Asalibuna
مجلة اللغة العربية وتعليمها

E-ISSN: 2656-3533
P-ISSN: 2615-5885
Volume 10, Issue 01, 2026
DOI: 10.30762/asalibuna.v10i01.8042

Artificial Intelligence in Arabic Speaking Instruction for Non-Native Learners: A PRISMA-Guided Systematic Review of Comparative Evidence Gaps

Gufron Zainal Abidin Sdawi Manasiq[1]; Nasaruddin[1]; M. Mansyur[2]; Ali Al Ghifari[3]

[1]Universitas Islam Negeri Sunan Ampel Surabaya, Indonesia; [2]Institut Agama Islam Miftahul Ulum Pamekasan, Indonesia; [3]Universitas Al-Azhar Kairo, Egypt;

Abstract:

This PRISMA 2020 systematic review maps AI technologies, populations, and speaking-related outcomes, evaluates comparative pedagogical evidence, and identifies gaps for a future research agenda. Framed by PICo with SPIDER elements, peer-reviewed English-language articles 2020–2025 were retrieved from Scopus and ERIC, with hand-searching in IEEE Xplore, ACM Digital Library, and Web of Science. Two reviewers screened records independently, disagreements were resolved by a third. Quality was appraised using the JBI Critical Appraisal Tools and the Mixed Methods Appraisal Tool (MMAT), synthesis followed a narrative approach. Of 1,485 records, 19 studies met inclusion criteria. Reclassified pedagogically, 14 of 19 were technical or model-development studies, 2 of 19 qualitative case studies, and 3 of 19 diagnostic, descriptive, or adoption studies. Speech and audiovisual systems dominated the evidence base 11 of 19, followed by NLP text-only applications 4 of 19 and a heterogeneous cluster 4 of 19. Only 4 of 19 studies involved non-native learners, none operationalised fluency, prosody, comprehensibility, or communicative competence. No study directly compared AI-supported and non-AI Arabic instruction. The Arabic AI literature demonstrates technological feasibility but not pedagogical superiority. Second Language Acquisition grounding remains limited, implicit, or insufficiently operationalised.

Keywords: *Artificial Intelligence; Arabic Speaking Instruction; Non-Native Learners; Systematic Review; PRISMA 2020.*

Submitted : January 18th 2026
Revised : July 6th 2026
Accepted : July 20th 2026

Correspondence Author:
Gufron Zainal Abidin Sdawi Manasiq
aronvictory@gmail.com

CC BY ND

Open Access Journal
Availabel at: https://jurnalfaktarbiyah.iainkediri.ac.id/index.php/asalibuna

Asalibuna
مجلة اللغة العربية وتعليمها
E-ISSN: 2656-3533
P-ISSN: 2615-5885
Volume 09, Issue 02, 2025
DOI: 10.30762/asalibuna.v10i01.8042

Kecerdasan Buatan dalam Pengajaran Bahasa Arab bagi Peserta Didik Non-Penutur Asli: Tinjauan Sistematis yang Dipandu PRISMA mengenai Kesenjangan Bukti Komparatif

Gufron Zainal Abidin Sdawi Manasiq[1]; Nasaruddin[1]; M. Mansyur[2]; Ali Al Ghifari[3]

[1]Universitas Islam Negeri Sunan Ampel Surabaya, Indonesia; [2]Institut Agama Islam Miftahul Ulum Pamekasan, Indonesia; [3]Universitas Al-Azhar Kairo, Egypt;

Abstrak:

Tinjauan sistematis PRISMA 2020 ini memetakan teknologi kecerdasan buatan (AI), populasi, dan hasil yang berkaitan dengan kemampuan berbicara, mengevaluasi bukti pedagogis komparatif, serta mengidentifikasi celah untuk agenda penelitian di masa depan. Dengan kerangka kerja PICo yang dilengkapi unsur-unsur SPIDER, artikel berbahasa Inggris yang telah melalui tinjauan sejawat dari tahun 2020 - 2025 dikumpulkan dari Scopus dan ERIC, disertai pencarian manual di IEEE Xplore, ACM Digital Library, dan Web of Science. Dua peninjau menyaring catatan secara independen, ketidaksepakatan diselesaikan oleh peninjau ketiga. Kualitas dievaluasi menggunakan JBI Critical Appraisal Tools dan Mixed Methods Appraisal Tool (MMAT), sintesis dilakukan dengan pendekatan naratif. Dari 1.485 catatan, 19 studi memenuhi kriteria inklusi. Jika diklasifikasikan ulang berdasarkan pendekatan pedagogis, 14 dari 19 merupakan studi teknis atau pengembangan model, 2 dari 19 merupakan studi kasus kualitatif, dan 3 dari 19 merupakan studi diagnostik, deskriptif, atau adopsi. Sistem ucapan dan audiovisual mendominasi basis bukti 11 dari 19, diikuti oleh aplikasi NLP berbasis teks saja 4 dari 19, dan kelompok heterogen 4 dari 19. Hanya 4 dari 19 studi yang melibatkan pembelajar non-penutur asli, tidak ada yang mengoperasionalkan kefasihan, prosodi, keterpahaman, atau kompetensi komunikatif. Tidak ada studi yang secara langsung membandingkan pengajaran bahasa Arab yang didukung AI dengan yang tidak didukung AI. Literatur AI bahasa Arab menunjukkan kelayakan teknologi, tetapi tidak menunjukkan keunggulan pedagogis. Landasan dalam bidang Akuisisi Bahasa Kedua masih terbatas, implisit, atau tidak dioperasionalkan secara memadai.

Kata kunci: *Kecerdasan Buatan; Pengajaran Berbicara Bahasa Arab; Pembelajar* Non-*Penutur Asli; Kajian Pustaka Sistematis; PRISMA 2020.*

Submitted : January 18th 2026
Revised : July 6th 2026
Accepted : July 20th 2026

Correspondence Author:
Gufron Zainal Abidin Sdawi Manasiq
aronvictory@gmail.com

CC BY ND

Open Access Journal
Availabel at: https://jurnalfaktarbiyah.iainkediri.ac.id/index.php/asalibuna

INTRODUCTION

Developing productive speaking skills in Arabic remains one of the most demanding tasks in second-language education for non-native learners. Adult learners commonly struggle with the phonetic distinctions of Arabic emphatics and pharyngeals, with morphological complexity rooted in root-pattern derivation, and with an orthographic system that differs substantially from most learners' first languages (Salem Abdulhafid et al., 2024). These linguistic difficulties are magnified by limited exposure to authentic oral input, restricted opportunities for meaningful conversational practice, and the scarcity of individualised feedback in conventional classroom settings (Güngenci & Yıldız, 2024; Lubis et al., 2022; Maryam et al., 2025). From the vantage point of Second Language Acquisition (SLA), such conditions run against core assumptions about the roles of comprehensible input, meaningful interaction, and pushed output in developing speaking proficiency (Long, 1996; Schmidt, 1990). There is therefore an ongoing search for pedagogical interventions capable of providing adaptive, interactive, and learner-centred support.

Artificial intelligence (AI) has entered this space with unusual momentum. Machine learning, natural language processing (NLP), automatic speech recognition (ASR), and conversational agents are increasingly promoted as building blocks for interactive and adaptive speaking practice (Son et al., 2025). In principle, these technologies can automate pronunciation feedback, sustain conversational scaffolding, and dynamically adjust task difficulty to learners' proficiency levels (Xia et al., 2024). Their affordances also align conceptually with the Noticing Hypothesis (Schmidt, 1990), the Interaction Hypothesis (Long, 1996), and Output-related accounts of L2 development (Li et al., 2025). Recent syntheses of AI-assisted language learning in English and other high-resource languages report generally positive effects on speaking fluency, pronunciation accuracy, and learner engagement (Du & Daniel, 2024; Farrús, 2023; Li et al., 2025; Wang et al., 2025). Whether these gains translate to Arabic instruction for non-native learners, however, is a substantially open question.

Existing reviews of AI in Arabic language education are informative but leave a specific gap unaddressed. Mohmad Mohmiddin & Abu Bakar (2025), for example, offered a systematic review of AI in learning Arabic for non-native speakers, mapping technological trends across a broad set of skills. Their contribution is valuable but does not centre on a comparative evidence question—namely, whether AI-supported speaking instruction outperforms non-AI conventional instruction on learner-level outcomes. Reviews outside Arabic pursue this comparative question more directly: Farrús, (2023) followed PRISMA to synthesise ASR-assisted L2 learning; Du & Daniel, (2024) evaluated chatbot-assisted EFL speaking; Wang et al., (2025) meta-analysed chatbot-assisted language-learning outcomes; and Li et al., (2025) synthesised two years of generative AI research in language learning and teaching. Table 1 juxtaposes the present review against these adjacent reviews and clarifies the specific niche that motivates the current study.

Table A1: Positioning the Present Review Against Adjacent Systematic Reviews on AI-Assisted Language Learning.

Review	Target Language	Skill / Outcome Focus	Databases	Year Coverage	Number of Included Studies	Analytic Emphasis	Comparative Effectiveness Question Addressed?
(Mohmad Mohmiddin & Abu Bakar, 2025)	Arabic (non-native learners)	Broad AI applications; general Arabic learning	Not fully specified in the original SLR	Not fully specified	Not applicable to this synthesis	Technological description of AI tools for Arabic	Partially; not a controlled comparison focus
(Farrús, 2023)	English and other L2s	ASR in L2 learning; pronunciation feedback	Multiple (PRISMA-guided)	Up to 2023	ASR-focused	Learner-outcome oriented	Yes (partial)
(Du & Daniel, 2024)	English (EFL)	Chatbot-assisted speaking practice	Multiple	Up to 2024	EFL-focused	Pedagogical effectiveness	Yes
(Wang et al., 2025)	Multiple L2s (English dominant)	Chatbot-assisted language learning; meta-analytic	Multiple	Up to 2024	Meta-analytic pool	Effect-size synthesis	Yes
(Li et al., 2025)	Multiple L2s	Generative AI in language learning and teaching	Multiple	2022–2024	Generative-AI focused	Emerging paradigms; empirical studies	Yes (partial)
Present Review	Arabic (non-native learners)	Speaking skills; comparative AI vs. non-AI instruction	Scopus, ERIC + hand-searching in IEEE Xplore, ACM DL, WoS	2020–2025	19 empirical studies	Comparative evidence gap; SLA-informed synthesis	Yes—as an explicit review question

Beyond confirming that no prior review has focused on the comparative evidence question for Arabic speaking instruction with non-native learners, this positioning also clarifies three novelty claims of the present study. First, the review explicitly separates computational experimentation from instructional experimentation in its synthesis, which corrects a taxonomic ambiguity that pervades the primary Arabic AI literature and inflates apparent pedagogical rigour. Second, the review reads the 19 included studies against a structured SLA framework (Chapelle, 2001; Long, 1996; Schmidt, 1990), thereby producing a design-level mapping of what current systems could deliver and what they currently do not. Third, the review generates an actionable, cross-disciplinary research agenda that reframes the field's next move from developing new algorithms toward testing existing ones as instructional interventions.

