



Arabic Large Language Models: A Scoping Review of Model Development, Evaluation Benchmarks, Dialectal Coverage, Cultural Alignment, and Research Gaps (2021–2026)

¹Abdul Basid, ²Nur Faizin, ³Habib Muharrom Sudarmawan, ⁴Moh. Fauzan, ⁵Abdul Adzim, ⁶Ahmad Munjin Nasih, ⁷Ali Ma'sum, ⁸Lilik Nur Kholidah, ⁹Achmad Sultoni, ¹⁰Mochammad Rizal Ramadhan, ¹¹Nur Mutamakkin, ¹²Mohammad Ahsanuddin, ¹³Nur Anisah Ridwan, ¹⁴Faris Khoirul Anam, ¹⁵Fery Fauzi, ¹⁶Muhammad Lukman Arifianto, ¹⁷Titit Thoriquattyas

Universitas Negeri Malang, Indonesia

Corresponding Author: abdul.basid.fs@um.ac.id

Abstract

Arabic large language models are evaluated using heterogeneous resources that differ in their linguistic coverage, provenance and interpretation. This paper maps primary research on model development and Arabic-language benchmarking published between 2021 and October 2026. Using a Population–Concept–Context framework, thirteen original research records identified through targeted scholarly discovery were provisionally charted. The mapped studies include generative models, knowledge and reasoning benchmarks, dialectal evaluations, cultural alignment audits, grounded dialogue and factuality assessment. ArabicMMLU contains 14,575 multiple-choice questions, whereas other evaluations focus on authentic dialectal response quality or context-dependent judgments. The evidence highlights differences between benchmark scale and construct validity, including human assessment and contamination controls. Because systematic database searches and independent full-text screening have not been completed, the results remain provisional. Future studies should report data origin, dialect representation, annotation reliability and use-specific error analysis.

Keywords: Arabic large language models; dialectal Arabic; evaluation benchmarks; cultural alignment

مستخلص البحث

تتناول هذه الدراسة نماذج اللغة الكبيرة للغة العربية وطرائق تقييمها في ضوء تنوع الفصحى واللهجات والسياقات الثقافية. وتهدف إلى رسم خريطة أولية للأبحاث الأصلية المتعلقة بتطوير النماذج ومعايير الاختبار المنشورة بين عامي 2021 و2026. اعتمدت الدراسة إطار السكان والمفهوم والسياق، وحلت بصورة تمهيدية ثلاث عشرة دراسة تم العثور عليها من خلال البحث الموجه في مواقع النشر العلمي. وتشمل الدراسات نماذج التوليد واختبارات المعرفة والاستدلال وتقييم اللهجات والمواءمة الثقافية ودقة المعلومات والحوار المرتبط بالمصادر. وتبين المقارنة أن زيادة عدد أسئلة الاختبار لا تكفي لإثبات القدرة اللغوية العامة، وأن صلاحية القياس تتأثر بمصدر البيانات وطبيعة المهمة ومعايير التحكيم البشري. ولأن البحث الشامل في قواعد البيانات وفحص النصوص الكاملة بصورة مستقلة لم يكتمل بعد، تعد النتائج أولية. وتوصي الدراسة بتوثيق مصدر البيانات وتمثيل اللهجات وفحص التلوث بين التدريب والاختبار وتحليل الأخطاء بمشاركة خبراء اللغة. الكلمات المفتاحية: نماذج اللغة الكبيرة؛ اللهجات العربية؛ معايير التقييم؛ المواءمة الثقافية

1. Introduction

1.1 Background and significance

Arabic large language models (LLMs) have become an important test case for whether contemporary artificial intelligence can support linguistically diverse populations rather than simply transfer performance estimates from English benchmarks. The Arabic writing system, rich morphology, variation in vocalization and the coexistence of Modern Standard Arabic (MSA) with regionally situated everyday varieties create distinct conditions for training, adaptation, and evaluation. Model-building and benchmark research therefore need to be considered together: an architecture can be computationally powerful while its apparent language competence reflects narrowly selected tasks.

