Arabiyatuna
Jurnal Bahasa Arab

Vol. 10. No. 2, November 2026, 779-800
P-ISSN: 2580-5045, E-ISSN: 2580-5053
DOI: http://dx.doi.org/10.29240/jba.v10i2.16736

AI in Arabic Language Learning: Evidence, Practices and Research Gaps in Pesantren-A Systematic Literature Review

Ahmad Farhan Rif'an Najib[1*], Ahmad Fataa Nuzilal Furqon[2] Muhammad Izzatul Haq[3], Muhammad Hasbi Abrori[4]
Universitas Islam Indonesia Yogyakarta, Indonesia[1]
Universitas Darussalam Gontor, Indonesia[2,4]
Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia[3]

[*]Corresponding Author*: 26917024@students.uii.ac.id*

Cite this article:
Najib, A. F. R., Fataa Nuzilal Furqon, A., Izzatul Haq, M., & Hasbi Abrori, M. (2026). AI in Arabic Language Learning: Evidence, Practices and Research Gaps in Pesantren-A Systematic Literature Review. Arabiyatuna: Jurnal Bahasa Arab, 10(2), 779–800. https://doi.org/10.29240/jba.v10i2.16736

Received: 20-06-2026	Revised: 10-09-2026	Accepted: 02-11-2026

Abstract

Artificial intelligence (AI) is increasingly used in Arabic language teaching, yet little is known about how it is used in pesantren, where Arabic instruction is tied closely to oral transmission and the study of classical religious texts. This study examines how AI is used in Arabic language learning in pesantren and what remains unresolved in this research area. Using a PRISMA-guided systematic literature review, we analyzed 13 primary studies published between 2020 and 2026, drawn from Scopus, SINTA, and Google Scholar. Ten studies (77%) appeared between 2024 and 2026, following the wider availability of generative AI. Methodologically, 8 studies (62%) are qualitative, 3 (23%) use an experimental or quasi-experimental design, and the remaining 2 use a survey or an evaluation of AI output quality. ChatGPT/GPT-4 appears in 5 studies, purpose-built applications such as TalkPal, Suno AI, and Nemo AI in 4, and the rest describe AI in general terms. By skill, speaking and listening account for 5 studies combined, translation and writing 2 each, reading 1, general proficiency 3, and teacher perspective 1, while curriculum readiness is barely represented. Reported outcomes are mostly gains in motivation and confidence; one quasi-experiment reported scores rising from 72.00 to 88.86. Reported challenges include uneven access, limited teacher readiness, and doubts about the accuracy of AI output on classical texts. Read alongside five

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

earlier systematic reviews, the evidence remains thin: no study compares AI-assisted with non-AI instruction under matched conditions, most designs are short and single-site, and only one study has checked AI accuracy against a classical religious text. These findings suggest that AI currently supports motivation and practice more reliably than productive skills such as writing and translation. Future research needs longer, comparative designs and Expert Review of Islamic Educational Content

Keywords: Artificial Intelligence; Arabic Language Learning; Islamic Boarding School; Chatbot; Systematic Literature Review.

Introduction

Interest in using artificial intelligence for Arabic language teaching has grown alongside two broader trends: the rapid development of generative AI and natural language processing, and a longstanding push in language pedagogy toward more personalized, student-cantered instruction (Fadlelmula & Qadhi, 2024)(Yadav, 2025). In practice this has meant using tools that range from general conversational agents such as ChatGPT to applications built specifically for language practice, such as TalkPal, or for producing audio content, such as Suno AI. These tools are typically introduced to raise engagement and motivation, and to give students access to practice that can be repeated and adjusted to their own pace (Haniah et al., 2026). A handful of recent studies indexed in Scopus, SINTA, and Google Scholar report gains in Arabic learning outcomes following AI use, though the outcomes reported tend to depend heavily on the institution and how the tool was actually implemented(Islami et al., 2026)(Hayad et al., 2025).

Two theoretical frameworks are commonly used to make sense of this trend: Computer-Assisted Language Learning (CALL), particularly its more recent form sometimes labelled Intelligent CALL (ICALL), and constructivist learning theory (Katinskaia, 2025)(Sa'diyah & Nisa', 2023). CALL/ICALL treats AI mainly as a delivery mechanism for adaptive practice, instant feedback, and simulated conversation(Blake, 2023). Constructivist theory, meanwhile, treats learning as something learners build through experience and interaction rather than something transmitted to them(Liangshi, 2024). Taken together, these two perspectives suggest that AI can support Arabic language learning by providing students with more opportunities to engage in dialogue, receive individualized feedback, and construct their understanding through repeated practice, rather than merely consuming learning materials passively(Kuddus, 2022). Whether this is actually happening in *pesantren* classrooms, and to what extent, is the question this review tries to answer.

