodological requirements. Second, we conduct a detailed analysis of existing evaluation datasets, identifying common pitfalls and areas for improvement. Third, we introduce the Arabic Depth Mini Dataset (ADMD), a specialized evaluation tool designed to assess both technical and cultural competencies across diverse domains.

The ADMD represents a significant advancement in the evaluation of Arabic LLM, featuring carefully curated questions that demand deep understanding rather than surface-level pattern matching. By evaluating leading language models using this dataset, we provide insights into current model capabilities and limitations, particularly in handling complex Arabic queries that require cultural awareness and specialized knowledge.

This paper is organized as follows: Section 2 reviews related work in Arabic LLM evaluation, Section 3 presents our theoretical guidelines, Section 4 analyzes existing evaluation datasets, Section 5 introduces the ADMD and presents evaluation results, and Section 6 discusses limitations and future work directions.

Dataset	Reviewed	Handwritten	Generated	Translated
GPTArEval (Khondaker et al., 2023)	×	✓	×	×
Ghafa (Almazrouei et al., 2023)	×	✓	×	✓
ArabicMMLU (Koto et al., 2024)	✓	×	×	✓
AraDICE (Mousi et al., 2025)	×	✓	×	×
ArSTEM (Mustapha et al., 2024)	✓	×	✓	×
Aya Expanse (Dang et al., 2024)	×	×	✓	✓
AraTrust (Alghamdi et al., 2024)	✓	✓	×	×
ILMAAM (Nacar et al., 2025)	✓	×	✓	×
ADMD (Ours)	✓	✓	×	×

Table 1: Comparison of Arabic LLM Evaluation Datasets based on annotation type and content origin.

2 Related Works

Recent advancements in Arabic large language model (LLM) evaluation have produced several notable datasets and benchmarks. GPTArEval (Khondaker et al., 2023) focuses on natural language understanding and generation tasks, incorporating ORCA (Elmadany et al., 2023) datasets. While Ghafa (Almazrouei et al., 2023) utilizes translated content with native speaker revisions and ArabicMMLU (Koto et al., 2024) covers diverse topics, both datasets face challenges with linguistic accuracy and comprehensive domain coverage. Cultural representation in Arabic LLMs has been examined through specialized datasets like AraDICE (Mousi et al., 2025) for dialectal and cultural evaluation, and ArSTEM (Mustapha et al., 2024) for scientific knowledge assessment. Notable Arabic LLM development teams have contributed evaluation methodologies, with Jais (Sengupta et al., 2023) and Allam (Bari et al., 2024) employing varied approaches to dataset curation. The Aya Expanse model (Dang et al., 2024) notably utilized translated and generated content, while maintaining transparency about GPT-generated materials. Critical analysis by (Nacar et al., 2025) has identified significant limitations in existing benchmarks and created a dataset generated from trusted books, particularly in ArabicMMLU (OpenAI, 2024), ranging from linguistic inaccuracies to methodological flaws. In response, AraTrust (Alghamdi et al., 2024) was developed to address these challenges and enhance LLM reliability assessment (look Table 1.)

In the scope of the ongoing efforts to establish robust Arabic evaluation datasets, our work provides a theoretical foundation and empirical assessment of three key evaluation datasets: Ghafa, ArabicMMLU, and INCLUDE. Furthermore, we introduce the Arabic Depth Mini Dataset (ADMD) as a foundational resource for developing a more extensive and specialized Arabic QnA dataset, addressing gaps in evaluating deep domain knowledge.

3 Theoretical Guidelines

This section outlines the theoretical standards and instructions necessary for building an Arabic evaluation dataset, ensuring linguistic, cultural, and methodological soundness. The guidelines are categorized into the following areas (look figure 1) and was inspired by our work in (Nacar et al., 2025):

3.1 Linguistic Standards

This section outlines the essential guidelines for ensuring high-quality and accurate translations, emphasizing linguistic precision, consistency, and contextual appropriateness in Arabic.