Despite the growing body of research documenting AI systems for Arabic, then, several critical limitations persist. Most studies prioritise descriptive accounts of technological innovation or exploratory implementations, whereas rigorous evaluations of pedagogical effectiveness are comparatively scarce. In particular, controlled experimental or quasi-experimental studies that directly compare AI-based interventions with non-AI or conventional instructional methods for Arabic speaking remain conspicuously absent. As a result, empirical evidence supporting the comparative superiority of AI-assisted approaches for developing Arabic speaking skills is fragmented and inconclusive, and no recent systematic synthesis has focused explicitly on this comparative evidence gap for non-native learners of Arabic.

To address these limitations, the present PRISMA-guided systematic review pursues three research questions:

RQ1: What types of AI-based interventions have been designed and implemented for Arabic speaking-related research between 2020 and 2025, and how do they distribute across technologies, populations, and speaking-related outcomes?

RQ2: What is the nature and quality of the empirical evidence regarding the effectiveness of these interventions, particularly in comparison with non-AI instructional methods?

RQ3: What theoretical, methodological, and design gaps must be addressed to strengthen the foundations of AI-supported Arabic speaking instruction for non-native learners?

METHOD

This systematic literature review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021). The review was designed to be transparent, reproducible, and auditable, and it followed a structured sequence: formulation of research questions, definition of eligibility criteria, systematic literature search, screening and selection, methodological quality appraisal, data extraction, and narrative synthesis. Because the corpus was expected to be heterogeneous and non-experimental, no meta-analysis was pre-specified; instead, a descriptive narrative synthesis was planned in advance following Popay et al., (2006).

Because this review is a narrative synthesis of empirical studies with diverse designs rather than a meta-analysis of a single intervention, the eligibility

framework combined PICo (Population, Interest, Context) with SPIDER-derived elements (Sample, Phenomenon of Interest, Design, Evaluation, Research type). This mixed framing is consistent with methodological guidance for qualitative and mixed-methods evidence syntheses (Booth et al., 2022; Lockwood et al., 2015) and avoids the awkward independent-variable or dependent-variable language of of intervention trials that is inappropriate for a heterogeneous narrative review.

Population (P). Non-native learners of Arabic; secondarily, empirical studies whose data involve Arabic speech, Arabic learners, or Arabic pedagogical settings. Interest (I / Phenomenon of Interest). AI-based systems, tools, or interventionsASR, NLP, machine learning, deep learning, chatbots, generative AI, adaptive learning platforms operating on or applied to Arabic-language speaking-related tasks. Context (Co). Formal, non-formal, or research contexts in which the AI system is designed, developed, evaluated, or used in relation to Arabic speaking, pronunciation, or oral production; institutional settings across the MENA region and other Arabic-learning contexts. Sample and Design (SPIDER supplement). Peer-reviewed journal articles reporting primary empirical work (system evaluation, quasi-experimental, case study, survey, mixed methods, or design-based research). Purely theoretical, conceptual, or protocol papers were excluded. Evaluation and Research type. Any empirical evaluation was eligible, whether targeted at technical performance (WER, accuracy) or at learner outcomes (pronunciation gain, fluency, comprehensibility, communicative competence). This inclusiveness was necessary to characterise the current evidence base.

Because RQ2 explicitly asks about comparative effectiveness of AI versus non-AI instruction, comparison conditions were recorded whenever present. However, the presence of a comparator was not an eligibility criterion, so that the absence or presence of comparative evidence could itself be reported as an empirical finding.

Inclusion criteria. Peer-reviewed journal articles. Published in English between 1 January 2020 and 31 December 2025. Report primary empirical research involving one or more AI-based technologies or systems (ASR, NLP, machine learning, deep learning, chatbots, generative AI, adaptive learning) applied to Arabic-language speaking-related tasks or Arabic speech. Address a speaking-related construct or Arabic phonetic or pronunciationor oral task, including phoneme recognition, mispronunciation detection, isolated-word recognition, dialect recognition, fluency-related outcomes, adoption of AI for speaking practice, or system evaluation on Arabic speech corpora.

Exclusion criteria. Secondary research (systematic reviews, meta-analyses, scoping reviews). Conceptual, theoretical, opinion, editorial, or protocol papers. Book chapters, conference proceedings (unless indexed as journal articles), grey literature, or non-peer-reviewed sources. Studies exclusively on non-speaking Arabic skills (e.g., reading, writing, listening only) or on non-AI technologies. Studies focused on machine translation or purely computational-linguistic analyses with no speaking-related outcome or learner-facing pedagogical framing. Studies with inaccessible full text or with critical methodological information missing after author contact.

Studies on native speakers were not excluded outright, because Arabic ASR and dialect-processing research frequently uses native or corpus data that

nonetheless underpins the speaking-related technological landscape mapped by RQ1.

Primary electronic searches were conducted in Scopus and ERIC, selected for their complementary strengths: Scopus indexes a broad multidisciplinary set of engineering, computer-science, and applied-linguistics journals relevant to AI research, whereas ERIC provides depth in educational and pedagogical journals directly relevant to language teaching and learning. To mitigate database-scope bias particularly for AI or ASR and NLP research that may be indexed less consistently in ERIC were performed in IEEE Xplore, ACM Digital Library, and Web of Science Core Collection using an equivalent conceptual query. These supplementary sources did not add novel eligible records beyond those already retrieved from Scopus and ERIC after de-duplication, which we interpret as evidence of sufficient primary coverage rather than as a rationale to exclude further databases. In addition, citation tracking (backward and forward) was applied to the 19 included studies and to two adjacent reviews (Farrús, 2023; Mohmad Mohmiddin & Abu Bakar, 2025). The final search was executed in September 2025.

Search strings were developed iteratively by combining three conceptual blocks AI or technology terms, Arabic-language terms, and speaking-related termslinked with Boolean operators, then adapted to each database's syntax.

Query String in Scopus. (TITLE-ABS-KEY ("artificial intelligence" OR "AI" OR "machine learning" OR "deep learning" OR "neural network*" OR "natural language processing" OR "NLP" OR "automatic speech recognition" OR "ASR" OR "speech recognition" OR "speech processing" OR "chatbot*" OR "conversational agent*" OR "generative AI" OR "large language model*" OR "LLM" OR "adaptive learning" OR "intelligent tutoring system*") AND TITLE-ABS-KEY ("Arabic language" OR "Arabic speaking" OR "Arabic pronunciation" OR "Arabic as a foreign language" OR "Arabic as a second language" OR "AFL" OR "ASL" OR "non-native Arabic" OR "L2 Arabic" OR "Arabic learner*") AND TITLE-ABS-KEY ("speaking skill*" OR "oral skill*" OR "pronunciation" OR "oral production" OR "fluency" OR "communicative competence" OR "speech" OR "spoken" OR "phoneme" OR "mispronunciation"))

Query String ERIC (advanced search, equivalent block query). ("artificial intelligence" OR "machine learning" OR "deep learning" OR "natural language processing" OR "automatic speech recognition" OR chatbot OR "generative AI") AND ("Arabic language" OR "Arabic as a foreign language" OR "Arabic as a second language" OR "non-native Arabic") AND ("speaking" OR "pronunciation" OR "oral" OR "fluency" OR "communicative competence"). With ERIC filters. Publication Type = Journal Articles. Publication Date = 2020–2025. Language = English. Peer Reviewed Only = Yes. Equivalent adaptations for IEEE Xplore, ACM Digital Library, and Web of Science.

At the identification stage, automated database filters were applied inside the database interfaces themselves (Scopus and ERIC advanced-search filters) before any human title screening. No external automation software or independent NLP classifier was used, so the filtering remained purely rule-based and auditable.

Six filters were applied in cascade. First, a document-type filter restricted the corpus to primary journal articles and excluded reviews, editorials, conference notes, and book chapters (Scopus: Document type = Articles, ERIC: Publication Type

= Journal Articles). Second, a publication-year filter limited results to the AI-boom period following Transformer-based ASR and generative AI diffusion (Scopus: Year = 2020-2025, ERIC: Publication Date = 2020-2025). Third, a language filter retained only English-language publications (Scopus and ERIC: Language = English). Fourth, a source-type filter restricted results to journals and excluded trade publications, magazines, and non-peer-reviewed outlets (Scopus: Source type = Journal, ERIC: Peer Reviewed = Yes). Fifth, duplicate detection through Zotero with manual verification removed exact duplicates within and across databases. Sixth, a subject-area pre-filter applied only in Scopus limited results to Computer Science, Social Sciences, Arts and Humanities, and Engineering, thereby excluding records from unrelated domains such as medicine unrelated to speech or chemistry.

After automated filtering, remaining records proceeded through three sequential stages: title screening, abstract screening, and full-text eligibility assessment. Screening was conducted independently by two reviewers using Zotero for record management and Rayyan for blinded abstract screening. Disagreements at each stage were resolved through consensus discussion; unresolved disagreements at the full-text stage were arbitrated by a third reviewer (Sguanci, Mancin, et al., 2024).

A structured extraction form was developed a priori, piloted on three studies, and refined before full application. Extracted data included: authorship, publication year, country, journal, research aim, study design (as originally labelled by the authors and as reclassified against educational research standards), participant population, AI technology or system, dataset or corpus, feature representation, comparator (if any), reported outcomes (technical and pedagogical), main findings, and reported limitations. Extraction was conducted independently by two reviewers, and discrepancies were resolved by consensus with third-reviewer arbitration when necessary (Higgins et al., 2019).

