The Need for Robust and Inclusive Benchmarks in Evaluating LLMs on Arabic Text

Lubana Al Rayes
Department of Computer Science
University of Sharjah
Sharjah, UAE
lrayes@sharjah.ac.ae

Ashraf Elnagar
Department of Computer Science
University of Sharjah
Sharjah, UAE
ashraf@sharjah.ac.ae

Abstract

The widespread success of large language models (LLMs) has prompted increasing interest in their evaluation across diverse linguistic settings, yet systematic assessments for Arabic remain underexplored. This survey presents a structured taxonomy of benchmarks specifically designed to evaluate LLMs on Arabic text. It critically reviews existing benchmarks, highlighting their coverage across multiple domains, including general single-task and multi-task scenarios, knowledge and reasoning tasks, and domain-specific applications. Finally, it identifies key methodological limitations and proposes future research directions to facilitate the development of more robust, inclusive, and culturally aligned evaluation frameworks for LLMs.

1 Introduction

Large Language Models (LLMs) have become a cornerstone of modern natural language processing (NLP), demonstrating remarkable performance across a wide spectrum of tasks such as machine translation (MT), sentiment analysis, dialogue generation, and reasoning (Yang et al., 2025). Their broad generalization capabilities have positioned them as foundational tools in diverse domains, ranging from healthcare and law to education and creative writing (Bommasani et al., 2021). However, their widespread deployment necessitates rigorous evaluation frameworks to ensure reliability, fairness, and robust performance in complex reasoning, factual consistency, and linguistic competence, particularly in low-resource languages like Arabic.

Although Arabic is among the most widely spoken languages globally, it is significantly underrepresented in the training data of many multilingual large language models (MLLMs), where English typically accounts for over 90% of the corpus and Arabic often constitutes less than 1% (Xu et al., 2025; Qian et al., 2024). Consequently, many Arabic-centric or multilingual models struggle to maintain consistent performance across dialects, linguistic styles, and culturally grounded tasks (Magdy et al., 2025; Alwajih et al., 2025). To address this gap, an increasing number of benchmarks have been proposed to evaluate LLMs on Arabic tasks. These benchmarks span a variety of domains and evaluation objectives, including general multi-task performance, commonsense and factual reasoning, domain-specific applications (e.g., legal and healthcare), and fine-grained single-task assessments. Despite this growing body of work, there is no unified or comprehensive framework that consolidates these efforts to guide comparative evaluation or diagnostic analysis.

This paper addresses these challenges by offering a structured survey focused exclusively on benchmarks used to evaluate LLMs on Arabic text. It systematically reviews existing benchmarks and organizes them into a unified taxonomy based on task type and domain focus. The paper also identifies common methodological gaps and proposes directions for future research. This survey serves as a foundational resource for researchers and practitioners seeking to understand the current landscape, design more inclusive benchmarks, or select appropriate evaluation frameworks for their models.

2 Related Work

Several recent surveys have synthesized progress in LLM development and evaluation, yet none have specifically focused on existing benchmarks for evaluating LLMs on Arabic text.

One of the most relevant works is by Mashaabi et al. (2025). This survey provides an overview of Arabic LLMs across different architectures (encoder-only, decoder-only, encoder-decoder), linguistic forms (Modern Standard Arabic (MSA), Classical Arabic, Dialectal Arabic), and pretraining datasets. It also evaluates the openness of these models and their performance across downstream

NLP tasks. However, the work does not systematically survey evaluation benchmarks used to assess these models. Benchmarks are only briefly mentioned in the context of task-based performance. In related efforts, benchmarks focusing specifically on Arabic word embeddings and contextualized embeddings have been proposed, including those by Yagi et al. (2023) and Elnagar et al. (2023), providing comprehensive evaluation frameworks for these foundational models. Furthermore, studies examining Arabic punctuation and its linguistic characteristics have offered insights into its rule-governed nature (Yagi et al., 2024).

Similarly, Rhel and Roussinov (2025) offer a general overview of Arabic LLMs. While the paper reflects on limitations in Arabic resources and the application of LLMs to Arabic NLP tasks, its focus is not on benchmarking. Instead, it summarizes the adoption of LLMs in Arabic contexts and briefly lists common datasets, without detailed analysis or categorization of benchmarks used across tasks. On a related note, cross-lingual models integrating Arabic language with images have recently been developed, such as the AraCLIP framework by Al-Barham et al. (2025), which explores novel approaches to Arabic vision-language understanding.

Outside the Arabic context, Laskar et al. (2024) presented a systematic review of LLM evaluation pipelines, identifying challenges such as reproducibility, dataset contamination, and fairness across benchmarks. Their work offers a robust foundation for understanding the complexities of LLM evaluation but focuses primarily on English and multilingual settings. Likewise, Lai et al. (2023) analyzed the multilingual performance of ChatGPT across 37 languages, including Arabic, through zero-shot evaluations on tasks like summarization and Part of Speech (POS) tagging. While their work evaluated Arabic among other languages, it did not aim to survey benchmarks nor did it focus on Arabic text.

To the best of our knowledge, this paper is the first to focus specifically on the evaluation benchmarks used to assess LLMs on Arabic text, rather than surveying Arabic LLMs themselves. While prior surveys have examined Arabic language models in terms of architecture, datasets, and application domains, none have systematically analyzed the benchmarks that underpin their evaluation. This distinction allows our work to fill a critical gap by offering a structured overview of the evaluation landscape and identifying methodological shortcomings in current benchmarking practices.

3 Methodology

A total of 26 relevant studies were included in this survey paper. All studies were published between 2022 and 2025. The search window spanned 2020 to 2025, and the methodology followed a systematic approach structured into three main phases.

3.1 Literature search

To identify relevant research, a comprehensive literature search was conducted across multiple scientific databases, including Google Scholar, Elsevier, and IEEE Xplore. The search queries used combinations of keywords such as "Large Language Models", "Benchmark", "Evaluation", and "Arabic Text". This process yielded a total of 42 records.

