Evaluating Large Language Models in Arabic Tasks: A Survey of Benchmarks, Methods, and Gaps

Anonymous ACL submission

Abstract

This survey provides the first systematic review of Arabic benchmarks that target the evaluation of LLMs with Arabic support. It showcases an analysis of 40+ Arabic evaluation benchmarks across NLP tasks, knowledge domains, cultural understanding, and specialized capabilities. A taxonomy organizing benchmarks into four categories is proposed: Knowledge, NLP Tasks, Culture and Dialects, and Target-Specific evaluations. The analysis reveals significant progress in Arabic benchmark diversity while identifying critical gaps: limited temporal evaluation, insufficient multi-turn dialogue assessment, and cultural misalignment in translated datasets. Three primary approaches are examined: native collections, translated collections, and synthetically generated collections. Their trade-offs regarding authenticity, scale, and cost are discussed. This work serves as a comprehensive reference for Arabic NLP researchers, providing insights into benchmark methodologies, reproducibility standards, and evaluation metrics while offering recommendations for future development.

1 Introduction

In recent years, Large Language Models (LLMs) have achieved major advances in natural language understanding and reasoning, moving closer to the vision of AGI (Naveed et al., 2025). Since the transformer era, LLMs have demonstrated strong multilingual capabilities beyond English (Huang et al., 2025). Arabic, spoken by nearly 500 million people worldwide[1] , underscores the importance of multilingual evaluation. Consequently, both Arabic-specific and multilingual LLMs with Arabic support have been released in open- and closed-source settings (Al-Khalifa et al., 2025). Robust benchmarks are essential for systematic evaluation, and in parallel with progress in Arabic-capable LLMs, substantial effort has been devoted to Arabic benchmarking.

Arabic benchmark development faces distinct challenges. Data scarcity and limited diversity in Arabic web content (Al-Khalifa et al., 2025) increase the cost of dataset creation. To address this, researchers rely on translation from English, synthetic data generation using LLMs, and native Arabic data collection. Each approach involves trade-offs: translation often leads to cultural misalignment, while synthetic data risks bias and circular evaluation. Both require extensive human validation and remain vulnerable to cultural misalignment (Nacar et al., 2025).

These challenges are amplified by Arabic's linguistic diversity across Modern Standard Arabic and numerous regional dialects (Keleg et al., 2025). Hence, Arabic demands dedicated benchmarking analysis due to its rich morphology, more than 20 dialects that function almost as separate languages, strong cultural sensitivity that limits translation-based evaluation, persistent data scarcity, and a fragmented evaluation landscape absent in English-centric benchmarks.

While recent surveys have analyzed Arabic-capable LLMs from various perspectives (Mashaabi et al., 2024; Al-Khalifa et al., 2025; Rhel and Roussinov, 2025), none provides a focused and comprehensive treatment of Arabic benchmarks. This survey fills that gap by systematically reviewing Arabic evaluation datasets and methodologies. Our contributions are threefold: (1) we introduce a taxonomy organizing benchmarks into four categories; (2) we analyze over 40 Arabic benchmarks; and (3) we examine evaluation practices, tools, trends, and critical gaps. We release a companion repository[2] consolidating benchmark datasets, code, and frameworks.

[1] World Bank (2024). Population, Total – Arab World. https://data.worldbank.org/indicator/SP.POP.TOTL?locations=1A

[2] Anonymous-github-Link

The paper is organized as follows. Section 2 reviews Arabic-supporting LLMs and related surveys. Section 3 introduces the taxonomy, with Sections 4–7 detailing each category. Section 8 discusses open challenges, and Section 9 concludes.

2 Background

This section provides the necessary context for understanding the Arabic benchmarking landscape for LLMs. We first categorize existing LLMs that support Arabic or are Arabic-specific, by their training approaches, then review related work to position our contribution within the broader literature.

2.1 LLMs with Arabic Support

State-of-the-art LLMs with Arabic support can be categorized into three types. **Native models** are trained from scratch exclusively on Arabic data, enabling Arabic-only interaction. Examples include Jais (Sengupta et al., 2023), ArabianGPT (Koubaa et al., 2024), and AraGPT (Antoun et al., 2021). **Multilingual models** support multiple languages including Arabic, such as Qwen3 (Yang et al., 2025), Gemma3 (Team et al., 2025b), and Llama (Grattafiori et al., 2024), as well as closed-source models like ChatGPT (OpenAI, 2023) and Claude (Anthropic, 2024). **Adapted Arabic models** apply continued pretraining or supervised fine-tuning to existing multilingual models to enhance Arabic performance, exemplified by AceGPT (Huang et al., 2024), SILMA (Team, 2024), Fanar (Team et al., 2025a), and Falcon-Arabic (TII, 2025).

2.2 Related Work

Several surveys have examined LLMs that have Arabic capabilities. Mashaabi et al. (2024) provided an in-depth discussion of pretraining and fine-tuning data for LLMs with Arabic support, including dialectal coverage, and listed available models with details on accessibility and reproducibility. Rhel and Roussinov (2025) surveyed pretrained Arabic LLMs with focus on classical NLP applications and benchmarks. Most recently, Al-Khalifa et al. (2025) presented the historical evolution of Arabic NLP, common pretraining and fine-tuning strategies, and current research trends and challenges. While Al-Khalifa et al. (2025) briefly discussed a subset of existing benchmarks, no comprehensive survey of Arabic benchmarks exists. This work fills that gap by systematically reviewing evaluation techniques and benchmarking datasets for LLMs with Arabic support.

3 Benchmarks Taxonomy

We reviewed 40+ existing Arabic benchmarks and constructed a taxonomy that captures various themes and categories, as depicted in Figure 1. Before introducing the taxonomy, we present the research methodology and the criteria used for benchmark inclusion.

3.1 Inclusion Criteria

While surveying the current state of benchmarks used to evaluate LLMs in Arabic tasks, we devised inclusion criteria that determined which benchmarks would be discussed in this survey paper. All existing works mentioned fall under one of the following categories, which we discuss below:

- Existing Arabic benchmarks introduced in academic papers that evaluate specific Arabic capabilities in LLMs. Public availability of the benchmark dataset or evaluation pipeline is not a requirement for inclusion.
- Arabic benchmarks that lack a detailed technical report but are well established in the field, as evidenced by their use in evaluating released LLMs or their presence in well-known leaderboards.
- Arabic subsets of established multilingual benchmarks, such as MMLU and EXAMS.

3.2 Taxonomy

Our proposed taxonomy organizes benchmarks into four categories:

Knowledge includes benchmarks evaluating general knowledge and STEM capabilities, along with domain-specific benchmarks in fields such as law and medicine.

Natural Language Processing (NLP) encompasses early task-specific benchmarks and comprehensive multi-task benchmarks, reflecting the evolution from narrow task evaluation to unified assessment across diverse dialects and domains.

Culture and Dialects groups benchmarks assessing cultural knowledge and dialect understanding, addressing the essential property of cultural awareness in LLMs with Arabic capabilities.

Target-Specific covers benchmarks designed to assess particular LLM properties such as safety, hallucination detection, instruction-following, and vision capabilities.

This taxonomy emerged from analyzing common patterns across benchmarks and reflects the

evolution from task-specific evaluation to comprehensive assessment. In the following sections, we describe each category and we provide a thorough table (Tables 6) listing all discussed benchmarks with their characteristics.

