Arabiyatuna
Jurnal Bahasa Arab

Vol. 10. No. 2, November 2026, 579-596
P-ISSN: 2580-5045, E-ISSN: 2580-5053
DOI: http://dx.doi.org/10.29240/jba.v10i2.16372

AI in Arabic Language Learning: A Critical Review of Pedagogical Opportunities and Challenges

Segaf Baharun[1], Muhamad Solehudin[2], Moh. Tohiri Habib[3], Muhammad Khoirul Umam[4], Iqbal Fathi Izzudin[5]
Universitas Islam Internasional Darullughah Wadda'wah Pasuruan, Indonesia[1,2,3,4]
Lancaster University, United Kingdom[5]

*Corresponding Author: *Khoirulumam@uiidalwa.ac.id*

Cite this article:
Baharun, S., Solehudin, M., Tohiri Habib, M., Khoirul Umam, M., & Fathi Izzudin, I. (2026). Artificial Intelligence in Arabic Language Learning: Pedagogical Implications. Arabiyatuna: Jurnal Bahasa Arab, 10(2), 559–596. https://doi.org/10.29240/jba.v10i2.16372

Received: 30-07-2026	Revised: 21-09-2026	Accepted: 01-11-2026

Abstract

The use of artificial intelligence (AI) in language education has expanded faster than the Arabic-specific evidence needed to guide classroom decisions. This critical narrative review examines what existing studies can and cannot establish about AI-supported Arabic learning and develops an Arabic-specific human–AI pedagogical integration framework. A revision-stage bibliographic audit screened the 39 references in the submitted manuscript for topic–claim alignment, retained seven sources for the analytical corpus, and added nine verifiable, directly relevant studies or linguistic reviews (16 sources in total). The review is a bounded qualitative synthesis, not a systematic review; the original database retrieval logs and screening counts were unavailable. Cross-study comparison indicates plausible advantages for practice frequency, writing feedback, supported dialogue, translation and differentiated learning, but outcome evidence is heterogeneous and generally short-term. For example, one Arabic-classroom study reported perceived communication benefits among 34 pupils, whereas an expert-assessed study of 30 Arabic texts found fully acceptable AI feedback for only 50.76% of identified errors. Linguistic risks are concentrated in root-and-pattern morphology, undiacritized forms, iʿrāb, Standard Arabic–dialect variation, pragmatic register and culturally situated meaning. The proposed framework assigns AI repetitive, adaptive and first-pass analytic tasks while requiring

Copyright © 2026 **Author'(s)**
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

teachers to validate linguistic output, scaffold interaction, assess communicative competence and protect ethical use. Its contribution is a testable allocation of instructional responsibilities rather than an unqualified claim of AI effectiveness. Future Arabic-classroom studies should evaluate accuracy, transfer, fairness and learning outcomes across learner levels and language varieties. The analysis differentiates direct Arabic classroom findings from comparative evidence in other languages, avoiding unsupported generalization across distinct linguistic systems. It also identifies practical safeguards for classroom adoption, including independent learner attempts, transparent verification of automated feedback, and assessment without AI assistance. These recommendations are intended for empirical testing rather than presented as validated outcomes.

Keywords: Arabic language learning; generative artificial intelligence; critical narrative review; human–AI collaboration; teacher mediation.

Introduction

Artificial intelligence (AI) now offers language learners rapid feedback, individualized prompts, conversational simulations and access to linguistic resources. Reviews of second-language learning describe these possibilities but warn that research has often concentrated on adoption, perceptions and short interventions rather than durable proficiency outcomes (Li et al., 2024; Yang & Li, 2024). Positive results from other languages can inform instructional design, but they do not automatically validate AI output in Arabic. Arabic learning encompasses not only vocabulary and grammar but also the interpretation of vowelization, morphological derivation, varieties of Arabic and socially appropriate discourse.

The empirical record illustrates both opportunity and caution. In an Arabic classroom, Fridayanti and Khusna (2025) reported that 84% of 34 participating pupils viewed ChatGPT-mediated question-and-answer activities as helping comfort and confidence; this self-report does not establish measured oral proficiency gains. A pre-experimental Arabic writing study with 45 learners reported improvement during chatbot-assisted translanguaging tasks, but its design limits causal attribution (Rahmawati Zaimah et al., 2024). By contrast, Zaid et al. (2026) assessed the usefulness of ChatGPT feedback against 132 identified errors from 30 student texts and found that only 50.76% of feedback was fully acceptable to expert raters. These studies differ in outcome measures and educational setting; they should therefore be compared critically rather than pooled as equivalent proof of effectiveness.