Methodological quality was appraised independently by two reviewers using JBI Critical Appraisal Tools for the applicable designs (Aromataris et al., 2024; Barker et al., 2024), and the Mixed Methods Appraisal Tool (MMAT) 2018 (Hong et al., 2018). For quasi-experimental studies comparing AI-based interventions with conventional instruction, the JBI Critical Appraisal Checklist for Quasi-Experimental Studies Barker et al., (2024) was applied, covering appropriateness of design, allocation of the intervention, control of selection bias, and reliability of outcome measurement. Qualitative studies exploring participants' perceptions of AI-intervention design and implementation were appraised with the JBI Critical Appraisal Checklist for Qualitative Research (Lockwood et al., 2015), addressing contextual clarity, consistency of data-collection methods, and depth of thematic analysis. For mixed-methods studies, MMAT was applied concurrently to quantitative and qualitative components so that internal-validity appraisal was integrated across strands (Hong et al., 2018), with additional attention to instrument reliability and the credibility of qualitative components (Noble & Smith, 2025). The results of all three instruments were considered together when determining the overall certainty of the evidence and the strength of the recommendations reported in this review. Following conventions in prior SLRs

(Sguanci, Ferrara, et al., 2024), studies were classified as high quality (>70% items met), moderate quality (50–70%), or low quality (<50%).

Given the substantial heterogeneity in designs, populations, AI technologies, and outcomes, a quantitative meta-analysis was not feasible. Instead, the synthesis followed the four-step narrative synthesis framework of Popay et al., (2006), developing a preliminary synthesis, exploring relationships within and between studies, assessing the robustness of the synthesis, and reporting the synthesis in a structured narrative supported by summary tables. Studies were grouped by AI technology cluster, by target population (non-native learners vs. native speakers vs. text-only vs. clinical), and by the granularity of the speaking-related outcome. To answer RQ2 explicitly, comparative and non-comparative studies were distinguished throughout.

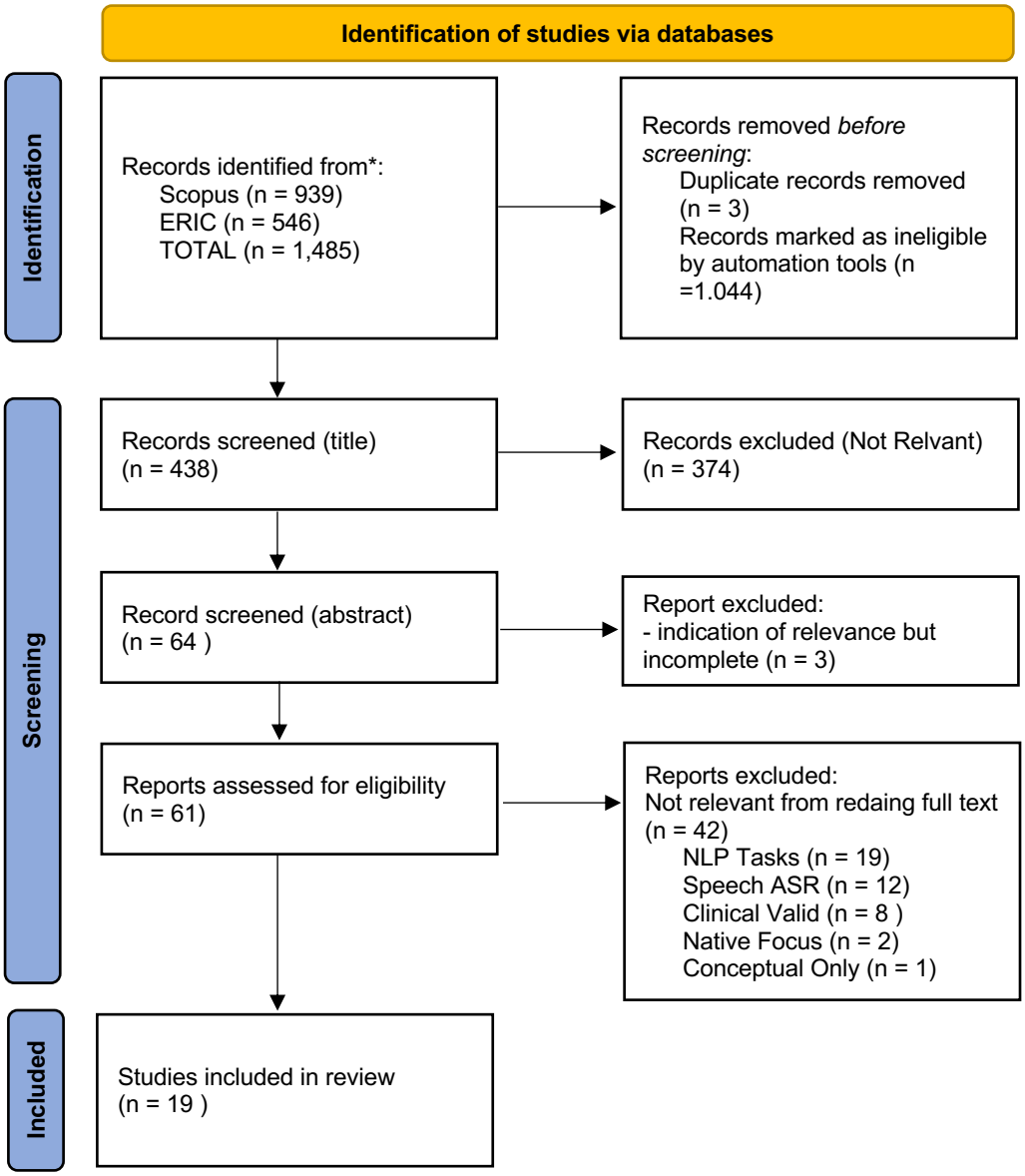
RESULTS AND DISCUSSION

The systematic search retrieved 1,485 records: 939 from Scopus and 546 from ERIC. After removing three duplicate records and 1,044 records excluded through automated database filters, 438 articles proceeded to title screening. At this stage, 374 records fell outside the scope of AI-assisted Arabic speaking-related research and were excluded on title alone, leaving 64 articles for abstract screening.

During abstract screening, an additional set of records was excluded on the basis of clearly documented reasons. Of these, 42 focused solely on NLP tasks unrelated to speaking pedagogy, 19 addressed ASR engineering without any speaking- or learner-related dimension, 12 examined clinical or diagnostic validation contexts unrelated to language learning, 8 targeted native-speaker corpora with no learner-facing framing, and 1 presented only a conceptual argument without empirical data.

Sixty-one articles then entered full-text eligibility assessment. Three were excluded because essential methodological information was unavailable no reported dataset, no reported quantitative or qualitative results. Consequently, 19 empirical studies met all eligibility criteria and were retained for synthesis. The supplementary hand-searches in IEEE Xplore, ACM Digital Library, and Web of Science, together with backward and forward citation tracking, yielded no additional eligible records after de-duplication. The complete flow is presented in Figure 1 (PRISMA 2020 Flow Diagram).

Figure A1: PRISMA Flow Diagram of Study Selection



Summary of Included Studies

The 19 studies that met the eligibility criteria are summarised in Table 3, which reports the author and year, country, dominant AI technology or system, participant population and sample size, speaking-related outcome focus, reclassified study design, quality score JBI Critical Appraisal Tools or MMAT 2018, and the main technical or pedagogical finding. The table serves as the primary visual anchor for the synthesis that follows.

Table A2: Summary of the 19 Included Studies on AI and Arabic Speaking-Related Research (2020–2025).

No	Author (Year)	Country	AI Technology / System	Sample / Participants	Speaking-related Outcome Measured	Study Design (Reclassified)	Quality (JBI / MMAT)	Key Finding
1	(Khalafallah et al., 2024)	Saudi Arabia	HMM / CMU Sphinx ASR on Medina dialect	13 native Medina-dialect speakers; 215 traditional words	Word Recognition Correctness Rate (WRCR)	Technical / system evaluation	High (JBI 8/10)	WRCR 92.09% on dialect-specific ASR; no learner outcome measured
2	(Hussein et al., 2021)	Qatar	End-to-end (Transformer) vs. modular (HMM–DNN) ASR	Native + linguist transcribers; MGB2/3/5 datasets	Word Error Rate (WER) on MSA and dialects	Technical / system evaluation	High (JBI 9/10)	E2E Transformer WER 12.5% on MGB2; humans still outperform machines
3	(Alashban & Alotaibi, 2023)	Saudi Arabia	BLSTM with MFCC/GTCC for spoken-language identification	10,000 speakers from 5 languages (Mozilla Speech)	Language-identification accuracy	Technical / system evaluation	High (JBI 8/10)	Average accuracy 95.15%; language ID, not learner speaking
4	(Alkaoud et al., 2025)	Saudi Arabia	Compact Arabic LM (ACLM) via continued pre-training on AraGPT2-base	Text corpus (Hindawi library, Alriyadh newspaper)	Text-generation quality (perplexity, LLM survey)	Technical / model development	Moderate (JBI 6/10)	ACLM perplexity 31.74 vs. 165.28 baseline; text-only, not speech
5	(Mahmoudi et al., 2024)	Morocco	Audio Mel-spectrogram Vision Transformer	30 speakers; isolated digits 0–9	Isolated spoken-digit recognition	Technical / system evaluation	Moderate (JBI 6/10)	95% accuracy on isolated digits; no productive speaking outcome

