



Artificial Intelligence for Arabic Natural Language Processing: A Scoping Review of Applications, Resources, Evaluation Practices, and Evidence Gaps

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 natural language processing (NLP) increasingly uses transformer models and generative artificial intelligence, yet linguistic variety and evaluation heterogeneity challenge interpretation. This scoping-review manuscript maps applications, datasets, assessment practices and gaps across Arabic text studies from 2021 to October 2026. A Population–Concept–Context framework and PRISMA-ScR informed the planned search and charting approach. An exploratory collection of 23 original journal articles documents dialect identification, classification, translation, question answering, summarization, error correction and culturally situated evaluation. Across these studies, model scores depend strongly on dialect, genre, annotation protocols and benchmark design. The principal implication is the need for dialect-disaggregated tests, transparent dataset provenance, stronger external validation and qualified human assessment. Comprehensive database retrieval and independently documented screening remain necessary before this evidence map can be reported as a completed scoping review.

Keywords: Arabic NLP; artificial intelligence; Arabic dialects; language resources

Introduction

Arabic-language AI is not a single technical domain. It involves written Modern Standard Arabic, contemporary colloquial varieties, historical and religious registers, code-switching, and heterogeneous orthographic conventions. Systems trained or evaluated on one linguistic setting cannot be assumed to generalize to another. A Moroccan named entity recognition study (Subait et al., 2025), a Classical Arabic NER evaluation (Salah et al., 2026), and a Saudi cultural benchmark (Ayash et al., 2025) pose fundamentally different modeling and validity questions despite their shared association with Arabic text.

Recent journal studies report increasingly sophisticated models, including contextual embeddings, transformer representations, transfer learning, and generative large language models (LLMs). Yet the literature is dispersed by task: Mohammed et al. (2023) study dialect identification, Faheem et al. (2024) target Egyptian-to-MSA translation, while Beidas et al. (2025) assess cross-dialectal outputs of general-purpose LLMs. The coexistence of mature classification benchmarks with emerging culturally grounded evaluations means that simply ranking models by published accuracy would conceal major differences in study design.

The evidence base also illustrates a tension between scale and relevance. New Arabic corpora include the 2,011-document Arabic Research Papers Dataset (Almutairi et al., 2025), a 5,009-question MSA benchmark (Abdelaziz et al., 2025), and a multi-topic stance dataset of 14,477 tweets (Alkhathlan et al., 2025). These

resources reduce particular bottlenecks but do not, by themselves, settle the question of representativeness across dialects, genres, or social contexts. Moreover, a model that performs well at classification may still fail when required to preserve cultural meaning or produce reliable domain-specific text.

A scoping review is appropriate when the central questions concern what has been studied, with which resources and measures, and where gaps in coverage remain, rather than estimating one pooled treatment effect. PRISMA-ScR provides a reporting framework for transparent evidence mapping (Tricco et al., 2018), and the distinction between scoping and focused effectiveness synthesis is emphasized by Munn et al. (2018).

This article therefore asks: (RQ1) Which Arabic text-processing applications are represented in recent original journal studies? (RQ2) Which Arabic varieties, data sources, and resource-building practices underpin the research? (RQ3) How are models evaluated, and what aspects of reliability, transfer, interpretability, and cultural competence are examined? (RQ4) Which evidence gaps should structure future Arabic AI research? The anticipated contribution is a task–resource–evaluation map, not a ranking of all Arabic language models.

2. Methods

2.1. Review design and scope

The review question is formulated using PCC: Population/field, computational studies of Arabic-language text; Concept, AI-enabled language understanding, generation, and language resources; Context, Modern Standard Arabic, regional dialects, Classical Arabic, and applied settings involving Arabic text. Eligible sources are peer-reviewed original journal articles published between 1 January 2021 and 10 October 2026. Text-based NLP, text-related datasets, model evaluation, and human assessment are in scope; speech-only recognition, optical character recognition without textual NLP, non-Arabic-only investigations, preprints without a journal record, editorial commentary, and review articles are outside the charted primary-study corpus.

2.2. Proposed database strategy

For Scopus, the planned search is: TITLE-ABS-KEY(("Arabic" OR "Arabic NLP" OR "Modern Standard Arabic" OR "dialectal Arabic" OR "Arabic dialect*" OR "Classical Arabic") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR transformer* OR "large language model*" OR "generative AI") AND ("natural language processing" OR "text classification" OR "language generation" OR "machine translation" OR "named entity recognition" OR "sentiment analysis" OR "question answering" OR corpus OR dataset)) AND PUBYEAR > 2020 AND PUBYEAR < 2027 AND DOCTYPE(ar) AND LANGUAGE(english). The Web of Science Core Collection analogue will use TS=(...) with the corresponding topic terms and filters for publication year, article document type, and English. Search refinements, exact query timestamps, database coverage, and exported records must be archived before these strings can be described as executed searches.