A second problem is that Arabic linguistic complexity changes which AI functions are dependable. Root-and-pattern morphology, optional diacritics, case and mood endings, orthographic ambiguity, and the distance between Modern Standard Arabic (MSA) and everyday varieties create error patterns that cannot be reduced to generic language-technology concerns (Alayba, 2025). Research on Arabic diglossia additionally identifies educational challenges linked to the relationship between standard and spoken varieties, requiring deliberate curricular mediation (Taha Thomure et al., 2025). Computational access to a form is not identical to a defensible pedagogical explanation of that form.

Existing language-AI reviews catalogue uses of ChatGPT, chatbot-supported interaction and perceived benefits (Li et al., 2024; Yang & Li, 2024). Arabic-specific reports increasingly examine speaking practice, grammar learning and feedback (Fridayanti & Khusna, 2025; Hakim et al., 2026; Zaid et al., 2026). Yet these strands are rarely articulated as a single instructional allocation that specifies when teachers should intervene and how the adequacy of AI output should be assessed in relation to Arabic language structures, communicative situations and learner needs. The gap addressed here is conceptual and pedagogical, not a claim that no Arabic-specific empirical studies exist. Comparative and contextual scholarship also informs this interpretation (Cong-Lem et al., 2025); (Wiboolyasarin et al., 2025); (Lo et al., 2024); (Li et al., 2025).

This article asks three questions: (1) What forms of AI assistance are supported by the available Arabic-specific and comparative L2 evidence, and how strong is that support? (2) Which Arabic-language characteristics and contextual conditions qualify the expected pedagogical benefits? (3) How can teachers operationalize an Arabic-specific human–AI workflow with observable learning outcomes? The study undertakes literature analysis only; it does not collect interviews, administer interventions or measure new learner outcomes. Its novelty lies in the proposed task-allocation and verification framework, which can be tested in subsequent empirical research.

Method

This study is a critical, structured narrative review based on library research and an explicit bibliographic audit. It is not presented as a systematic review, meta-analysis or PRISMA-compliant search. This distinction is necessary because the original manuscript describes searching Scopus, Google Scholar and accredited journal portals but contains no archived query logs, database-specific hit totals, exported records, deduplication history or independent reviewer screening records. Reconstructing those data retrospectively as if they had been recorded would create false methodological precision.

On 23 October 2026, the revision process re-examined the 39 bibliographic entries in the two supplied copies of the same manuscript. Titles,

substantive topics and available publisher/journal metadata were checked against claims in the draft. Seven sources with an appropriate direct AI-language relationship were retained for the primary synthesis: Abduh et al. (2025), Karataş et al. (2024), Kim and Su (2024), Li et al. (2024), Metwally et al. (2024), Wang et al. (2024), and Yang and Li (2024). The other 32 were not used as principal support because they were largely about unrelated educational topics, lacked an AI-language focus, or had been cited for claims their research did not investigate. Exclusion here means exclusion from the analytical evidence base, not a judgment that a study has no intrinsic merit.

Targeted supplementary discovery using combinations of 'Arabic language', 'ChatGPT', 'AI', 'writing feedback', 'speaking', 'grammar', 'teacher', 'diglossia' and 'Arabic NLP' identified nine additional articles on journal or publisher landing pages: Alayba (2025), Fridayanti and Khusna (2025), Hakim et al. (2026), Qomari et al. (2026), Rahmawati Zaimah et al. (2024), Rambe et al. (2025), Rini et al. (2026), Taha Thomure et al. (2025), and Zaid et al. (2026). These were selected purposively to fill documented Arabic-specific evidence and linguistic-context gaps. The resulting bounded corpus comprises 16 publications: seven Arabic-specific empirical investigations (including the study of sermon translation), three Arabic-focused conceptual or linguistic reviews, and six comparative or broader L2 AI studies. Database-level retrieval totals and duplicate counts for an exhaustive search remain unavailable and are not asserted.

For future replication rather than as a claim about historical search execution, suitable search templates are: (Arabic OR 'Arabic language') AND ('artificial intelligence' OR ChatGPT OR chatbot) AND (teaching OR learning OR writing OR speaking); ('Arabic morphology' OR diglossia OR 'Arabic dialect') AND (AI OR 'language technology'); and ('generative AI' OR chatbot) AND ('second language' OR L2) AND (feedback OR assessment). A subsequent systematic review would need exact interface-specific strings, date stamps, exports, duplicate removal, screening logs and a prospective protocol. The 23 October 2026 date refers only to this revision-stage audit.

Sources were eligible for the analytical corpus if they (a) addressed AI-mediated language learning or an Arabic linguistic constraint material to AI-assisted pedagogy, (b) had identifiable authors, publication metadata and a retrievable DOI or journal record, and (c) offered empirical, review-based or conceptual information linked to one of the three research questions. Sources whose actual topic concerned, for example, educational management, Islamic history media, general student motivation without AI, or non-AI curriculum policies were not used to substantiate AI efficacy claims. General EFL or Korean-language evidence was retained only as comparative, not Arabic-direct, evidence.