4 Knowledge

Paper	Primary Topic (Type)	Total
MMLU_ar	General Knowledge (MCQ)	15k
EXAMS_ar	General Knowledge (MCQ)	0.5k
ArabicMMLU	General Knowledge (MCQ)	14.57k
AraSTEM	STEM (MCQ)	11.63k
GAT	Linguistic Ability (MCQ)	0.56k
Qiyas	Linguistics & Math (MCQ)	2.4k
ArabLegalEval	Law (MCQ, GEN)	26.4k
AraMed	Medical (GEN)	270k
MizanQA	Law (MCQ)	1.7k
Fann or Flop	Poetry (GEN)	6.9k
GATmath, GATLc	Linguistics & Math (MCQ)	9k
3LM	STEM & Coding (MCQ)	3.1k
Arabic-GSM8K	Mathematics (GEN)	8.9k
MedArabiQ	Medical (MCQ)	0.7k
Hajj-FAQ	Religious QA (QA)	1.2k

Table 1: Arabic knowledge benchmarks summarized at a high level.

This section examines benchmarks evaluating LLMs' acquired knowledge and reasoning capabilities, covering both general and STEM topics as well as specialized domains such as law, medicine, and poetry, as summarized in Table 1.

4.1 General and STEM

Multilingual efforts produced early general knowledge benchmarks with Arabic components. MMLU_ar (Hendrycks et al., 2020) comprises 14,079 human-translated MCQs spanning 57 subjects across difficulty levels, while EXAMS_ar (Hardalov et al., 2020) contains 562 high-school exam questions covering physics, chemistry, and biology. However, these suffer from translation concerns or limited scale.

ArabicMMLU (Koto et al., 2024) addressed these limitations with 14,575 native Arabic MCQs curated from educational exams across Arab countries, covering all school levels plus university, spanning STEM, social sciences, humanities, and Arabic language understanding. AraSTEM (Mustapha et al., 2024) further specialized in STEM with 11,637 native MCQs which is 7,000 more STEM samples than ArabicMMLU though it remains unpublished despite evaluation results on open-weight models.

The 3LM suite (Boussaha et al., 2025) combines three benchmarks totaling 3,151 questions: 3LM_nat from Arabic STEM exams, 3LM_syn synthetically generated via Yourbench's pipeline (Shashidhar et al., 2025) using Qwen3-235B-A22B, and 3LM_code comprising Arabic translations of HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) instructions and comments.

Several benchmarks leverage Saudi Arabia's General Aptitude Test (GAT), which assesses verbal abilities (reading comprehension, contextual errors, sentence completion) and quantitative skills (arithmetic, algebra, geometry, data analysis). Early efforts (Alkaoud, 2024; Al-Khalifa and Al-Khalifa, 2024) used 456 and 2,407 GAT samples respectively but lacked reproducibility and scale, evaluating only GPT-3.5 and GPT-4 with limited shots. GATmath and GATLc (AlBallaa et al., 2025) addressed these issues with 7k and 9k samples respectively (16k total), publicly released with 5-shot evaluation on diverse Arabic and bilingual LLMs. Arabic-GSM8K (Omartificial-Intelligence-Space, 2025) provides human-validated translations of the established GSM8K benchmark (Cobbe et al., 2021), assessing middle-school mathematical reasoning through 5-shot settings.

4.2 Domain Knowledge

Legal Domain. ArabLegalEval (Hijazi et al., 2024) pioneered Arabic legal LLM evaluation using documents scraped from Saudi Arabia's Ministry of Justice and Board of Experts websites. The benchmark employs three approaches: synthetic MCQ generation using GPT-4 and Claude-3-opus with in-context examples from ArabicMMLU's law section, QA pairs from governmental FAQ sections, and machine-translated datasets from Legalbench (Guha et al., 2023) verified by legal experts. MIZANQA (Bahaj and Ghogho, 2025) extends legal evaluation to Moroccan law using MCQs from various law exams.

Poetry and Linguistics. Fann or Flop (Alghallabi et al., 2025) assesses poetry understanding through 6,984 poem-explanation pairs, evaluating metaphorical and figurative comprehension. Evaluation uses BLEU, chrF(++), BERTScore, and mDeBERTaV3 for character-level overlap and semantic alignment, with GPT-4o as judge for faithfulness, grammatical correctness, and interpretive depth.

Medical Domain. Al-Majmar et al. (2024) introduced 808k medical QA samples from the AlTibbi patient-doctor forum[3]. AraMed (Alasmari et al., 2024) refined this to 270k high-quality samples

[3] https://altibbi.com/

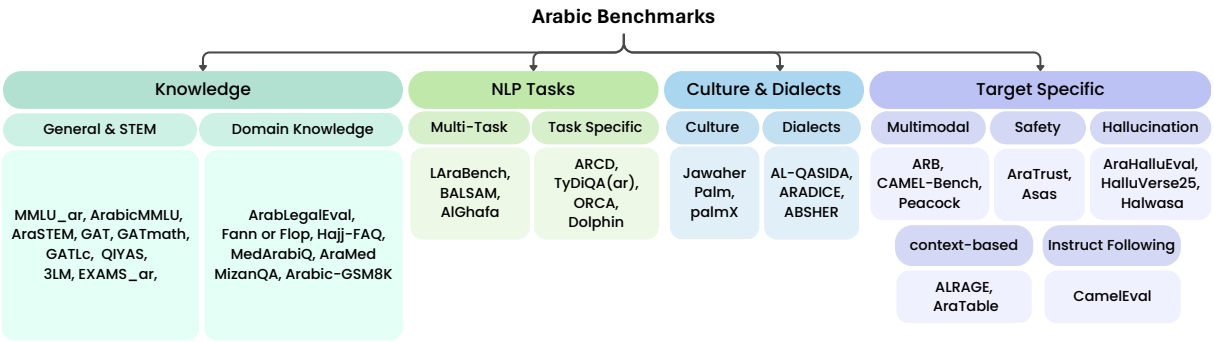

Figure 1: Taxonomy of Arabic Benchmarks.

based on vote counts. MedArabiQ (Daoud et al., 2025) further curated 100 AraMed samples, enhanced them through grammatical correction and LLM modification, and added medical exam questions (MCQs and fill-in-the-blank), yielding 700 samples evaluating Arabic medical knowledge.

Religious Domain. Hajj-FAQ (Aleid and Azmi, 2025) addresses religious knowledge through question-answering on Hajj fatwas (Islamic legal rulings), providing a benchmark for LLMs' understanding of Islamic jurisprudence and pilgrimage-related guidance. This specialized dataset evaluates models' ability to handle religiously and culturally sensitive content requiring both linguistic competence and domain-specific religious knowledge.

5 NLP Tasks

Paper	Primary Task (Type)	Total
ARCD	Reading Comprehension (MCQ)	1.4k
TyDiQA (Ar)	Reading Comprehension (GEN)	14k
ORCA	Multi-task NLP (BENCH)	588k
Dolphin	Multi-task NLP (BENCH)	2M
Alghafa	Reading Comprehension (MCQ)	33.2k
LAraBench	Multi-task NLP (BENCH)	296k
BALSAM	Multi-task NLP (BENCH)	52k

Table 2: Arabic NLP benchmarks covering core language understanding and generation tasks.

LLM evaluation on fundamental NLP tasks has evolved from narrow, task-specific datasets to comprehensive, multi-dimensional benchmarks. Early efforts such as ARCD (Mozannar et al., 2019) for reading comprehension and TyDiQA (Clark et al., 2020) for question answering established foundational paradigms for the pre-LLM era. The landscape has since shifted toward unified benchmarks that assess generalist capabilities across understanding, generation, and reasoning tasks, addressing morphological complexity, dialectal variation, and domain diversity Table 2.