The preliminary corpus reported here was assembled through targeted discovery of publisher journal pages and bibliographic records, supplemented by direct DOI checks. This exploratory process is not equivalent to database-wide retrieval. No numerical Scopus/Web of Science search results or deduplication totals are asserted. Reference lists of relevant publications should be hand-searched, and all newly identified records added to an auditable screening log.

2.3. Screening and charting protocol

Two reviewers should independently screen titles and abstracts, resolve discrepancies by consensus or a third reviewer, then screen full texts using recorded exclusion reasons. Duplicate detection should prioritize DOI, followed by normalized title–year–author matching. The proposed charting form captures citation, year, publication type, task, Arabic variety, genre, dataset construction and availability, model family, training/test split, evaluation metrics, numerical results where directly comparable, human evaluation, limitations, and claims relevant to RQ1–RQ4. A pilot coding exercise and iterative calibration are advisable before complete extraction.

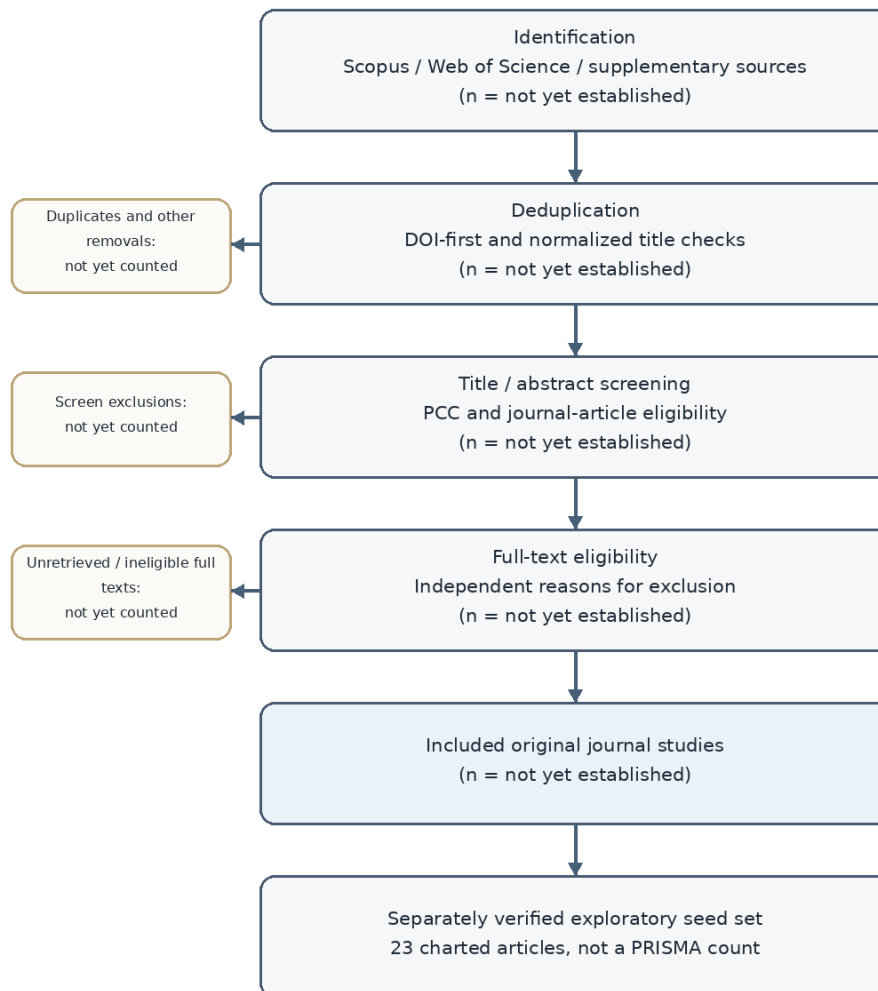
For synthesis, studies will be grouped by task and linguistic setting, and mapped descriptively by resource type and evaluation method. Since metrics such as exact match, macro-F1, accuracy, BLEU, ROUGE, and human preference refer to different targets, the review will not pool them as a single measure of ‘Arabic AI performance’. Critical appraisal is optional in the mapping design, but validity flags—sampling bias, leakage controls, dataset overlap, and external validation—should be charted when reported.