- **Translation Quality:**
  - Ensure that all terms are translated accurately; untranslated terms must be transliterated if necessary (and the non-Arabic word could be mentioned between brackets).
  - Avoid literal translations by focusing on contextual adaptation, ensuring natural and consistent rendering.
  - Review machine translations thoroughly and ensure alignment across multiple uses of the same term (e.g., consistency in letter choices for the answers like listing the Answers either in A,B,C or in Arabic أ، ب، ج.
- **Linguistic Accuracy:**

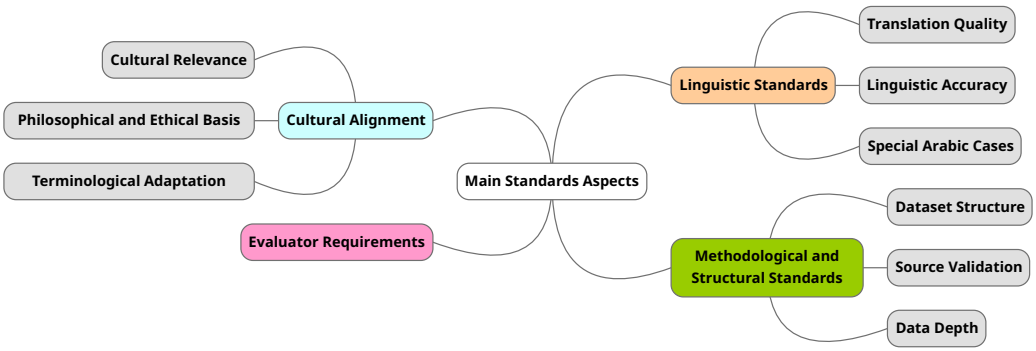

Figure 1: Mindmap Representation

- Adhere strictly to Arabic grammar, morphology, syntax, and orthographic rules.
- Avoid weak linguistic structures even if grammatically correct.
- Ensure stylistic adequacy and use expressions that match the intended purpose and context.

- **Special Cases:**
  - Write poetry accurately, maintaining its structure and prosody.
  - Write mathematical notations either in Arabic form or provide clear rules for using Latin symbols.
  - Ensure consistent orthographic representation of dialects by adhering to a standard framework, for example, (Habash et al., 2018).

3.2 Cultural Alignment

This subsection emphasizes aligning content with Arabic cultural contexts, adapting philosophical concepts, and using culturally appropriate terminology.

- **Cultural Relevance:**
  - Ensure questions, examples, and references align with the cultural, historical, and social contexts of the Arabic-speaking world.
  - Avoid introducing examples or entities that are disconnected from Arab culture, such as irrelevant or Western-specific references.

- **Philosophical and Ethical Basis:**
  - Refrain from presenting Western philosophical or ethical concepts as universal truths without explanation or adaptation.
  - Avoid using expressions or examples that conflict with Arab cultural context or not used.

- **Terminological Adaptation:**
  - Replace Westernized terms with culturally and linguistically appropriate Arabic terms (in standard Arabic or in dialects).
  - Provide Arabic equivalents or transliterations where necessary, maintaining cultural integrity.

3.3 Methodological and Structural Standards

This subsection defines standards for organizing datasets, validating sources, and ensuring data depth and inclusivity.

- **Dataset Structure:**
  - Organize questions logically, ensuring they are placed in their relevant categories.
  - Avoid redundancy or confusion by grouping related queries appropriately.
  - Ensure the information is current and includes accurate dates.

- **Source Validation:**
  - Attribute knowledge and data to original Arabic primary sources, including books, studies, and statistical studies that are connected to Arabic societies.
  - Avoid over-reliance on non-Arabic secondary references when constructing Arabic datasets.

- Writing Quranic texts with complete accuracy using the Uthmanic script.

- **Data Depth:**
  - Ensure the dataset reflects depth and richness, avoiding straightforward, shallow, or overly simplistic questions and answers.
  - Incorporate diverse perspectives within the Arabic-speaking world for inclusivity.

3.4 Evaluator Requirements

Evaluators must demonstrate proficiency in Arabic, including linguistic nuances and cultural contexts, alongside strong subject matter expertise. To enhance evaluation efficiency, a Python library utilizing the Claude Sonnet model was developed post-analysis. This library, available on GitHub[1], automates dataset evaluation based on theoretical guidelines.