Pesantren present a specific setting in which to ask that question. Arabic instruction there is closely tied to the direct study of classical texts (*kitāb*) and to a mode of transmission built around the relationship between kyai or ustadz and santri(Hamid et al., 2024), so introducing AI is not simply a matter of technology adoption. Infrastructure varies widely between institutions, and there are open questions about whether AI-generated content, especially when it touches classical Arabic religious texts (*kutub turāts*), can be trusted for accuracy (Ma'arif & Afyudiin, 2025). This review uses a PRISMA-based systematic literature review to map what has actually been studied so far: which AI tools are used, for which skills, with what reported results, and what has not yet been examined(Page et al., 2021).

A quick look at recent studies gives a sense of where the field currently stands. Haniah et al.(Haniah et al., 2026) ran a two-cycle classroom action research project in which postgraduate students used tools such as TalkPal and Duolingo, and reported improvement in listening and speaking. Zahro et al.(Zahro et al., 2026) found that using ChatGPT was associated with more interest in reading Arabic text and better recall of vocabulary among students at a *pesantren*-affiliated institute. On the speaking side, Zaimah and Hartanto (Zaimah & Hartanto, 2026), observed that a ChatGPT chatbot set up to respond empathetically produced longer and more spontaneous conversation over time among learners from pesantren backgrounds, and a separate study on TalkPal use among *pesantren* students reported similar gains in speaking and vocabulary (Moenir, 2025). For listening and vocabulary, Solehudin, Hanifansyah, and Izzuddin (Solehudin et al., 2024), reported that AI-generated music made with Suno AI helped santri retain vocabulary and stay engaged during listening practice. On the writing side, Mukmin et al.(Mukmin et al., 2025) and Masnun (Masnun, 2025), found that ChatGPT and similar tools helped with insya' (composition) practice, though how much students actually improved depended on their starting level. A different kind of concern was raised by Ma'arif, Salma, and Afyuddin (Ma'arif & Afyudiin, 2025), who checked how accurately ChatGPT/GPT-4 translated the classical text Matan al-Ghāyah wa at-Taqrīb and found the results only partly reliable, a reminder that fluency in output is not the same as correctness when religious texts are involved. At the institutional level, Julaeha and Nurcahyaningtias (Julaeha & Nurcahyaningtias, 2025) and Salam, Fitri, and Zulan (Salam et al., 2024), looked at how teachers were using AI alongside more traditional teaching methods. Studies like these are still few, and questions about curriculum readiness are addressed even less often.

There have also been attempts to review this literature at a broader level. Hanun (Hanun & Wassalwa, 2025), reviewed how AI is being discussed in *pesantren*-based Arabic learning, covering its potential, its risks, and ethical questions, though this was a conceptual discussion rather than a synthesis of primary data. Two other reviews, by Supriyanto and Toifah (Supriyanto &

Toifah, 2024), and in Online Learning in Educational Research (Hady et al., 2025), found that AI use is generally reported as helpful across reading, writing, listening, and speaking, but neither review was specifically about *pesantren*. A PRISMA-guided review of AI in Arabic speaking instruction for non-native learners (Manasiq et al., 2026) pointed out that even though AI tools are technically workable, the literature has not really shown that they outperform non-AI methods, at least not with strong comparative evidence. Almelhes (Almelhes, 2026) reached a broadly similar conclusion in a separate review and meta-analysis. Read together, these reviews suggest that AI-Arabic learning research is expanding, but no one has yet pulled together the primary studies specifically situated in *pesantren* into a single, structured picture of what AI tools are used for, how they are studied, and what still needs testing. That is the gap this article tries to address.

The contribution of this article lies not in proposing a new theory of AI-assisted Arabic language learning, but in providing a systematic synthesis and critical evaluation of the primary empirical evidence currently available in the *pesantren* context. Previous studies have tended to examine individual AI tools separately or address the broader ethical, pedagogical, and conceptual implications of AI integration in education. Although Hanun's review has specifically addressed the *pesantren* context, its discussion has remained largely conceptual. As a result, the empirical evidence on how AI is actually being used in Arabic language learning within *pesantren* remains fragmented and insufficiently examined. In particular, the existing literature has not adequately brought together the available primary studies into a unified analytical framework to identify their collective findings, methodological strengths and weaknesses, and areas that remain underexplored.

Addressing this gap is important because the integration of AI into Arabic language learning cannot be adequately understood merely by examining the technological capabilities of individual tools. It also requires consideration of how these technologies interact with the pedagogical practices, institutional characteristics, and educational values of *pesantren*. By systematically examining the available empirical studies, this review identifies the ways AI has been employed in Arabic language instruction, the learning opportunities and outcomes reported, and the limitations that constrain the interpretation and generalizability of existing findings. Furthermore, it highlights the extent to which current research has addressed essential dimensions of Arabic language learning, including speaking practice, individualized feedback, learner engagement, and the development of language competencies.

The significance of this study therefore extends beyond summarizing previous research: it provides an evidence-based account of the current state of knowledge, identifies specific empirical and methodological research gaps, and establishes priorities for future investigations. Ultimately, the findings are

expected to inform more context-sensitive and pedagogically appropriate approaches to AI integration in Arabic language education within *pesantren*, while ensuring that technological innovation remains aligned with the educational purposes and values of Islamic boarding schools.