2.4. Search transparency and PRISMA status

Figure 1 adopts the identification–screening–eligibility–inclusion structure of the supplied PRISMA template. Its boxes deliberately show that database-level counts have not yet been established. The 23 verified journal studies charted below are a purposive, exploratory seed set, not the final number of studies selected through comprehensive PRISMA screening. Accordingly, frequency statements refer only to this seed set and must not be generalized to the total literature.

PRISMA-ScR WORKFLOW — COUNTS PENDING

Planned exhaustive screening | exploratory evidence map is separate



No database-screening denominators are claimed in this draft.

Figure 1. PRISMA-ScR screening plan adapted from the supplied flow-diagram framework. Numerical database and selection counts are pending; the 23 verified records constitute an exploratory seed set, not the final included-study denominator.

3. Results: Exploratory evidence map

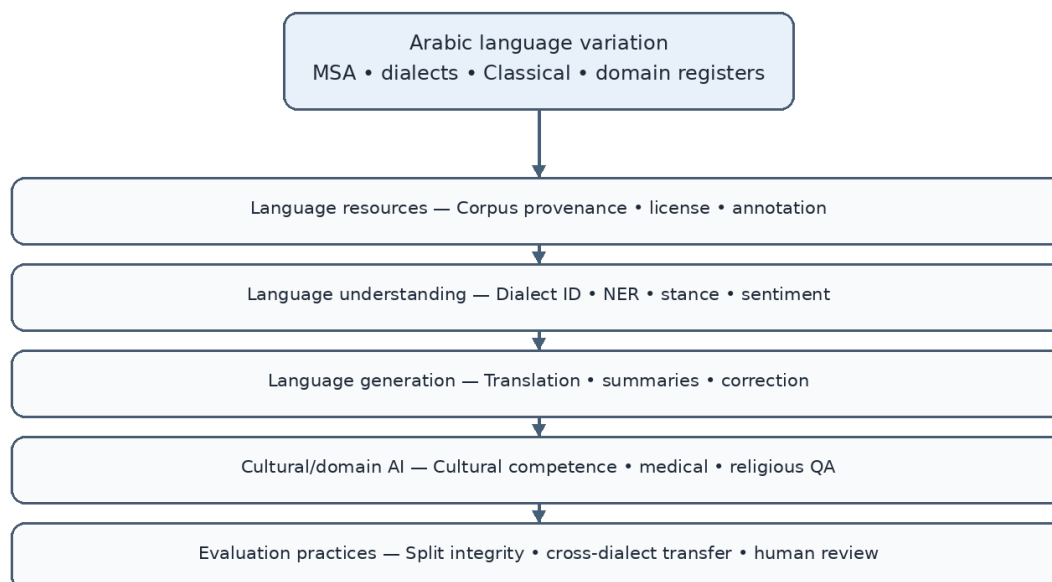
3.1. Application clusters

The seed literature can be organized into overlapping applications: (i) dialect identification and translation; (ii) classification of sentiment, stance, misinformation, and entities; (iii) question answering and language-resource development; (iv) text generation, summarization, and grammatical error correction; and (v) culturally situated evaluation. Figure 2 illustrates these domains

as connected layers of Arabic-language variation, resources, tasks, and evaluation rather than mutually exclusive subfields.

Tables 1 and 2 should be read together: Table 1 describes the evidence settings, while Table 2 distinguishes source-reported findings from transferability cautions. Figure 2 helps explain why identical metric names do not imply comparable populations.

A FRAMEWORK FOR MAPPING ARABIC-LANGUAGE AI EVIDENCE



Interpretive taxonomy; categories overlap and do not represent study counts.

Figure 2. Task–resource–evaluation taxonomy developed for the present evidence map. The domains overlap and do not imply comparative publication counts.

Dialect work highlights why Arabic should not be treated as a single homogeneous input distribution. Mohammed et al. (2023) evaluate representations on MADAR, NADI, and QADI; Alsuwaylimi (2024) introduces a 121,289-comment resource spanning four dialect groupings; Chabane et al. (2025) focus on low-label Algerian Arabic; and Faheem et al. (2024) address Egyptian-to-MSA neural translation. Each tackles a different grain of linguistic variation, reducing the interpretability of direct score comparisons across papers.

Classification-focused studies similarly cover heterogeneous outcomes. Aboulola and Umer (2024) report 98.42% test accuracy for Arabic fake-news classification under their study design, whereas Alkhathlan et al. (2025) report 0.67 overall F1 for multi-topic stance classification on ArabicStanceX. These numbers do not establish that one task or method is easier: annotation targets, class balance, text sources, and train/test arrangements differ. Jaber et al. (2025) explicitly contrast balanced and imbalanced sentiment collections; Ullah et al. (2025) provide within-study comparisons between multilingual transformers for Arabic sentiment.