4 Review of famous Arabic Evaluation datasets

We selected three datasets for evaluation, sampled them, and analyzed the issues based on our proposed theory. The evaluation criteria were aligned with the key concepts discussed in the previous section, including **Language Rules, Scientific Writing, Cultural Values,** and **Information Correctness**. Each criterion was scored on a scale of 1 to 10. The chosen datasets are as follows: (1) the Ghafa dataset (Almazrouei et al., 2023), (2) the ArabicMMLU dataset (OpenAI version) (OpenAI, 2024), selected for being the latest version of MMLU and reviewed by native Arabic speakers, and (3) the Cohere "Include" dataset (Romanou et al., 2024).

4.1 Al Ghafa Dataset

From this dataset (Almazrouei et al., 2023), we sampled 100 examples, which were reviewed by a native Arabic speaker according to the evaluation criteria outlined previously. The dataset received the following scores:

and we decided that the sample which is under 5 evaluation is a 'wrong sample' and we extracted: 50 wrong samples from language rules, 42 from Scientific Writing, 60 from Cultural Values, and 26 from Information Correctness Below are examples

[1] https://github.com/serrysibaee/EAED

Criterion	Score */10*
Language Rules	4.5
Scientific Writing	4.6
Cultural Values	3.9
Information Correctness	6.1

Table 2: Evaluation Scores for Al Ghafa Dataset

of evaluated samples, along with their identified issues:

1. صيام يوم الشك سنة **Translation:** Fasting on the day of doubt is a Sunnah.
   - **Issue:** The answer is inconsistent—its ruling is either obligatory or forbidden, depending on the disagreement.

2. ((١٨) سَنَدْعُ الدْبَانِيَةَ **Translation:** Allah said, "So let him call his associates (17), We will call the guards of Hell (18)."
   - **Issue:** Incorrect transcription of the Quranic text, including errors in diacritics. The correct form is الزَّبانِيَة.

3. الْـ ١٣ عَامْ بِيْتَرْ لِينْزْ **Translation:** Thirteen years old Peter Linz.
   - **Issue:** Grammatical error—the correct form is الْـ ١٣ عَامًا.

4. كَمَا يَعْتَقِدُونَ فِيهِ العِصم ن هُوَ سَيِّدُ الرِّجَالِ؟ **Translation:** As they believe in his infallibility, is he the master of men?
   - **Issue:** Spelling and typographical error. The correct form is الْعِصْمَةُ.

4.2 ArabicMMLU

The Arabic MMLU Benchmark (OpenAI, 2024), derived from the original English version (Hendrycks et al., 2020), exists in two translations: one by GPT-3.5 Turbo and another by native Arabic translators. Despite its widespread adoption for Arabic LLM evaluation, the benchmark exhibits significant limitations in cultural adaptation and translation quality. Empirical analysis revealed three primary deficiencies:

(1):Linguistic Fidelity following Arabic Grammar and quality Translating, **(2):Cultural Alignment:** variant western focus with no Arabic alignment and **(3): Structural Integrity:** Suboptimal

organization and insufficient Arabic source attribution. with scores Table 3

Criterion	Score */10*
Language Rules	6.5
Scientific Writing	5.5
Culture Values	3.4
Information Correctness	6.5

Table 3: Evaluation Scores for ArabicMMLU Dataset

Below are three representative examples of identified issues:

1. المضاعفات الفسيولوجية Translate: Physiological complications
   - **Issue:** did not translate the word الفسيولوجية which has Arabic term الجسمورية، وظائف الأعضاء.
2. المبادئ التوجيهية للجنة تكافؤ Translate: Guidelines of the Equality Committee
   - **Issue:** The reliance on Western laws and regulations without providing Arabic contextual alternatives.
3. No mention of any Arabic society studies or statistics.