A transparent three-level interpretive appraisal was applied to individual claims: direct Arabic empirical observations (A), Arabic-specific linguistic or conceptual scholarship (B), and non-Arabic L2 comparison studies (C). These labels indicate relevance and design proximity, not standardized risk-of-bias grades. Within each level the discussion distinguishes self-report, qualitative observations, non-randomized pre/post evaluation, expert error judgments and broader review evidence; potential design weaknesses and the absence of long-term outcomes are explicitly noted. Titles/abstracts and accessible journal descriptions underlie several appraisals; the source checks are not equivalent to complete independent full-text risk-of-bias assessment.

The analytic trace is organized as source → documented observation or argument → initial code (e.g., feedback accuracy, participation, morphological ambiguity) → category (technology affordance, learner outcome, linguistic risk, teacher function) → theme. The resulting five synthesis themes were (1) differentiated practice, (2) writing and translation reliability, (3) speaking and affective engagement, (4) Arabic-specific validity, and (5) mediated implementation and ethics. Table 2 records representative evidence-to-inference links. No unverifiable claim of independent duplicate coding or adjudicated inter-coder agreement is made; before journal submission, the authors should validate the source-level decisions and provide archived extraction sheets if available.

Table 1. Transparent reconstruction of the literature base and reporting limits.

Review element	Procedure or eligibility criterion	Documented result / limitation
Original manuscript	Bibliography in the two uploaded reviewer versions was checked as one underlying manuscript.	39 unique reference entries identified.
Topical audit	Keep verifiable work whose topic supports AI-related pedagogy or specific Arabic linguistic constraints.	7 original sources retained; 32 not used as principal evidence.
Supplementary discovery	Targeted searches and verification at journal/publisher landing pages during revision.	9 relevant sources added on 23 October 2026.
Final analytical corpus	Arabic empirical, Arabic conceptual/linguistic and comparative L2 evidence.	16 publications: 7 Arabic empirical, 3 Arabic contextual/review and 6 comparative L2.
Historical searches	Scopus and Google Scholar were named in the original draft, but logs/exports were absent.	Original hits, duplicate count, excluded full texts and exact original search dates: unavailable.
Evidence	Classify each claim by Arabic	A = direct Arabic empirical; B =

Review element	Procedure or eligibility criterion	Documented result / limitation
appraisal	relevance, study design, observation and limitation.	Arabic context/review; C = comparison L2.
Interpretive coding	Map source claim → code → category → cross-source theme; display an audit trail.	Five themes synthesized; independent dual-coder agreement was not documented.
Review status	Purposive, bounded critical narrative synthesis.	No comprehensiveness, prevalence estimates or meta-analytic effect size claimed.

Findings and Discussion

Overview: What can be inferred from the Available Studies?

The 16-source corpus offers evidence for instructional possibilities rather than a single generalizable effect size. The strongest Arabic-specific evidence in this selection concerns writing feedback, grammatical support, mediated interaction and teacher constraints. Self-reported enthusiasm and immediate activity performance are not interchangeable with objectively assessed proficiency, and studies from EFL or Korean language learning offer design analogies rather than estimates of Arabic outcomes. For that reason, claims that adaptive learning is the 'most dominant' application, or that AI invariably improves achievement, have been removed. The synthesis instead separates demonstrated observations, plausible mechanisms and instructional proposals.

Differentiated Practice and Learner Agency

AI can quickly generate graded examples, vocabulary prompts, practice items and initial error explanations. This is pedagogically useful when students arrive with different levels of Arabic literacy or struggle with particular morphological constructions. In a qualitative Arabic grammar study of 30 university students, Hakim et al. (2026) described the use of generative tools alongside translation and digital grammar platforms and reported favorable student engagement; however, its descriptive design and the absence of a clean AI-only comparator preclude attribution of course performance to AI. Comparatively, Abduh et al. (2025) examined generative AI in Indonesian EFL settings and described autonomy and participation benefits conditional on access and teachers' critical guidance. The transferable lesson concerns the design of scaffolded activities, not identical linguistic performance across languages. Comparative and contextual scholarship also informs this interpretation (Guo & Li, 2024); (Tasdelen & Bodemer, 2025); (Shadiev et al., 2023).

A defensible Arabic task therefore starts with a diagnostic demand: vocabulary recall may justify fast drill generation, while case-governed parsing of authentic texts requires teacher evaluation. Learners should attempt an answer before requesting machine feedback, compare alternatives, record disagreements and explain their final choice. This sequencing protects the reasoning involved in *Sarf* and *nahw* from being bypassed by simply copying generated solutions.