Among earlier Arabic generation-oriented advances, AraT5 introduced three T5-style systems and the ARGEN suite of seven generation tasks, with its authors reporting gains over mT5 in 52 of 59 test sets (Nagoudi et al., 2022). AceGPT subsequently combined Arabic continued pretraining, instruction tuning and AI feedback with an emphasis on local values (Huang et al., 2024). AlGhafa contributed multiple-choice evaluation and decoder-only Arabic models, while including a toxicity analysis (Almazrouei et al., 2023). These research directions illustrate how performance, localization and evaluation resources have advanced along partially independent paths.

1.2 Earlier research and assessment traditions

ArabicMMLU addressed knowledge-intensive evaluation with 14,575 MSA multiple-choice questions in 40 subject tasks and comparisons across 35 models (Koto et al., 2024). GATmath and GATLc assembled 7,016 mathematical and 9,036 linguistic test questions; the authors reported a highest accuracy of 66.9% and 64.3% respectively within the evaluated model set (AlBallaa et al., 2025). BALSAM proposed a substantially broader community platform: 78 NLP tasks across 14 categories, with 52,000 examples partitioned between development and blind test components (Al-Matham et al., 2025).

A second strand places dialect and culture at the centre of evaluation. AraDiCE comprises seven synthetic dialect datasets combined with human post-editing and cultural probes (Mousi et al., 2025). TounsiBench includes 744 human-written reference responses to Tunisian Arabic instructions and analyses agreement between people and automated judging (Ben Hassine et al., 2025). DialectalArabicMMLU adapts 3,000 source questions into Syrian, Egyptian, Emirati, Saudi and Moroccan Arabic, providing 15,000 dialectal question–answer pairs (Altakrori et al., 2026).

1.3 Knowledge gap and rationale for a scoping review

Taken together, these publications do not yet form a single comparable test of Arabic LLM capability. Generation benchmarks, school-examination knowledge tests, dialectal instructions, cultural ratings and factuality judgments operationalize different constructs. A score from one source cannot support a claim about another domain without additional validation. For example, a high MSA multiple-choice accuracy says little directly about conversational appropriateness in Tunisian Arabic, while a fluent written response is not necessarily factually correct.

Halwasa illustrates this separation by annotating 10,000 Arabic sentences for factuality and reporting that one quarter were factually incorrect within that dataset

(Mubarak et al., 2024). An expert panel in a cultural-alignment study reviewed more than 2,500 questions for appropriateness, bias and related dimensions (Nacar et al., 2025). The coexistence of such varied evaluation mechanisms calls for evidence mapping before any justified attempt at effect synthesis or model ranking. The preliminary review thus focuses on measurement coverage, not the question of which model is universally best.

1.4 Review objectives and potential contribution

The review asks: What original research has developed or evaluated Arabic language models since 2021, how are Arabic varieties and cultural contexts represented, which evaluation constructs and metrics are used, and what gaps constrain defensible interpretation? The objectives are to chart model-building and benchmark designs; map representation of MSA and regional dialects; identify how human and automated judgments are used; and articulate gaps in data provenance, leakage controls, fairness and deployment validity.

The contribution is an integrative conceptual framework linking data origins, model adaptation, benchmark tasks, evaluation judgments and the uses to which results are put. Figure 2 presents this chain explicitly. Rather than equating more test items with better measurement, the review emphasizes transparency about the phenomenon that each dataset can and cannot measure.

1.5 Organization of the article

Section 2 specifies the proposed review protocol and transparently separates completed targeted discovery from database-search stages not yet performed. Section 3 presents provisional source-by-source charts; Section 4 interprets methodological tensions and a research agenda. Section 5 states limitations, Section 6 provides concluding implications, and supplementary verification documentation is prepared separately.

2. Methods

2.1 Design, PCC, and eligibility

The manuscript uses a Population–Concept–Context (PCC) framework consistent with scoping-review methodology. The population or evidence objects are Arabic-centric and multilingual text-based language models evaluated on Arabic input or output; the concept is development, localization, benchmarking and claims concerning Arabic capability; and the context includes MSA, written dialects and culturally contextualized applications. Scoping rather than comparative-effect synthesis is appropriate because studies use divergent models, task distributions, reference answers, metrics and reporting protocols (Munn et al., 2018). PRISMA-ScR is designated for final reporting (Tricco et al., 2018).