3.2. Language-resource diversification

The source domains are extending beyond general news and social media. Almutairi et al. (2025) release Arabic academic papers covering seven subject fields. Abdelaziz et al. (2025) introduce AAFAQ, a question-classification dataset with multiple linguistic and cognitive attributes. Aleid and Azmi (2025) assemble an annotated Hajj-fatwa question-answering resource, and Lee et al. (2025) investigate

knowledge distillation for Arabic medical summaries using synthetic dialogues. These contributions broaden the empirical repertoire, but differ in whether the data are naturally occurring, generated, expert annotated, or domain restricted.

The contrast across corpora also motivates a stronger conception of resource quality. Public availability alone is insufficient: documentation should specify sampling frame, dialectal distribution, annotation expertise, restrictions on sensitive text, versioning, licensing, and reproducible splits. In domains involving religious interpretation or clinical communication, formal task accuracy should be supplemented by domain-expert review before deployment-oriented conclusions are drawn.

3.3. Evidence charting

Table 1 charts every study in the exploratory seed set, with its linguistic setting and methodological object. Table 2 provides a companion one-row-per-study synthesis of the directly supported contribution and an explicit generalizability or reporting caution. This pairing makes the boundary between what a source reports and what the review infers auditable.

Table 1. Exploratory study characteristics: original journal articles (one row per verified study; n = 23).

ID	Study	Application and Arabic variety	Method / model family	Data or benchmark
S01	Mohammed et al., 2023	Dialect identification Multiple dialects	Three-stage LSTM/Transformer dialect representation	MADAR, NADI and QADI benchmarks
S02	Faheem et al., 2024	Dialect-to-MSA translation Egyptian → MSA	Low-resource neural MT with non-parallel corpora	Egyptian dialect and MSA materials
S03	Alsuwaylimi, 2024	Dialect identification Egyptian, Jordanian, Gulf, Yemeni	CAMeLBERT/BiLSTM and ALBERT/BiLSTM	121,289 social-media comments
S04	Aboulola & Umer, 2024	Misinformation classification Arabic news text	ELMo embeddings; CNN/LSTM voting; LIME	Arabic fake-news benchmark
S05	Miah et al., 2024	Multilingual sentiment Arabic plus three other languages	Translation to English; transformer/LLM ensemble	Multilingual sentiment corpora
S06	Chabane et al., 2025	Dialect identification Algerian Arabic	CAMeLBERT/FastText transfer from MSA	Algerian dialect low-label setting
S07	Almutairi et al., 2025	Language resources Academic Arabic	Corpus construction, document categorization	ARPD: 2,011 documents, seven fields
S08	Abdelaziz et al., 2025	Language resources / QA Modern Standard Arabic	Question taxonomy, annotated benchmark and classification	AAFAQ: 5,009 MSA questions
S09	Aleid & Azmi, 2025	Domain question answering Religious/legal Arabic	Manually annotated benchmark; model evaluation	Hajj fatwa questions and answers
S10	Ayash et al., 2025	Cultural LLM evaluation Saudi regional language and culture	Native-context benchmark; six LLMs	Five Saudi regions plus general-cultural questions
S11	Beidas et al., 2025	LLM dialect translation MSA and multiple dialects	Comparisons of GPT family and Bard/Gemini	QADI and MADAR datasets
S12	Alkhathlan et al., 2025	Stance classification Saudi social-media Arabic	MARBERTv2 stance detection; zero/few-shot	14,477 tweets across 17 topics
S13	Jaber et al., 2025	Sentiment classification MSA and Syrian Arabic tweets	Word embeddings and voting ensembles	ArTwitter and Syria_Tweets
S14	Ouchene & Bessou, 2025	Dialect sentiment Algerian Arabic	Benchmarking sentiment methods	Algerian X/Twitter posts