Table 1. Research Questions

#	Research Questions	Motivation / purpose / benefits
RQ1	What are the trends, characteristics, and research designs of studies on AI use in Arabic language learning in pesantren?	To map the current state of research and its methodological patterns.
RQ2	What do these studies report about the effect of AI use on Arabic learning outcomes, including competence, motivation, and engagement, compared with non-AI methods?	To see what the existing evidence actually supports, rather than assume AI is effective.
RQ3	What types of AI tools are used, and how are they being implemented in practice?	To give educators a concrete picture of the tools already in use and how they were applied.
RQ4	What challenges are reported in integrating AI into Arabic learning in pesantren, including ethical and religious-content concerns?	To surface practical obstacles that future implementation needs to account for.

This section describes how the review was conducted, so that the process can be followed and, in principle, repeated.

Method

This section describes how the review was conducted so that the process can be followed and repeated. The study follows a systematic literature review (SLR) approach based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines. The aim was to identify, evaluate, and pull together prior research on AI use in Arabic language learning in pesantren, so as to get a clearer sense of what has already been studied, what the reported effects have been, and what is still missing(Cabrera & Cabrera, 2023).

Four sources were searched: Scopus, ScienceDirect, and SpringerLink and Google Scholar. The same combination of terms joined with AND/OR

was used in all sources: ("Arabic language learning" OR "Arabic language teaching" OR "Arabic as a foreign language") AND ("pesantren" OR "Islamic boarding school") AND ("artificial intelligence" OR "AI" OR "chatbot" OR "ChatGPT" OR "AI-based learning"). The search was limited to publications from 2020 to 2026, and was run on 14 September 2026, using titles, abstracts, and keywords where these search fields were available on each platform(Mangaroo-Pillay & Coetzee, 2022).

Eligibility followed a population, intervention, outcome, and design logic. Articles were included if they discussed AI or AI-based/chatbot tools in Arabic language learning, were available in full text, were published between 2020 and 2026, were empirical primary studies (qualitative, quantitative, mixed, or evaluations of AI output), and were set in, or clearly relevant to, a pesantren or comparable Islamic education context. Context was coded in three tiers and analyzed separately: Tier 1, studies conducted in a pesantren or in a school run by one; Tier 2, studies of pesantren-affiliated institutions, learners with pesantren backgrounds, or pesantren curricular content; and Tier 3, comparable Islamic education settings that are not pesantren. Articles were excluded if they dealt with digital media that did not involve AI, if they were reviews, conceptual papers, bibliometric studies, book chapters, or non-peer-reviewed pieces, or if they were not specifically about Arabic language learning.

Selection followed the standard PRISMA stages: identification, screening, eligibility, and inclusion. Records were first collected from the four sources. Of 274 records, 51 duplicates were removed and 223 titles and abstracts were screened; 173 were excluded. Fifty full texts were then assessed against the criteria above, and 38 were excluded, leaving 13 studies(Hinderks et al., 2020).

The 13 articles were analyzed descriptively using thematic grouping. Each study was coded for the AI tool it used, the language skill it focused on, its research method, what it reported as outcomes, what challenges it noted, and what limitations or gaps the authors themselves acknowledged. These codes were then compared across studies to identify recurring patterns and, just as importantly, areas where very little had been studied(Palomino et al., 2019)(Salmona & Kaczynski, 2024).

To keep the selection process consistent, only articles from Scopus and Sinta-indexed sources, and the same inclusion criteria were applied throughout. The same coding categories were used for every article, which helps keep the comparison across studies fair.