3.2 Inclusion and exclusion criteria

The retrieved records were screened for eligibility using predefined criteria. Studies were included if they evaluated LLMs on Arabic text, regardless of whether other languages were involved, provided Arabic evaluation was a core component. Studies that focused exclusively on non-textual modalities (e.g., images, audio, video) or did not contain Arabic content were excluded.

Duplicates were identified and removed (n = 2), resulting in 40 records screened. Of these, 4 were excluded during initial screening due to irrelevance, and 10 more were excluded after full-text assessment. No records were missing or unretrievable. In total, 26 unique studies met the inclusion criteria and were included in the final review. A visual summary of this process is shown in Figure 1.

3.3 Taxonomy

The selected studies were organized using a structured taxonomy designed to categorize LLM evaluations on Arabic text. Each study was assigned to a distinct subcategory under one of four main categories, based on its primary objective and evaluation scope. The taxonomy comprises:

- General Multi-Task Evaluation Benchmarks
- Knowledge and Reasoning Benchmarks
- Domain-Specific Benchmarks
- Focused Single-Task Evaluations

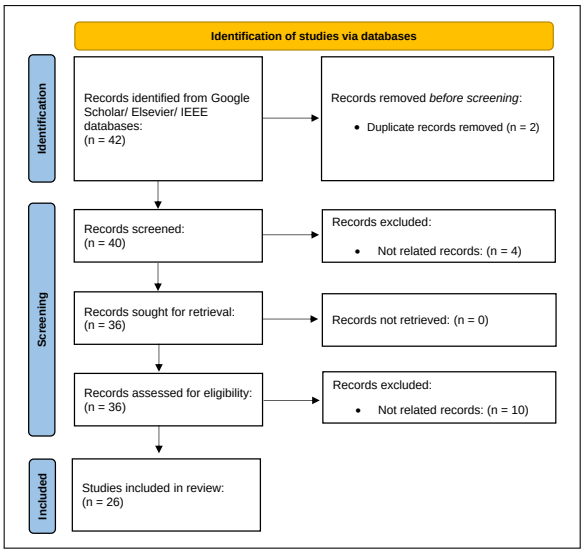

Figure 1: Flow-chart for study inclusion

Subcategories reflect the evaluation scope, task specificity, and domain orientation of each study, as illustrated in Figure 2 and detailed in the following sections.

While our taxonomy was initially designed around the specific context of LLM evaluations on Arabic text, its fundamental structure is language-agnostic and can be generalized across diverse linguistic contexts, potentially serving as a broader blueprint for evaluating LLMs.

4 Taxonomy for Evaluating LLMs on Arabic Text

This section presents the taxonomy used to classify benchmarks for evaluating LLMs on Arabic text. The taxonomy is divided into four major categories: (1) General Multi-Task Evaluation Benchmarks, (2) Knowledge and Reasoning Benchmarks, (3) Domain-Specific Benchmarks, and (4) Focused Single-Task Evaluations. Each category captures distinct evaluation objectives, methodological designs, and linguistic considerations.

Detailed characteristics of each benchmark are summarized in Appendix A.

4.1 General Multi-Task Evaluation Benchmarks

General multi-task evaluation benchmarks are designed to assess LLMs on a broad range of NLP tasks that combine natural language understanding (NLU) and generation (NLG). Within this category, we distinguish between two subcategories: Multi-Task Mixed NLU/NLG Benchmarks and NLG-Focused Multi-Task Benchmarks.

The first subcategory, Multi-Task Mixed NLU/NLG Benchmarks, includes benchmarks that evaluate LLMs across diverse general-domain tasks. One example is the AraT5/ARGEN benchmark (Elmadany et al., 2022), which adopts a text-to-text format to uniformly structure input and output for eight tasks, including sentiment analysis, classification, Named Entity Recognition (NER), extractive QA, summarization, and paraphrasing. The benchmark tests models like AraT5, mT5, and mBART in zero- and few-shot settings, using task-appropriate metrics such as F1, BLEU, and ROUGE. Despite its extensive task coverage, the benchmark is primarily based on MSA, with minimal attention to dialectal Arabic. This limits its applicability in real-world scenarios involving linguistic variation.

Another benchmark in this subcategory is GPTAraEval (Khondaker et al., 2023), which assesses ChatGPT-3.5 and GPT-4 across 44 tasks drawn from 60 datasets, encompassing classification, paraphrase detection, QA, and NER. The benchmark operates exclusively in zero-shot mode to reflect typical usage of proprietary LLMs. While GPT-4 demonstrates superior performance over its predecessor, the study's focus on only two models introduces bias and excludes insights from Arabic-centric or fine-tuned models.

LAraBench (Abdelali et al., 2023) expands multi-task evaluation by including speech-related tasks, such as automatic speech recognition (ASR) and text-to-speech (TTS), in addition to standard NLP tasks. It covers 33 tasks across 61 datasets and evaluates models including GPT-4, Jais, and Whisper. The benchmark shows that even the strongest LLMs face difficulties with syntactic and sequence tagging tasks. These issues are partly due to the lack of Arabic-specific pretraining and inconsistent output formatting. The models also perform poorly across different Arabic language varieties, which can be attributed to the lack of dialectal data.