Writing, Morphology and Translation: Speed Versus Reliability

Arabic writing shows why feedback speed cannot be treated as feedback validity. Rahmawati Zaimah et al. (2024) found favourable outcomes in a 45-participant pre-experimental chatbot/translanguaging activity, but group design and measurement boundaries restrict strong causal conclusions. Zaid et al. (2026) directly evaluated AI correction in 30 student texts and documented 132 errors; expert raters found AI feedback fully acceptable for 50.76% of these errors. The difference is not necessarily a contradiction: the first investigates learner performance under a teaching design, whereas the second assesses the accuracy and usefulness of particular corrective responses. Together they support supervised drafting and revision, not unmonitored automated grading. Comparative and contextual scholarship also informs this interpretation (Li & Han, 2024); (Li, 2024); (Mohammed & Khalid, 2025).

For translation, the risk extends beyond grammatical form to religious and pragmatic meaning. Metwally et al. (2024) identified lexical, register and meaning-related weaknesses in ChatGPT translation of the Farewell Sermon. This is directly relevant to Arabic religious or heritage texts, where the lexical rendering of a term and the interpersonal stance of the original are integral to appropriate comprehension. Automated translation can serve as a provisional comparison text; learners should justify translations against contextual, morphological and specialist reference evidence. Claiming that machine translation ensures accurate understanding of classical texts would exceed the reviewed evidence. Comparative and contextual scholarship also informs this interpretation (Xu et al., 2025); (Qobti & Almohaimeed, 2024).

Conversational AI: Opportunity with Limited Transfer Evidence

Chatbots may create low-pressure spaces for rehearsing formulaic exchanges and negotiating word choice. Fridayanti and Khusna (2025) documented favorable perceptions among 34 school pupils, but an 84% satisfaction-style response is not an observed proficiency effect. Rini et al. (2026) similarly report perceived opportunities for self-directed practice and immediate feedback from interviews, observation and questionnaires with two lecturers and 13 students in a single-site Arabic program; the authors explicitly caution that reliable morphology and measurable learning effects require further work. A conceptually oriented account by Rambe et al. (2025) likewise describes

the pedagogical potential of dialogue practice rather than testing a controlled intervention.

Broader studies show why transfer must be qualified. Kim and Su (2024) reported improved willingness to communicate in a Korean-as-a-foreign-language chatbot condition, while Wang et al. (2024) compared conversational GenAI conditions for Chinese learners of English and examined anxiety, communicative self-perceptions and speaking performance. These studies support hypotheses about reduced speaking anxiety and more frequent practice, but Arabic pronunciation and pragmatic appropriateness demand additional assessment. A learner who can produce a fluent-looking MSA dialogue on a chatbot may still be unable to handle an appropriate greeting, turn-taking choice or register shift in a real interaction. Comparative and contextual scholarship also informs this interpretation (Klimova et al., 2024).

Arabic-Specific Constraints as Instructional Problems

Morphology and vocalization. Arabic root-and-pattern derivation and inflection require learners to distinguish lexical identity from grammatical function. Undiacritized words are often underdetermined without syntax and context; an automatically generated vowelization or *iʿrāb* explanation may therefore sound convincing while selecting an implausible interpretation. Alayba (2025) describes morphology, diacritics and ambiguity as continuing Arabic NLP challenges; Zaid et al. (2026) provide classroom-level evidence that feedback reliability weakens for morphologically complex and context-dependent cases. Teachers should make ambiguity visible through contrastive tasks rather than treating a single generated analysis as final.

Diglossia and dialects. MSA and spoken varieties differ in lexical usage, grammatical patterns and social distribution. Taha Thomure et al. (2025) synthesize education research identifying important learning consequences of this variation. Language models may silently blend an MSA lesson with colloquial forms or generate dialect language without clear regional labelling. Pedagogically, prompts and assessments must declare the intended variety, target level and communicative setting; human speakers or curated recordings should establish whether the selected register is appropriate. These criteria are especially important when learners transfer from textbook Arabic to authentic conversation.

Pragmatics and culture. A syntactically plausible response can still be unsuitable for an age relationship, institutional setting or culturally sensitive topic. Metwally et al. (2024) document translation problems involving religious voice and register, which illustrate that semantic adequacy is not captured by word-for-word equivalence. Teachers should use role-based dialogue, cultural comparison, peer interaction and meaning-justification to evaluate whether students understand why an expression is used. Accuracy needs to be defined as

formal correctness plus contextual appropriateness, rather than matching a model's preferred text.