5.1 Comprehensive Multi-Task Benchmarks

LAraBench (Abdelali et al., 2024) introduced one of the first systematic benchmarking suites for Arabic NLP and speech, evaluating general-purpose LLMs against task-specific models across 33 tasks and 61 datasets. By integrating earlier resources such as ARCD and dialect identification corpora, it bridges pre-LLM and LLM evaluation. Results from zero- and few-shot evaluation of GPT-3.5-turbo, GPT-4 (OpenAI et al., 2024), BLOOMZ (Muennighoff et al., 2023), Jais-13b-chat (Sengupta et al., 2023), and Whisper (Radford et al., 2023) show that specialized models outperform LLMs in zero-shot settings, while larger LLMs substantially narrow the gap under few-shot prompting.

BALSAM (Al-Matham et al., 2025) provides a community-driven benchmark emphasizing instruction-following across 78 tasks and 14 categories (52K examples), spanning core NLP tasks such as summarization, QA, translation, and reasoning. It mitigates contamination via blind test sets, offers an integrated leaderboard, and demonstrates that LLM-as-judge aligns more closely with human judgments than traditional metrics. Drawing on mixed natural, translated, and synthetic data from sources including xP3 (Muennighoff et al., 2023), PromptSource (Bach et al., 2022), SuperNaturalInstructions (Wang et al., 2022), TruthfulQA (Lin et al., 2022), and newly curated datasets, BALSAM shows that large closed-source models outperform smaller Arabic-centric models, with performance influenced by tokenization, data scale, and Arabic-specific tuning. Identified limitations include residual cultural misalignment and limited evaluation of multi-turn dialogue and hallucination.

5.2 Task-Specific Benchmarks

Understanding and generation capabilities require evaluation beyond broad benchmark suites. ORCA (Elmadany et al., 2023) targets Arabic natural language understanding by consolidating 60 datasets into seven task categories, covering classification, sequence labeling, semantic similarity, inference, disambiguation, and question answering across MSA and dialects. It evaluates 18 multilingual and Arabic-specific models using task-appropriate metrics and provides a public leaderboard with detailed metadata, supporting reproducible comparison and tracking progress toward modern LLM evaluation.

Dolphin (Nagoudi et al., 2023) focuses on Arabic natural language generation across nine tasks, including generation, paraphrasing, summarization, transliteration, and grammatical error correction, spanning MSA and dialects. Aggregating 15 datasets, it evaluates encoder–decoder and decoder-only models using standard generation metrics and task-specific measures. Results highlight persistent gaps between general-purpose multilingual models and Arabic-finetuned baselines, though Dolphin predates instruction-tuned LLMs and relies on zero-shot evaluation, likely underestimating current performance.

Together with LAraBench and BALSAM, ORCA and Dolphin form complementary benchmarks covering understanding, generation, reasoning, and instruction-following across dialects and domains, providing a foundation for future Arabic LLM evaluation.

6 Culture and Dialects

Paper	Primary Focus (Type)	Total
Jawaher	Cultural Proverbs (GEN)	10k
PALM	Cultural Language Use (GEN)	17.4k
PalmX	Cultural Knowledge (MCQ)	6.4k
Commonsense	Cultural Commonsense (MCQ)	3.5k
AraDiCE	Dialect & Culture (BENCH)	81.8k
AL-QASIDA	Dialectal Analysis (BENCH)	–
Absher	Saudi Dialect & Culture (MCQ)	18k

Table 3: Benchmarks targeting Arabic culture, dialects, and region-specific knowledge.

Evaluation of Arabic cultural and dialectal understanding has evolved incrementally, with successive benchmarks addressing limitations of earlier efforts, as summarized in Table 3. Jawaher (Magdy et al., 2025) introduced one of the earliest culturally grounded resources by compiling 10,037 Arabic proverbs annotated with dialectal origin and idiomatic meaning, enabling the evaluation of figurative reasoning in dialectal Arabic. However, its scope was limited to proverbial knowledge.

PALM (Alwajih et al., 2025a) expanded cultural coverage through a large-scale, community-driven effort spanning all 22 Arab countries, covering diverse genres and topics and evaluating cultural inclusivity via perplexity and generation quality. PalmX (Alwajih et al., 2025b) further targeted deep understanding of Arabic and Islamic culture through MCQ-based evaluation in MSA, covering traditions, history, and religious practices. Arab-specific commonsense reasoning, absent from prior benchmarks, was later addressed by Commonsense Reasoning in Arab Culture (Sadallah et al., 2025), which introduced culturally grounded inference tasks.

Dialectal evaluation has followed a parallel trajectory. AraDiCE (Mousi et al., 2025) assessed dialect identification, generation, and cognitive reasoning in Egyptian, Levantine, and Gulf Arabic, largely via dialectal adaptations of existing benchmarks. Alqasida et al. (Robinson et al., 2025) proposed a broader evaluation framework spanning identification, comprehension, generation quality, and dialect–MSA translation. NADI 2024 (Abdul-Mageed et al., 2024) was employed for dialect identification and translation-based chat evaluation, while Absher (Al-Monef et al., 2025) focused specifically on Saudi dialect vocabulary, phrases, and proverbs.

Overall, despite meaningful progress, benchmarks targeting Arabic cultural alignment and dialectal diversity remain limited in scope and coverage. Future efforts must broaden dialectal representation and deepen cultural reasoning evaluation to better reflect real-world Arabic language use.

7 Target-Specific Tasks

Paper	Primary Focus (Type)	Total
CamelEval	Instruction Following (GEN)	1.6k
Halwasa	Hallucination Detection (GEN)	10k
Henna	Multimodal Understanding (GEN)	1.1k
CAMEL-Bench	Multimodal (BENCH)	29k
AraTrust	Safety & Trust (MCQ)	0.52k
Arabic Safeguard	Safety Evaluation (GEN)	5.8k
ALRAGE	Retrieval-Augmented QA (GEN)	21.2k
AraTable	Tabular Reasoning (GEN)	0.6k
AraHalluEval	Hallucination Detection (GEN)	1.5k
HalluVerse25	Hallucination Classification (CL)	0.8k
ARB	Multimodal Reasoning (GEN)	1.36k
ASAS	Safety Benchmark (GEN)	0.8k

Table 4: Target-specific Arabic benchmarks for safety, multimodality, reasoning, and robustness.

Beyond general NLP capabilities, LLMs with Arabic capabilities require evaluation on specialized tasks reflecting real-world deployment scenarios and LLM-specific challenges, as summarized in Table 4.

Instruction-Following. CamelEval (Qian et al., 2024) evaluates conversational abilities and instruction-following through 1,610 generative questions: 805 human-validated translations from AlpacaEval (Dubois et al., 2023) and 805 synthetically generated using GPT-4 from culturally grounded textbooks. Evaluation uses LLM-as-judge computing win rates for open-ended generation.

Context-Based Reasoning. ALRAGE (El Filali et al., 2025) targets RAG evaluation through 2.12K question-answer-context trios from 40 Arabic books, synthetically generated via Meta-Llama-3.1-70B and validated by native speakers. AraTable (Alshaikh et al., 2025) addresses structured tabular data through 41 tables (15 QA pairs each) evaluating direct answering, fact verification, and complex reasoning. Tables sourced from Wikipedia, government portals, and GPT-4o were verified by human experts, with evaluation combining accuracy and the Assisted Self-Deliberation (ASD) framework employing judge LLMs and human evaluation.