The second subcategory, NLG-Focused Multi-Task Benchmarks, centers specifically on generative language capabilities. The Dolphin benchmark (Elmadany et al., 2023) exemplifies this by focusing exclusively on Arabic NLG tasks, including summarization, storytelling, dialogue, and data-to-text generation. Comprising 200,000 completions across 20,000 prompts, Dolphin evaluates LLMs like GPT-4, Falcon, and ChatGPT using both hu-

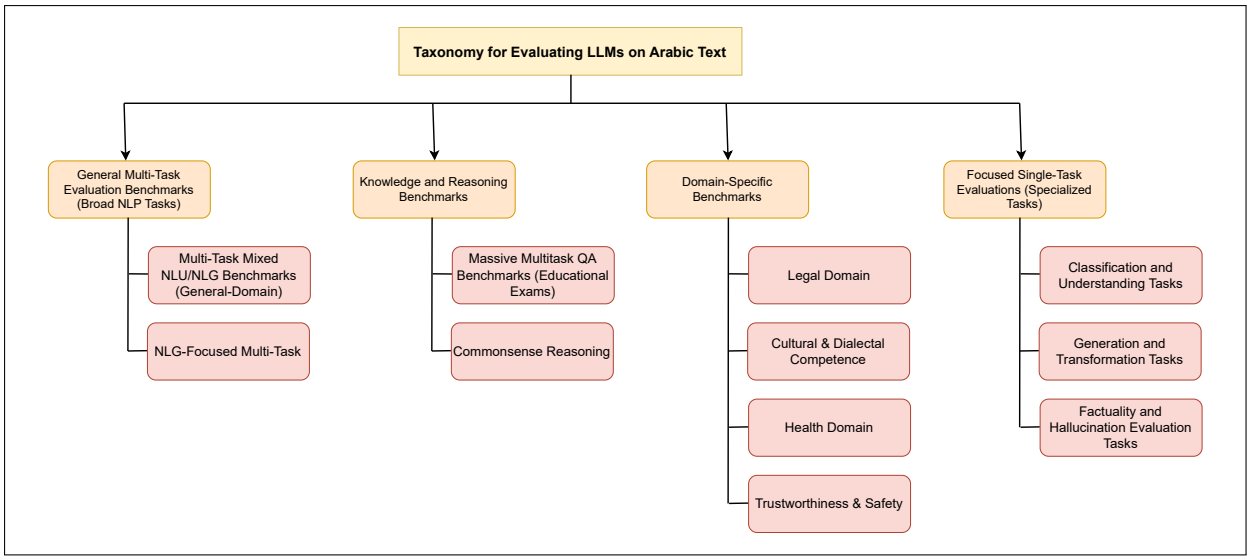

Figure 2: Taxonomy for Evaluating LLMs on Arabic Text

man judgments (e.g., grammaticality, coherence) and automatic metrics (e.g., BLEU, ROUGE-L, COMET). While Dolphin provides a rich resource for assessing generative fluency and factuality, a limitation of considering only NLG is that it overlooks other critical language understanding capabilities, such as reasoning, retrieval, and classification.

Benchmarks under the general multi-task category offer foundational insights into the capabilities of LLMs in Arabic across diverse tasks. However, limitations such as restricted dialectal coverage, model scope, and narrow task focus indicate a need for more comprehensive, balanced, and culturally representative evaluation frameworks.

4.2 Knowledge and Reasoning Benchmarks

Knowledge and reasoning benchmarks aim to assess the depth of logical inference and factual understanding of LLMs beyond basic comprehension. These are typically structured as multi-choice questions (MCQs) or explanatory tasks designed to simulate complex, real-world problem-solving situations.

A primary subcategory is Massive Multitask QA Benchmarks, which assess a model's breadth of knowledge across subjects. For example, ArabicMMLU (Koto et al., 2024) covers 14,575 MCQs across 40 tasks, drawing from real-world school exams in various Arabic-speaking regions. Similarly, AlGhafa (Almazrouei et al., 2023) includes 7,226 MCQs across 45 tasks, categorized into reasoning, knowledge, reading comprehension, and math. Another example is AraSTEM (Mustapha et al., 2024), which focuses on STEM subjects with over 11,000 questions ranging from primary school to college-level. Finally, the Qiyas Benchmark (Al-Khalifa and Al-Khalifa, 2024) evaluates models using questions from the Saudi General Aptitude Test, covering both verbal and mathematical reasoning. These benchmarks offer broad task coverage, but their formats rely entirely on MCQs, which simplify the task structure and may inflate performance by enabling guessing (Koto et al., 2024; Almazrouei et al., 2023; Mustapha et al., 2024; Al-Khalifa and Al-Khalifa, 2024). Such format constraints can limit a model's opportunity to demonstrate deeper reasoning or generative capabilities. Additionally, evaluating only a narrow set of models restricts the ability to offer a comprehensive view of performance across the broader LLM landscape (Al-Khalifa and Al-Khalifa, 2024), including emerging or open-source models. Most benchmarks are also confined to MSA, excluding dialects, informal text, or culturally specific content (Koto et al., 2024; Mustapha et al., 2024; Al-Khalifa and Al-Khalifa, 2024).

The second subcategory, Commonsense Reasoning Benchmarks, evaluates a model's intuitive understanding of everyday scenarios. ArabicSense (Lamsiyah et al., 2025) is a newly proposed benchmark that assesses commonsense validation, explanation selection, and generative explanation. The dataset is synthetically generated and covers a range of reasoning skills. However, it remains limited in scope, focusing only on three task types and lacking the diversity of real-world language use.

Additionally, its synthetic nature may introduce biases or overfitting tendencies not representative of actual human-authored content.

4.3 Domain-Specific Benchmarks

Domain-specific benchmarks are designed to evaluate LLMs on tasks rooted in real-world applications and specialized knowledge areas. Unlike general-purpose benchmarks, which assess broad linguistic competence, these benchmarks target specific domains, such as law, health, cultural reasoning, and safety, to assess how well models handle context-sensitive, factual, and domain-relevant language use. This subsection is organized into four subcategories of domain-specific benchmarks: legal, cultural and dialectal competence, health, and trustworthiness and safety.

In the legal domain, the ArabLegalEval benchmark (Hijazi et al., 2024) provides a multi-task framework designed to evaluate Arabic LLMs' legal reasoning capabilities. It includes over 15,000 instances covering MCQs, open-ended QA, and carefully translated items from the English-language LegalBench dataset. These tasks primarily draw from Saudi legal sources, such as regulations on consumer contracts and privacy policies. While ArabLegalEval provides a rigorous and diverse evaluation setting, its heavy reliance on Saudi legal texts may limit its applicability across broader Arabic legal systems.