Planned eligibility criteria include peer-reviewed primary model-development and evaluation studies or original conference research published from 1 January 2021 to 10 October 2026. Papers must report extractable Arabic-language model evaluations, original benchmark construction, or training with direct Arabic evaluation. Reviews, editorials, unrelated Arabic speech-recognition-only studies, studies assessing only English, and duplicate preprint-proceedings versions are excluded. The language of the research report is unrestricted if sufficient content can be assessed.

An important boundary is that development articles cannot automatically be compared with independent benchmark evaluations. A study that produces a model and tests it on its own tasks contributes evidence about an architecture and experimental conditions; a study that audits contamination, dialectal balance or safety contributes evidence about measurement. Both meet the PCC but will be coded separately.

2.2 Proposed Boolean searches and databases

Proposed Scopus query: TITLE-ABS-KEY((Arabic OR "Modern Standard Arabic" OR "dialectal Arabic" OR "Arabic dialect*") AND ("large language model*" OR LLM* OR "language model*" OR "generative language model*") AND (benchmark* OR evaluat* OR fine-tun* OR train* OR hallucinat* OR alignment OR dialect* OR cultural)) AND PUBYEAR > 2020 AND PUBYEAR < 2027. Equivalent Web of Science TS= groups, plus ACL Anthology, ACM Digital Library, IEEE Xplore and publisher-site queries are planned. Arabic-script retrieval terms should include اللهجات العربية and نماذج اللغة الكبيرة. These strings constitute a proposed—not retrospectively asserted executed—database strategy.

For every future search, reviewers should preserve the exact syntax, date, interface, exported record file, database yield and deduplication decision. Citation tracking should be reported separately to avoid double counting. The initial discovery used direct scholarly-index and publisher-page inspection for the 13 named candidate studies; this does not establish database sensitivity or completeness.

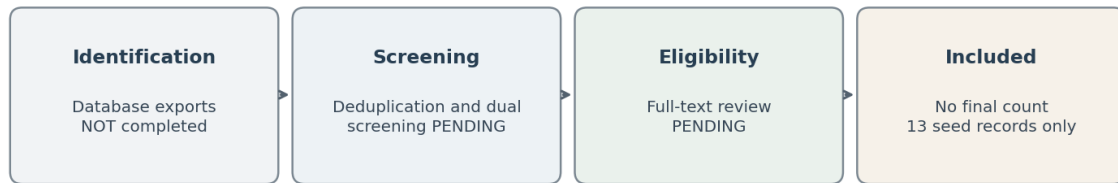
2.3 Selection, data charting and review integrity

At least two reviewers are required to screen title/abstract records independently, retrieve potentially eligible full texts, log reasons for exclusion and reconcile differences. The extraction form should record model training data, architecture, parameter scale where reported, dialect categories, benchmark item origin, annotation strategy, prompt or few-shot settings, sampling temperature, metric, human evaluation criteria, contamination safeguards and reporting limitations. A reference-level claims ledger is supplied as a separate document.

No completed Scopus or Web of Science export, independent screening log or auditable excluded-record denominator was provided with the source manuscript. Figure 1 therefore labels the 13 sources as a targeted seed corpus rather than fabricating a final PRISMA-ScR inclusion number. The case mapping is descriptive and no statistical pooled estimate is attempted.

2.4 Synthesis logic

Source records are charted at the study level (Table 1) and against the kinds of evidence needed for their stated use (Table 2). The analysis is organized across model development, benchmark coverage, dialect representation, cultural judgments, factuality and reproducibility. Numeric data are quoted only when the publisher-facing study description provides them; differences in item format, domain and model selection preclude naïve comparison across studies.

Figure 1. PRISMA-ScR accounting status: incomplete stages explicitly identified.**Figure 1. PRISMA-ScR workflow status**

Seed records are not final PRISMA-included studies; no screening totals have been inferred.

Figure 1 must be replaced with authentic database counts after screening. This status diagram deliberately does not transform the 13 seed records into included studies.