ID	Study	Application and Arabic variety	Method / model family	Data or benchmark
S15	Essa et al., 2025	Abstractive summarization Written Arabic	NER-enhanced text generation	EASC dataset
S16	El-Sayed et al., 2025	Error analysis / XAI Arabic emotion tweets	Human-in-the-loop error analysis; LIME/SHAP	Arabic emotion-classification dataset
S17	Lee et al., 2025	Domain summarization Arabic medical dialogue	Knowledge distillation to Arabic clinical summarizer	Synthetic Arabic medical dialogues
S18	Ullah et al., 2025	Multilingual sentiment Arabic among eight languages	Multilingual transformers and prompt tuning	Sentiment datasets in eight languages
S19	Almarzoqi & Alsuhaibani, 2025	Cultural text / translation Quranic source and translations	Emotion detection with LLMs	Quran translation material
S20	Mahmoud et al., 2026	Grammatical error correction Modern Standard Arabic	Unified correction and explanation with synthetic data	ExplAGEC ~21.8 million pairs; QALB benchmarks
S21	Salah et al., 2026	Classical Arabic NER Classical Arabic	Rules, genetic feature selection and Naive Bayes	CANER with leakage-controlled split
S22	Al-Monef et al., 2026	Dialect/cultural LLM evaluation Saudi regional dialects	Dialect-word/proverb multiple-choice LLM benchmark	Absher: >18,000 questions in six categories
S23	Subait et al., 2025	Dialect named entity recognition Moroccan Arabic	AI-based named entity recognition	Moroccan-dialect corpus

4. Discussion

4.1. Benchmark success is not evidence of cross-variety competence

The core pattern of the mapped evidence is one of increasingly specialized success under increasingly different validation settings. A dialect identifier, a Classical Arabic named-entity recognizer, and a culturally oriented LLM benchmark do not measure a common capability. Salah et al. (2026) report micro-F1 of 93.0% and macro-F1 of 88.7% on a Classical Arabic named-entity corpus; Subait et al. (2025) instead examine Moroccan Arabic entity recognition. Their populations, target labels, linguistic conventions, and application domains differ. A useful field-level synthesis should therefore chart task-specific evidence maturity and external validity, not rank headline scores.

The same difficulty occurs within sentiment and translation. Miah et al. (2024) assess a translation-mediated multilingual sentiment ensemble, so its reported overall accuracy is not an Arabic-only estimate. Ullah et al. (2025) compare multilingual encoders under a common experiment and report an Arabic F1 of 0.60 for XLM-RoBERTa versus 0.48 for multilingual BERT in the standard fine-tuning setting. Within-study comparisons are informative; cross-paper score ordering without aligned corpora, metrics, and protocols is generally not.

4.2. The tension between linguistic coverage and evaluation convenience

As illustrated in Figure 3, regional, historical and sensitive-domain coverage should be assessed separately. Figure 1 remains a protocol schematic rather than a report of completed screening, and its counts require auditable database exports.

Several seed studies construct new benchmarks, but their coverage is naturally constrained by the source setting. ArabicStanceX concerns Saudi social media (Alkhathlan et al., 2025); the Algerian sentiment benchmark concerns X posts (Ouchene & Bessou, 2025); the Hajj-FQA resource targets a specific religious-information task (Aleid & Azmi, 2025). These are valuable, sharply defined contributions. It would be unwarranted, however, to assume that results transfer to other Arab societies, historical registers, or settings where errors carry more serious consequences.

A promising development is a shift toward explicit cultural and regional evaluations. Ayash et al. (2025) show weaker regional than general cultural accuracy among tested models; Al-Monef et al. (2026) examine comprehension of Saudi dialect words, phrases, and proverbs with more than 18,000 questions. Beidas et al. (2025) study variation in LLM translation across dialects and prompting settings. Together, these studies imply that lexical fluency or general task accuracy is an incomplete proxy for dialect fidelity and cultural competence. The priority is not merely to add more Arabic examples, but to make variation, provenance, and evaluation personnel visible.

4.3. Explainability, data-centric methods, and specialized generation

The seed map includes studies that treat data quality and interpretability as empirical objects. El-Sayed et al. (2025) use a human-in-the-loop framework and explanation methods to investigate anomaly patterns in Arabic emotion classification. Jaber et al. (2025) examine ensemble methods across data distributions with markedly different class balance. Aboulola and Umer (2024) integrate LIME explanations into misinformation classification. These designs demonstrate that auditability cannot be reduced to a single post-hoc explanation figure; it requires transparent data generation, label checks, error decomposition, and clearly described model comparisons.

For generation-oriented systems, there is a parallel need to separate surface-quality metrics from semantic and domain-specific validity. Essa et al. (2025) evaluate

NER-enhanced abstractive summarization with ROUGE and additional coherence measures. Lee et al. (2025) report automated and Arabic-speaking evaluator judgments for domain-specific clinical summarization. Mahmoud et al. (2026) unify Arabic grammatical correction with explanatory outputs, using a large synthetic corpus. Such examples suggest that Arabic generation research is moving from task completion toward richer output audits, although synthetic-data bias, factual fidelity, and expert validation must remain explicit evaluation dimensions.