Cultural and dialectal competence has emerged as a critical dimension in evaluating LLMs on Arabic text due to the region's linguistic diversity. AraDiCE (Mousi et al., 2025) benchmarks dialectal and cultural understanding across Egyptian, Gulf, Levantine, and MSA. It spans dialect identification, misinformation detection, and cultural reasoning. However, it primarily relies on synthetic data generated via machine translation with post-editing, which may introduce unnatural phrasing.In addition, the omission of key dialects such as Maghrebi limits its regional coverage. The Palm benchmark (Alwajih et al., 2025) offers 17,411 annotated instruction–response pairs covering ten dialects across 20 culturally salient domains. Despite its breadth, Palm exhibits skewed country-level representation. Similarly, the SaudiCulture benchmark (Ayash et al., 2025) evaluates LLMs on region-specific cultural questions within Saudi Arabia, capturing intranational differences across five regions. Nonetheless, its geographic scope limits generalizability to broader Arab cultural contexts. Jawaher benchmark (Magdy et al., 2025) targets proverb translation and explanation in 20 dialects, exposing the limitations of current LLMs in handling idiomatic, figurative, and culturally grounded expressions. However, its evaluation is affected by the use of English-only prompts, which limits the assessment of models' native Arabic comprehension. Lastly, the culturally aligned benchmark (Nacar et al., 2025) critiques the Western bias of traditional evaluation frameworks and introduces ILMAAM, a curated leaderboard tailored to Arabic sociocultural contexts. It improves cultural appropriateness.

In the health domain, the Health Claims benchmark (obaid Alharbi et al., 2025) evaluates GPT-4's ability to classify and verify health-related claims across Saudi, Egyptian, Lebanese, and Moroccan dialects. The study utilizes 329 expert-verified claims from AraFacts and ArCOV19-Rumors, generating 6,520 dialect-specific queries with varying presupposition levels. It applies a novel Cultural Sensitivity Score to measure context-aware accuracy. The benchmark is limited by its evaluation of only a single model (GPT-4), which restricts its comparative utility, and by its narrow dialectal coverage that excludes other widely spoken Arabic varieties.

The domain of trustworthiness and safety is addressed by AraTrust (Alghamdi et al., 2025), which includes 522 multiple-choice questions evaluating LLMs on ethics, legality, offensiveness, and privacy. It introduces evaluations across several prompting settings, including chain-of-thought reasoning. However, the benchmark's exclusive use of multiple-choice formats restricts deeper assessment of models' ethical reasoning in open-ended contexts.

4.4 Focused Single-Task Evaluations

Benchmarks in this category are designed to evaluate LLMs on narrowly defined tasks that test specific competencies in Arabic. Unlike multi-task benchmarks, these evaluations isolate a single task, such as sentiment classification, machine translation, or hallucination detection, allowing for more fine-grained assessment of model performance. This category comprises three major subcategories of tasks: classification and understanding, generation and transformation, and factuality and hallucination detection.

Classification and Understanding Tasks target the ability of LLMs to label and disambiguate text based on semantic, syntactic, or pragmatic cues. In the domain of sentiment classification, Al-Thubaity et al. (2023) evaluated GPT-3.5, GPT-4, and PaLM 2 (Bard AI) using the Saudi Dialect Twitter Corpus, covering a small-scale dataset of 2,690 tweets labeled as positive, negative, or neutral. The benchmark revealed close performance between GPT-4 and fine-tuned BERT baselines, yet it is restricted to a single dialect (Saudi). A benchmark for Cross-Lingual NER was proposed by Al-Duwais et al. (2024) to test six multilingual LLMs using seven datasets across domains like news and social media. The benchmark revealed strong performance by encoder-based models such as XLM-R and mBERT. Abdel-Salam (2024) introduced a benchmark for Word Sense and Location Mention Disambiguation using SALMA and IDRISI-D datasets. While demonstrating LLM competence in controlled zero-shot setups, the benchmark omits dialectal variations and depends heavily on short contexts and English translations for retrieval.

Generation and Transformation Tasks evaluate how well models perform structured text transformation, such as translation or correction. A machine translation benchmark was proposed by Kadaoui et al. (2023) using 1,000 dialectal Arabic sentences across ten varieties from the MADAR corpus. While the benchmark spans several dialects, it includes only two LLMs, ChatGPT and Bard. Kwon et al. (2023) benchmarked LLMs on Arabic grammatical error correction (AGEC) using QALB datasets, evaluating performance with prompting strategies such as zero-shot, few-shot, and instruction tuning. The benchmark highlights LLM underperformance on semantic errors and lacks dialectal diversity. Another example in this group is the punctuation restoration benchmark (Al Wazrah et al., 2025), which uses a curated dataset of 10,046 paragraphs to test seven LLMs and a fine-tuned AraBERT. GPT4-o performed best overall. The benchmark suffers from skewed punctuation distributions.

Factuality and hallucination evaluation tasks assess LLMs' ability to distinguish between true and false claims or to avoid generating fabricated content. Gupta et al. (2025) developed a fact-checking benchmark using 771 claims from the X-Fact dataset, focusing on binary classification with English reasoning strategies applied to Arabic input. The dataset is heavily skewed toward false claims and excludes recent advanced models like GPT-4, which limits longitudinal comparisons. In the area of hallucination detection, the Halwasa benchmark (Mubarak et al., 2024) evaluates Arabic hallucinations using 10,000 synthetic factual sentences generated by LLMs. The dataset was created using 1,000 randomly selected words from the SAMER Arabic readability lexicon. For each word, both GPT-3.5 and GPT-4 were prompted to generate ten factual Arabic sentences. After filtering out duplicates and invalid outputs, five unique sentences were retained per model, resulting in 5,000 sentences from each and a total of 10,000 sentences. Each sentence was manually annotated by trained human annotators across four dimensions: (1) whether it makes a verifiable factual claim, (2) whether the claim is factually correct, (3) whether the sentence follows proper Arabic grammar, and (4) the reference sources used for factual verification. A key limitation of this benchmark is its exclusive focus on just two models, GPT-3.5 and GPT-4, which restricts its comparative scope across a broader range of Arabic or multilingual LLMs. Similarly, HalluVerse25 (Abdaljalil et al., 2025) is a multilingual hallucination detection benchmark that includes 828 Arabic sentence pairs focused on biographical content. While it supports cross-lingual comparison, the benchmark inherits potential biases from Wikidata and the use of GPT-generated data, constraining its generalizability beyond the biographical domain.