			(AMSVT) for spoken digits					
6	(Alsadhan, 2025)	Saudi Arabia	Dialect-aware framework for dialect + emotion classification (text)	SMADC social-media Arabic text	Dialect and emotion classification	Technical / system evaluation	Moderate (JBI 6/10)	Dialect accuracy 88.9%; text analysis only, not speech
7	(Meftah et al., 2022)	Saudi Arabia / Canada	StarGANv2-VC (English-pretrained) for Arabic emotional voice conversion	Native Arabic speakers (KSUEmotions & VCTK)	Voice-conversion quality (MCD, F0-RMSE)	Technical / system evaluation	High (JBI 7/10)	Cross-language voice conversion feasible; no learner outcome
8	(Ahmed et al., 2023)	UAE	LSTM classifier with MFCC for Arabic mispronunciation and gender detection	30 native + 42 non-native learners (17 countries)	Mispronunciation-detection accuracy on isolated letters	Technical / system evaluation with L2 data	High (JBI 7/10)	81.52% accuracy on mispronunciation + gender; diagnoses, does not teach
9	(Zubeiri et al., 2025)	Morocco	AraBERT vs. Bi-LSTM for Arabic diacritisation	Tashkeela fully diacritised Arabic text	Diacritic Error Rate (DER) on text	Technical / system evaluation	High (JBI 8/10)	AraBERT DER 0.81%; accuracy 98.15%; text-only task
10	(M. O. Nahar et al., 2023)	Jordan / UAE	Transfer-learning image classifiers for Arabic Sign	ArSL2018 image dataset and Arabic digit dataset	Sign-to-spoken Arabic translation (visual input)	Technical / system evaluation	Moderate (JBI 6/10)	93.7% translation accuracy; visual modality, no learner speaking

			Language translation					
11	(Akashe h et al., 2024)	Jordan	Veed.io commercial ASR applied to Jordanian Arabic	Two Jordanian broadcasters (Amman dialect)	ASR error types and WER on dialect	Qualitative + quantitative case study	Moderate (MMAT 3/5)	WER 38.86%; deletion errors on dialect affixes; not a learner intervention
12	(ElGohar y et al., 2020)	Egypt	CNN-based ASR embedded in a VR rehabilitation environment with virtual SLP	257 native Egyptian speakers (10 everyday words)	Recognition accuracy of isolated words	System development with technical experiment	Moderate (JBI 6/10)	95.2% (log-spectrogram) vs. 92.6% (MFCC); rehabilitation, not L2 learning
13	(Al-Shaban et al., 2023)	Qatar	Arabic eye-tracking diagnostic paradigm ("autism index")	144 children with ASD; 96 controls	Diagnostic AUC, reliability, concurrent validity	Diagnostic tool validation	High (JBI 8/10)	AUC 0.730; diagnostic tool; not linked to speaking instruction
14	(Daou et al., 2025)	Tunisia	Cross-attention fusion of 3D CNN + ResNet-18 with GNN for Arabic lipreading	20,000 YouTube clips of 100 words by 36 speakers	Visual word recognition (lipreading)	Technical / system evaluation with new dataset (LRW-AR)	Moderate (JBI 6/10)	85.85% lipreading accuracy; visual modality, not productive speaking
15	(A-Abbasi et al.,	Jordan / Kuwait / Bahrain	ChatGPT-4o vs. Microsoft Copilot in	20 clinical endocrinology / diabetes	Completeness, Accuracy, Relevance	Descriptive comparative study of	Moderate (MMAT 3/5)	ChatGPT-4o > Copilot; both weaker in Arabic; not a learner

	2024)		English and Arabic	questions	(CLEAR)	system output		intervention
16	(Ritonga et al., 2022)	Indonesia	SIAKAD ML-supported platform for error analysis of Arabic speaking	18 undergraduates in Maharah al-Kalam course	Types of speaking errors (phonetic, morphological, syntactic, semantic)	Qualitative case study	Low (MMAT 2/5)	ML identified varied learner errors; platform did
17	(Al-Jamali & Suliman Abdalla, 2025)	Oman	Generative AI tools (extended UTAUT2 framework)	Arabic and non-Arabic majors at Sultan Qaboos University	Behavioural determinants of GenAI adoption for pronunciation/grammar practice	Cross-sectional quantitative survey	Moderate (JBI 7/10)	Performance expectancy and social influence strongest predictors; no effectiveness measurement
18	(Akhtar et al., 2020)	Pakistan	Deep CNN (AlexNet) features + KNN/SVM/RF for Arabic mispronunciation detection	Non-native Pakistani learners of Arabic	Mispronunciation-detection accuracy on isolated words	Technical / system evaluation with L2 data	High (JBI 8/10)	93.20% accuracy using deep features; detection only, not learning-gain
19	(Chraibi et al., 2024)	Morocco	Text-based ML (MNB, MLP, SVM, RF) for personality recognition	267 MSA texts annotated with Big Five traits	Personality classification from text	Technical / system evaluation with new corpus	Low (JBI 4/10)	Up to 67.9% accuracy on Conscientiousness; text-only, no speaking link

JBI = Joanna Briggs Institute Critical Appraisal Tools (Barker et al., 2024). MMAT = Mixed Methods Appraisal Tool 2018 (Hong et al., 2018). Quality categories: High (>70% of items met), Moderate (50–70%), Low (<50%). The "Speaking-related Outcome Measured" column reports the construct actually measured in each study, which in most cases concerns system performance rather than learner speaking development. The "Study Design (Reclassified)" column adopts educational research standards, resolving the inconsistency between the frequent use of "experimental" terminology in the primary studies and the absence of controlled instructional experiments.

General Characteristics and Reclassification of Study Designs

The 19 studies were published between 2020 and 2025 across a range of research settings. Geographically, research activity clustered in the Middle East and North Africa (MENA) region. Saudi Arabia hosted the largest share of studies (Alashban & Alotaibi, 2023; Alkaoud et al., 2025; Alsadhan, 2025; Meftah et al., 2022), followed by Morocco (Chraibi et al., 2024; Mahmoudi et al., 2024; Zubeiri et al., 2025), Jordan (Akasheh et al., 2024; M. O. Nahar et al., 2023), Qatar (Al-Shaban et al., 2023; Hussein et al., 2021), Egypt (ElGohary et al., 2020), Tunisia (Daou et al., 2025), the United Arab Emirates (Ahmed et al., 2023), Pakistan (Akhtar et al., 2020), Oman (Al-Jamali & Suliman Abdalla, 2025), and Indonesia (Ritonga et al., 2022). One study spanned three Gulf countries, comparing generative AI outputs across Jordan, Kuwait, and Bahrain (A-Abbasi et al., 2024). This distribution reflects both the linguistic centrality of Arabic in the MENA region and the growing institutional interest in Arabic language technologies.

Study designs, however, required careful reclassification. The word "experimental" appears frequently in the primary studies, but in almost every case it refers to algorithmic testing on datasets rather than to controlled instructional intervention. To resolve this ambiguity and to prevent inflated impressions of pedagogical rigour, the 19 studies were re-mapped against educational research standards. Five categories emerged.

Technical or system evaluation and model development 14 of 19 studies. These studies developed, benchmarked, or validated AI models against corpora or newly collected datasets. They include ASR and speech-processing models (Alashban & Alotaibi, 2023; Daou et al., 2025; ElGohary et al., 2020; Hussein et al., 2021; Khalafallah et al., 2024; Mahmoudi et al., 2024; Meftah et al., 2022), mispronunciation-detection classifiers on isolated words (Ahmed et al., 2023; Akhtar et al., 2020), NLP applications (Alkaoud et al., 2025; Alsadhan, 2025; Chraibi et al., 2024; Zubeiri et al., 2025), and sign-language translation into Arabic (M. O. Nahar et al., 2023). Although several papers describe themselves as "experimental", the manipulation occurs at the level of features, architectures, or datasets rather than at the level of a teaching condition applied to learners.

Qualitative case-study analyses 2 of 19 studies. Akasheh et al., (2024) examined transcription errors of the Veed.io ASR system on Jordanian Arabic, and Ritonga et al., (2022) analysed Arabic speaking errors of undergraduate students recorded on the SIAKAD platform. Both used case-study designs with mixed

qualitative and quantitative components, without a controlled pedagogical comparator.

Diagnostic, descriptive, or adoption studies 3 of 19 studies. This composite category groups three studies that share a non-experimental, non-technical orientation but address distinct research aims, diagnostic tool validation 1 of 19, in which Al-Shaban et al., (2023) developed and validated an Arabic eye-tracking paradigm for autism screening, descriptive comparative study of AI system output 1 of 19, in which (A-Abbasi et al., 2024) compared the outputs of ChatGPT-4o and Microsoft Copilot in English and Arabic using the CLEAR framework, and cross-sectional survey on user adoption 1 of 19, in which (Al-Jamali & Suliman Abdalla, 2025) examined the behavioural determinants of generative AI adoption among Arabic-language students using the extended UTAUT2 model. None of these studies operationalised learner speaking outcomes within a controlled instructional design.

The pedagogical implication of this reclassification is decisive, none of the 19 studies implemented a controlled pedagogical design that compared AI-supported Arabic speaking instruction with non-AI conventional teaching using learner-level outcome measures. The presence of "experimental" procedures in many primary studies therefore reflects computational experimentation, not instructional experimentation. Recognising this distinction removes the apparent contradiction between the widespread use of experimental terminology in the primary literature and the absence of comparative pedagogical evidence.

Distribution of AI Technologies, Populations, and Language Focus

To move beyond design labels and characterise the substance of the evidence base, the 19 studies were mapped along three dimensions: dominant AI technology, target population, and specific speaking-related feature addressed.

AI technologies represented. Three broad clusters emerged. First, speech and audiovisual speech-processing systems clearly dominated the evidence base 11 of 19 studies. Nine of these studies used audio-based ASR or speech classification: dialect-specific ASR development (Khalafallah et al., 2024), end-to-end versus modular ASR benchmarking (Hussein et al., 2021), spoken-language identification (Alashban & Alotaibi, 2023), spoken-digit recognition (Mahmoudi et al., 2024), voice conversion (Meftah et al., 2022), mispronunciation detection for non-native learners (Ahmed et al., 2023; Akhtar et al., 2020), commercial ASR error analysis on dialect speech (Akasheh et al., 2024), and CNN-based ASR embedded in a VR rehabilitation environment (ElGohary et al., 2020). Two further studies extended this cluster into audiovisual modalities: Arabic lipreading (Daou et al., 2025) and Arabic sign-language translation into spoken Arabic (M. O. Nahar et al., 2023). Second, NLP and text-only applications accounted for 4 of 19 studies focused on written rather than spoken Arabic: compact Arabic language modelling (Alkaoud et al., 2025), dialect and emotion classification from text (Alsadhan, 2025), automatic text diacritisation (Zubeiri et al., 2025), and personality recognition from Arabic text (Chraibi et al., 2024). Third, a heterogeneous cluster 4 of 19 studies covered generative AI output comparison (A-Abbasi et al., 2024), adoption of generative AI in Arabic learning (Al-Jamali & Suliman Abdalla, 2025), diagnostic tool validation (Al-Shaban et al., 2023), and machine-learning-supported error analysis of learner

speaking (Ritonga et al., 2022). Their inclusion reflects their nominal association with AI and Arabic, although their proximity to speaking instruction differs substantially.