4.4. A cross-cutting agenda for Arabic AI

Three priorities follow from the charted evidence. First, benchmark protocols should be disaggregated by Arabic variety, genre, and data provenance, with held-out dialect or domain tests where feasible. Second, reporting must separate overall accuracy from macro-level and minority-class performance; the contrasting task-level metrics collected in Table 2 exemplify why standardized reporting descriptors matter. Third, culturally sensitive and high-stakes settings warrant linguistically qualified human assessments in addition to automated scores. Figure 3 organizes these priorities as an interpretive gap matrix; the matrix is a synthesis of concerns illustrated by the seed studies, not a statistical estimate of field-wide deficiencies.

EVIDENCE-GAP MATRIX — INTERPRETIVE (NON-QUANTITATIVE)

Evidence setting	Model / data coverage	Evaluation design	Priority implication
MSA formal text	Relatively visible	Task metrics emphasized	Add external-domain checks
Regional dialects	Uneven by region	Cross-dialect tests sparse	Dialect-disaggregated reports
Classical Arabic	Specialized corpora	Limited genre transfer	Historical-genre evaluation
Cultural contexts	New benchmarks	Human nuance inconsistent	Native-speaker evaluation
Sensitive domains	Emerging QA / medical	Safety and deployment scarce	Human-in-loop validation

Qualitative interpretation of the exploratory map; not an estimate of prevalence.

Figure 3. Interpretive evidence-gap matrix derived from the provisional study map. Terms describe thematic issues, not measured prevalence or definitive absence of research.

Taken together, the discussion reframes progress as an interaction between model capability and evaluation validity. A strong model result is meaningful only in relation to the linguistic population, source dataset, task objective, and error tolerance it represents. This argument explains why comparisons across SaudiCulture, ArabicStanceX, AAFAQ, and domain-specific summarization systems should be organized by evidence function rather than interpreted as a single league table of Arabic AI.

Table 2. Reported contribution, scope caveat, and evidence-mapping implication (one row per verified study; n = 23).

ID	Source-supported observation	Interpretation boundary
S01	Reports improvements on multiple dialect benchmarks by identifying dialect-relevant information.	Cross-benchmark comparability and dialect coverage need audit.
S02	Explores non-parallel data for Egyptian-to-MSA translation under resource constraints.	Domain and dialect specificity.
S03	The CAMeLBER-BiLSTM hybrid reports 87.67% accuracy on the study's dialect classification setting.	Only four dialect groupings; social-media source.
S04	Reports 98.42% accuracy under its own benchmark and cross-validation protocol.	High test accuracy does not establish transfer to unseen events.
S05	Reports over 86% accuracy overall for a translation-mediated multilingual ensemble.	Aggregate cross-lingual figure is not an Arabic-only score.
S06	Investigates cross-lingual transfer for dialect identification with few labels.	Transfer effectiveness may vary by dialectal distance.
S07	Provides an open corpus of Arabic research papers across seven academic categories.	Academic-register coverage is narrower than spoken dialects.
S08	Releases a question dataset labeled across linguistic and cognitive dimensions.	MSA-only resource; underrepresented content categories.
S09	Introduces human-annotated domain-specific Arabic QA data and evaluates language models.	Specialized domain limits unqualified generalization.
S10	Finds regional cultural accuracy below general-cultural accuracy for tested models.	Region-bound cultural tasks; scoring/rubric detail matters.
S11	Evaluates model variation across dialects and prompt conditions using multiple translation metrics.	Automatic metrics do not fully capture contextual appropriateness.
S12	Creates a multi-topic dataset and reports an overall stance F1 of 0.67 in the reported setting.	Single platform and country; neutral category difficult.
S13	Reports 90.22% accuracy for a specified ensemble on ArTwitter.	Results differ across balanced and imbalanced sets.
S14	Provides an Algerian-dialect sentiment benchmark comparison.	Platform/topic-specific data and dialect coverage.
S15	Reports a ROUGE-1 of 74% for the specific evaluated summarization system.	ROUGE alone insufficient to establish factual faithfulness.

ID	Source-supported observation	Interpretation boundary
S16	Identifies spurious correlations and potential dataset anomalies through explanations.	Single task case study; XAI patterns need human verification.
S17	Reports domain-specific AraSum evaluation relative to JAIS on automated and human measures.	Synthetic-data and clinical external-validation concerns.
S18	Reports Arabic F1 values of 0.60 for XLM-RoBERTa and 0.48 for multilingual BERT under its standard fine-tuning setting.	Within-study comparison, not universal Arabic model ranking.
S19	Studies preservation and detection of emotion in Quran translations using language models.	Translation and expert-interpretation sensitivity.
S20	Reports F0.5 of 80.02 and 81.73 on QALB-2014/2015 with its AraT5 variant.	Synthetic-data reliability and benchmark transfer.
S21	Reports micro-F1 of 93.0% and macro-F1 of 88.7% on its evaluated corpus.	Classical-Arabic setting distinct from modern dialect tests.
S22	Reports gaps in proverb/context understanding across Arabic and multilingual LLMs.	Benchmark artifacts and regional representativeness.
S23	Investigates named-entity recognition in Moroccan Arabic.	Dialect and genre transfer require independent testing.