4.3 INCLUDE dataset

INCLUDE (Romanou et al., 2024) is a multilingual benchmark evaluating knowledge and reasoning across 44 languages. The Arabic subset (551 MCQs) exhibited significant quality issues: (1) Poor Quality – 70% contained severe spelling errors, and 80% required major revisions in structure and content. (2) Incorrect Answers – Notably in Islamic studies, where precision is critical. (3) Misinformation – Some questions conveyed ambiguous or incorrect meanings, particularly in religious contexts. Table 4 presents the dataset evaluation (excluding culture-related data[2]).:

Below are examples of evaluated samples along with their identified issues:

1. **Spelling Errors:**

   Original: المنشؤة على تعّد Translation: was constructed on.
   - **Issue:** spelling mistake the correct is المنشأة على تعد.

[2]No culture data was in the dataset

Criterion	Score */10*
Language Rules	4.5
Scientific Writing	3.5
Cultural Values	-
Information Correctness	7.0

Table 4: Evaluation Scores for INCLUDE Dataset.

2. **Misleading Questions:**

   Original: صوم رمضان سنة **Translation:** Fasting Ramadan is not mandatory.
   - **Issue:** Ramadan Fasting in Islam is mandatory.

5 MiniDataset

We developed a compact yet highly challenging Arabic dataset[3] consisting of 490 carefully curated questions sourced from diverse books and references. The dataset spans ten major domains, covering general science, Islamic studies, Arabic language, and cultural topics (detailed in Appendix A). Unlike conventional benchmarks that rely on automated statistical analysis, our evaluation methodology is based on thorough manual review[4].

To assess the ability of language models to handle complex Arabic inquiries with precision and depth, we conducted extensive testing on leading models, including GPT-4, Sonnet Claude 3.5[5], Gemini Flash 1.5, CommandR 100B[6], and Qwen-Max 2.5[7]. The primary results are presented in Figure 2, with key insights discussed in the following section.

5.1 main insights

The human evaluation results reveal significant performance differences among language models in handling complex Arabic questions[8]. **Claude 3.5 Sonnet** achieved the highest accuracy, correctly answering 147 questions (30%), with notable strength in **Mathematics & Computational Sciences (50%)**, **Philosophy & Logic (50%)**, and

[3]uploaded to Hugging Face: https://huggingface.co/datasets/riotu-lab/ADMD

[4]After several experiments, we found that the most effective way to automate the evaluation is by using a judge LLM

[5]claude-3-5-sonnet-20241022

[6]https://huggingface.co/CohereForAI/c4ai-command-r-plus

[7]https://qwenlm.github.io/blog/qwen2.5-max/

[8]True means the model answered correctly and False is not-correct. Partially-True it answered 60-80% correct, Partially-False the answer is 20-30% correct.