Hallucination Detection. HalluVerse25 (Abdaljalil et al., 2025) provides fine-grained multilingual evaluation across entity, relation, and sentence hallucination types, offering 828 Arabic samples from Wikidata autobiographies where GPT-4 injected false hallucinations validated by human judges. AraHalluEval (Alansari and Luqman, 2025) focuses on Arabic through QA (300 samples from TyDiQA-GoldP-AR and translated TruthfulQA) and summarization (100 XLSum instances), distinguishing factuality and faithfulness hallucinations through manual annotation. Halwasa (Mubarak et al., 2024) provides larger-scale evaluation with 10K samples where ChatGPT and GPT-4 generated factual sentences from 1,000 SAMER Arabic Lexicon words, with hallucination indicators annotated by 200 annotators enabling sentence-level analysis.

Safety and Trustworthiness. Ashraf et al. (2025) introduced 5,799 questions spanning direct attacks, indirect attacks, and harmless requests with sensitive words, employing dual-perspective evaluation from governmental and oppositional viewpoints. ASAS (aiastrolabe, 2025) provides the first human-rated Arabic safety index for red-teaming frontier models, exposing persistent weaknesses in top-performing systems. AraTrust (Alghamdi et al., 2024) assesses trustworthiness across nine dimensions through 522 human-written MCQs.

Multimodal Capabilities. Peacock (Alwajih et al., 2024) pioneered Arabic multimodal evaluation with Henna benchmark combining standard VQA/OCR prompts with focus on culture understanding. CAMEL-Bench (Ghaboura et al., 2024) provides large-scale evaluation across eight domains and 38 subdomains, revealing promising performance but substantial weaknesses in specialized domains requiring precise vision-text alignment. ARB (Ghaboura et al., 2025) advances step-by-step multimodal reasoning, emphasizing logical integration of visual and textual inputs over simple captioning.

8 Discussion

Category	Coverage
Q&A / Reading Comprehension	Moderate (5 datasets)
Translation & Multitask Generation	Moderate (4 datasets)
Reasoning & Multi-step Thinking	**Limited** (3 datasets)
STEM / Academic Evaluation	Strong (8 datasets)
Law / Legal Reasoning	**Limited** (2 datasets)
Poetry / Literature / Arts	**Limited** (3 datasets)
Cultural Alignment & Dialect Evaluation	Strong (7 datasets)
Commonsense & Cultural Reasoning	**Limited** (3 datasets)
Hallucination / Truthfulness	**Limited** (3 datasets)
Retrieval-Augmented / Contextual Tasks	**Limited** (3 datasets)

Table 5: Summary of Arabic NLP dataset coverage and key gaps. Coverage indicates the number of datasets per task category, while Key Gaps outline remaining needs such as standardization, scale, and cultural balance.

This section discusses key observations regarding reproducibility practices, community initiatives, methodological inconsistencies, and critical gaps requiring future attention.

8.1 Reproducibility and Evaluation Benchmarks

Reproducibility is fundamental to scientific progress, yet our survey reveals significant heterogeneity in benchmark accessibility and transparency. We classify benchmarks into three categories based on their openness:

Private Benchmarks release neither datasets nor evaluation pipelines, severely limiting community impact and making verification impossible. Examples include AraSTEM (11,637 samples) and Halwasa (10K samples), which remain unavailable despite published results.

Partially-Public Benchmarks release datasets but withhold evaluation code, hindering exact reproduc-

tion due to ambiguities in preprocessing, prompting, and metric computation.

Public Benchmarks release both datasets and complete pipelines, either through dedicated repositories (LAraBench, 3LM) or integration with mainstream frameworks like lighteval (Habib et al., 2023) and lm-eval-harness (Gao et al., 2024).

Our analysis reveals that approximately 25% of surveyed benchmarks remain private or partially public, limiting their impact towards Arabic NLP community. We strongly advocate for complete transparency as the community standard, recognizing that data contamination concerns can be addressed through alternative mechanisms such as blind test sets (as employed by BALSAM), periodic dataset refreshment, or held-out evaluation techniques rather than complete privatization.

8.2 Common Metrics

Over the course of this survey, we identified the most commonly used metrics, which we list and discuss below:

Accuracy This was a common approach in MCQ benchmarks, with most inferring accuracy based on the log-probabilities of choice indices. Other benchmarks used joint log-probabilities over the actual choice text, which typically leads to lower scores. Accuracy based on log-probabilities has been widely used since the rise of LLMs, but it is less applicable to instruction-tuned models and cannot be used to evaluate closed-source models.

LLM-as-Judge For generative tasks, several existing Arabic benchmarks rely on a closed-source model acting as an expert to evaluate LLM outputs, typically by prompting the model to score responses using a predefined prompt. Despite being adopted by nine benchmarks, this approach suffers from high API costs, expert bias, inconsistent standards, and limited reproducibility. To mitigate some of these issues, a small number of benchmarks incorporated human validation on a limited subset to increase trust in the used expert.

Human-as-Judge In at least two of the reported benchmarks, a Human-as-Judge approach was adopted, in which human evaluators manually assess model responses. This approach represents the most reliable method for obtaining qualitative signals about LLM performance. However, it is both intensive in terms of time and labor, making it difficult to apply for large scale datasets.

Apart from the aforementioned metrics, commonly known NLP metrics were used such F1, EM, ROUGE, and BLEU.

8.3 Methodological Inconsistencies and Quality Issues

Through detailed examination of released benchmarks, we identified several concerning inconsistencies and quality issues that affect evaluation validity and cross-benchmark comparability.

Multiple-Choice Formatting. Benchmarks employ inconsistent option labeling: some use Latin letters while others use Arabic letters. This inconsistency affects model performance as LLMs may be more familiar with Latin alphabetic indices from English pretraining. Standardization is needed for fair cross-benchmark comparison.

Prompting Variations. Few-shot evaluation ranges from 0-shot to 5-shot with limited systematic investigation of optimal settings per task type. Prompt phrasing varies from formal MSA to conversational styles, potentially affecting responses.

Quality Control Issues. Manual inspection revealed typos, grammatical errors, and formatting inconsistencies in several released datasets. Some benchmarks contain culturally inappropriate content despite claims of cultural alignment. These quality issues undermine benchmark validity and highlight the need for rigorous review processes.

LLM-as-Judge Validation. While BALSAM reports 0.824-0.977 correlation with human judgments, most benchmarks adopting LLM-as-judge lack validation specifically in Arabic contexts. Judge model selection, prompt sensitivity, and potential circular evaluation when judges resemble evaluated models require systematic investigation.

8.4 Arabic LLM Leaderboards

Centralized leaderboards serve critical functions: establishing performance baselines, enabling fair model comparisons, tracking field progress over time, and guiding model selection for practitioners. Several initiatives have emerged to fulfill these roles for LLMs with Arabic support, each with distinct philosophies and design choices.

The Open Arabic LLM Leaderboard (OALL) (El Filali et al., 2024a) pioneered open-source rankings, initially using Alghafa, EXAMS_ar, and MMLU_ar. OALL v2 (El Filali et al., 2025) transitioned to native benchmarks (ArabicMMLU, ALRAGE, AraTrust, MadinahQA), reflecting community consensus against translated content. BALSAM (Al-Matham et al., 2025) offers comprehensive evaluation across 78 tasks (52K samples)

with private test sets preventing contamination, including both closed and open-source models. ILMAAM (Nacar et al., 2025) specializes in culturally aligned evaluation using refined ArabicMMLU by ensuring religious sensitivity and social norms. The AraGen benchmark (El Filali et al., 2024b) and its associated leaderboard adopt a 3C3H evaluation metric (Correctness, Completeness, Conciseness, Helpfulness, Honesty, Harmlessness), using LLM-as-judge, and combine it with a dynamic, blind-testing approaches to push for robust and fair benchmarking of Arabic capabilities in LLMs.