Infrastructure and equity. AI-supported practice is dependent on device access, connectivity, teachers' confidence in evaluating output, and students' privacy literacy. Qomari et al. (2026), in a survey of 46 Arabic teachers, identify an adoption gap between positive perceptions or knowledge and classroom practice, with training and unequal access remaining important barriers. Because such a survey does not demonstrate a universal pattern, institutions should conduct their own access audit and provide non-AI alternatives for learners who cannot participate on equivalent terms. Comparative and contextual scholarship also informs this interpretation (Cong-Lem et al., 2025); (Muhdi et al., 2024); (Zhou et al., 2024).

Table 2. Cross-study evidence matrix (claims are restricted to the studies' reported designs and outcomes).

AI practice / skill	Evidence source and design	Observed opportunity	Limitations and necessary teacher role
Chatbot dialogue / speaking	Fridayanti & Khusna (2025), Arabic school-based qualitative inquiry; n=34.	84% reported greater comfort/confidence with ChatGPT-assisted questions and answers.	Self-report, small sample, not a proficiency test; teacher must verify response relevance, pronunciation and real-life transfer.
Chatbot/translanguaging / writing	Rahmawati Zaimah et al. (2024), Arabic pre-experiment; n=45.	Reported writing-comprehension improvement during AI-supported translanguaging.	Limited causal identification; assess independent writing and compare without AI.
AI correction / Arabic writing	Zaid et al. (2026), expert appraisal of 30 student texts, 132 errors.	Rapid candidate explanations and corrections; 50.76% of feedback fully acceptable.	Substantial unvalidated feedback; teacher checks ṣarf, syntax and contextual appropriateness.
Digital grammar resources / iʿrāb	Hakim et al. (2026), Arabic descriptive classroom study; n=30.	Perceived scaffolding, resource access and learner engagement.	Multiple co-used platforms and no isolating comparator; teacher requires independently justified parsing.
AI translation / religious Arabic	Metwally et al. (2024), case-	Fast draft translation for critical	Register and lexical/meaning

AI practice / skill	Evidence source and design	Observed opportunity	Limitations and necessary teacher role
	based evaluation of Farewell Sermon.	comparison.	errors; cultural and disciplinary specialists validate final meaning.
Chatbot participation / Arabic	Rini et al. (2026), case study, two lecturers and 13 students.	Perceived autonomy and immediate-form feedback possibilities.	Single-site and perception-based evidence; teacher verifies morphology and learning gains.
AI adoption / teacher development	Qomari et al. (2026), survey of 46 Arabic teachers.	Identifies support needed to turn favorable perceptions into teaching practice.	Context-specific, correlational survey; staff development plus access audit required.
L2 speaking comparisons	Kim & Su (2024) Korean; Wang et al. (2024) EFL.	Reports stronger willingness to communicate or examined changes in communicative affect/performance.	Not Arabic outcome measures; teachers must test dialect, pronunciation and authentic transfer.
Language AI evidence synthesis	Li et al. (2024); Yang & Li (2024), multi-study reviews.	Maps prompting, content generation, feedback and self-directed learning uses.	Heterogeneous designs and relatively short evidence base; no basis for one pooled Arabic effectiveness claim.
Arabic linguistic context	Alayba (2025) NLP review; Taha Thomure et al. (2025) diglossia review.	Specifies orthographic, morphological and variety-related risks and instructional needs.	Computational or general literacy reviews, not tests of the proposed AI classroom workflow.

Integrative Pedagogical Framework: Arabic-Specific Human–AI Co-Regulation

The review's central conceptual contribution is an Arabic-specific human–AI co-regulation framework. It links language-specific task demands to learner needs, assigns AI only functions supported by its affordances, and makes teacher validation obligatory where a mistake would distort linguistic reasoning, cultural meaning or evaluation. 'Co-regulation' denotes supervised orchestration of practice and feedback, not equal authority between machine and teacher. The

model is a proposed design hypothesis, not an empirically validated intervention.

1 ARABIC TASK DEMAND →	2 LEARNER DIAGNOSIS →	3 AI SUPPORT →	4 TEACHER VALIDATION →	5 VERIFIED OUTCOME
MSA/dialect; morphology and iʿ rāb; unvowelized reading; contextual register.	Prior competence, desired variety, task difficulty, access and confidence.	Repeated exercises, provisional parsing, dialogue rehearsal, draft feedback.	Check accuracy and sources; explain ambiguity; scaffold culture and peer interaction.	Independent oral/written performance, justified revisions and responsible AI use.

Figure 1. Editable conceptual flow of the proposed instructional model.

Feedback loop: learners justify or reject AI responses → teacher reviews error patterns → next task/prompt is adjusted → independent assessment checks transfer.