3. Results: Provisional Evidence Map

3.1 Study families and evidence objects

The 13 located candidate records span Arabic generation and model localization, broad knowledge and reasoning benchmarks, dialectally targeted resources, cultural assessment, factuality and grounded dialogue. This is a purposeful seed set and cannot be interpreted as the annual or geographical distribution of all Arabic LLM research. Tables 1 and 2 specify the relevant studies individually, preventing broad claims from becoming detached from underlying sources.

Figure 1 describes the absence of completed screening denominators; the methods limitation directly constrains interpretation here. A PRISMA-ScR flow diagram reporting search yields and eligibility decisions can only replace it after independent review.

Table 1. Provisional chart of 13 original Arabic LLM studies

One row per identified original record. Evidence descriptions are grounded in official publisher records.

Study	Research object	Design / data	Publisher-supported finding
Nagoudi et al. (2022)	Arabic generative model	Three Arabic T5-style models evaluated using ARGEN	ARGEN covers seven generation tasks; reported improvement on 52/59 test sets versus mT5.
Almazrouei et al. (2023)	Arabic MCQ benchmark + models	Arabic multiple-choice evaluation and decoder-only training	Describes a model family including a 14B-parameter variant and examines toxicity.
Huang et al. (2024)	Arabic instruction-tuned model	Arabic continued pretraining, SFT and AI-feedback alignment	Describes Arabic training and culture-sensitive reward alignment.
Koto et al. (2024)	MSA knowledge benchmark	14,575 MCQs, 40 tasks; 35 models evaluated	School-examination questions from multiple Arabic-speaking regions; top reported Arabic-centric model score of 62.3%.
Mubarak et al. (2024)	Arabic factuality	10,000 generated sentences manually annotated for correctness	Reports 25% of sampled generated sentences as factually incorrect.
Mousi et al. (2025)	Dialect/culture benchmarks	Seven synthetic dialect datasets plus culture-focused probes, post-edited by humans	About 45,000 post-edited samples covering dialectal and cultural capabilities.
Nacar et al. (2025)	Cultural appropriateness	11 experts annotate more than 2,500 benchmark questions	Judged fluency, adequacy, cultural appropriateness, bias, religion and social norms.
AlBallaa et al. (2025)	Math + comprehension reasoning	7,016 mathematics and 9,036 linguistic questions; 7 models	Maximum accuracy in reported tests 66.9% (GATmath) and 64.3% (GATLc).
Ben Hassine et al. (2025)	Tunisian response generation	744 instructions; human gold replies and model-as-judge comparison	Assesses ten LLMs on quality, correctness, relevance and dialectal adherence.
Al-Matham et al. (2025)	Broad-task blind platform	78 tasks; 14 categories; 52K examples, blind evaluation	37K test and 15K development examples, with community evaluation platform.

Study	Research object	Design / data	Publisher-supported finding
Zeinalipour et al. (2025)	Grounded multilingual-variety dialogue	30,000 six-turn Wikipedia-grounded conversations, three Arabic varieties	10,000 conversations per variety; model generation and human correction.
Altakrori et al. (2026)	Five-dialect knowledge benchmark	3,000 items adapted into five dialects; 15K pairs and 19 evaluated models	Syrian, Egyptian, Emirati, Saudi and Moroccan; 32 domains.
Al-Monef et al. (2026)	Saudi regional words/proverbs	Over 18,000 MCQs from five major Saudi regions; six tested models	Across six task categories, results vary by dialect; context-rich cultural questions remain difficult.

Table 1 separates development, benchmarking and diagnostic evaluation. It lists charted observations rather than pooled cross-benchmark rankings.

Table 2. Interpretive cautions and verification targets for each study

Study-specific strengths and unresolved issues; limitations are reviewer interpretations, not necessarily the authors' own wording.

Study	Evidence domain	Primary interpretive caution	Verification status
Nagoudi (2022)	Arabic generative model	Historical generation baseline, not by itself a contemporary instruction-following benchmark.	Official article/abstract record checked; full-text procedural cross-check still required.
Almazrouei (2023)	Arabic MCQ benchmark + models	Benchmark items and toxicity analysis capture different aspects of quality; neither validates all real-world uses.	Official article/abstract record checked; full-text procedural cross-check still required.
Huang (2024)	Arabic instruction-tuned model	Authors' comparative performance claims need task-wise evaluation and independent replication.	Official article/abstract record checked; full-text procedural cross-check still required.
Koto (2024)	MSA knowledge benchmark	Predominantly written MSA multiple-choice knowledge, not spontaneous dialect dialogue.	Official article/abstract record checked; full-text procedural cross-check still required.
Mubarak (2024)	Arabic factuality	Dataset-specific factual-error proportion, not a population-wide hallucination rate.	Official article/abstract record checked; full-text procedural cross-check still required.