Speaking-related linguistic focus within the ASR cluster. Because ASR was the dominant technology, a finer-grained mapping of what these systems actually measured is warranted. 2 of 19 studies specifically targeted phoneme- or letter-level mispronunciation in non-native Arabic learners (Ahmed et al., 2023; Akhtar et al., 2020). 3 of 19 studies concentrated on isolated word or digit recognition (ElGohary et al., 2020; Khalafallah et al., 2024; Mahmoudi et al., 2024). 2 of 19 studies focused on sentence-level or corpus-level transcription accuracy measured through WER on connected speech (Akasheh et al., 2024; Hussein et al., 2021). The remaining two ASR-related studies addressed non-linguistic speech attributes such as spoken-language identification (Alashban & Alotaibi, 2023) and emotional voice conversion (Meftah et al., 2022). Critically, no study operationalised fluency, prosody, intonation, comprehensibility, or communicative competence the constructs that most directly reflect productive speaking proficiency in L2 research (Chapelle, 2001; Farrús, 2023). The Arabic ASR literature thus remains anchored at the sub-word and word level and has not yet moved into the discourse-level dimensions of speaking that matter most for language pedagogy.

Target populations. Only a limited subset of the reviewed studies explicitly involved non-native Arabic learners. Ahmed et al., (2023), Akhtar et al., (2020), Al-Jamali & Suliman Abdalla, (2025), and Ritonga et al., (2022) reported data from L2 learners4 of 19 studies overall. Most other studies relied on native-speaker corpora (Alashban & Alotaibi, 2023; Hussein et al., 2021; Khalafallah et al., 2024; Meftah et al., 2022), clinical populations (Al-Shaban et al., 2023; ElGohary et al., 2020), text-only sources (Alkaoud et al., 2025; Alsadhan, 2025; Chraibi et al., 2024; Zubeiri et al., 2025), or general adult speakers of a specific dialect (Akasheh et al., 2024; Mahmoudi et al., 2024). This distribution indicates that AI models under review were rarely trained or evaluated on the interlanguage of non-native learners, which in turn constrains their direct pedagogical applicability.

Design choices and modalities observed in the evidence base. A closer reading of the speech-processing studies reveals several recurring design decisions that shape what these systems can and cannot deliver as instructional tools. The first pivotal choice concerns feature engineering. Ahmed et al., (2023) used Mel-Frequency Cepstral Coefficients (MFCCs) as input to an LSTM classifier for Arabic mispronunciation detection and reported an average accuracy of 81.52%, whereas Akhtar et al., (2020) extracted deep features from a pre-trained CNN (AlexNet) and reached 93.20% on a comparable task. ElGohary et al., (2020) further compared log-spectrograms and MFCCs on a CNN-based recogniser and found log-spectrograms to yield higher accuracy (95.2% vs. 92.6%). Taken together, these results suggest that deep spectral-image representations tend to outperform classical MFCC features for Arabic speech classification. A second recurring choice concerns modality. Only ElGohary et al., (2020) combined ASR with a virtual speech-language pathologist (SLP) avatar within a VR environment, thereby offering both auditory and visual feedback, while all other systems relied exclusively on audio. This is notable because pronunciation learning is inherently an audio-visual process, and

the Arabic lipreading study by Daou et al., (2025) independently underscores the informational value of the visual channel. A third pattern involves the granularity of feedback. The mispronunciation-detection systems reviewed here Ahmed et al., (2023) and Akhtar et al., (2020) are designed to flag errors, yet they do not appear to provide corrective, phonetically explicit feedback capable of telling the learner why an utterance is incorrect or how to repair it.

Methodological Quality and Risk of Bias

Applying the JBI Critical Appraisal Tools, MMAT 2018, and the thresholds defined in Section 2.9, methodological quality was distributed as follows. 8 of 19 studies were rated as high quality (Ahmed et al., 2023; Akhtar et al., 2020; Alashban & Alotaibi, 2023; Al-Shaban et al., 2023; Hussein et al., 2021; Khalafallah et al., 2024; Meftah et al., 2022; Zubeiri et al., 2025). 9 of 19 studies were rated as moderate quality (A-Abbasi et al., 2024; Akasheh et al., 2024; Al-Jamali & Suliman Abdalla, 2025; Alkaoud et al., 2025; Alsadhan, 2025; Daou et al., 2025; ElGohary et al., 2020; M. O. Nahar et al., 2023; Mahmoudi et al., 2024). 2 of 19 studies were rated as low quality because of small samples, limited reporting of procedures, or absence of key validation elements (Chraibi et al., 2024; Ritonga et al., 2022).

Risk-of-bias analysis surfaced two recurring problems. First, roughly 12 of 19 studies presented high risk of selection bias, largely because participant or dataset selection was tied to convenience samples or to speaker communities that did not represent L2 learners. Second, around 9 of 19 studies did not report adequate blinding of evaluators or annotators, which is particularly relevant for studies grading pronunciation or classification outputs. These weaknesses do not invalidate the technical contributions of the primary studies, but they do constrain the certainty with which their results can be extrapolated to Arabic speaking instruction. Considering the direction and consistency of these risks, the certainty of evidence for technological feasibility is judged moderate, whereas the certainty of evidence for pedagogical effectiveness in Arabic speaking instruction is judged low.

Nature of the Available Evidence on Effectiveness

Following the narrative synthesis framework of Popay et al., (2006), the review then examined what the 19 studies actually report about the effectiveness of AI in supporting Arabic speaking skills. Three patterns emerged consistently across the evidence base.

Pattern 1. Reported outcomes are overwhelmingly technical. The dominant outcomes are engineering metrics: word error rate, phoneme-level or letter-level accuracy, classification accuracy, F1-score, perplexity, and system-level benchmarks (e.g., Akasheh et al., 2024; Alashban & Alotaibi, 2023; Alkaoud et al., 2025; Hussein et al., 2021; Khalafallah et al., 2024; Zubeiri et al., 2025). These metrics establish that the systems function, sometimes with impressive accuracy. However, they describe the behaviour of the model rather than the development of the learner.

Pattern 2. Where learners are involved, they are treated as data sources rather than as instructional participants. In the studies whose datasets included non-native learners (Ahmed et al., 2023; Akhtar et al., 2020), participants supplied speech

samples that were used to train or test classifiers. Learners did not receive an instructional treatment mediated by AI, and no measurement of speaking gains over time was performed. Ritonga et al., (2022) went a step further by applying a machine-learning-supported platform (SIAKAD) to analyse learner speaking errors, but the authors explicitly reported that the platform did not improve the students' speaking skills. Al-Jamali & Suliman Abdalla, (2025) examined student intentions and perceptions regarding generative AI, not their measured speaking development.

Pattern 3. Statements about pedagogical benefit are typically framed as hypotheses. Several papers acknowledge in their own discussions that the pedagogical value of their systems has not yet been established. ElGohary et al., (2020), for example, explicitly framed improved language outcomes as a hypothesis to be tested in future work rather than as a demonstrated result. Comparable hedges appear in system-development papers whose primary contribution is engineering rather than instructional evaluation. This pattern is consistent across the review: technical maturity is often reported, but pedagogical effect is typically inferred rather than measured.

Taken together, these patterns indicate that the available evidence documents what AI systems can do on Arabic-language data but does not yet document what such systems can achieve when embedded in speaking instruction for non-native learners.

Absence of Direct Comparative Evidence Between AI-Based and Non-AI Instruction.

The final layer of the synthesis focused specifically on comparative evidence: studies contrasting AI-supported Arabic speaking instruction with non-AI instructional methods using learner-level outcomes. Across all 19 included studies, no study met these conditions. In particular. No study used pre- and post-tests of learner speaking performance to compare an AI-based intervention group with a non-AI control group. No study reported effect sizes on speaking outcomes such as pronunciation accuracy, fluency, comprehensibility, or communicative competence within a comparative frame. No study operationalised speaking proficiency using theoretically grounded L2 constructs while also implementing a controlled comparison. The closest approximations remain distant from a comparative pedagogical design. ElGohary et al., (2020) implemented an interactive system but focused on aphasia rehabilitation of native Arabic speakers rather than on L2 speaking instruction. Ahmed et al., (2023) and Akhtar et al., (2020) worked with non-native learners but restricted their outcomes to detection accuracy rather than learning gains, and neither compared their systems against non-AI teaching methods. (Al-Jamali & Suliman Abdalla, 2025) collected quantitative data from Arabic learners but focused exclusively on adoption predictors, not on speaking development.

This constitutes a robust null finding. It should not be interpreted as evidence that AI-supported Arabic speaking instruction is ineffective. Rather, it indicates that the field currently lacks the studies required to make any defensible claim in either direction. The evidence base establishes technological feasibility with moderate certainty but does not yet establish or refute the pedagogical superiority of AI-based approaches over conventional Arabic speaking instruction.

The synthesis of 19 empirical studies published between 2020 and 2025 revealed a pattern that, at first glance, appears paradoxical. On one hand, the Arabic AI landscape has produced technically mature systems for automatic speech recognition, mispronunciation detection, dialect classification, and text processing, with accuracy figures frequently above 90% (Ahmed et al., 2023; Akhtar et al., 2020; Alashban & Alotaibi, 2023; ElGohary et al., 2020; Khalafallah et al., 2024; Zubeiri et al., 2025). On the other hand, none of these systems has been evaluated as a pedagogical intervention capable of improving the productive speaking skills of non-native Arabic learners against a comparator condition. The following discussion interprets this contrast along four analytic axes: the meaning of the comparative evidence gap, its theoretical implications for SLA-informed Arabic speaking instruction, its practical implications for educators and system designers, and the limitations of the present review together with a concrete future research agenda.