In practice the sequence begins with an explicit target and a non-AI baseline. The teacher selects a bounded tool function (for example, generating alternative sentences for a single syntactic pattern). Learners produce an independent first attempt, use AI for examples or preliminary feedback, verify disputed segments against a grammar explanation or trusted source, and discuss consequential corrections with a teacher or peers. Assessment concludes with an unaided communicative or analytic task; automated fluency or a chatbot's agreement must not substitute for independent demonstrated competence. Learners log the prompt, output, detected error, final decision and supporting reason.

An operational example concerns the unvowelized phrase كتب الطالب الدرس. Instead of accepting one machine-provided vowelization, the learner identifies possible lexical and syntactic analyses, explains why a contextual reading is preferred, and checks case marking with the teacher. A second task asks the learner to compose a polite response for a formal academic interaction using the specified Arabic variety. AI supplies first-pass options, whereas teacher-guided discussion assesses appropriateness and pragmatic meaning. In both cases the observable learning outcome is justified independent performance, not the quality of generated text alone.

Table 3. Practical implementation and assessment indicators for the framework.

Instructional phase	Human–AI distribution of work	Observable indicator / safeguard
Preparation	Teacher specifies Arabic variety, proficiency target, access conditions and acceptable use; AI may propose differentiated item banks.	Pre-task diagnostic; lesson objectives; non-AI pathway offered.
Guided generation	Learner drafts first; AI gives prompts, repetition and provisional explanations.	Original learner attempt stored; prompts and tool/version recorded.
Linguistic validation	Teacher/peers check diacritics, agreement, case endings, word sense, dialect consistency and sources.	Error log with corrected form and justification; high-stakes cases manually checked.
Communication transfer	Human interlocutors or authentic resources test register, turn taking, repair and sociocultural fit.	Unaided role-play or composition rubric rather than AI self-rating.
Reflection and ethics	Teacher trains verification, attribution, privacy protection and bias detection.	Student declaration of AI use; no sensitive student material entered without authorization.
Improvement cycle	Teacher uses error categories and outcome evidence to revise task design; AI assists low-stakes item regeneration.	Compare pre/post unaided performance across task types and Arabic varieties.

Limits of the Evidence and Implications for Research

The literature base is a purposive 16-publication analytical corpus built from an audit of a prior manuscript and supplementary discovery, not an exhaustive database search. The original source logs and numerical screening history cannot be recovered from the supplied documents; thus percentages of all published work, frequency rankings of AI applications and pooled learning gains cannot be inferred. The 16 publications also differ by year, setting, language, AI tool, task and outcome measure. Several Arabic-specific studies rely on qualitative observations, self-reports or nonrandomized designs; technical reviews specify linguistic risks without directly testing classroom interventions. The quality appraisal applied here is a claim-level interpretive assessment and cannot replace a full protocol-driven risk-of-bias process.

Future evaluation should prospectively register search and inclusion procedures, make source-extraction sheets accessible and examine the framework through multi-site pre/post designs with valid comparators.

Outcomes should include unaided writing accuracy, morphology and diacritic interpretation, comprehension, pronunciation, sustained interaction, pragmatic appropriateness and delayed retention. Researchers should stratify analyses by proficiency level, MSA versus named dialects, and text genre, then audit teacher feedback time, access inequities, privacy risks and the rate of consequential AI errors. A comparison of AI-alone, teacher-alone and human–AI mediated conditions would be more informative than favorable perceptions alone. Comparative and contextual scholarship also informs this interpretation (Liu et al., 2025); (Maghfirah, 2025); (Safrudin et al., 2024).

Conclusion

The reviewed sources support a conditional, not universal, case for AI in Arabic learning. Rapid generation and repeated practice can be useful, and early Arabic-classroom studies report engagement and selected writing benefits; however, Arabic writing-error appraisal and translation evaluation also reveal substantial reliability problems. The quality and transferability of evidence vary considerably. Root-and-pattern morphology, omitted diacritics, Arabic varieties, pragmatic choices and culturally sensitive meanings make unsupervised AI explanations especially risky. The proposed Arabic-specific human–AI co-regulation framework therefore assigns AI provisional practice and analytic functions while keeping pedagogical authority, contextual interpretation, ethical safeguards and authentic assessment with educators and learners. The framework is a transparent conceptual contribution that now requires prospective classroom validation.

Acknowledgment

The authors acknowledge the scholarly environments of Universitas Islam Internasional Darullughah Wadda'wah and International Islamic University Malaysia, as stated in the submitted manuscript. The reviewer feedback was instrumental in sharpening the methodological reporting, source alignment and conceptual contribution.

Author Contributions Statement

S.B.: conceptualization, research design, supervision, and manuscript review. A.R.: literature investigation, data curation, and formal analysis. M.K.U.: literature synthesis, methodological refinement, and writing—original draft. N.A.B.: literature verification, language review, and writing—review and editing. This statement is retained from the submitted manuscript; author approval of the revised contributions should be confirmed before submission.