Study	Evidence domain	Primary interpretive caution	Verification status
Mousi (2025)	Dialect/culture benchmarks	Machine-translation generation and post-editing affect linguistic naturalness.	Official article/abstract record checked; full-text procedural cross-check still required.
Nacar (2025)	Cultural appropriateness	Expert ratings describe sampled questions, not exhaustive cultural or religious consensus.	Official article/abstract record checked; full-text procedural cross-check still required.
AlBallaa (2025)	Math + comprehension reasoning	An official correction was published in Dec 2025; recheck corrected version for precise reuse.	Official article/abstract record checked; full-text procedural cross-check still required.
Ben Hassine (2025)	Tunisian response generation	Automatic-judge calibration within this benchmark cannot be assumed transferable.	Official article/abstract record checked; full-text procedural cross-check still required.
Al-Matham (2025)	Broad-task blind platform	Dataset overlap and changes in platform version require continued audit.	Official article/abstract record checked; full-text procedural cross-check still required.
Zeinalipour (2025)	Grounded multilingual-variety dialogue	Model-generated corpus, even when reviewed, differs from naturally observed conversation.	Official article/abstract record checked; full-text procedural cross-check still required.
Altakrori (2026)	Five-dialect knowledge benchmark	Manual adaptation improves fit but does not prove representative real-world dialect proficiency.	Official article/abstract record checked; full-text procedural cross-check still required.
Al-Monef (2026)	Saudi regional words/proverbs	Saudi-focused geographical coverage; should not be generalized to all Arabic dialects.	Official article/abstract record checked; full-text procedural cross-check still required.

Table 2 shows why neither dataset size nor headline accuracy should substitute for a validity argument. Its final column explicitly limits the scope of checking.

3.2 Quantitative but non-equivalent benchmark coverage

ArabicMMLU's 14,575 items and 40 tasks sample formal knowledge and written comprehension; GATmath and GATLc sample Arabic reasoning and language comprehension from aptitude-exam materials; and BALSAM broadens NLP-task coverage through blind evaluation (Koto et al., 2024; AlBallaa et al., 2025; Al-Matham et al., 2025). These datasets differ not only in size but in who produced the items, the possibility of prior exposure, and what successful responses mean. Larger datasets can reduce sampling noise without curing construct undercoverage or annotation error.

AlBallaa et al. (2025) evaluate seven models and report highest observed accuracies of 66.9% on GATmath and 64.3% on GATLc under their testing protocol. A correction to this PLOS ONE article appeared in December 2025, making use of its amended version essential. The results should not be juxtaposed as if equivalent to the ArabicMMLU 62.3% value or to BALSAM evaluation scores because question distributions and protocols differ.

3.3 Dialectal and cultural evaluation

AraDiCE's roughly 45,000 human-post-edited dialect examples, DialectalArabicMMLU's 15,000 dialectal items across 32 domains, and TounsiBench's 744 native reference responses are three distinct ways to bridge MSA-centred evaluation (Mousi et al., 2025; Altakrori et al., 2026; Ben Hassine et al., 2025). AraDiCE blends machine translation and revision; DialectalArabicMMLU adapts an MMLU-style question set; TounsiBench elicits contextually grounded Tunisian instructions and includes human judgments. A claim of dialectal competence needs to specify which form of linguistic evidence was actually sampled.

Absher studies culturally situated Saudi dialect items drawn from words, phrases and proverbs across five regions, totaling more than 18,000 multiple-choice questions. The authors evaluate six LLMs on meaning, contextual usage and cultural dimensions and report heterogeneous performance across varieties (Al-Monef et al., 2026). That geographical depth is valuable, but it does not establish national or pan-Arabic representativeness beyond the sampled material.