5 Critical Analysis of Existing Arabic LLM Evaluations

Despite significant advancements in evaluating LLMs for Arabic text, existing benchmarks reveal several critical challenges. Specifically, there is a pronounced absence of separate intrinsic and extrinsic evaluations. Currently, benchmarks frequently blend these tasks into general multi-task evaluations, making it difficult to comprehensively assess specific competencies such as linguistic understanding, factual reasoning, and cultural awareness. This methodological conflation fails to provide a clear diagnostic of a model's performance, particularly in distinguishing whether success is driven by deep comprehension or surface-level task handling.

Another considerable limitation lies in the limited scope of model evaluations. Most benchmarks evaluate only a narrow set of LLMs, predominantly

focusing on well-known models such as GPT variants, neglecting emerging or specialized Arabic-centric models. Consequently, this narrow selection restricts the ability to address crucial comparative questions, such as identifying which models excel in specific tasks. Moreover, the scarcity of comparative analyses across a broader spectrum of models limits insights into model scalability and adaptability in diverse Arabic linguistic environments.

Additionally, the prevalent reliance exclusively on MCQs in many benchmarks represents another critical limitation. Solely using MCQs inherently simplifies evaluation tasks, potentially inflating model performance by allowing for guessing and limiting the ability to assess more sophisticated generative or explanatory capabilities.

In parallel with these methodological considerations, it is equally important to situate Arabic within the broader multilingual evaluation landscape. While this survey focuses on Arabic benchmarks for evaluating LLMs, understanding how Arabic is represented across cross-lingual benchmarks provides valuable context. Several cross-lingual benchmarks, such as XTREME (Hu et al., 2020), XGLUE (Liang et al., 2020), Blend (Myung et al., 2024), and (Chollampatt et al., 2025), include Arabic alongside other languages, often as a representative of Semitic or low-resource linguistic groups. However, these benchmarks typically offer limited task coverage for Arabic and rarely account for the linguistic diversity within the language, such as dialectal variation or cultural specificity. In contrast, Arabic-specific benchmarks provide more fine-grained evaluations tailored to the complexities of Arabic, including dialect identification, cultural reasoning, and script variants. Moreover, while cross-lingual benchmarks are valuable for assessing generalization and transfer learning, they often rely on translated or parallel data that may not reflect authentic language use. Arabic-centric benchmarks, by contrast, frequently involve native-authored content and culturally grounded tasks, offering a more accurate assessment of LLM performance on Arabic.

In addition to broadening evaluation contexts, this survey primarily focuses on benchmarking coverage and evaluation frameworks, we acknowledge the importance of analyzing bias in LLMs more explicitly. Several Arabic benchmarks, such as AraTrust and Palm, begin to address dimensions of bias related to ethics, offensiveness, and regional representation. However, most existing datasets lack systematic annotations for sensitive attributes like gender, dialect, or sociopolitical context, making it difficult to assess fairness across subpopulations. Furthermore, benchmarks that rely on machine-translated or synthetic data may introduce unintended cultural or linguistic biases.

6 Conclusion and Future Directions

This survey has provided a comprehensive overview of existing benchmarks for evaluating LLMs on Arabic text, highlighting both significant progress and critical gaps. While current benchmarks offer valuable insights across various linguistic tasks and domains, they often conflate intrinsic and extrinsic evaluations, focus narrowly on a limited set of popular models, and rely heavily on simplified formats such as multiple-choice questions. Moreover, the underrepresentation of Arabic dialects and cultural nuances limits the applicability of these evaluations to the diverse Arabic language landscape. Bias and fairness considerations remain insufficiently addressed in most datasets, posing challenges for equitable model assessment.

To advance the field, future research should explicitly differentiate intrinsic language-specific evaluations (e.g., syntactic parsing, semantic understanding, morphological analysis) from extrinsic task-based assessments focused on real-world applications such as healthcare, law, and education. Expanding the range of evaluated models to include emerging, open-source, and Arabic-centric LLMs will enhance comparative analyses and foster innovation tailored to Arabic's unique linguistic characteristics.

Future benchmarks must incorporate diverse, realistic datasets reflecting dialectal variety and cultural context to improve real-world relevance. The growing importance of prompt engineering calls for systematic exploration of prompt formulations in both Arabic and English to optimize model performance and reliability. Additionally, incorporating bias-sensitive design principles and targeted fairness metrics is essential to ensure equitable evaluation across dialects, regions, and sociolinguistic groups.

Overall, addressing these methodological and practical gaps will deepen understanding of how LLMs perform on Arabic text and guide the development of more robust, culturally aware, and

effective language technologies.

Limitations

This survey is limited by its exclusive focus on publicly documented academic benchmarks, omitting proprietary or industrial evaluations that may provide additional perspectives.

References

Samir Abdaljalil, Hasan Kurban, and Erchin Serpedin. 2025. Halluverse25: Fine-grained multilingual benchmark dataset for llm hallucinations.

Reem Abdel-Salam. 2024. rematchka at arabicnlu2024: Evaluating large language models for arabic word sense and location sense disambiguation. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 383–392.

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, et al. 2023. Larabench: Benchmarking arabic ai with large language models. *arXiv preprint arXiv:2305.14982*.