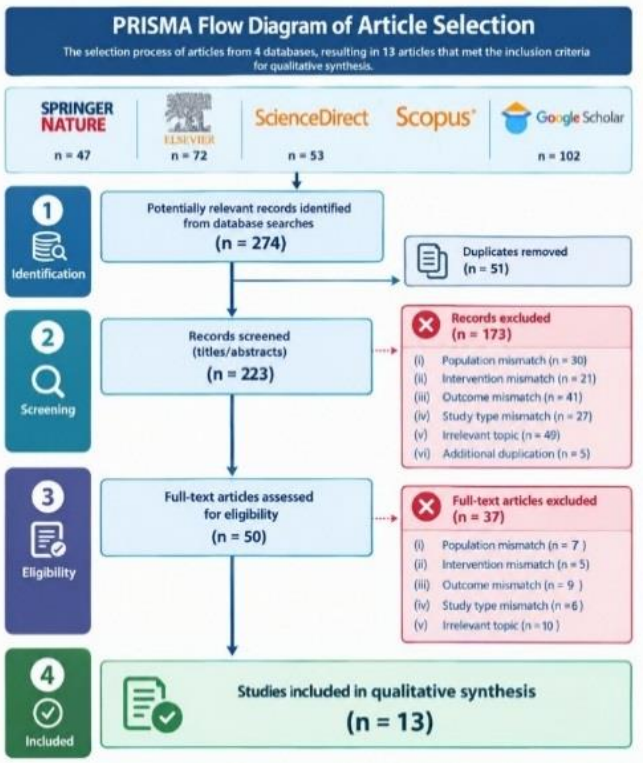

Figure 1. PRISMA Flow of Research Article Selection

The image shows the article selection process using the PRISMA flow which transparently explains how the number of studies was reduced from 274 to 20 final articles. At the identification stage, 274 articles were found from four databases (Springer Nature, Elsevier, ScienceDirect, Scopus, and Google Scholar). After removing duplication of 51 articles, 223 articles remained for the screening stage. At this stage, titles and abstracts were selected so that 173 articles were eliminated for various reasons such as population mismatch (n=30), intervention (n=21), outcome (n=41), type of study (n=27), irrelevant topics (n=49), and additional duplication (n=5). Furthermore, 50 articles passed the full-text assessment stage, but at the eligibility stage as many as 37 articles were again removed because they did not meet the same criteria (e.g. not relevant to the context of the pesantren or did not focus on learning media). Finally, only 13 articles met all the inclusion criteria and were used in the qualitative synthesis, so the decrease in the number of articles reflects a rigorous, systematic, and clearly criteria-based selection process according to PRISMA standards(Page et al., 2021)(Sarkis-Onofre et al., 2021).

All 13 articles come from Scopus- or SINTA-indexed journals. Each was checked against five basic quality points: whether the research aim was clearly stated, whether the method was sound, whether the sample was adequate

for the claims made, whether the findings were supported by the data presented, and whether the study was relevant to the pesantren context. Studies were scored on a simple 1 to 3 scale, and only those rated moderate or higher were kept in the final set.

Findings and Discussion

The selection process described above left 13 studies on AI use in Arabic language learning in pesantren, published between 2020 and 2026. Each was checked for topical relevance, methodological clarity, explicit use of an AI tool, and a reported link to learning outcomes or classroom practice. Their main characteristics are summarized in Table 2.

Table 2. Synthesis of Characteristics of 13 Selected Studies on the Implementation of Learning Media in Pesantren

Code	Year	Researcher	Title / Focus	AI Tool	Skill Focus	Method & Data	Reported Outcome
S1	2026	Haniah et al.	Improving Arabic Language Proficiency and Competence Using AI Technology	AI apps (TalkPal, Duolingo, etc.)	Istimā' + Kalām	PTK, 2 cycles, 11 students	Reported gain in proficiency and competence
S2	2026	Zahro et al.	The Use of ChatGPT in Enhancing Arabic Text Reading Interest	ChatGPT	Qirā'ah / Vocabulary	Qualitative descriptive; interviews & documentation	Reported increase in reading interest, vocabulary, motivation, confidence
S3	2026	Ma'arif, Salma & Afyuddin	Quality of ChatGPT's Arabic-Indonesian Translation	ChatGPT / GPT-4	Translation (classical text)	Evaluation of AI output quality	Mixed results on accuracy, acceptability,

			of Matan al-Ghāyah wa at-Taqrīb				readability, and semantic /theological precision
S4	2026	Zaimah & Hartanto	Empathy-Driven Conversation with ChatGPT	ChatGPT (conversational chatbot)	Kalām / Speaking	Observation & chatbot interaction	Reported gain in conversational output and emotional support
S5	2025	Julaeha & Nurcahyaningtias	Creativity of Islamic Boarding School Teachers in Combining Tradition and Technology	AI (general)	Teacher perspective	Qualitative descriptive; interview, observation, documentation	Described teacher creativity in combining tradition with AI use
S6	2025	Mukmin et al.	Implementing ChatGPT in Teaching Insyā' to Intermediate Students	ChatGPT	Kitābah / Writing	Qualitative; interview, observation, FGD	Reported perceived improvement in writing and feedback quality
S7	2024	Ilham et al.	Use of AI Technology in Improving Arabic Language Ability at PPM Rahmatul Asri	AI (general, unspecified)	Arabic proficiency (general)	Quasi-experiment, two-group pretest-posttest, 56 students	Mean score rose from 72.00 to 88.86
S8	20	Salam,	Educational	AI-	Gener	Quantitat	Reported

	24	Fitri & Zulan	Technology Based on AI in Supporting Distance Learning	based educational technology	al	ive survey, 43 students	positive perception of AI-supported learning
S9	2024	Moenir	Use of TalkPal to Improve Arabic Speaking Skills at MTs Darussalam Tangerang	TalkPal	Kalām / Speaking	Pre-experimental, one-group pretest-posttest + questionnaire	Reported gain in speaking fluency, vocabulary, confidence
S10	2024	Solehudin, Hanifansyah & Izzuddin	Enhancing Arabic Listening and Vocabulary Acquisition through AI-Powered Music (Suno AI)	Suno AI	Istimā' + Vocabulary	Qualitative, 3 weeks, 120 students	Reported gain in listening comprehension, vocabulary retention, motivation
S11	2024	Nurfaiza	Effect of ChatGPT Use in Indonesian-Arabic Translation Learning	ChatGPT	Translation	Qualitative, triangulation	Reported gains in translation speed and independent learning, with noted limitations
S12	2023	Masnun	Integrating AI in Arabic Writing Proficiency	AI writing tools	Kitābah / Writing	Qualitative case study	Reported changes in grammatical accuracy,

							coherence, lexical variety
S13	2020	Shiddiq	The Application of "Nemo" AI in Arabic Language Learning	Nemo AI	Arabic ability (general)	Qualitative descriptive	Described use and perceived usefulness of the tool