The cultural benchmark by Nacar et al. (2025) demonstrates another measurement dimension: eleven experts assessed over 2,500 questions for categories including cultural appropriateness, religion and bias. Cultural disagreements cannot always be reduced to a single correct answer. Annotation protocols must disclose who judged the materials, whether disagreements were retained, and whether plural regional perspectives were represented.

3.4 Truthfulness, dialogue and measurement validity

Halwasa's 10,000 generated Arabic sentences and reported 25% factually incorrect rate are an example of error-based assessment rather than general language proficiency testing (Mubarak et al., 2024). The dataset-specific proportion does not estimate hallucination rates in every Arabic model or use case. Grounded tasks require reference documents, clearly defined temporal context and a distinction among factual error, unsupported assertion and linguistically malformed expression.

Shawarma Chats builds 30,000 six-turn conversations—10,000 per MSA, Egyptian and Maghrebi variety—by using generative models, human comparison and a

repair process (Zeinalipour et al., 2025). The approach offers a sizeable resource, but the origin of the dialogue matters: machine-produced, repaired exchanges are not equivalent to naturally occurring user conversations. The contrast reinforces Figure 2, which links dataset provenance to the validity of a deployment-oriented claim.

Figure 2. Analytic chain connecting Arabic-language data to score interpretation.

Figure 2. From training resources to defensible evaluation claims



Conceptual synthesis drawn by the authors; arrows denote review logic, not a tested causal pathway.

Figure 2 should be read alongside Tables 1 and 2: a test score is informative only when the intended construct and deployment context match the sampled tasks.

4. Discussion

4.1 Central argument: a benchmark score is not a general Arabic capability score

The candidate literature suggests a persistent construct-to-claim problem. Accuracy on written MSA school-examination questions is evidence about the sampled knowledge tasks, not evidence of robust dialectal conversational interaction. Performance on a culturally sensitive multiple-choice item is not necessarily evidence of safe counseling or reliable document-grounded advice. The crucial question is therefore not only whether a model ranks first but whether the assessment supports the proposed use of its score.

This problem is visible by comparing Table 1 with Table 2. Studies construct data through examinations, machine translation, human authoring, expert cultural coding or model-generated conversations. Each pathway produces different benefits and possible biases. In an evidence review, methodological diversity must be explained rather than eliminated through an arbitrary unified performance index.

4.2 Arabic varieties and the danger of benchmark translation effects

Conventional benchmark expansion often proceeds by translating or adapting tasks from another language or from MSA. That process can preserve task comparability but may import culturally anomalous scenarios, unnatural word choice or register shifts. AraDiCE's post-editing and DialectalArabicMMLU's manual adaptation attempt to mitigate such issues, yet their outputs still have designed rather than population-sampled provenance (Mousi et al., 2025; Altakrori et al., 2026).

Authenticity should be demonstrated through native annotator variation, geographic sampling, agreement statistics and error taxonomies for code-switching, borrowed vocabulary, spelling variation and diacritics. Tunisian Arabic and five-dialect evaluations are therefore complementary rather than redundant. No single dataset captures the full diversity of Arabic-speaking communities. Figure 3 displays these non-equivalent domains and Table 2 records corresponding threats.

Figure 3. Four complementary, non-interchangeable domains of Arabic model evaluation.

Figure 3. Non-interchangeable dimensions of Arabic LLM performance



Each dimension needs fit-for-purpose tasks and human validity checks.

Figure 3 underscores that linguistic accuracy, cultural fit and factuality require different reference evidence and error analyses.

4.3 Culture, fairness and safety require human-centered evidence

A preference-optimized model can still fail on community-specific norms or factually consequential questions. AceGPT includes an alignment component intended to account for Arabic cultural values (Huang et al., 2024), while expert assessment work demonstrates why culturally contextualized judgments remain necessary (Nacar et al., 2025). Such work should avoid presenting one set of locally collected labels as culturally universal.

Model-based grading may scale evaluation but adds another system whose reliability must itself be tested. TounsiBench's comparison of human and LLM-based judgments makes this concern visible (Ben Hassine et al., 2025). Future work should report inter-rater reliability, disagreement reasons, audit samples, calibration across dialects and harmful-error categories before conclusions about safety or fairness are generalized.