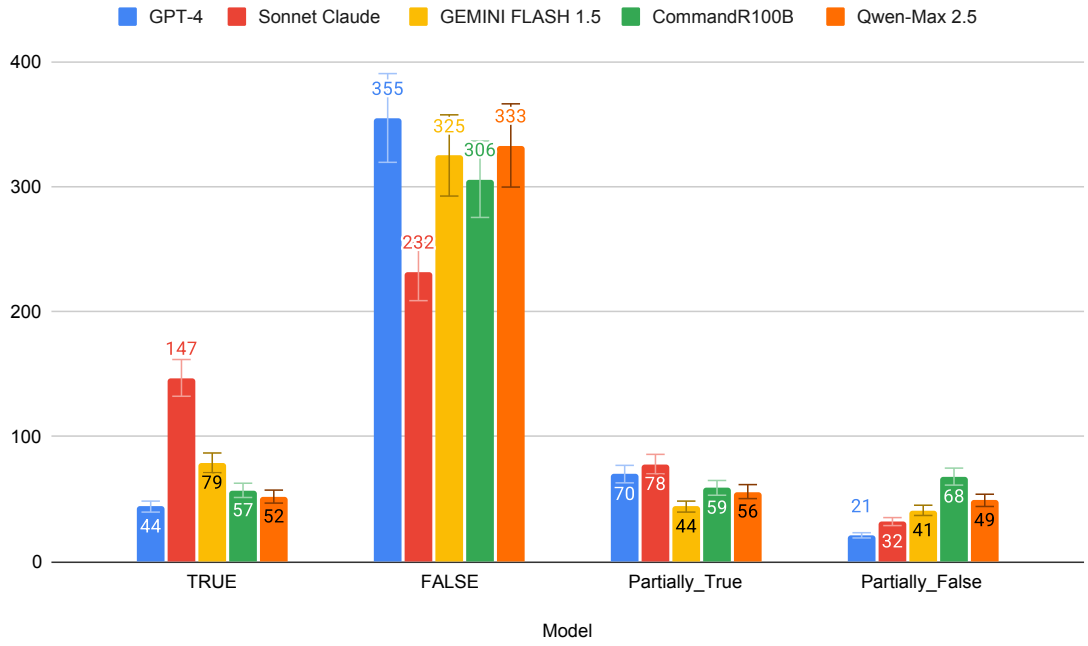

Figure 2: Models Results: True means the model answered correctly and False is not-correct. Partially-True it answered 60-80% correctly, Partially-False the answer is 20-30% correct.

General & Miscellaneous Sciences (51.67%), as shown in Table 11. In **Natural Sciences**, it exhibited a balanced mix of **True (45%)** and **Partially-True (45%)** responses.

GPT-4 had the weakest performance, with only 44 correct answers and the highest incorrect count (355) (Table 7), indicating difficulty in nuanced Arabic queries. **Gemini Flash 1.5** and **CommandR100B** showed moderate performance but high false rates (Table 10, Table 9). **Qwen-Max** had one of the lowest **True** counts (52) while being competitive in **Partially-True** responses (Table 8), reflecting weaknesses in factual reasoning.

Islamic & Religious Studies and **Linguistics & Literature** had the highest **false rates**, with **Claude 3.5 Sonnet** performing relatively better (41.82% False vs. over 80% for other models). These results highlight the models' struggles with nuanced interpretation. Future improvements should focus on reducing **False** responses while refining **Partially-True** classifications to enhance factual accuracy.

We can see from Table 5 while comparing it to Table 6[9] (El Filali et al., 2025) the big difference in the evaluations.

[9]from huggingface https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard

Model	T (%)	F (%)	PT (%)	PF (%)
Sonnet	33.5	43.5	18.2	4.8
R+	15.0	54.0	15.6	15.4
Gemini	22.1	56.2	12.0	9.7
GPT-4	11.8	67.3	17.3	3.5
Qwen	13.1	57.4	17.8	11.7

Table 5: Model Performance Metrics (average for each model on the categories). T:True, F:False, PT:Partially-True, PF:partially-False

MN	AIGhafa	arMMLU	madQA	AraTrust
[1]	78.17	75.84	75.15	89.65
[2]	80.36	69.76	72.91	88.95
[3]	78.10	78.85	68.57	89.96
[4]	78.34	78.60	58.43	89.22
[5]	78.22	71.43	58.11	89.21

Table 6: Model Performance Comparison. The numbers [1], [2], [3], [4], and [5] correspond to Ultiima-72B, Llama-3.3-70B-Instruct, calme-2.1-qwen2.5-72b, calme-2.2-qwen2.5-72b, and Qwen2.5-72B-Instruct, respectively.

6 Limitations and Future Work

This section outlines the limitations of the current study and suggests directions for future research to improve model robustness and evaluation.

6.1 Limitations

The study faces several limitations, including the scalability challenge of manual evaluation and limited query diversity per topic. Key subjects such as Physics, Chemistry, and advanced mathematics were excluded, alongside minimal expertise in specialized fields like Medicine. Subjective topics (e.g., Psychology, Cosmology) complicate assessment, and dataset evaluation remains time-intensive. Additionally, the exclusion of several Arabic models restricts the breadth of comparative analysis.

6.2 Future Work

Future work will focus on expanding the dataset to cover more topics and question types, including MCQs and logic-based questions, to enhance evaluation comprehensiveness. Additional models, such as Jais, Allam, Fanar, Aya, and DeepSeek, will be assessed for broader comparison. Moreover, optimized prompting strategies will be explored to improve response accuracy and quality.