Grouped by the skill or focus each study addresses, the pattern looks like this:

Table 3. Studies Grouped by Skill / Focus

Skill / Focus	**Existing Studies**
Speaking / Kalām	Moenir; Zaimah & Hartanto
Writing / Kitābah / Insyā'	Masnun; Mukmin et al.
Reading / Qirā'ah	Zahro et al.
Listening / Istimā'	Solehudin et al.
Translation	Nurfaiza; Ma'arif et al.
Vocabulary	Solehudin et al.; Zahro et al.; TalkPal
Arabic proficiency (general)	Ilham et al.; Shiddiq; Haniah et al.
AI integration / teacher perspective	Julaeha & Nurcahyaningtias; Salam et al.
Curriculum readiness	Karsayuda et al.

Karsayuda et al., who studied curriculum readiness for AI integration, is included in this grouping for completeness even though the article falls outside the 13-study dataset analyzed in Table 2; it is mentioned here because it is one of the very few studies touching on institutional readiness, an area otherwise almost absent from the literature.

Looking at the 13 studies over time, the earliest one Shiddiq(Shiddiq, 2020) is a standalone case, and most of the activity clusters from 2024 onward, roughly matching the point at which ChatGPT and similar tools became widely accessible. The tools used fall into three loose categories. Some studies use AI in a fairly generic sense, without naming a specific model or application(MR et al., 2024)(Salam et al., 2024)(Haniah et al., 2026). Others use a tool built for one particular purpose, such as TalkPal for speaking practice, Suno AI for listening and vocabulary through music, or Nemo AI for general vocabulary work. The rest, a slightly larger group by 2025 and 2026, use ChatGPT or GPT-4 directly, applying it across reading, speaking, writing, and translation tasks. This last group is worth noting mainly because it shows researchers treating a general-

purpose chatbot as usable across several skills at once, rather than needing a dedicated app for each one.

As Table 3 shows, speaking and translation each have three or four associated studies, listening and vocabulary are mostly represented by a single study on Suno AI, and reading rests on one study using ChatGPT. Writing has only two entries Mukmin et al.(Mukmin et al., 2025); Masnun(Masnun, 2025), which is a smaller number than might be expected given how central kitābah is to pesantren instruction, including letter-writing, essays, and commentary on classical texts. Teacher perspective and curriculum readiness are addressed in two or three studies at most. None of this means AI is not used for writing or discussed by teachers in practice; it means these areas have not yet been studied much in the published literature that this review could locate.

Research Methods Used in These Studies

The 13 studies mostly use qualitative approaches: interviews, observation, documentation, focus-group discussion, or case study. This applies to eight of them (Zahro et al.; Zaimah & Hartanto; Julaeha & Nurcahyaningtias; Mukmin et al.; Solehudin, Hanifansyah & Izzuddin; Nurfaiza; Masnun; and Shiddiq). Three studies use an experimental or quasi-experimental design, including classroom action research and a one-group pretest-posttest model (Haniah et al.; Ilham et al.; Shiddiq). One uses a quantitative survey (Salam, Fitri & Zulan), and one evaluates the quality of AI-generated output directly rather than measuring student learning (Ma'arif, Salma & Afyuddin).

This mix says something about where the field is right now. Most studies describe what happened when a particular group of students used a particular tool, which is useful for understanding how AI is actually being used, but only Ilham et al. compare an AI-using group against a separate group, and even there the AI tool itself is not named. No study in this set puts AI-assisted instruction head-to-head with non-AI instruction under otherwise matched conditions. That gap lines up with what the Asalibuna review on AI in Arabic speaking instruction found more broadly: the literature has not yet produced strong comparative evidence that AI performs better than non-AI methods. It may well do so; the point is that this particular set of studies does not yet show it either way. Sample sizes are also small and tied to a single institution in every case, from 11 postgraduate students to 120 santri in the largest study, so how far these findings travel to other pesantren is an open question.

What the Studies Report about AI's Effect on Learning

It helps to separate what the 13 studies report into three rough categories: gains in competence, gains in motivation or confidence, and changes in productive output such as writing or translation.