4.4 Technical reproducibility and benchmark governance

Publicly available benchmarks permit repeated comparison but may later enter model training data. BALSAM's blind-test orientation is an important practical response to this risk (Al-Matham et al., 2025). Independent decontamination analyses, refreshed hidden evaluations, versioned data splits and auditable scoring scripts would strengthen both credibility and comparability. These measures are not simply engineering conveniences: they protect the interpretation of apparent model improvement.

Standardized reporting should include tested model revision, system prompting, number of examples, response generation settings, handling of refusals, scoring scripts, dataset license and disclosed uncertainty. Figure 4 converts these requirements into an editorial and research design agenda. Without such details, replication of a benchmark paper may be limited even when model weights or some sample tasks are shared.

Figure 4. Proposed research and reporting priorities for responsible Arabic LLM evaluation.

Figure 4. A reproducibility-focused validation agenda



Future research priorities proposed from cross-study differences; not prevalence estimates.

Figure 4 proposes an audit sequence useful for researchers and reviewers; validation should not end at leaderboard reporting.

4.5 Implications for language education and applications

Arabic educational uses offer particularly direct examples of why metrics need interpretation. A model might answer grammar questions accurately but give misleading explanations, overlook dialect–MSA distinctions or produce culturally inappropriate role-play dialogue. Practical validation should therefore include teacher or learner ratings, longitudinal evaluation when educational outcomes are claimed, and accessible deployment procedures.

Researchers should avoid importing ranking results into policy without documenting target users, expected tasks, potential harm, and consequences of incorrect answers. A future systematic review could focus on a narrower, comparable decision problem—for example, reliably evaluating Arabic writing feedback—but the present evidence map does not justify such a pooled conclusion.

5. Limitations

5.1 Review limitations

The current article remains a substantial editorial revision of a targeted seed review, not an independently completed scoping review. The source manuscript presented 13 identified records but did not provide Scopus, Web of Science, ACL export logs, duplicate counts, two-reviewer full-text screening or exclusion reasons. It would be methodologically incorrect to portray those records as a final PRISMA-included population. Moreover, core original studies were checked through official publisher records and abstracts; not every experimental detail has been independently extracted from full texts.

Search-language and venue choices may privilege highly indexed English-language accounts of Arabic NLP. Preprints and technical reports not yet peer reviewed, closed industrial evaluations and non-indexed regional work may add substantial evidence. Quantitative figures from original abstracts are kept within their reported study settings.

5.2 Actions needed before a submission claim

Finalization requires executing dated searches, screening and charting with at least two assessors, resolving each reference and claims ledger against complete articles, and generating an auditable PRISMA-ScR diagram. The results and conclusions must be recalibrated if new studies substantially change coverage. The manuscript also needs confirmed author metadata, declarations, and explicit journal checks.

6. Conclusion

The provisional record shows a shift from Arabic generative model construction to evaluation ecosystems that increasingly consider general reasoning, dialects, culturally contextualized language and factuality. ArabicMMLU, BALSAM, TounsiBench and related resources illuminate different abilities rather than interchangeable dimensions of one universal language score. A defensible evaluation needs clear data provenance, a fit-for-purpose construct, transparent prompting and scoring, human scrutiny and explicitly bounded inference.

The priority for further work is to connect benchmark design to authentic use, independent contamination auditing, dialect-specific evaluation, fairness across speaker communities and rigorous factuality analysis. These recommendations are an agenda derived from charted source differences and should be revisited after an exhaustive PRISMA-ScR screening process.