5. Limitations and priorities for the final review

This manuscript is a substantive first draft, not a completed systematic evidence selection exercise. The 23 mapped original articles were confirmed individually through publisher or bibliographic records, but the retrieval strategy has not been executed in subscription databases, independent screening has not been performed, and full-text extraction has not been replicated by a second coder. Inference is therefore restricted to an exploratory map. The desired final sample of approximately 40–60 eligible papers is not a target to be achieved by relaxed criteria: the eventual number must be determined by documented search and screening.

Additional risks concern publication and source selection bias, coverage of Arabic-language articles where the current journal-only/English limits may suppress locally relevant research, incomplete model-version documentation, possible data overlap between benchmarks, and rapidly evolving commercial model behavior. Full database exports and deduplication logs should be retained. A PRISMA-ScR flow diagram must replace Figure 1 with actual identification, removal, screening, retrieval, and inclusion counts before submission.

Future empirical research should prioritize longitudinal benchmark maintenance, code-switching and underserved dialects, datasets originating beyond high-volume social media sources, domain-shift tests, explicit leakage checks, multilingual-versus-Arabic-specific comparisons with identical data, compute/energy reporting, and carefully designed native-speaker and domain-expert evaluation. These are research implications from the present illustrative coverage, not claims that no prior studies exist on these subjects.

6. Conclusion

The preliminary map of 23 DOI-verified original journal studies depicts Arabic AI research as a varied evidence landscape spanning classification, generative text processing, resource creation, and culturally situated evaluation. The most defensible synthesis is not that any architecture is universally superior, but that claims of progress depend on precisely described Arabic varieties, data provenance, training and testing procedures, and evaluation constructs. An exhaustive database search, independent screening, and final PRISMA-ScR reporting are required to transform this first draft into a publication-ready scoping review. The task–resource–evaluation framework provided here offers a structure for completing that work without overstating the available evidence.

References

- Abdelaziz, M. E., Deif, M. A., Algamdi, S. A., & Elgohary, R. (2025). A benchmark Arabic dataset for Arabic question classification using AAFAQ framework. *Scientific Data*, 12, Article 1444. <https://doi.org/10.1038/s41597-025-05688-0>
- Aboulola, O. I., & Umer, M. (2024). Novel approach for Arabic fake news classification using embedding from large language features with CNN-LSTM ensemble model and explainable AI. *Scientific Reports*, 14, Article 30463. <https://doi.org/10.1038/s41598-024-82111-5>
- 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>
- Aleid, H. A., & Azmi, A. M. (2025). Hajj-FQA: A benchmark Arabic dataset for developing question-answering systems on Hajj fatwas. *Journal of King Saud University – Computer and Information Sciences*, 37, Article 135. <https://doi.org/10.1007/s44443-025-00128-w>