Muhammad Al-Barham, Imad Afyouni, Khalid Almubarak, Ayad Turky, Ibrahim Abaker Targio Hashem, Ali Bou Nassif, Ismail Shahin, and Ashraf Elnagar. 2025. Unlocking language boundaries: Araclip-transforming arabic language and image understanding through cross-lingual models. *Engineering Applications of Artificial Intelligence*, 151:110577.

Mashael Al-Duwais, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2024. A benchmark evaluation of multilingual large language models for arabic cross-lingual named-entity recognition. *Electronics*, 13(17):3574.

Shahad Al-Khalifa and Hend Al-Khalifa. 2024. The qiyas benchmark: Measuring chatgpt mathematical and language understanding in arabic. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 343–351.

Abdulmohsen Al-Thubaity, Sakhar Alkhereyf, Hanan Murayshid, Nouf Alshalawi, Maha Omirah, Raghad Alateeq, Rawabi Almutairi, Razan Alsuwailem, Manal Alhassoun, and Imaan Alkhanen. 2023. Evaluating chatgpt and bard ai on arabic sentiment analysis. In *Proceedings of ArabicNLP 2023*, pages 335–349.

Asma Ali Al Wazrah, Afrah Altamimi, Hawra Aljasim, Waad Alshammari, Rawan Al-Matham, Omar Elnashar, Mohamed Amin, and Abdulrahman AlOsaimy. 2025. Evaluation of large language models on arabic punctuation prediction. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 144–154.

Emad A Alghamdi, Reem Masoud, Deema Alnuhait, Afnan Y Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. 2025. Aratrust: An evaluation of trustworthiness for llms in arabic. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8664–8679.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, et al. 2023. Alghafa evaluation benchmark for arabic language models. In *Proceedings of ArabicNLP 2023*, pages 244–275.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, Abdelrahim A Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, et al. 2025. Palm: A culturally inclusive and linguistically diverse dataset for arabic llms. *arXiv preprint arXiv:2503.00151*.

Lama Ayash, Hassan Alhuzali, Ashwag Alasmari, and Sultan Aloufi. 2025. Saudiculture: A benchmark for evaluating large language models cultural competence within saudi arabia. *arXiv preprint arXiv:2503.17485*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Shamil Chollampatt, Minh Quang Pham, Sathish Reddy Indurthi, and Marco Turchi. 2025. Cross-lingual evaluation of multilingual text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7766–7777.

AbdelRahim Elmadany, Muhammad Abdul-Mageed, et al. 2022. Arat5: Text-to-text transformers for arabic language generation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long papers)*, pages 628–647.

Abdelrahim Elmadany, Ahmed El-Shangiti, Muhammad Abdul-Mageed, et al. 2023. Dolphin: A challenging and diverse benchmark for arabic nlg. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1404–1422.

Ashraf Elnagar, Sane Yagi, Youssef Mansour, Leena Lulu, and Shehdeh Fareh. 2023. A benchmark for evaluating arabic contextualized word embedding models. *Information Processing & Management*, 60(5):103452.

Ayushman Gupta, Aryan Singhal, Thomas Law, Veekshith Rao, Evan Duan, and Ryan Luo Li. 2025. Can llms verify arabic claims? evaluating the arabic fact-checking abilities of multilingual llms. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 104–113.

Faris Hijazi, Somayah Alharbi, Abdulaziz AlHussein, Harethah Shairah, Reem Alzahrani, Hebah Alshamlan, George Turkiyyah, and Omar Knio. 2024. Arablegaleval: A multitask benchmark for assessing arabic legal knowledge in large language models. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 225–249.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

Karima Kadaoui, Samar M Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties. *arXiv preprint arXiv:2308.03051*.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *arXiv preprint arXiv:2305.14976*.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, et al. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *arXiv preprint arXiv:2402.12840*.

Sang Yun Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Beyond english: Evaluating llms for arabic grammatical error correction. *arXiv preprint arXiv:2312.08400*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189.

Salima Lamsiyah, Kamyar Zeinalipour, Matthias Brust, Marco Maggini, Pascal Bouvry, Christoph Schommer, et al. 2025. Arabicsense: A benchmark for evaluating commonsense reasoning in arabic with large language models. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 1–11.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, et al. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pretraining, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Samar M Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. Jawaher: A multidialectal dataset of arabic proverbs for llm benchmarking. *arXiv preprint arXiv:2503.00231*.

Malak Mashaabi, Shahad Al-Khalifa, and Hend Al-Khalifa. 2025. A survey of large language models for arabic language and its dialects.

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.

Hamdy Mubarak, Hend Al-Khalifa, and Khaloud Suliman Alkhalefah. 2024. Halwasa: Quantify and analyze hallucinations in large language models: Arabic as a case study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8008–8015.

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. *arXiv preprint arXiv:2501.00559*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Omer Nacar, Serry Taiseer Sibaee, Samar Ahmed, Safa Ben Atitallah, Adel Ammar, Yasser Alhabashi, Abdulrahman S Al-Batati, Arwa Alsehibani, Nour Qandos, Omar Elshehy, et al. 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.

Abdulsalam obaid Alharbi, Abdullah Alsuhaibani, Abdulrahman Abdullah Alalawi, Usman Naseem, Shoaib Jameel, Salil Kanhere, and Imran Razzak. 2025. Evaluating large language models on health-related claims across arabic dialects. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 95–103.

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

Haneh Rhel and Dmitri Roussinov. 2025. Large language models and arabic content: a review. *arXiv preprint arXiv:2505.08004*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Sane Yagi, Ashraf Elnagar, and Shehdeh Fareh. 2023. A benchmark for evaluating arabic word embedding models. *Natural Language Engineering*, 29(4):978–1003.

Sane Yagi, Shehdeh Fareh, Ashraf Elnagar, Mariam Balajeed, Abdalla El-mneizel, and Mohammad Al-Badawi. 2024. Is arabic punctuation rule-governed? *Cogent Arts & Humanities*, 11(1):2303818.