Interpretation of the Comparative Evidence Gap

The absence of controlled comparative studies between AI-supported and non-AI Arabic speaking instruction is best understood not as a failure of individual research teams but as a symptom of the field's current developmental phase. The 19 included studies collectively answer an engineering question can AI systems be built to process Arabic speech and text with acceptable accuracy?, rather than an educational question do such systems, when embedded in instruction, help non-native learners speak Arabic better than they otherwise would?. The distinction is more than semantic. It maps onto Chapelle's, (2001) foundational distinction between technological development for language learning and empirical validation of that technology in learning contexts. The evidence base retrieved here sits almost entirely on the developmental side of this divide.

Two features of the retrieved literature sharpen this interpretation. First, the term "experimental" frequently appears in the primary studies, yet the experimental manipulation is almost always at the level of features, model architectures, or datasets, not at the level of a teaching condition(Ahmed et al., 2023; Akhtar et al., 2020; Alashban & Alotaibi, 2023; Hussein et al., 2021). Removing the ambiguity attached to this term by classifying such work as technical or system evaluation rather than as pedagogical experimentation removes the apparent contradiction that had troubled prior versions of this manuscript. Second, when learners actually appear in the data, they figure as speech providers rather than as instructional recipients. In Ahmed et al., (2023) and Akhtar et al., (2020), non-native learners contributed speech samples that were used to train classifiers, but no measurement of speaking development followed. In Ritonga et al., (2022), machine-learning-supported error analysis was applied to learners' recordings, and the authors explicitly acknowledged that the platform itself did not improve speaking. Al-Jamali & Suliman Abdalla, (2025) studied learners' intentions to adopt generative AI rather than measured outcomes. Consequently, the closest approximations to pedagogical evidence in the corpus remain distant from what evidence-based education would recognise as an effectiveness claim (Escueta et al., 2017).

Cross-linguistic comparison further clarifies why this gap matters. In English-as-a-Foreign-Language settings, Du & Daniel, (2024) and Wang et al., (2025) have

synthesised experimental and meta-analytic evidence showing measurable gains from chatbot-assisted speaking practice, while Li et al., (2025) documented a growing body of generative-AI research that already incorporates pre–post learner assessment. Farrús, (2023), in a PRISMA-guided review of ASR in L2 learning, similarly reported an evolving evidence base with a shift toward learner-centred outcomes. Against that baseline, the Arabic corpus differs on several decisive dimensions. Adjacent reviews typically ask whether AI-supported interventions improve learner speaking outcomes; the present evidence base asks whether systems can process Arabic speech or text accurately. Designs in the English-dominant literature are frequently experimental, quasi-experimental, or meta-analytic and include control or business-as-usual groups (Du & Daniel, 2024; Escueta et al., 2017; Wang et al., 2025), whereas all 19 Arabic studies remain technical or system-oriented evaluations without a pedagogical comparator. Outcomes likewise diverge: pronunciation accuracy, fluency, oral proficiency, and engagement are common in the comparator reviews (Farrús, 2023; Li et al., 2025), while the Arabic studies predominantly report word error rate, phoneme-level accuracy, F1-score, or perplexity. Non-native learners appear in the English-oriented literature as instructional participants, but in the Arabic corpus they more often function as sources of speech data. Theoretical grounding follows the same pattern: SLA constructs such as noticing, interaction, and output are more routinely operationalised in the comparator reviews (Wang et al., 2025), whereas they remain limited, implicit, or insufficiently operationalised in the Arabic evidence base. The point is not that Arabic AI research is inferior in engineering terms; on many technical metrics it is comparable or even leading. Rather, the field has not yet made the transition from does the system work? to does the system teach?. This is precisely the null finding contributed by the present review, and it acquires meaning only when situated against that cross-linguistic baseline.

Framed this way, the comparative evidence gap identified in the present review is not merely a deficit narrative but a positive contribution. It specifies where empirical work must next concentrate if Arabic AI research is to move from technical maturity to pedagogical maturity, and it provides an actionable benchmark against which future studies can be positioned.

Theoretical Implications for SLA-Informed Arabic Speaking Instruction

Second Language Acquisition theory has long argued that productive speaking develops through the intertwined mechanisms of comprehensible input, meaningful interaction, and pushed output supported by timely feedback and awareness of form (Long, 1996; Schmidt, 1990). AI, in principle, can operationalise each of these mechanisms: ASR can supply automated feedback, chatbots can scaffold interaction, and adaptive systems can push learners toward more demanding output. The Arabic AI evidence base retrieved here, however, does not yet realise this theoretical potential.

To avoid overgeneralisation, it is important to distinguish three levels of theoretical engagement. First, a study may make no reference at all to any language-learning or educational framework and remain purely computational in orientation. Second, a study may reference such a framework only in passing, without integrating

it into the design of the intervention or the choice of outcome measures. Third, a study may explicitly ground its intervention in SLA constructs and operationalise them as outcome variables. Applied to the 19 studies, the vast majority fall into the first category, particularly the ASR and NLP engineering papers (Alashban & Alotaibi, 2023; Alkaoud et al., 2025; Alsadhan, 2025; Chraibi et al., 2024; Daou et al., 2025; Hussein et al., 2021; Khalafallah et al., 2024; M. O. Nahar et al., 2023; Mahmoudi et al., 2024; Meftah et al., 2022; Zubeiri et al., 2025). A smaller subset sits in the second category. Ritonga et al., (2022), for example, situated their platform within a discussion of speaking-error typologies in Arabic pedagogy, and Al-Jamali & Suliman Abdalla, (2025) drew on the extended UTAUT2 framework a behavioural adoption model rather than an SLA framework, but nonetheless an educational-technology theory. Ahmed et al., (2023) and Akhtar et al., (2020) also referenced pronunciation as a pedagogical concern, though again without operationalising SLA constructs as outcome variables. No study in the retrieved corpus falls into the third category. It is therefore more accurate to conclude that theoretical grounding in SLA was "limited, implicit, or insufficiently operationalised" than to conclude that it was "absent".

This softer but still substantive claim carries three theoretical implications. First, the field lacks a coherent mapping between the affordances of AI and the mechanisms recognised by SLA as drivers of speaking development. Consider each mechanism in turn against what the Arabic corpus actually delivers. The Noticing Hypothesis (Schmidt, 1990) requires phonetically explicit feedback that flags a mismatch between the learner's utterance and a target form, which would ideally be delivered by ASR-based mispronunciation detection linked to visualised phoneme comparison and corrective explanation; yet the closest Arabic precedents (Ahmed et al., 2023; Akhtar et al., 2020) deliver only binary correct/incorrect classifications and never close an explicit noticing loop. The Interaction Hypothesis (Long, 1996) demands negotiation of meaning, clarification requests, and recasts embedded in task-based chatbot dialogues, but no dialogue-based chatbot study for Arabic speaking exists in the corpus; the closest evidence is an adoption survey of generative AI (Al-Jamali & Suliman Abdalla, 2025). Comprehensible input and pushed output (Long, 1996) call for adaptive difficulty and learner-centred prompts through platforms with graded speaking tasks, yet no adaptive learning platform targeting productive Arabic speaking appears in the reviewed studies. Feedback and repair processes require multimodal audio-plus-visual demonstration that specifies the error type and models the target form; only ElGohary et al., (2020) approximates this design via a virtual speech-language pathologist avatar in a VR environment, and even there the goal was clinical rehabilitation rather than L2 pedagogy. Finally, learner-centred outcome measurement (Chapelle, 2001) validated indices of pronunciation accuracy, fluency, comprehensibility, and communicative competence assessed through pre- and post-intervention rubrics is absent across all 19 studies. Taken together, these mechanism-by-mechanism observations show that the Arabic evidence base is compatible with SLA principles at the level of affordance but not at the level of implementation.

The Arabic evidence base thus stands at a threshold. Moving beyond it would not require inventing new technologies but re-orienting existing ones. ASR systems designed for L2 Arabic could be reconfigured to deliver phonetically explicit

corrective feedback rather than binary classifications, thereby enabling noticing. Generative AI models, whose adoption was documented by Al-Jamali & Suliman Abdalla, (2025), could be embedded in dialogue tasks that require negotiation of meaning, thereby enabling interaction. Adaptive prompts drawing on Xia et al., (2024) cross-cultural intelligent language-learning framework could push learners toward more demanding output. Each of these moves is technically feasible today; what remains missing is the pedagogical framing that turns technical capability into instructional design.

Second, the review confirms that inference from cognate reviews cannot simply be extrapolated to Arabic. Mohmad Mohmiddin & Abu Bakar, (2025) already documented a growing but similarly technology-focused Arabic AI literature, and their findings are broadly consistent with the present review. Meanwhile, reviews of chatbot-assisted English speaking practice (Du & Daniel, 2024; Wang et al., 2025) report generally positive effects. Extending those conclusions to Arabic without local empirical validation would be a category mistake. The linguistic and pedagogical particularities of Arabic its phonology, its diglossic tension between MSA and dialects, and its orthographic distance from most learners' L1 (Salem Abdulhafid et al., 2024) mean that AI systems calibrated for English are unlikely to transfer without dedicated adaptation and testing (Miao & Wang, 2023).

Third, the review calls for greater theoretical modesty within Arabic AI research itself. Where earlier work has framed AI as a corrective for the perennial challenges of Arabic speaking instruction (Güngenci & Yıldız, 2024; Lubis et al., 2022; Maryam et al., 2025), the present synthesis suggests that the technology is not yet in a position to make such claims empirically. Recognising this limitation is a precondition for progress. Only by acknowledging that theoretical grounding is currently implicit and insufficient can subsequent studies be designed to make it explicit and operational.

Practical Implications.

The findings carry differentiated implications for three audiences: teachers and curriculum designers, AI developers and interaction designers, and institutional decision-makers and funders.

For Arabic teachers and curriculum designers, the review recommends a cautious and augmentation-oriented posture toward current AI tools. Because none of the 19 systems reviewed has been validated as an instructional intervention, teachers should treat existing AI applications particularly ASR-based mispronunciation-detection tools (Ahmed et al., 2023; Akhtar et al., 2020) and generative AI applications explored by Al-Jamali & Suliman Abdalla, (2025) as diagnostic aids that operate under teacher supervision, rather than as autonomous instructional replacements. A binary flag of "mispronounced" provides useful triage but does not, by itself, teach the learner what to do next; the corrective feedback that would close the loop still resides with the instructor. Curriculum designers can nonetheless integrate such tools into blended designs where AI-generated diagnostics feed into teacher-led remediation, thereby reducing the load of manual error detection while preserving pedagogical judgement.