8.5 Critical Gaps

Despite substantial progress, several critical areas remain underexplored or entirely absent from current Arabic benchmarking efforts. Table 5 summarizes the coverage of datasets per category as well as key gaps.

Underexplored Areas. Current benchmarks lack temporal evaluation, multi-turn dialogue assessment, code-switching evaluation (Arabic-English/French mixing), low-resource dialect coverage (Sudanese, Mauritanian), pragmatic understanding (sarcasm, indirect speech), and specialized domains beyond law/medicine (education, journalism, technical documentation).

Methodological Challenges. Data contamination threatens validity as training corpora expand. Static benchmarks fail to capture temporal degradation. Cultural alignment lacks standardized metrics. Heavy reliance on LLM-as-judge requires more rigorous Arabic-specific validation.

Dataset Size Matters. As Figure 2 shows, benchmark sizes range from under 500 samples (Qiyas, EXAMS_ar) to over 2 million (Dolphin). This dramatic variability is consequential: small benchmarks provide weaker statistical signals and are more susceptible to overfitting, while aggregate scores averaging across diverse benchmark sizes can be misleading if not properly weighted. Evaluation interpretation must account for dataset scale.

8.6 Recommendations

To address the identified gaps, future Arabic benchmarks should:

Prioritize native Arabic data with mandatory cultural and dialectal review to reduce translation bias.

Ensure full reproducibility by releasing evaluation code, prompts, and scoring procedures within standardized frameworks.

Control contamination using blind test sets and automated overlap detection.

Standardize evaluation protocols by fixing prompts, answer formats, and decoding settings to enable fair comparison.

Validate automated judges through calibration against diverse Arabic human annotations.

Expand dialectal coverage by defining target dialect sets and minimum sample sizes.

Strengthen dataset quality control via multi-pass annotation and transparent documentation.

Incorporate temporal evaluation through versioned or rolling benchmarks to measure performance drift.

These steps are essential for building Arabic benchmarks that are culturally grounded, reproducible, and robust over time.

9 Conclusion

This survey comprehensively reviews 40+ Arabic benchmarks, providing the first systematic taxonomy across NLP tasks, knowledge domains, cultural understanding, and specialized capabilities. The field has evolved from task-specific datasets (ARCD, TyDiQA) to comprehensive benchmarks (BALSAM, LAraBench, ArabicMMLU) addressing understanding, generation, and reasoning.

Progress includes increased dialectal coverage, cultural grounding, and domain diversity. However, persistent challenges remain: data contamination, cultural misalignment in translations, insufficient coverage of temporal reasoning and multi-turn dialogue, and inconsistent reproducibility. Our recommendations emphasize reproducibility, cultural authenticity, dialectal inclusivity, and methodological rigor. As Arabic capabilities in LLMs advance, evaluation must authentically capture Arabic's linguistic complexity and cultural richness. This survey serves as a comprehensive reference, guiding development of more robust, culturally aligned LLM evaluation on Arabic capabilities.

Limitations

espite our comprehensive review, this survey has limitations. Rapidly evolving Arabic LLM benchmarking means recent or proprietary benchmarks may be excluded, particularly those without public documentation. Our review covers benchmarks through early 2025, so some information may quickly become outdated. Several benchmarks

(e.g., AraSTEM, Halwasa, Qiyas) remain unavailable, limiting analysis to published descriptions.

We focus on evaluation benchmarks rather than training datasets, emphasizing text-based and multimodal tasks while excluding speech-only benchmarks. Assessments of benchmark quality and cultural alignment reflect our perspectives; native speakers from other regions may judge differently. Finally, our recommendations are informed opinions, not community consensus, and classifications of benchmarks as public or private reflect the information available at the time of writing.

References

Samir Abdaljalil, Hasan Kurban, and Erchin Serpedin. 2025. Halluverse25: Fine-grained multilingual benchmark dataset for llm hallucinations. *Preprint*, arXiv:2503.07833.

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520, St. Julian's, Malta. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. Nadi 2024: The fifth nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2407.04910*.

aiastrolabe. 2025. Redteaming frontier llms with ai astrolabe arabic safety index (asas). Accessed: Oct. 3, 2025.

Shahad Al-Khalifa and Hend Al-Khalifa. 2024. The qiyas benchmark: Measuring chatgpt mathematical and language understanding in arabic. *arXiv preprint arXiv:2407.00146*.

Shahad Al-Khalifa, Nadir Durrani, Hend Al-Khalifa, and Firoj Alam. 2025. The landscape of arabic large language models. *Communications of the ACM*.

Nashwan Ahmed Al-Majmar, Hezam Gawbah, and Akram Alsubari. 2024. Ahd: Arabic healthcare dataset. *Data in Brief*, 56:110855.

Rawan Al-Matham, Kareem Darwish, Raghad Al-Rasheed, Waad Alshammari, Muneera Alhoshan, Amal Almazrua, Asma Al Wazrah, Mais Alheraki, Firoj Alam, Preslav Nakov, Norah Alzahrani, Eman alBilali, Nizar Habash, Abdelrahman El-Sheikh, Muhammad Elmallah, Haonan Li, Hamdy Mubarak, Mohamed Anwar, Zaid Alyafeai, and 24 others. 2025. Balsam: A platform for benchmarking arabic large language models. *Preprint*, arXiv:2507.22603.

Renad Al-Monef, Hassan Alhuzali, Nora Alturayeif, and Ashwag Alasmari. 2025. Absher: A benchmark for evaluating large language models' understanding of saudi dialects. *Preprint*, arXiv:2507.10216.

Aisha Alansari and Hamzah Luqman. 2025. Arahallueval: A fine-grained hallucination evaluation framework for arabic llms. *Preprint*, arXiv:2509.04656.

Ashwag Alasmari, Sarah Alhumoud, and Waad Alshammari. 2024. Aramed: Arabic medical question answering using pretrained transformer language models. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024*, pages 50–56.

Safa AlBallaa, Nora AlTwairesh, Abdulmalik AlSalman, and Sultan Alfarhood. 2025. Gatmath and gatlc: Comprehensive benchmarks for evaluating arabic large language models. *PLoS One*, 20(9):e0329129.

Hayfa A Aleid and Aqil M Azmi. 2025. Hajj-fqa: A benchmark arabic dataset for developing question-answering systems on hajj fatwas: H. aleid and a. azmi. *Journal of King Saud University Computer and Information Sciences*, 37(6):135.

Wafa Alghallabi, Ritesh Thawkar, Sara Ghaboura, Ketan More, Omkar Thawakar, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Fann or flop: A multigenre, multiera benchmark for arabic poetry understanding in llms. *arXiv preprint arXiv:2505.18152*.

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.

Mohamed Alkaoud. 2024. A bilingual benchmark for evaluating large language models. *PeerJ Computer Science*, 10:e1893.

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.

Rana Alshaikh, Israa Alghanmi, and Shelan Jeawak. 2025. Aratable: Benchmarking llms' reasoning and understanding of arabic tabular data. *Preprint*, arXiv:2507.18442.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer

Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025a. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

Fakhraddin Alwajih, Abdellah El Mekki, Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed, and Muhammad Abdul-Mageed. 2025b. Palmx 2025: The first shared task on benchmarking llms on arabic and islamic culture. *Preprint*, arXiv:2509.02550.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of arabic multimodal large language models and benchmarks. *Preprint*, arXiv:2403.01031.