AI Writing Statement

Generative AI was used during revision to assist with language editing, manuscript restructuring, consistency checks and the organization of a literature-based synthesis. Source metadata and claim–citation correspondence were reviewed against accessible academic and publisher records. The authors must independently check the final bibliography, analytic interpretations and author-contribution statements and retain full responsibility for scientific accuracy, originality, scholarly judgments and the submitted version.

References

Abduh, A., Rosmaladewi, R., Arham, M., Asnur, M. N. A., & Dalle, A. (2025). Empowering EFL learners through generative AI: A qualitative exploration of inclusive and autonomous learning in multicultural contexts. International Journal of Language Education, 9(1), 168–179. https://doi.org/10.26858/ijole.v1i1.74760

Alayba, A. M. (2025). Arabic natural language processing (NLP): A comprehensive review of challenges, techniques, and emerging trends. Computers, 14(11), 497. https://doi.org/10.3390/computers14110497

Cong-Lem, N., Bui, C. N. V., Nguyen, N. P. N., & Huynh, M. Q. (2025). Bridging or breaking? A systematic review of how generative AI shapes equity in foreign language education. International Journal of Applied Linguistics. https://doi.org/10.1111/ijal.70052

Fridayanti, A. A., & Khusna, Q. K. (2025). Implementation of the ChatGPT-based question and answer method in improving student Arabic communication skills. Maharaat: Jurnal Pendidikan Bahasa Arab, 6(2), 101–114. https://doi.org/10.18196/mht.v6i2.24388

Guo, K., & Li, D. (2024). Understanding EFL students' use of self-made AI chatbots as personalized writing assistance tools: A mixed methods study. System, 124, 103362. https://doi.org/10.1016/j.system.2024.103362

Hakim, R. I., Darman, I. H., Deviana, A. D., Qomariyah, L., Rusuly, U., & Gusman, E. (2026). AI-assisted learning in Arabic grammar: Investigating the use of ChatGPT and digital platforms in Arabic learning. Al-Ta'rib: Jurnal Ilmiah Program Studi Pendidikan Bahasa Arab, 14(1), 1–16. https://doi.org/10.23971/altarib.v14i1.11513

Karataş, F., Abedi, F. Y., Ozek Gunyel, F., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. Education and Information Technologies, 29, 19343–19366. https://doi.org/10.1007/s10639-024-12574-6

Kim, A., & Su, Y. (2024). How implementing an AI chatbot impacts Korean as a foreign language learners' willingness to communicate in Korean. System, 122, 103256. https://doi.org/10.1016/j.system.2024.103256

Klimova, B., Pikhart, M., & Al-Obaydi, L. H. (2024). Exploring the potential of ChatGPT for foreign language education at the university level. Frontiers in Psychology, 15, 1269319. https://doi.org/10.3389/fpsyg.2024.1269319

Li, B., Lowell, V. L., Wang, C., & Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education: Examining research on ChatGPT's use in language learning and teaching. Computers and Education: Artificial Intelligence, 7, 100266. https://doi.org/10.1016/j.caeai.2024.100266

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

Li, M. (2024). Leveraging ChatGPT for second language writing feedback and assessment. International Journal of Computer-Assisted Language Learning and Teaching, 14(1). https://doi.org/10.4018/IJCALLT.360382

Li, M., & Han, J. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. System, 126, 103502. https://doi.org/10.1016/j.system.2024.103502

Liu, J., Sihes, A. J. B., & Lu, Y. (2025). How do generative artificial intelligence (AI) tools and large language models (LLMs) influence language learners' critical thinking in EFL education? A systematic review. Smart Learning Environments, 12, 48. https://doi.org/10.1186/s40561-025-00406-0

Lo, C. K., Yu, P. L. H., Xu, S., Ng, D. T. K., & Jong, M. S. (2024). Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies. Smart Learning Environments, 11, 50. https://doi.org/10.1186/s40561-024-00342-5

Maghfirah, T. (2025). A decade of Arabic language education research in Indonesian Islamic higher education: A bibliometric mapping of theses and dissertations (2015–2024). ELOQUENCE: Journal of Foreign Language, 4(3), 683–704. https://doi.org/10.58194/ELOQUENCE.V4I3.3022

Metwally, A. A., Bin-Hady, W. R. A., & Asiri, E. (2024). Assessing AI-powered translation quality: Insights from the translation of the Farewell Sermon.