Wenli Yang, Lilian Some, Michael Bain, and Byeong Kang. 2025. A comprehensive survey on integrating large language models with knowledge-based methods. *Knowledge-Based Systems*.

A Overview of Arabic LLM Benchmarks

Table 1: Overview of Arabic LLM Benchmarks

Benchmark	Year	LLMs Evaluated	Task(s)	Dataset(s) Description
AraT5 (Elmadany et al., 2022)	2022	AraT5 (Small, Base, Large, XL), mT5, mBART, AraGPT2, MARBERT	8 tasks: Text Classification, Sentiment Analysis, Named Entity Recognition (NER), Extractive Question Answering (QA), Paraphrasing, Summarization, Headline Generation, Text Simplification	Data collected from 8 diverse Arabic sources including Arabic Wikipedia, OSCAR, OPUS, Tashkeela, SLSA, and others; resulting in a corpus of 200M sentences (50GB); preprocessed into a text-to-text format.
Beyond English (Kwon et al., 2023)	2023	GPT-4, ChatGPT-3.5 Turbo, LLaMA-7B, Vicuna-13B, Bactrian-Xbloom-7B, Bactrian-Xllama-7B	Grammatical Error Correction (GEC)	QALB-2014 (L1), QALB-2015 (L1 & L2)
Dolphin (Elmadany et al., 2023)	2023	Falcon-40B-Instruct, Falcon-180B-Chat, GPT-3.5-Turbo, GPT-4, ChatGPT	10 NLG tasks: dialogue generation, question answering, data-to-text, storytelling, summarization, translation, paraphrasing, definition generation, classification, correction/refinement (includes code-switching and Arabizi)	20K Arabic prompts with 200K completions, covering diverse topics and language forms, including Modern Standard Arabic, dialects, code-switched inputs, and Arabizi; prompts created by native speakers and aligned with high-quality completions
Evaluating ChatGPT and Bard AI on Arabic Sentiment Analysis (Al-Thubaity et al., 2023)	2023	GPT-3.5, GPT-4, Bard AI (PaLM 2)	Sentiment Analysis (Classification & Generation)	Saudi Dialect Twitter Corpus (SDTC): 2,690 used (558 positive, 1,632 negative, 500 neutral)
Evaluation of Bard and ChatGPT on MT (Kadaoui et al., 2023)	2023	ChatGPT (GPT-3.5-turbo), Bard	Machine Translation	1,000 sentences from 10 Arabic dialects (100 per dialect) from the MADAR corpus, with corresponding MSA and English translations
GPTAraEval (Khondaker et al., 2023)	2023	ChatGPT-3.5, ChatGPT-4	Text classification, natural language inference (NLI), question answering (QA), paraphrase identification, sentiment analysis, named entity recognition (NER), topic classification, hate speech detection, offensive language detection, dialect identification, translation, coreference resolution, headline generation, text summarization	60 Arabic datasets covering Modern Standard Arabic and multiple Arabic dialects; varying in domain, size, and complexity; formatted for zero-shot prompt-based evaluation
AlGhafa (Almazrouei et al., 2023)	2023	AraT5, CAMeLBERT, mBERT, mGPT, GPT-3.5-turbo, AraGPT2-Mega, Noor-10B, Jais-13B, lGhafa-1B/3B/7B/14B	45 tasks across 4 categories: knowledge, reasoning, reading comprehension, math & coding	7,226 multiple-choice questions from diverse Arabic sources across linguistic and domain topics
A Benchmark Evaluation of Multilingual LLMs for Arabic Cross-Lingual NER (Al-Duwais et al., 2024)	2024	mBERT, XLM-R, BERTIN, ByT5, BLOOM, mT0	NER	7 Arabic NER datasets: ANERcorp, AQMAR, CAMeL, WikiFANE, Winerz, Arman, ArapTweet; domains: news, Wikipedia, social media
ArabicMMLU (Koto et al., 2024)	2024	GPT-3.5, GPT-4, BLOOMZ, mT0, LLaMA2, Falcon, XGLM, AraT5, AraGPT2, AceGPT, Jais (total 35 models)	Knowledge tasks	14,575 Arabic multiple-choice questions from school exams in 8 Arabic-speaking countries;
ArabLegalEval (Hijazi et al., 2024)	2024	GPT-4, GPT-4o, GPT-3.5, Claude-3 Opus, Command R, Command R Plus, Llama3 (8B & 70B), Aya-101, Jais	MCQs, Open-ended QA, LegalBench QA (Consumer Contracts, Contracts, Privacy QA/Entailment)	10,583 Arabic MCQs (from MoJ & BoE), 492 Najiz FAQs, 15,804 translated LegalBench samples, ArabicMMLU subset for legal reasoning benchmarking.
ARADICE (Mousi et al., 2025)	2025	Jais-13B, AceGPT-13B, Llama-3-8B, Mistral-7B, Fanar-8.7B, Qwen2.5-7B, Gemma2-9B, Aya-8B	Dialect Identification, Dialect Generation, Machine Translation, Commonsense Reasoning, World Knowledge, Reading Comprehension, Misinformation Detection, Cultural Understanding	45K+ post-edited examples across QADI, ADI, ADD, MADAR, ArabicMMLU, PIQA, OBQA, Winogrande, BoolQ, Belebele, TruthfulQA, and AraDiCE-Culture
AraSTEM (Mustapha et al., 2024)	2024	AraT5, AraGPT2, MT0 (Small, Base, Large), XGLM (1.7B–7.5B), Bloomz (560M–7B1), AceGPT (7B, 13B), LLaMA (2 & 3.1), Falcon (7B, 40B), Jais (13B, 30B)	Zero-shot multiple-choice answering	11,637 Arabic MCQs in STEM (math, biology, physics, IT, chemistry, pharmacy, medicine, dentistry); levels: primary, secondary, college; sourced via scraping, manual extraction, OCR from PDFs; annotated with source traceability
AraTrust (Alghamdi et al., 2025)	2025	GPT-3.5 Turbo, GPT-4, AceGPT 7B, AceGPT 13B, Jais 13B	Trustworthiness evaluation	522 multiple-choice questions across 8 categories (Truthfulness, Ethics, Physical Health, Mental Health, Unfairness, Illegal Activities, Privacy, Offensive Language) and 34 subcategories, all human-written
Halwasa (Mubarak et al., 2024)	2024	GPT-3.5, GPT-4	Factual sentence generation to evaluate models' hallucinations	10K Arabic sentences (5K/model) generated using 1,000 random words from the SAMER corpus, annotated for factuality, correctness, linguistic errors, and references
LAraBench (Abdelali et al., 2023)	2023	GPT-3.5-turbo, GPT-4, BLOOMZ, Jais-13b-chat, Whisper, USM	33 tasks across NLP and Speech	61 publicly available datasets; 296K samples; 46h speech; 30 TTS sentences; covers MSA and dialects, across genres like news, tweets, telephony