Anthropic. 2024. Introducing the next generation of claude.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Yasser Ashraf, Yuxia Wang, Bin Gu, Preslav Nakov, and Timothy Baldwin. 2025. Arabic dataset for llm safeguard evaluation. *Preprint*, arXiv:2410.17040.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *Preprint*, arXiv:2108.07732.

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, and 8 others. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *Preprint*, arXiv:2202.01279.

Adil Bahaj and Mounir Ghogho. 2025. Mizanqa: Benchmarking large language models on moroccan legal question answering. *arXiv preprint arXiv:2508.16357*.

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Preprint*, arXiv:2003.05002.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E Shamout. 2025. Medarabiq: Benchmarking large language models on arabic medical tasks. *arXiv preprint arXiv:2505.03427*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Preprint*, arXiv:2305.14387.

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

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

Ali El Filali, Neha Sengupta, Abouelseoud, Preslav Nakov, and Clémentine Fourrier. 2024b. Rethinking llm evaluation with 3c3h: Aragen benchmark and leaderboard. `https://huggingface.co/blog/leaderboard-3c3h-aragen`.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Orca: A challenging benchmark for arabic language understanding. *Preprint*, arXiv:2212.10758.

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, and 5 others. 2024. The language model evaluation harness.

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. *Preprint*, arXiv:2410.18976.

Sara Ghaboura, Ketan More, Wafa Alghallabi, Omkar Thawakar, Jorma Laaksonen, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Arb: A comprehensive arabic multimodal reasoning benchmark. *Preprint*, arXiv:2505.17021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

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. *arXiv preprint arXiv:2011.03080*.

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*.

Faris Hijazi, Somayah AlHarbi, Abdulaziz AlHussein, Harethah Abu Shairah, Reem AlZahrani, Hebah AlShamlan, Omar Knio, and George Turkiyyah. 2024. Arablegaleval: A multitask benchmark for assessing arabic legal knowledge in large language models. *arXiv preprint arXiv:2408.07983*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. Acegpt, localizing large language models in arabic. *Preprint*, arXiv:2309.12053.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2025. A survey on large language models with multilingualism: Recent advances and new frontiers. *Preprint*, arXiv:2405.10936.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2025. Revisiting common assumptions about arabic dialects in nlp. *arXiv preprint arXiv:2505.21816*.

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.

Anis Koubaa, Adel Ammar, Lahouari Ghouti, Omar Najar, and Serry Sibaee. 2024. Arabiangpt: Native arabic gpt-based large language model. *Preprint*, arXiv:2402.15313.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. JAWAHER: A multidialectal dataset of Arabic proverbs for LLM benchmarking. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12320–12341, Albuquerque, New Mexico. Association for Computational Linguistics.

Malak Mashaabi, Shahad Al-Khalifa, and Hend Al-Khalifa. 2024. A survey of large language models for arabic language and its dialects. *arXiv preprint arXiv:2410.20238*.

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.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem M. Hajj. 2019. Neural arabic question answering. *CoRR*, abs/1906.05394.

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, Torino, Italia. ELRA and ICCL.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. 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. *arXiv 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.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Ahmed Oumar El-Shangiti, and Muhammad Abdul-Mageed. 2023. Dolphin: A challenging and diverse benchmark for Arabic NLG. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1404–1422, Singapore. Association for Computational Linguistics.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Omartificial-Intelligence-Space. 2025. Arabic gsm8k: Arabic grade school math dataset. `https://huggingface.co/datasets/Omartificial-Intelligence-Space/Arabic-gsm8k-v2`.

OpenAI. 2023. Chatgpt. `https://chat.openai.com/chat`. Accessed: 2 October 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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

Nathaniel R. Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. Al-qasida: Analyzing llm quality and accuracy systematically in dialectal arabic. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. Commonsense reasoning in Arab culture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7695–7710, Vienna, Austria. 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, and 13 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Sumuk Shashidhar, Clémentine Fourrier, Alina Lozovskia, Thomas Wolf, Gokhan Tur, and Dilek Hakkani-Tür. 2025. Yourbench: Easy custom evaluation sets for everyone. *Preprint*, arXiv:2504.01833.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025a. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025b. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Silma Team. 2024. Silma.

TII. 2025. Falcon-arabic: A breakthrough in arabic language models.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, and 21 others. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *Preprint*, arXiv:2204.07705.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Summary Table

Table 6 provides a comprehensive overview of all the Arabic benchmarks reviewed in this survey. This appendix clarifies abbreviations and highlights key insights for interpreting the data.

A.1 Column Definitions

Sample Composition Columns:

- **NAT:** Natively authored Arabic samples (originally created in Arabic by native speakers)
- **SYN:** Synthetically generated samples (created using LLMs like GPT-4, Claude, or Llama)
- **TRAN:** Translated samples (translated from English or other languages to Arabic)
- **Total:** Total benchmark size

Evaluation Details:

- **Metrics:** Evaluation metrics employed (e.g., Accuracy, F1, BLEU, ROUGE, LLM-as-Judge)
- **Type:** Question format - **MCQ** (Multiple-choice), **GEN** (Generative/open-ended),**BENCH** (Multi-task suite), **CL** (Classification), **RS** (Reasoning Steps)

Accessibility:

- **PD (Public Dataset):** Yes (publicly available), No (private/request-only), Partial (subset public)
- **PR (Public Repository):** Yes (complete evaluation code available), No (no public repository)

Notation: "–" (not available/applicable), "k" (thousands), "m" (millions), "DS" (datasets)

A.2 Key Insights from the Table

Dataset Composition Trends:

- **Native vs. Translated:** Early benchmarks (2020) relied heavily on translation (MMLU_ar: 100% translated), while recent benchmarks (2024-2025) prioritize native content (ArabicMMLU, GATmath: 100% native), reflecting community consensus on cultural authenticity.

- **Synthetic Generation:** Recent benchmarks increasingly employ synthetic generation (ALRAGE: 21.2k synthetic, ArabLegalEval: 10.58k synthetic), balancing scale with cost. However, purely native benchmarks like Jawaher (10k native) and Absher (18k native) demonstrate that large-scale native collection remains feasible.
- **Hybrid Approaches:** BALSAM exemplifies effective hybrid strategy (26k native, 1.7k synthetic, 24k translated), combining strengths of multiple approaches.

Scale Distribution:

- **Large-Scale:** Dolphin (2m samples) and LAraBench (296k) represent comprehensive multi-task benchmarks.
- **Medium-Scale:** Most benchmarks range from 1k-20k samples, balancing quality and coverage.
- **Specialized:** Domain-specific benchmarks (legal, medical) tend toward smaller, higher-quality datasets (1-7k samples).

Reproducibility Concerns:

- **Fully Accessible:** Only 21 of 41 benchmarks provide both public datasets and repositories.
- **Dataset-Only:** 8 benchmarks release datasets without evaluation code, hindering exact reproduction.
- **Private:** 9 benchmarks remain completely private such as AraST EM, Halwasa, CamelEval and Qiyas, including some with substantial contributions (AraST EM: 11.6k samples), severely limiting community impact.
- **Critical Gap:** The high proportion of inaccessible benchmarks impedes scientific progress and independent validation.

Evaluation Methodology Evolution:

- **Traditional Metrics:** Early benchmarks use standard metrics (Accuracy, F1, BLEU, ROUGE).
- **LLM-as-Judge:** Recent benchmarks increasingly adopt LLM-as-judge (BALSAM, ALRAGE, ArabLegalEval), particularly for generative tasks, though validation against human judgments remains limited.
- **Human Evaluation:** Hallucination benchmarks (AraHalluEval, Halwasa) maintain human-as-judge given the critical nature of the task, despite higher costs.