- Alkathlan, A., Alahmadi, F., Kateb, F., & Al-Khalifa, H. (2025). Constructing and evaluating ArabicStanceX: A social media dataset for Arabic stance detection. *Frontiers in Artificial Intelligence*, 8, Article 1615800. <https://doi.org/10.3389/frai.2025.1615800>
- Almarzoqi, A., & Alsuhaibani, M. (2025). Leveraging large language models for detecting and preserving emotions in Quran translations. *Journal of King Saud University – Computer and Information Sciences*, 37, Article 271. <https://doi.org/10.1007/s44443-025-00269-y>
- Almutairi, T. M., Saifuddin, S. R., Alotaibi, R. M., & Sarhan, S. (2025). Open source Arabic research paper dataset for natural language processing. *Scientific Reports*, 15, Article 31631. <https://doi.org/10.1038/s41598-025-16647-5>
- Alsawaylimi, A. A. (2024). Arabic dialect identification in social media: A hybrid model with transformer models and BiLSTM. *Heliyon*, 10(17), Article e36280. <https://doi.org/10.1016/j.heliyon.2024.e36280>
- Ayash, L., Alhuzali, H., Alasmari, A., & Aloufi, S. (2025). SaudiCulture: A benchmark for evaluating large language models' cultural competence within Saudi Arabia. *Journal of King Saud University – Computer and Information Sciences*, 37, Article 123. <https://doi.org/10.1007/s44443-025-00137-9>
- Beidas, A., Mohi, K., Ghaddar, F., Ahmad, I., & Abed, S. (2025). Cross-dialectal Arabic translation: Comparative analysis on large language models. *Frontiers in Artificial Intelligence*, 8, Article 1661789. <https://doi.org/10.3389/frai.2025.1661789>
- Chabane, M., Harrag, F., & Shaalan, K. (2025). Advancing low-resource dialect identification: A hybrid cross-lingual model leveraging CAMELBERT and FastText for Algerian Arabic. *Expert Systems with Applications*, 284, Article 127816. <https://doi.org/10.1016/j.eswa.2025.127816>
- El-Sayed, A., Nasr, A., Mohamed, Y., Alaaeldin, A., Ali, M., Salah, O., Khalid, A., & Lazem, S. (2025). A data centric HitL framework for conducting a systematic error analysis of NLP datasets using explainable AI. *Scientific Reports*, 15, Article 30406. <https://doi.org/10.1038/s41598-025-13452-y>
- Essa, N., El-Gayar, M. M., & El-Daydamony, E. M. (2025). Enhanced model for abstractive Arabic text summarization using natural language generation and named entity recognition. *Neural Computing and Applications*, 37, 7279–7301. <https://doi.org/10.1007/s00521-024-10949-x>
- Faheem, M. A., Wassif, K. T., Bayomi, H., & Abdou, S. M. (2024). Improving neural machine translation for low resource languages through non-parallel corpora: A case study of Egyptian dialect to modern standard Arabic translation. *Scientific Reports*, 14, Article 2265. <https://doi.org/10.1038/s41598-023-51090-4>
- Jaber, A., Bahati, I., & Martínez, P. (2025). Leveraging pre-trained embeddings in an ensemble machine learning approach for Arabic sentiment analysis. *Frontiers in Artificial Intelligence*, 8, Article 1653728. <https://doi.org/10.3389/frai.2025.1653728>
- Lee, C., Kumar, S., Vogt, K. A., Munshi, M., Tallapudi, P., Vogt, A., Awad, H., & Khan, W. (2025). Democratizing cost-effective, agentic artificial intelligence to multilingual medical summarization through knowledge distillation. *Scientific Reports*, 15, Article 27619. <https://doi.org/10.1038/s41598-025-10451-x>
- Mahmoud, Z., Zappatore, M., Kryvinska, N., Abdalsalam, M., Alfatemi, A., & Solyman, A. (2026). MTAGEC: Multi-task Arabic grammatical error correction as a sequence generation with synthetic data. *Journal of King Saud University –*

- Computer and Information Sciences, 38, Article 35. <https://doi.org/10.1007/s44443-025-00354-2>
- Miah, M. S. U., Kabir, M. M., Sarwar, T. B., Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports*, 14, Article 9603. <https://doi.org/10.1038/s41598-024-60210-7>
- Mohammed, A., Zheng, J., & Murtadha, A. (2023). A three-stage neural model for Arabic dialect identification. *Computer Speech & Language*, 80, Article 101488. <https://doi.org/10.1016/j.csl.2023.101488>
- Ouchene, L., & Bessou, S. (2025). Benchmarking sentiment analysis of Algerian Arabic dialect on X (Twitter). *Journal of King Saud University – Computer and Information Sciences*, 37, Article 358. <https://doi.org/10.1007/s44443-025-00295-w>
- Salah, R., Mukred, M., Zakaria, L. Q. B., Al-Yarimi, F. A. M., Khasro, I. T. N., & Sadiq, A. S. (2026). Resource efficient hybrid baseline for named entity recognition in classical Arabic. *Scientific Reports*, 16, Article 570. <https://doi.org/10.1038/s41598-025-30171-6>
- Subait, W. B., Ahmad, N., Alzaidi, M. S. A., Alqahtani, M., Alhashmi, A. A., Aljehane, N. O., Alshahrani, M., & Alotaibi, S. D. (2025). Unveiling Arabic named entity recognition using natural language processing with artificial intelligence approach on Moroccan dialect. *Scientific Reports*, 15, Article 30882. <https://doi.org/10.1038/s41598-025-05940-y>
- Ullah, F., Faizullah, S., Khan, I. U., Alghamdi, T., Syed, T. A., Alkhodre, A. B., Ayub, M. S., & Karim, A. (2025). Prompt-based fine-tuning with multilingual transformers for language-independent sentiment analysis. *Scientific Reports*, 15, Article 20834. <https://doi.org/10.1038/s41598-025-03559-7>
- 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, Article 143. <https://doi.org/10.1186/s12874-018-0611-x>
- 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>