Table 1 – continued from previous page

Benchmark	Year	LLMs Evaluated	Tasks	Dataset(s) Description
Arabic Word/Location Sense Disambiguation (Abdel-Salam, 2024)	2024	LLama3, LLama3-Instruct, WizardLM-2, AceGPT-7B, OpenChat	Word Sense Disambiguation (WSD), Location Mention Disambiguation (LMD)	WSD: SALMA corpus (100 train, 1,340 test); LMD: IDRISI-D (2,170 train, 333 val, 791 test)
The Qiyas Benchmark (Al-Khalifa and Al-Khalifa, 2024)	2024	ChatGPT-3.5-turbo, ChatGPT-4, Gemini-pro (partial)	Mathematical reasoning and Language understanding	2,407 multiple-choice questions derived from Saudi Arabia's Qiyas GAT. Includes math, geometry, algebra, statistics, and five types of verbal tasks
Jawaher (Magdy et al., 2025)	2025	Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct, Gemma-2-9B-it, GPT-4o, Gemini 1.5 Pro, Claude 3.5 Sonnet, Cohere Command R+	Translation, Explanation	10,037 Arabic proverbs from 20 dialects with idiomatic/literal English translations, Arabic and English explanations.
ArabicSense (Lamsiyah et al., 2025)	2025	Gemma, LLaMA-3, Mistral-7b	Commonsense Validation, Multiple-Choice Explanation, Generative Explanation	3954 train, 848 val, 848 test samples per task from Arabic Wikipedia, generated using GPT-4
Arabic Fact-Checking (Gupta et al., 2025)	2025	Llama 3 8B, Llama 3 70B, GPT-3.5-turbo, Gemini 1.0 Pro	Arabic fact-checking (binary classification: true/false)	771 Arabic claims from X-Fact dataset (filtered for 'true' or 'false' only; 730 false, 41 true)
Health-Related Claims Across Arabic Dialects (obaid Alharbi et al., 2025)	2025	GPT-4	Health claim verification across dialects	329 claims (191 from AraFacts + 138 from ArCOV19-Rumors), categorized as true, false, mixed
Evaluation of LLMs on Arabic Punctuation Prediction (Al Wazrah et al., 2025)	2025	GPT4-o, Gemini 1.5, JAIS-13B, AceGPT-13B, SILMA-9B, ALLaM-1, CommandR+, AraBERT	Arabic punctuation prediction	10,046 annotated Arabic paragraphs from 25 books, manually cleaned and tokenized, covering six punctuation marks; split into training, validation, and test sets
HalluVerse25 (Abdaljalil et al., 2025)	2025	GPT-4o, GPT-4o-mini, phi-4, PaLM 2, Mistral-7b, Qwen-2.5 (7b, 72b), LLaMA-3.3, Gemini, Gemma	Hallucination Detection	3116 factual + hallucinated pairs (biography-based) in English, Arabic, Turkish
Palm (Alwajih et al., 2025)	2025	GPT-4o, Claude-3.5-Sonnet, Command R+ (104B), Qwen2.5-72B, Qwen2.5-7B, Qwen2.5-3B, Qwen2.5-1.5B, JAIS-13B, AceGPT-v2-32B, AceGPT-v2-8B, LLaMA-3.1-70B, LLaMA-3.1-8B, LLaMA-3.2-3B, LLaMA-3.2-1B, Gemma-2-27B, Gemma-2-9B, Gemma-2-2B, Phi-3.5-mini (18 models)	To benchmark LLMs' capabilities in culturally-aware and dialect-specific instruction following across the Arab world. It evaluates LLMs' ability to understand and generate culturally relevant, linguistically appropriate responses in Arabic dialects and MSA.	17,411 human-authored Arabic instruction–response pairs (MSA and 10 dialects) across 22 Arab countries and 20 cultural domains; includes train, public test, and private test splits
SaudiCulture (Ayash et al., 2025)	2025	GPT-4, Llama 3.3, FANAR, Jais, AceGPT	Cultural understanding, QA (open-ended, single-answer, and multi-answer formats)	SaudiCulture: 441 questions across 5 Saudi regions + general, covering 8 cultural domains (food, clothing, celebrations, etc.) in open-ended, single-answer, and multi-answer formats
Towards Inclusive Arabic LLMs (Nacar et al., 2025)	2025	Qwen2.5-72B-Instruct, CohereForAI/aya-expanse-32b, Qwen2.5-32B-Instruct, Google/Gemma-2, SILMA-9B, FreedomIntelligence/AceGPT, JAIS-family, LLaMA models	Multitask Language Understanding	Refined Arabic MMLU benchmark with over 14,000 questions, including 2,466 culturally sensitive questions and 766 culturally enriched additions (e.g., Islamic Religion, Islamic Ethics, Old Arab History).