For AI developers and interaction designers, the review yields more concrete guidance. Three design decisions recur across the retrieved literature and each

carries actionable lessons. First, on feature engineering, deeper spectral representations tend to outperform classical MFCC features on Arabic speech-classification tasks: log-spectrograms outperformed MFCCs in the CNN-based recogniser of ElGohary et al., (2020), and deep CNN features AlexNet outperformed MFCCs in the mispronunciation-detection classifier of Akhtar et al., (2020) relative to Ahmed et al.'s (2023) MFCC-with-LSTM baseline. Future Arabic speaking-support systems should therefore consider transfer-learning approaches over hand-crafted acoustic features, particularly when the target task involves fine-grained phonemic distinctions crucial for L2 pronunciation development. Second, on modality, the audio-only default that dominates the corpus is inconsistent with the fundamentally audio-visual nature of pronunciation learning. ElGohary et al.'s (2020) integration of a virtual speech-language pathologist avatar within a VR environment, together with the informational value of the visual channel documented by Daou et al., (2025) for lipreading, suggests that Arabic AI systems for speaking should routinely couple audio feedback with a visual articulatory model. Third, on feedback granularity, current mispronunciation-detection systems (Ahmed et al., 2023; Akhtar et al., 2020) diagnose errors but do not explain them. From an SLA perspective, this omission is decisive: noticing requires that the learner attend to a specific mismatch, and repair requires guidance on how to correct it. Interaction designers should therefore prioritise systems capable of returning phonetically explicit feedback "the tongue should approach the interdental position" rather than only categorical judgements.

For institutional decision-makers and funders, the review points to an interdisciplinary funding architecture. Because the present gap is not primarily a technological gap but a translational one, isolated computer-science investment is unlikely to close it. Funding calls that explicitly require the integration of Arabic language educators, applied linguists, and AI researchers in the same project along with pre-registered pedagogical designs are more likely to generate the kind of comparative evidence that the field needs. Cross-institutional infrastructures such as shared L2 Arabic learner speech corpora, ideally reflecting multiple dialect exposures and diverse L1 backgrounds, would address one of the most fundamental constraints identified in the synthesis: the absence of representative learner data for training and evaluation.

Limitations of the Review and Future Research Agenda

Several limitations of the present review should be acknowledged. First, although supplementary hand-searching was performed in IEEE Xplore, ACM Digital Library, and Web of Science, the primary systematic search was conducted in Scopus and ERIC and included only English-language, peer-reviewed journal articles published between 2020 and 2025. This choice, while consistent with PRISMA 2020 recommendations for transparency and reproducibility (Page et al., 2021), may have excluded relevant Arabic-language studies and grey literature, particularly conference proceedings from Arabic-speaking regions where much of the applied research is disseminated. Second, high methodological heterogeneity across the included studies precluded a quantitative meta-analysis; the descriptive narrative synthesis approach (Popay et al., 2006) was necessarily interpretive. Third, because the corpus contained no studies with a controlled pedagogical design, the certainty of

evidence for the pedagogical effectiveness question was rated low from the outset, and no amount of downstream synthesis can raise that certainty until new primary studies are conducted. Fourth, quality appraisal was conducted using the JBI Critical Appraisal Tools (Barker et al., 2024) and the MMAT (Hong et al., 2018) with fixed thresholds, and although this approach is defensible and widely used in comparable reviews (Aromataris et al., 2024; Sguanci, Ferrara, et al., 2024), it inevitably standardises studies that are heterogeneous in design. Fifth, validity and reliability in qualitative synthesis remain a general concern (Noble & Smith, 2025); the transparent audit trail provided in the Supplementary Materials is intended to mitigate this risk but cannot fully eliminate it.

Notwithstanding these limitations, the synthesis supports a concrete and prioritised research agenda organised around four axes: study design, participant sampling, outcome measurement, and cross-disciplinary integration. Together, these four axes translate the null finding of this review into a set of positive next steps.

The first axis concerns study design. Because no prior precedent exists among the 19 included studies, future work should conduct controlled experimental or quasi-experimental investigations that compare AI-supported and non-AI Arabic speaking instruction. Illustrative design elements include random assignment or matched control, pre- and post-intervention speaking assessment, and intention-to-treat analysis. The second axis addresses participant sampling. Rather than relying on native speakers or clinical populations, future studies should recruit non-native adult Arabic learners at defined proficiency levels, with multi-site recruitment across L1 backgrounds, documented proficiency baselines, and ethical release of learner speech data for shared corpora. Ahmed et al., (2023) and Akhtar et al., (2020) provide partial precedents through their use of non-native speech data, and Al-Jamali & Suliman Abdalla, (2025) offers an adoption survey model, though none reaches the level of a controlled pedagogical sample. The third axis targets outcome measurement. Speaking proficiency should be operationalised using validated L2 constructs pronunciation accuracy, fluency, comprehensibility, and communicative competence alongside technical metrics, with rater-based speaking scales, comprehensibility ratings, longitudinal assessment, and alignment with CEFR descriptors. This dimension is absent across the corpus and would draw its theoretical grounding from Chapelle, (2001) and Farrús, (2023). The fourth axis concerns cross-disciplinary integration. Design teams should combine AI engineers, applied linguists, Arabic pedagogy specialists, and instructional designers, working through SLA-informed feedback loops (noticing, interaction, output), multimodal audio-visual feedback, and corrective, phonetically explicit responses. Isolated precedents exist= the virtual speech-language pathologist avatar of ElGohary et al., (2020), the dialect-aware ASR of Khalafallah et al., (2024), and the dialect and emotion classification of Alsadhan, (2025), but none unites these elements within an L2-pedagogical intervention.

This agenda is deliberately conservative. It does not call for exotic new technologies. Instead, it asks the field to apply the standards of evidence-based educational research Escueta et al., (2017) to technologies that already exist. The next generation of Arabic AI-speaking research need not begin with a new algorithm; it can begin with a rigorously designed comparative study that treats existing tools as instructional interventions worth testing. In this sense, the present review does not

merely document a gap. It clarifies precisely where the field's next investment should land and provides a road map for how the transition from technical to pedagogical maturity can be planned, executed, and evaluated.

The four sections of this discussion refine rather than overturn the original null finding. AI-supported Arabic speaking instruction remains a promising but empirically underdetermined enterprise. What has changed, as a result of the more careful synthesis presented here, is the specificity with which the community can now describe where the gap lies, why it persists, and what would be required to close it.

CONCLUSION

This PRISMA-guided systematic review has synthesised 19 empirical studies (2020–2025) on artificial intelligence in Arabic speaking-related research, with a specific interest in non-native learners and in the comparative evidence question. The review shows that the Arabic AI evidence base is technically vibrant but pedagogically thin. Most studies 14 of 19 sit within technical/system evaluation or model development; only 4 of 19 explicitly involve non-native Arabic learners; and no study implements a controlled pedagogical comparison between AI-supported and non-AI Arabic speaking instruction with validated learner-level outcomes. This is a robust null finding, and it constitutes the review's primary empirical contribution.

The review makes three further contributions. Methodologically, by reclassifying study designs against educational research standards, the review resolves the taxonomic ambiguity that had allowed "experimental" computational work to be mistaken for pedagogical experimentation. Theoretically, by mapping SLA mechanisms to AI affordances and to observed design features, the review moves the field from a general claim that "SLA is absent" to a more defensible claim that SLA is limited, implicit, or insufficiently operationalised—and it specifies exactly where operationalisation must be added. Practically, by translating the null finding into a four-axis research agenda (design, sampling, outcomes, integration), the review reframes the community's next move: not the invention of new algorithms, but the rigorous instructional testing of existing ones with non-native Arabic learners.

For teachers and curriculum designers, the implication is a cautious, teacher-supervised adoption of AI as a diagnostic aid. For AI developers, the priority is to move beyond MFCC-only, audio-only, binary-feedback systems toward transfer-learning-based, audio-visual, and phonetically explicit corrective designs. For funders and policymakers, the priority is to support interdisciplinary teams and shared L2 Arabic learner speech corpora that can sustain the pedagogical validation the field now needs. Only through such coordinated moves can Arabic AI research complete the transition from technological maturity to pedagogical maturity for the benefit of non-native learners of Arabic.

REFERENCES

A-Abbasi, H., Al-Qudheeby, M., Kheyami, Z. A., Khalil, R., Khamees, N., Hijjawi, O., Sallam, M., & Barakat, M. (2024). Cross-Linguistic Evaluation of Generative AI Models for Diabetes and Endocrine Queries. *Jordan Medical Journal*. https://doi.org/10.35516/jmj.v58i4.3369

Ahmed, A., Bader, M., Shahin, I., Nassif, A. B., Werghi, N., & Basel, M. (2023). Arabic Mispronunciation Recognition System Using LSTM Network. *Information*, *14*(7), 413. https://doi.org/10.3390/info14070413

Akasheh, W. M., Haider, A. S., Al-Saideen, B., & Sahari, Y. (2024). Artificial intelligence-generated Arabic subtitles: Insights from Veed.io's automatic speech recognition system of Jordanian Arabic. *Texto Livre*, *17*, e46952. https://doi.org/10.1590/1983-3652.2024.46952

Akhtar, S., Hussain, F., Raja, F. R., Ehatisham-ul-haq, M., Baloch, N. K., Ishmanov, F., & Zikria, Y. B. (2020). Improving Mispronunciation Detection of Arabic Words for Non-Native Learners Using Deep Convolutional Neural Network Features. *Electronics*, *9*(6), 963. https://doi.org/10.3390/electronics9060963

Alashban, A. A., & Alotaibi, Y. A. (2023). A Deep Learning Approach for Identifying and Discriminating Spoken Arabic Among Other Languages. *IEEE Access*, *11*, 11613–11628. https://doi.org/10.1109/ACCESS.2023.3241855

Al-Jamali, S., & Suliman Abdalla, S. Z. (2025). Behavioral Determinants of AI-Driven Arabic Language Learning: Insights from the Extended UTAUT2 Model. *Educational Process International Journal*, *16*(1). https://doi.org/10.22521/edupij.2025.16.282