References

- Al-Matham, R., Darwish, K., Al-Rasheed, R., Alshammari, W., Alhoshan, M., Almazrua, A., Al Wazrah, A., Alheraki, M., Alam, F., Nakov, P., Alzahrani, N., AlBilali, E., Habash, N., El-Sheikh, A., Elmallah, M., Li, H., Mubarak, H., Anwar, M., Alyafeai, Z., ..., AlOsaimy, A. (2025). BALSAM: A Platform for Benchmarking Arabic Large Language Models. In Proceedings of The Third Arabic Natural Language Processing Conference (pp. 258–277). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.arabnlp-main.21>
- Al-Monef, R., Alhuzali, H., Alturayef, N., & Alasmari, A. (2026). From words to proverbs: Evaluating LLMs' linguistic and cultural competence in Saudi dialects with Absher. Alexandria Engineering Journal, 137, 25–41. <https://doi.org/10.1016/j.aej.2025.12.066>
- AlBallaa, S., AlTwaresh, N., AlSalman, A., & Alfarhood, S. (2025). GATmath and GATLc: Comprehensive benchmarks for evaluating Arabic large language models. PLOS ONE, 20(9), e0329129. <https://doi.org/10.1371/journal.pone.0329129>
- Almazrouei, E., Cojocaru, R., Baldo, M., Malartic, Q., Alobeidli, H., Mazzotta, D., Penedo, G., Campesan, G., Farooq, M., Alhammedi, M., Launay, J., & Noune, B. (2023). AlGhafa Evaluation Benchmark for Arabic Language Models. In Proceedings of ArabicNLP 2023 (pp. 244–275). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.arabnlp-1.21>
- Altakrori, M. H., Habash, N., Lynn, T., Samih, Y., Freihat, A. A., Chirkunov, K., AbuOdeh, M., Florian, R., Nakov, P., & Aji, A. F. (2026). DialectalArabicMMLU: Benchmarking Dialectal Capabilities in Arabic and Multilingual Language Models. In Proceedings of the Fifteenth Language Resources and Evaluation Conference (pp. 3199–3219). ELRA Language Resource Association. <https://doi.org/10.63317/3cy68duew55b>
- Ben Hassine, S., Arrak, A., Addhoum, M., & Wilson, S. R. (2025). TounsiBench: Benchmarking Large Language Models for Tunisian Arabic. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 34627–34642). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.emnlp-main.1756>
- Huang, H., Yu, F., Zhu, J., Sun, X., Cheng, H., Dingjie, S., Chen, Z., Alharthi, M., An, B., He, J., Liu, Z., Chen, J., Li, J., Wang, B., Zhang, L., Sun, R., Wan, X., Li, H., & Xu, J. (2024). AceGPT, Localizing Large Language Models in Arabic. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 8139–8163). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-long.450>
- Koto, F., Li, H., Shatnawi, S., Doughman, J., Sadallah, A., Alraeesi, A., Almubarak, K., Alyafeai, Z., Sengupta, N., Shehata, S., Habash, N., Nakov, P., & Baldwin, T. (2024). ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 5622–5640). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.334>
- Mousi, B., Durrani, N., Ahmad, F., Hasan, M. A., Hasanain, M., Kabbani, T., Dalvi, F., Chowdhury, S. A., & Alam, F. (2025). AraDiCE: Benchmarks for Dialectal and Cultural Capabilities in LLMs. In Proceedings of the 31st International Conference on Computational Linguistics (pp. 4186–4218). Association for Computational Linguistics. <https://aclanthology.org/2025.coling-main.283/>

- Mubarak, H., Al-Khalifa, H., & Alkhalefah, K. S. (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) (pp. 8008–8015). ELRA and ICCL. <https://aclanthology.org/2024.lrec-main.705/>
- Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, 18, 143. <https://doi.org/10.1186/s12874-018-0611-x>
- Nacar, O., Sibae, S. T., Ahmed, S., Atitallah, S. B., Ammar, A., Alhabashi, Y., Al-Batati, A. S., Alsehibani, A., Qandos, N., Elshehy, O., Abdelkader, M., & Koubaa, A. (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 (pp. 387–401). Association for Computational Linguistics. <https://aclanthology.org/2025.loreslm-1.29/>
- Nagoudi, E. M. B., Elmadany, A., & Abdul-Mageed, M. (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) (pp. 628–647). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.47>
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., ... Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467–473. <https://doi.org/10.7326/M18-0850>
- Zeinalipour, K., Saad, M. Z., Attafi, O., Maggini, M., & Gori, M. (2025). Shawarma Chats: A Benchmark Exact Dialogue & Evaluation Platter in Egyptian, Maghrebi & Modern Standard Arabic—A Triple-Dialect Feast for Hungry Language Models. In Proceedings of The Third Arabic Natural Language Processing Conference (pp. 472–524). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.arabicnlp-main.39>