7 Conclusion

This paper proposed a comprehensive framework for evaluating Arabic language models, addressing linguistic, cultural, and methodological aspects. Our analysis identified limitations in existing evaluation datasets, including linguistic inaccuracies and cultural misalignment. To bridge these gaps, we introduced the Arabic Depth Mini Dataset (ADMD) with 490 questions across ten domains. Model evaluations using ADMD revealed varied performance, with Claude 3.5 Sonnet excelling in Mathematics & Logic but all models struggling with culturally nuanced topics. These findings highlight the need for more refined evaluation methodologies to enhance Arabic NLP, ensuring both technical precision and cultural competence.

References

Emad A. Alghamdi, Reem I. Masoud, Deema Alnuhait, Afnan Y. Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. 2024. Aratrust: An evaluation of trustworthiness for llms in arabic. *Preprint*, arXiv:2403.09017.

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.

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

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Ali El Filali, Manel ALOUI, Tarique Husaain, Ahmed Alzubaidi, Basma El Amel Boussaha, Ruxandra Cojocaru, Clémentine Fourrier, Nathan Habib, and Hakim Hacid. 2025. Open arabic llm leaderboard 2. https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A challenging benchmark for Arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser

Zalmout, Sara Hassan, Faisal Al-Shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

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

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.

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

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.

Omer Nacar, Serry Taiseer Sibaee, Samar Ahmed, Safa Ben Atitallah, Adel Ammar, Yasser Alhabashi, Abdulrahman S. Al-Batati, Arwa Alsehibani, Nour Qandos, Omar Elshehy, Mohamed Abdelkader, and Anis Koubaa. 2025. Towards inclusive Arabic LLMs: A culturally aligned benchmark in Arabic large language model evaluation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 387–401, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI. 2024. Multilingual massive multitask language understanding (mmmlu). Accessed: 2025-01-14.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, et al. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

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

A Appendix

A.1 Topics of the ADMD

This dataset covers **42 topics** across various domains (each topic has 10 questions except general language and diversified science each has 50). The topics are categorized as follows:

- **Applied Sciences & Engineering:**
  - Mechanical Engineering
  - Computer Science
  - Medicine
  - Nutrition (include Health & Fitness)
  - Earth Science
- **Natural Sciences:**
  - Biology
  - Cosmology
- **Social Sciences & Humanities:**
  - Psychology
  - Sociology
  - Anthropology
  - Media & Communication
  - Economics
- **Islamic & Religious Studies:**
  - Quranic Exegesis (Tafsir)
  - Hadith
  - Hadith Terminology (Mustalah)
  - Islamic Jurisprudence (Fiqh)
  - Principles of Islamic Jurisprudence (Usul Al-Fiqh)
  - Inheritance Laws (Fara'id)
  - Islamic Creed (Aqeedah)
  - Quranic Recitation Rules (Tajweed)
  - Quranic Readings (Qira'at)
  - Biography of the Prophet (Seerah)
  - Biographies of Islamic Scholars (Tarajim Al-Rijal)
- **Linguistics & Literature:**
  - Arabic Grammar (Nahw)
  - Arabic Morphology (Sarf)
  - Rhetoric (Balagha)
  - Arabic Prosody & Poetic Meters (Arood & Qawafi)
  - Poetry
  - Arabic Language
  - General Linguistics
  - Arabic Linguistics
  - Arabic Logic
- **Philosophy & Logic:**
  - Philosophy
  - Arabic Logic
- **Culture & Arts:**
  - Music
  - Folklore & Cultural Studies
- **Mathematics & Computational Sciences:**
  - Mathematics
  - Machine Learning
- **General & Miscellaneous Sciences:**
  - General Sciences
  - Cooking
- **Historical & Genealogical Studies:**
  - Genealogy (Ansab)

This structured categorization ensures a well-organized representation of the dataset's diverse topics, making it suitable for evaluating Arabic LLMs across multiple domains.

A.2 Examples from the ADMD

In this section, we present examples from each topic in the ADMD dataset. Due to the length of these examples and technical issues related to handling long Arabic texts in the ACL format, we have opted to provide the examples in a more accessible format via a Google Sheet. This allows for easier reading and also includes their English translations.