Competence. Haniah et al.'s two-cycle action research reported improved listening and speaking after AI-assisted practice. Ilham et al.'s quasi-experiment reported the largest measured change in the dataset, with average scores moving from 72.00 to 88.86, though the specific AI tool used was not identified. Suno AI use was linked to better listening comprehension and vocabulary retention, and ChatGPT use to gains in reading-related vocabulary in Zahro et al.'s study.

Motivation and confidence. This is the outcome reported most consistently across the set. TalkPal use was associated with more confident speaking among santri. Zaimah and Hartanto found that a chatbot designed to respond empathetically seemed to give students a form of support they described positively, something not obviously available in a regular classroom setting. Suno AI use was linked to more engagement during listening practice, and Zahro et al. reported more confidence and more independent reading among ChatGPT users.

Writing and translation. Mukmin et al. and Masnun both reported some improvement in writing, covering grammatical accuracy, feedback, and vocabulary variety, though the extent varied by student level. Translation is where the picture is more mixed: Nurfaiza found that ChatGPT sped up translation practice and supported more independent work, but Ma'arif, Salma, and Afyuddin, who specifically checked translation quality against the classical text Matan al-Ghāyah wa at-Taqrīb, found the accuracy and semantic precision only partial. Taken together, these two studies point to a distinction worth keeping in mind for a pesantren setting: AI can make translation practice faster without necessarily making the output reliable enough to use on its own for religious texts

Challenges Reported in the Literature

Four kinds of challenges come up across the 13 studies, though not every study mentions all of them. Access and connectivity are uneven between institutions, which matters more for tools that require continuous interaction, such as chatbots. Teacher readiness to use generative AI in a planned, pedagogically grounded way is limited; the two studies that look at teachers directly (Julaeha & Nurcahyaningtias; Salam, Fitri & Zulan) suggest AI is often picked up informally rather than through an institutional strategy. Content accuracy is a concern specific to this setting: Ma'arif, Salma, and Afyuddin's findings suggest AI-generated translation of classical texts cannot simply be trusted at face value, which is a different kind of problem than what general CALL research usually deals with. And at the institutional level, curriculum policy for AI use appears to be mostly absent so far, since so few studies touch on it, meaning individual teachers and students are largely deciding on their own how, or whether, to use these tools.

Where the Research Still Falls Short

Putting the 13 primary studies together with the five earlier systematic reviews on AI and Arabic learning points to a few specific gaps.

Study design. Almost all 13 studies are short-term and based in a single institution. Longitudinal studies, replication across sites, and direct comparison between AI and non-AI instruction are all missing. This matches what the Asalibuna review found for the wider field: the comparative evidence simply is not there yet.

Skill coverage. Research so far leans toward speaking, listening/vocabulary, and translation. Writing gets less attention, and no study in this set looks at more than one or two skills together. The OLER review noted a similar unevenness in the broader, non-pesantren-specific literature.

A pesantren-specific synthesis. Hanun's review discussed AI and Arabic learning in pesantren conceptually, but no prior review, as far as we can tell, has assembled a dataset of primary studies from this setting the way this article does. That is the main thing this article adds to the existing reviews.

Religious-text accuracy. Only one study in the set, Ma'arif, Salma & Afyuddin, actually tested how accurate AI output is against a classical text. Given how central these texts are to pesantren education, this seems like an area that needs more attention than it has received so far.

Institutional readiness. Only two or three of the 13 studies look at teachers or curriculum planning. Whether pesantren as institutions are ready to adopt AI in a structured way, rather than through individual teachers experimenting on their own, is still largely unanswered.

Interpretation

A synthesis of the 13 studies reviewed here suggests that the use of artificial intelligence in Arabic language learning in pesantren has grown noticeably since 2024, moving from a single isolated case in 2020 toward a more varied set of tools by 2025 and 2026. Several studies report gains in motivation and engagement, particularly for listening and speaking, and one quasi-experimental study reported scores rising from 72.00 to 88.86. Even so, most of these studies remain limited in scope, whether in the skill they target, the length of implementation, or how deeply the AI tool is woven into actual teaching rather than used as a one-off activity.

These findings can be interpreted through Computer-Assisted Language Learning (CALL), particularly its development towards Intelligent Computer-Assisted Language Learning (ICALL), and constructivist learning theory. Compared with earlier CALL approaches, which primarily relied on structured exercises and predetermined feedback, generative AI tools such as ChatGPT

and TalkPal offer more flexible interaction, personalized feedback, and opportunities for communicative practice. This is consistent with previous research highlighting the potential of AI to facilitate language learning. However, the findings from the 13 studies reveal that these technological capabilities are not yet fully translated into adaptive and learner-centered pedagogical practices. Most studies still emphasize content generation and corrective feedback rather than sustained interaction and independent language development.

From a constructivist perspective, these findings also highlight the difference between interacting with AI and actively constructing knowledge through interaction. Although AI provides opportunities for dialogue, experimentation, and reflection, students are still frequently positioned as recipients of AI-generated information rather than active participants in the learning process. This suggests that meaningful AI-supported learning requires more than technological access; it requires instructional strategies that encourage students to evaluate feedback, revise their language production, and construct understanding independently. For future Arabic language education, particularly in *pesantren*, CALL and constructivist principles should therefore be integrated through teacher-guided AI activities that promote communicative competence, learner autonomy, and critical thinking. Future research should examine not only students' perceptions of AI but also its long-term effects on actual language proficiency and learning outcomes.