Task Coverage Patterns:

- **NLP Tasks:** Well-covered with multiple comprehensive benchmarks (ORCA, Dolphin, LAraBench, BALSAM).
- **STEM:** Substantial progress with native benchmarks (ArabicMMLU, AraST EM, GATmath) addressing earlier translation limitations.
- **Culture & Dialects:** Growing emphasis (7 benchmarks in 2025 alone), reflecting recognition of cultural alignment importance.
- **Target-Specific:** Emerging area with recent focus on hallucination detection and context-based reasoning.

Temporal Distribution: Most benchmarks were released in 2024-2025, indicating rapid field acceleration. However, this concentration also suggests potential redundancy in some areas (multiple GAT-based benchmarks) while other areas remain unexplored (temporal evaluation, code-switching, multi-turn dialogue).

Geographic and Cultural Bias: Several benchmarks focus on specific regional variants (Absher: Saudi dialect, MizanQA: Moroccan law, GATmath: Saudi standardized tests), highlighting both progress in regional representation and gaps in coverage of other Arabic-speaking regions (North Africa, Levant, Iraq).

B Arabic NLP Benchmarks Timeline

Figure 3 presents a timeline of Arabic benchmark releases from 2019 to 2025, categorized by our four-category taxonomy. The visualization reveals dramatic acceleration in benchmark development, with 82% of all benchmarks (34 of 41) released in 2024-2025 alone. This recent surge reflects growing recognition of Arabic LLM evaluation needs and demonstrates rapid field maturation across all categories, particularly in cultural/dialectal assessment where all 7 benchmarks emerged in 2025.

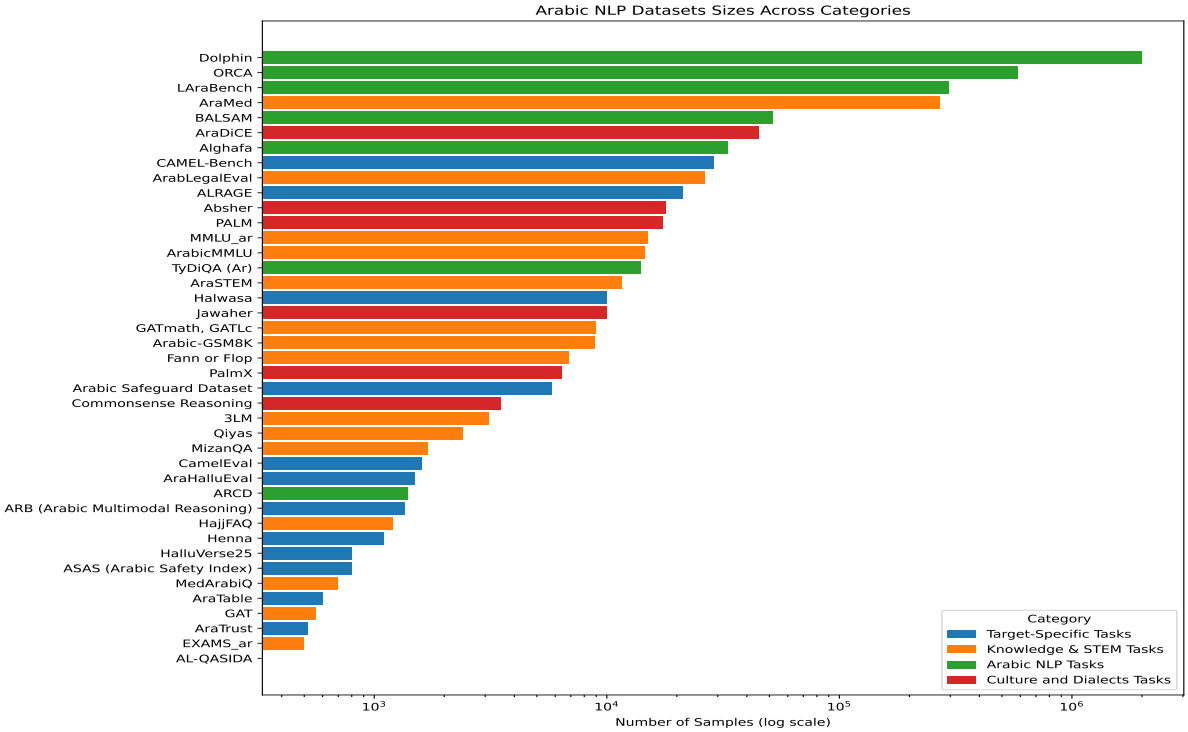

Figure 2: Arabic Datasets Sizes Across Categories. Size is log-scaled.

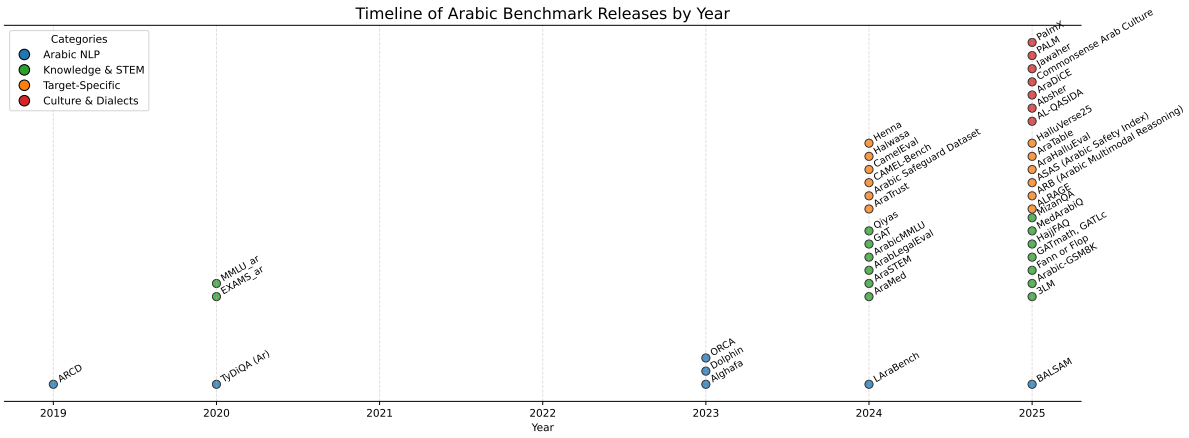

Figure 3: Timeline of Arabic benchmark releases (2019-2025) showing dramatic acceleration, with 82% of the benchmarks released in 2024-2025.