International Journal of Language Education, 8(4), 836–854. https://doi.org/10.26858/ijole.v8i4.70034

Mohammed, S. J., & Khalid, M. W. (2025). Under the world of AI-generated feedback on writing: Mirroring motivation, foreign language peace of mind, trait emotional intelligence, and writing development. Language Testing in Asia, 15, 7. https://doi.org/10.1186/s40468-025-00343-2

Muhdi, A., Yuslam, Y., Ngarifillaili, N., Iqbal, A., & Purnama, Y. (2024). Use of ICT in Arabic language education at high school level and its relationship with teacher skills. Journal Neosantara Hybrid Learning, 2(1), 284–301. https://doi.org/10.55849/jnhl.v2i1.746

Qobti, A., & Almohaimeed, S. (2024). Translation research in the Arabic language: A bibliometric study. International Journal of Linguistics, Literature and Translation, 7(5), 77–88. https://doi.org/10.32996/ijllt.2024.7.5.10

Qomari, N., Sa'diyah, K. K. H., Makki, A., Habibullah, M. R., & Babiker, G. A. B. M. (2026). Teachers' perceptions, knowledge, attitudes, and practices in integrating artificial intelligence into Arabic language teaching. Journal of Arabic Language Teaching, 6(1), 47–62. https://doi.org/10.35719/arkhas.v6i1.2461

Rahmawati Zaimah, N., Fatchiatuzahro, & Hartanto, E. B. (2024). Enhancing writing comprehension in L2 Arabic learners through AI-based translanguaging chatbots. Al-Mubin: Islamic Scientific Journal, 7(1), 21–34. https://doi.org/10.51192/almubin.v7i1.753

Rambe, M. S., Tambunan, I. A., Rahmani, N. S., Nazhifah, N., & Nurhidayati. (2025). Utilizing ChatGPT to develop Arabic speaking skills through deep learning approach. Al-Ma'rifah, 22(1), 129–138. https://doi.org/10.21009/almakrifah.22.01.10

Rini, R., Aflisia, N., Yasmar, R., Rahman, A. A., & Larasati, M. S. (2026). AI chatbots, learner perception, and linguistic reliability in Arabic language learning. Alsina: Journal of Arabic Studies, 8(2), 342–354. https://doi.org/10.21580/alsina.8.2.32844

Safrudin, R., Nandang, A., Siregar, S. D. P., Musthafa, I., Fauzi, M. F., Alby, M. H. F., & Suyono, S. (2024). Development of Arabic language learning research: A bibliometric study on Scopus (2009–2024). Al-Ta'rib, 12(2), 321–338. https://doi.org/10.23971/altarib.v12i2.8929

Shadiev, R., Wen, Y., Uosaki, N., & Song, Y. (2023). Future language learning with emerging technologies. Journal of Computers in Education, 10(3), 463–467. https://doi.org/10.1007/s40692-023-00285-9

Taha Thomure, H., Taha, H., Sabella, T., & Saleh, R. (2025). A systematic review of teaching and learning in the context of Arabic diglossia. Social Sciences & Humanities Open, 12, 102281. https://doi.org/10.1016/j.ssaho.2025.102281

Tasdelen, O., & Bodemer, D. (2025). Generative AI in the classroom: Effects of context-personalized learning material and tasks on motivation and performance. International Journal of Artificial Intelligence in Education, 35, 3049–3070. https://doi.org/10.1007/s40593-025-00491-9

Wang, C., Zou, B., Du, Y., & Wang, Z. (2024). The impact of different conversational generative AI chatbots on EFL learners: An analysis of willingness to communicate, foreign language speaking anxiety, and self-perceived communicative competence. System, 127, 103533. https://doi.org/10.1016/j.system.2024.103533

Wiboolyasarin, W., Wiboolyasarin, K., Tiranant, P., Jinowat, N., & Boonyakitanont, P. (2025). AI-driven chatbots in second language education: A systematic review of their efficacy and pedagogical implications. Ampersand, 14, 100224. https://doi.org/10.1016/j.amper.2025.100224

Xu, S., Su, Y., & Liu, K. (2025). Investigating student engagement with AI-driven feedback in translation revision: A mixed-methods study. Education and Information Technologies, 30, 16969–16995. https://doi.org/10.1007/s10639-025-13457-0

Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. System, 124, 103351. https://doi.org/10.1016/j.system.2024.103351

Zaid, A. H. bin, Muassomah, M., Ismail, M., Setiyadi, A. C., & Yasin, A. (2026). Beyond error correction: ChatGPT as formative feedback for Arabic writing. Jurnal Pendidikan Islam, 13(1), 13–24. https://doi.org/10.15575/jpi.v13i1.57550

Zhou, J., Shen, L., & Chen, W. (2024). How ChatGPT transformed teachers: The role of basic psychological needs in enhancing digital competence. Frontiers in Psychology, 15, 1458551. https://doi.org/10.3389/fpsyg.2024.1458551