What this study tries to do differently is treat AI not simply as an add-on tool but as something that has to sit within the broader pesantren learning environment, including its oral tradition and its reliance on classical texts. Where earlier studies tend to look at one tool or one skill in isolation, bringing the 13 studies together here shows more clearly how AI use is currently scattered: strong in speaking and listening, thinner in writing, and, in the case of translation, showing a real difference between how fast AI can produce output and how accurate that output actually is when checked against a classical text. This review does not offer a finished model for how AI should be integrated. What it offers is a clearer picture of how fragmented current practice still is, and where a more deliberate, theory-informed approach would be needed.

A further pattern worth noting is that gains reported in receptive or motivational terms do not appear to carry over easily into productive skills such as writing or reliable translation. This matter more in a pesantren setting than it might elsewhere, since Arabic here is tied closely to oral recitation, religious transmission, and increasingly to institutional expectations around digital competence. Bridging that gap, between what students seem to enjoy using and what they can actually produce afterward, looks like one of the more concrete directions this line of research could take.

Implications

The implications of these findings go beyond the classroom. For teachers, this points to a need for training that covers not just how to operate an AI tool but how to design tasks around it so that students are producing language, not just consuming AI output. For pesantren administrators, the finding that AI-generated translation of a classical text was only partly accurate suggests some form of review process is worth putting in place before such content is used with students. For developers building AI tools for this context, it suggests that fluency in Arabic is not enough; getting the religious and textual details right matters just as much, and probably requires input from people who know the texts, not only people who know the technology.

Looking ahead, studies that follow students over a longer period, that test AI-assisted instruction directly against non-AI instruction, and that pair AI use with more communicative, real-interaction-based teaching would help answer questions the current 13 studies leave open. Combining generative AI with a more structured approach to speaking practice and text-based accuracy checking seems like a reasonable next step, one that would let Arabic learning in pesantren draw on new tools without losing sight of what has always made pesantren education distinctive.

Conclusion

This review looked at 13 primary studies, published between 2020 and 2026, on AI use in Arabic language learning in pesantren, following a PRISMA-based SLR process. The tools used range from dedicated applications such as TalkPal, Suno AI, and Nemo AI to general-purpose conversational AI such as ChatGPT and GPT-4, with most activity concentrated from 2024 onward. Speaking, listening, vocabulary, and translation are the most commonly studied areas; writing, reading, teacher perspective, and curriculum readiness are studied less often.

Rather than claim these findings settle the question of whether AI improves Arabic learning in pesantren, this review is better read as a stocktake of what has actually been examined so far, and what has not. Reading the 13 studies alongside five earlier systematic reviews points to a consistent pattern: most studies are short, single-site, and descriptive rather than comparative; skill coverage is uneven; and, specific to this setting, only one study has checked whether AI output can be trusted with classical religious texts. Addressing these gaps, through longer studies, direct comparisons with non-AI instruction, and more attention to religious-text accuracy and institutional readiness, would give a much clearer answer to the question this field is currently only beginning to ask.

Acknowledgment

The authors would like to express their sincere appreciation to the previous researchers whose published studies provided the empirical foundation for this systematic literature review. Their valuable findings and scholarly contributions have helped the authors develop a comprehensive understanding of the applications of artificial intelligence in Arabic language learning within pesantren contexts. The authors also acknowledge the academic contributions of scholars whose work has enriched the discussion and supported the development of this research.

Author Contributions Statement

A.F.R.N. contributed to the conceptualization and design of the study, literature search and selection, data extraction, and preparation of the manuscript. A.F.N.F. contributed to the literature search and selection, data extraction, data analysis, and interpretation of findings. M.I.H. contributed to the conceptualization and design of the study, data analysis, interpretation of findings, and manuscript revision. M.H.A. contributed to the data extraction, interpretation of findings, and critical revision of the manuscript. All authors reviewed and approved the final version of the manuscript.

AI Writing Statement

The authors employed Claude (Anthropic) to support English proofreading, grammatical correction, and improvements in academic writing style during the manuscript revision process. Its use was limited exclusively to language refinement and did not influence the research design, data collection, analytical process, or interpretation of the results, which remain the original work of the authors. All AI-assisted revisions and references were thoroughly checked and verified by the authors, who assume full responsibility for the accuracy, academic integrity, and final content of the publication.

References

Almelhes, S. (2026). A Systematic Review and Meta-Analysis on Artificial Intelligence and Its Utilization in Arabic Learning. In A. Gokhale (Ed.), *Proceedings of The 11th International Conference on Frontiers of Educational Technologies 2025* (pp. 219–237). Springer Nature Singapore.