Alkaoud, M., Alsaqoub, M., Aljodhi, I., Alqadibi, A., & Altammami, O. (2025). ACLM: Developing a Compact Arabic Language Model. *The International Arab Journal of Information Technology*, *22*(3). https://doi.org/10.34028/iajit/22/3/9

Alsadhan, N. (2025). A Novel Dialect-Aware Framework for the Classification of Arabic Dialects and Emotions. *Journal of Computer Science*, *21*(1), 88–95. https://doi.org/10.3844/jcssp.2025.88.95

Al-Shaban, F. A., Ghazal, I., Thompson, I. R., Klingemier, E. W., Aldosari, M., Al-Shammari, H., Al-Faraj, F., El-Hag, S., Tolefat, M., Ali, M., Nasir, B., & Frazier, T. W. (2023). Development and validation of an Arabic language eye-tracking paradigm for the early screening and diagnosis of autism spectrum disorders in Qatar. *Autism Research*, *16*(12), 2291–2301. https://doi.org/10.1002/aur.3046

Aromataris, E., Lockwood, C., Porritt, K., Pilla, B., & Jordan, Z. (Eds.). (2024). *JBI Manual for Evidence Synthesis*. JBI. https://doi.org/10.46658/JBIMES-24-01

Barker, T. H., Habibi, N., Aromataris, E., Stone, J. C., Leonardi-Bee, J., Sears, K., Hasanoff, S., Klugar, M., Tufanaru, C., Moola, S., & Munn, Z. (2024). The revised JBI critical appraisal tool for the assessment of risk of bias for quasi-experimental studies. *JBI Evidence Synthesis*, *22*(3), 378–388. https://doi.org/10.11124/JBIES-23-00268

Booth, A., Sutton, A., Clowes, M., & Martyn-St James, M. (2022). *Systematic Approaches to a Successful Literature Review*. SAGE Publications Ltd. https://doi.org/10.4135/9781036232009

Chapelle, C. A. (2001). *Computer Applications in Second Language Acquisition* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139524681

Chraibi, K., Chaker, I., Dhassi, Y., & Zahi, A. (2024). MSAPersonality: A modern standard Arabic dataset for personality recognition. *International Journal of Electrical and Computer Engineering (IJECE)*, *14*(4), 4498. https://doi.org/10.11591/ijece.v14i4.pp4498-4507

Daou, S., Ben-Hamadou, A., Rekik, A., & Kallel, A. (2025). Cross-Attention Fusion of Visual and Geometric Features for Large-Vocabulary Arabic Lipreading. *Technologies*, *13*(1), 26. https://doi.org/10.3390/technologies13010026

Du, J., & Daniel, B. K. (2024). Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Computers and Education: Artificial Intelligence*, *6*, 100230. https://doi.org/10.1016/j.caeai.2024.100230

ElGohary, S. H., Lithy, A., Khamis, S., Ali, A., Alaa El-Din, A., & El-Azim, H. A. (2020). Interactive Virtual Rehabilitation for Aphasic Arabic-Speaking Patients. *Advances in Science, Technology and Engineering Systems Journal*, *5*(5), 1225–1232. https://doi.org/10.25046/aj0505148

Escueta, M., Quan, V., Nickow, A. J., & Oreopoulos, P. (2017). *Education Technology: An Evidence-Based Review* (No. W23744; p. w23744). National Bureau of Economic Research. https://doi.org/10.3386/w23744

Farrús, M. (2023). Automatic Speech Recognition in L2 Learning: A Review Based on PRISMA Methodology. *Languages*, *8*(4), 242. https://doi.org/10.3390/languages8040242

Güngenci, M. M., & Yıldız, M. (2024). *Challenges in Listening and Speaking Skills for Arabic Language Pre-Service Teachers: A Correlational Study*. https://doi.org/10.5281/ZENODO.13860910

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane Handbook for Systematic Reviews of Interventions* (1st ed.). Wiley. https://doi.org/10.1002/9781119536604

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, *34*(4), 285–291. https://doi.org/10.3233/EFI-180221

Hussein, A., Watanabe, S., & Ali, A. (2021). *Arabic Speech Recognition by End-to-End, Modular Systems and Human* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2101.08454

Khalafallah, H. B., Fattah, M. A., & Abdulrahman, R. (2024). Speech corpus for Medina dialect. *Journal of King Saud University - Computer and Information Sciences*, *36*(2), 101864. https://doi.org/10.1016/j.jksuci.2023.101864

Li, B., Tan, Y. L., Wang, C., & Lowell, V. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. *Computers and Education: Artificial Intelligence*, *9*, 100445. https://doi.org/10.1016/j.caeai.2025.100445

Lockwood, C., Munn, Z., & Porritt, K. (2015). Qualitative research synthesis: Methodological guidance for systematic reviewers utilizing meta-aggregation. *International Journal of Evidence-Based Healthcare*, *13*(3), 179–187. https://doi.org/10.1097/XEB.0000000000000062

Long, M. H. (1996). The Role of the Linguistic Environment in Second Language Acquisition. In *Handbook of Second Language Acquisition* (pp. 413–468). Elsevier. https://doi.org/10.1016/B978-012589042-7/50015-3

Lubis, I., Lessy, Z., & Hasan, M. N. (2022). STUDENTS' DIFFICULTIES IN PRACTICING ARABIC'S MAHARAH KALAM AT THE ARABIC LANGUAGE EDUCATION DEPARTMENT IAIN LANGSA. *El-Tsaqafah : Jurnal Jurusan PBA*, *21*(1), 1–14. https://doi.org/10.20414/tsaqafah.v21i1.4690

M. O. Nahar, K., Almomani, A., Shatnawi, N., & Alauthman, M. (2023). A Robust Model for Translating Arabic Sign Language into Spoken Arabic Using Deep Learning. *Intelligent Automation & Soft Computing*, *37*(2), 2037–2057. https://doi.org/10.32604/iasc.2023.038235

Mahmoudi, O., El Allali, N., & Bouami, M. F. (2024). AMSVT: Audio Mel-spectrogram vision transformer for spoken Arabic digit recognition. *Indonesian Journal of Electrical Engineering and Computer Science*, *35*(2), 1013. https://doi.org/10.11591/ijeecs.v35.i2.pp1013-1021

Maryam, S., Sutaman, S., Miftahul Huda, Mufti Rasyid, & Dissanayake, B. Y. (2025). Student Engagement with the Arabic Language on Social Media: A Media Exposure Approach to Language Learning. *Izdihar : Journal of Arabic Language Teaching, Linguistics, and Literature*, *8*(1), 87–104. https://doi.org/10.22219/jiz.v8i1.36201

Meftah, A. H., Alotaibi, Y. A., & Selouani, S.-A. (2022). Arabic Emotional Voice Conversion Using English Pre-Trained StarGANv2-VC-Based Model. *Applied Sciences*, *12*(23), 12159. https://doi.org/10.3390/app122312159

Miao, X., & Wang, P. (2023). A Literature Review on Factors Affecting Motivation for Learning Arabic as a Foreign Language. *Open Journal of Social Sciences*, *11*(06), 203–211. https://doi.org/10.4236/jss.2023.116014

Mohmad Mohmiddin, N. A., & Abu Bakar, R. (2025). The Role Of Artificial Intelligence In Learning Arabic For Non-Native Speakers: A Systematic Literature Review: دور الذكاء الاصطناعي في تعلم اللغة العربية للناطقين بغيرها: استعراض منهجي للدراسات السابقة. *Al-Qanatir: International Journal of Islamic Studies*, *34*(03), 198–216. https://doi.org/10.64757/alqanatir.2025.3403/1121

Noble, H., & Smith, J. (2025). Ensuring validity and reliability in qualitative research. *Evidence Based Nursing*, *28*(4), 206–208. https://doi.org/10.1136/ebnurs-2024-104232

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). *Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC Methods Programme*. Lancaster University. https://doi.org/10.13140/2.1.1018.4643

Ritonga, M., Zulmuqim, Z., Bambang, B., Kurniawan, R., & Pahri, P. (2022). SIAKAD machine learning for correcting errors in speaking Arabic. *World Journal on Educational Technology: Current Issues*, *14*(3), 768–780. https://doi.org/10.18844/wjet.v14i3.7214

Salem Abdulhafid, K. A., Mustapha, N. F., Hassan, A. R., & Ismail, M. Z. (2024). Improvement of Arabic Speaking Skills among Non-Arabic Students through Role-Play Activities. *International Journal of Academic Research in Progressive*

Education and Development, *13*(1), Pages 2299-2313. https://doi.org/10.6007/IJARPED/v13-i1/20760

Schmidt, R. W. (1990). The Role of Consciousness in Second Language Learning1. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Sguanci, M., Ferrara, G., Palomares, S. M., Parozzi, M., Godino, L., Gazineo, D., Anastasi, G., & Mancin, S. (2024). Dysgeusia and Chronic Kidney Disease: A Scoping Review. *Journal of Renal Nutrition*, *34*(5), 374–390. https://doi.org/10.1053/j.jrn.2024.04.005

Sguanci, M., Mancin, S., Piredda, M., & De Marinis, M. G. (2024). Protocol for conducting a systematic review on diagnostic accuracy in clinical research. *MethodsX*, *12*, 102569. https://doi.org/10.1016/j.mex.2024.102569

Son, J.-B., Ružić, N. K., & Philpott, A. (2025). Artificial intelligence technologies and applications for language learning and teaching. *Journal of China Computer-Assisted Language Learning*, *5*(1), 94–112. https://doi.org/10.1515/jccall-2023-0015

Wang, F., Cheung, A. C. K., Neitzel, A. J., & Chai, C. S. (2025). Does Chatting with Chatbots Improve Language Learning Performance? A Meta-Analysis of Chatbot-Assisted Language Learning. *Review of Educational Research*, *95*(4), 623–660. https://doi.org/10.3102/00346543241255621

Xia, Y., Shin, S.-Y., & Kim, J.-C. (2024). Cross-Cultural Intelligent Language Learning System (CILS): Leveraging AI to Facilitate Language Learning Strategies in Cross-Cultural Communication. *Applied Sciences*, *14*(13), 5651. https://doi.org/10.3390/app14135651

Zubeiri, I., Souri, A., & El Mohajir, B. E. (2025). Arabic text diacritization using transformers: A comparative study. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *14*(1), 702. https://doi.org/10.11591/ijai.v14.i1.pp702-711