Year	Paper	Topic	NAT	SYN	TRAN	Total	Metrics	Type	PD	PR
				Knowledge						
2020	MMLU_ar (Hendrycks et al., 2020)	Humanities • Social Science • STEM	-	-	15k	15k	Accuracy • Log-prob	MCQ	Yes	Yes
2020	EXAMS_ar (Hardalov et al., 2020)	Science • Social Science • Religion	0.5k	-	-	0.5k	Accuracy • Log-prob	MCQ	Yes	Yes
2024	ArabicMMLU (Koto et al., 2024)	Humanities • Social Science • STEM • Arabic Language	14.57k	-	-	14.57k	Accuracy • Log-prob	MCQ	Yes	Yes
2024	AraSTEM (Mustapha et al., 2024)	STEM	11.63k	-	-	11.63k	Accuracy • Log-prob	MCQ	No	No
2024	GAT (Alkaoud, 2024)	Linguistic Abilities	0.45k	-	-	0.56k	Accuracy	MCQ	No	No
2024	Qiyas (Al-Khalifa and Al-Khalifa, 2024)	Linguistic Abilities • Mathematics	2.4k	-	-	2.4k	Accuracy	MCQ	No	No
2024	ArabLegalEval (Hijazi et al., 2024)	Law	79	10.58k	15.8	26.4k	Accuracy • LLM-as-Judge	MCQ, GEN	Yes	Yes
2024	AraMed (Alasmari et al., 2024)	Medical Domain • Healthcare	270k	-	-	270k	Accuracy • QA Metrics	GEN	Yes	Yes
2025	MizanQA (Bahaj and Ghogho, 2025)	Law	1.7k	-	-	1.7k	Accuracy • F1 • ECE	MCQ	Yes	No
2025	Fann or Flop (Alghallabi et al., 2025)	Poetry	6.9k	-	-	6.9k	BLEU • chrf(++) • BERTScore	GEN	Yes	Yes
2025	GATmath, GATLc (AlBallaa et al., 2025)	Linguistic Abilities • Mathematics	9k	-	-	9k	Accuracy	MCQ	Yes	No
2025	3LM (Boussaha et al., 2025)	STEM • Coding	0.8k	1.74k	0.54k	3.1k	Accuracy • Log-prob	MCQ, GEN	Yes	Yes
2025	Arabic-GSM8K (Omartificial-Intelligence-Space, 2025)	Mathematics • Reasoning	-	-	8.9k	8.9k	Exact-Match	GEN	Yes	No
2025	MedArabiQ (Daoud et al., 2025)	Medical	-	0.7k	-	0.7k	Accuracy • BERTScore	MCQ	Yes	Yes
2025	Hajj-FAQ (Aleid and Azmi, 2025)	Religious Domain • Islamic Jurisprudence	1.2k	-	-	1.2k	Accuracy • F1	QA	No	No
				Arabic NLP Tasks						
2019	ARCD (Mozannar et al., 2019)	Reading Comprehension	1.4k	-	-	1.4k	EM • F1	MCQ	Yes	Yes
2020	TyDiQA (Ar) (Clark et al., 2020)	Reading Comprehension	14k	-	-	14k	EM • F1	GEN	Yes	Yes
2023	ORCA (Elmadany et al., 2023)	Sentiment • Text Classification • NER • QA • Paraphrase • NLI • Dialect ID • MT	-	-	-	588k	Accuracy • F1 • Pearson • EM	BENCH	Yes	No
2023	Dolphin (Nagoudi et al., 2023)	QA • Reading Comprehension • Reasoning • Math • Commonsense • Dialogue • Summarization • Paraphrasing • Writing	-	-	-	2m	BLEU • ROUGE • METEOR • BERTScore	BENCH	No	No
2023	Alghafa (Almazrouei et al., 2023)	Reading Comprehension • Sentiment Analysis • QA	33.2k	-	-	33.2k	Accuracy • Log-prob	MCQ	Yes	Yes
2024	LAraBench (Abdelali et al., 2024)	Sentiment • Topic Classification • NER • QA • Paraphrase • NLI • Dialect ID • MT • Reasoning • Summarization	296k	-	-	296k	F1 • BLEU • Task-specific	BENCH	Yes	Yes
2025	BALSAM (Al-Matham et al., 2025)	MT • Transliteration • DialectMT • Simplification • QRewrite • Paraphrase • Intent • GrammarCorr • GenderRewrite • TextClass • Sentiment • Sarcasm • DialectID • Command • Summarization • SubjectGen • AnswerExt • SeqTag	26k	1.7k	24k	52k	LLM-as-Judge • BLEU • ROUGE	BENCH	Partial	Yes
				Culture and dialects Tasks						
2025	Jawaher (Magdy et al., 2025)	Proverbs • Figurative QA	10k	-	-	10k	BLEURT • BERTScore • LLM-as-Judge • Human-as-Judge	GEN	Yes	No
2025	PALM (Alwajih et al., 2025a)	Prose • Dialogue • Cultural Expressions	17.4k	-	-	17.4k	LLM-as-Judge	GEN	Yes	Yes
2025	PalmX (Alwajih et al., 2025b)	Shared Task • Arabic Culture MCQs • Islamic Culture MCQs	-	6.4k	-	6.4k	Accuracy, Log-prob	MCQ	Yes	No
2025	Commonsense Reasoning in Arab Culture (Sadallah et al., 2025)	Cultural Commonsense QA	3.5k	-	-	3.5k	Accuracy • Log-prob	MCQ	Yes	Yes
2025	AraDiCE (Mousi et al., 2025)	Dialect Identification/Generation/Translation • Cognitive Abilities on Dialect * Culture	41.8k	-	45k	81,8k	Accuracy	BENCH	Yes	Yes
2025	AL-QASIDA (Robinson et al., 2025)	Dialectal Analysis	-	-	-	-	Accuracy • Dialectal Error Rate	BENCH	No	No
2025	Absher (Al-Monef et al., 2025)	Saudi Dialect • Saudi Culture	18k	-	-	18k	Accuracy • F1	MCQ	No	No
				Target-Specific Tasks						
2024	CamelEval (Qian et al., 2024)	Instruct-following	-	0.8k	0.8k	1.6k	Win-rate, LLM-as-Judge	GEN	No	No
2024	Halwasa (Mubarak et al., 2024)	Hallucination	-	10k	-	10k	Human-as-Judge	GEN	No	No
2024	Henna (Alwajih et al., 2024)	Multimodal	–	1.1k	–	1.1k	LLM-as-Judge	VQA, OCR, GEN	Yes	Yes
2024	CAMEL-Bench (Ghaboura et al., 2024)	Multimodal	29.0k	-	-	29.0k	Accuracy, PATS • LLM-as-Judge	VQA, OCR, RS, GEN	Yes	Yes
2024	AraTrust (Alghamdi et al., 2024)	Safety	0.52k	-	-	0.52k	Accuracy	MCQ	Yes	Yes
2024	Arabic Safeguard Dataset (Ashraf et al., 2025)	Safety	1.0k	-	4.8k	5.8k	LLM-as-Judge	GEN	Yes	Yes
2025	ALRAGE (El Filali et al., 2025)	Context-based (RAG)	-	21.2K	-	21.2K	LLM-as-Judge	GEN	Yes	Yes
2025	AraTable (Alshaikh et al., 2025)	Context-based (Tabular)	-	0.6k	-	0.6k	Accuray • Human-as-Juge	GEN	Yes	Yes
2025	AraHalluEval (Alansari and Luqman, 2025)	Hallucination	0.4k	-	0.8k	1.5k	Human-as-Judge	GEN	Yes	Yes
2025	HalluVerse25 (Abdaljalil et al., 2025)	Hallucination	-	0.8k	-	0.8k	Accuracy	CL	No	No
2025	ARB (Arabic Multimodal Reasoning) (Ghaboura et al., 2025)	Multimodal	1.36k	-	-	1.36k	BLEU • ROUGE • BERTScore • LaBSE • LLM-as-Judge	VQA, RS	Yes	Yes
2025	ASAS (Arab-ic Safety Index) (aiastrolabe, 2025)	Safety	0.8k	-	-	0.8k	LLM-as-Judge	GEN	No	No

Table 6: Summary of Arabic Benchmarks. **NAT:** # of native samples, **TRAN:** # of translated samples, **SYN:** # of synthetic samples, **PD:** dataset is public, and **PR:** repo is public.