You can access the examples and their translations through the following link:

https://docs.google.com/spreadsheets/d/1Nl9ZDzNK29yJPpFepx453Lhbwf6HAPSnc_K5sGIfZ7U/edit?usp=sharing

Field of Study	True (%)	False (%)	Partially-True (%)	Partially-False (%)
Applied Sciences & Engineering	22.00	42.00	28.00	8.00
Natural Sciences	20.00	35.00	45.00	0.00
Social Sciences & Humanities	12.00	56.00	26.00	6.00
Islamic & Religious Studies	0.91	80.91	10.00	8.18
Linguistics & Literature	1.82	94.55	2.73	0.91
Philosophy & Logic	10.00	80.00	10.00	0.00
Culture & Arts	10.00	75.00	10.00	5.00
Mathematics & Computational Sciences	25.00	45.00	25.00	5.00
General & Miscellaneous Sciences	16.67	65.00	16.67	1.67
Historical & Genealogical Studies	0.00	100.00	0.00	0.00

Table 7: Statistics for gpt-4 answers for the categories

Field of Study	True (%)	False (%)	Partially-True (%)	Partially-False (%)
Applied Sciences & Engineering	20.00	42.00	18.00	20.00
Natural Sciences	20.00	15.00	40.00	25.00
Social Sciences & Humanities	18.00	42.00	24.00	16.00
Islamic & Religious Studies	4.55	80.00	5.45	10.00
Linguistics & Literature	1.82	90.00	3.64	4.55
Philosophy & Logic	15.00	70.00	5.00	10.00
Culture & Arts	5.00	85.00	10.00	0.00
Mathematics & Computational Sciences	20.00	30.00	35.00	15.00
General & Miscellaneous Sciences	26.67	50.00	16.67	6.67
Historical & Genealogical Studies	0.00	70.00	20.00	10.00

Table 8: Statistics for (Qwen-Max)

Field of Study	True (%)	False (%)	Partially-True (%)	Partially-False (%)
Applied Sciences & Engineering	30.00	52.00	6.00	12.00
Natural Sciences	30.00	15.00	50.00	5.00
Social Sciences & Humanities	18.00	46.00	20.00	16.00
Islamic & Religious Studies	3.64	69.09	10.00	17.27
Linguistics & Literature	4.55	82.73	3.64	9.09
Philosophy & Logic	10.00	45.00	15.00	30.00
Culture & Arts	15.00	70.00	5.00	10.00
Mathematics & Computational Sciences	25.00	30.00	25.00	20.00
General & Miscellaneous Sciences	13.33	60.00	11.67	15.00
Historical & Genealogical Studies	0.00	70.00	10.00	20.00

Table 9: Statistics for (commandR_100B)

Field of Study	True (%)	False (%)	Partially-True (%)	Partially-False (%)
Applied Sciences & Engineering	24.00	46.00	24.00	6.00
Natural Sciences	40.00	15.00	20.00	25.00
Social Sciences & Humanities	38.00	32.00	14.00	16.00
Islamic & Religious Studies	0.00	88.18	5.45	6.36
Linguistics & Literature	2.75	84.40	4.59	8.26
Philosophy & Logic	15.00	60.00	5.00	20.00
Culture & Arts	10.00	70.00	10.00	10.00
Mathematics & Computational Sciences	45.00	30.00	25.00	0.00
General & Miscellaneous Sciences	36.67	56.67	1.67	5.00
Historical & Genealogical Studies	10.00	80.00	10.00	0.00

Table 10: Statistics for (gemini-1.5-flash)

Field of Study	True (%)	False (%)	Partially-True (%)	Partially-False (%)
Applied Sciences & Engineering	42.00	28.00	24.00	6.00
Natural Sciences	45.00	5.00	45.00	5.00
Social Sciences & Humanities	38.00	38.00	20.00	4.00
Islamic & Religious Studies	30.00	41.82	16.36	11.82
Linguistics & Literature	12.84	66.97	13.76	6.42
Philosophy & Logic	50.00	50.00	0.00	0.00
Culture & Arts	15.00	65.00	15.00	5.00
Mathematics & Computational Sciences	50.00	20.00	20.00	10.00
General & Miscellaneous Sciences	51.67	40.00	8.33	0.00
Historical & Genealogical Studies	0.00	80.00	20.00	0.00

Table 11: Statistics for (claude-3-5-sonnet)