Blake, J. (2023). Intelligent CALL: Individualizing Learning Using Natural Language Generation. In A. W. B. TSO, S. K. K. NG, L. LAW, & T. S. BAI (Eds.), *The Post-pandemic Landscape of Education and Beyond: Innovation and Transformation* (pp. 3–18). Springer Nature Singapore.

Cabrera, D., & Cabrera, L. L. (2023). The steps to doing a systems literature review (SLR). *Journal of Systems Thinking Preprints*.

Fadlelmula, F. K., & Qadhi, S. M. (2024). A systematic review of research on artificial intelligence in higher education: Practice, gaps, and future directions in the GCC. *Journal of University Teaching and Learning Practice*, *21*(6). https://doi.org/10.53761/pswgbw82

Hady, Y., Azkiyah, S. N., Muhbib, M., & Maswani, M. (2025). Artificial intelligence-based Arabic language learning: A systematic study of the development and challenges of pedagogical innovation. *Online Learning in Educational Research*, *6*(1), 49–63.

Hamid, M., Murtadho, M. A. C., Firdaus, A. Y., & Masturi, M. (2024). Language Environment and Acquisition Dynamics of Arabic in Pesantren: Perspectives on Islamic Education and Learning Tradition. *International Journal of Islamic Thought and Humanities*, *3*(2), 387–400.

Haniah, H., Ilyas, H., & Samae, H. (2026). Improving Arabic Language Proficiency and Competence of Arabic Language Education Using AI Technology. *Jurnal Al-Maqayis*, *13*(1), 1–22.

Hanun, A., & Wassalwa, A. (2025). Current Trends in Integrating Artificial Intelligence for Arabic Language Learning in Pesantren. *Arabia*, *17*(2), 293–306.

Hayad, Z., Pertiwi, R. W. L., Ayumagita, T., & Safirah, A. (2025). The Role, Effectiveness, and Challenges of Ai in Arabic Language Learning: A Systematic Literature Review. *PINBA*, *15*(1), 280–295.

Hinderks, A., Mayo, F. J. D., Thomaschewski, J., & Escalona, M. J. (2020). An SLR-tool: Search process in practice: A tool to conduct and manage systematic literature review (SLR). *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Companion Proceedings, ICSE '20*, 81–84. https://doi.org/10.1145/3377812.3382137

Islami, A., Ibrohim, B., Sudarmawan, H. M., Muttaqien, A., & Lidinillah, F. M. (2026). Technological Transformations in Arabic Language Learning: A Comprehensive Review of Emerging Tools and Techniques. *Arabiyatuna: Jurnal Bahasa Arab*, *10*(1), 227–244.

Julaeha, S., & Nurcahyaningtias, N. D. (2025). The Creativity of Islamic Boarding School Teachers in Combining Tradition and Technology: The Utilization of Artificial Intelligence (AI) as a Learning Innovation for Students. *Proceeding of Annual International Conference on Islamic Education (AICIED)*, *3*, 135–151.

Katinskaia, A. (2025). An overview of artificial intelligence in computer-assisted language learning. *ArXiv Preprint ArXiv:2505.02032*.

Kuddus, K. (2022). Artificial Intelligence in Language Learning: Practices and Prospects. In *Advanced Analytics and Deep Learning Models* (pp. 1–17). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119792437.ch1

Liangshi, Y. (2024). Constructivism Learning Theory. In Z. Kan (Ed.), *The ECPH Encyclopedia of Psychology* (pp. 311–313). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-7874-4_1077

Ma'arif, S., & Afyudiin, M. S. (2025). The Quality of ChatGPT's Arabic-Indonesian Translation of Matan al-Ghayah wa at-Taqrib: Implications for Arabic Language Learning in Pesantren: Kualitas Terjemahan Arab-Indonesia ChatGPT terhadap Matan al-Ghayah wa at-Taqrib: Implikasi bagi Pembelajaran B. *ATHLA: Journal of Arabic Teaching, Linguistic and Literature*, *6*(2), 81–97.

Manasiq, G. Z. A. S., Mansyur, M., & Al Ghifari, A. (2026). Artificial Intelligence in Arabic Speaking Instruction for Non-Native Learners: A PRISMA-Guided Systematic Review of Comparative Evidence Gaps. *Asalibuna*, *10*(01), 43–72.

Mangaroo-Pillay, M., & Coetzee, R. (2022). Lean frameworks: A systematic literature review (SLR) investigating methods and design elements. *Journal of Industrial Engineering and Management; Vol 15, No 2 (2022)*. https://doi.org/10.3926/jiem.3677

Masnun, M. (2025). Integrating artificial intelligence in Arabic writing proficiency: A case study in Islamic boarding schools. *ANCOLT: International Conference on Language Teaching*, *2*(01), 254–264.

Moenir, N. L. (2025). Pemanfaatan Media Pembelajaran Artificial Intelligence (AI) Aplikasi TalkPal Untuk Meningkatkan Keterampilan Berbicara Bahasa Arab Siswa MTs Darussalam Tangerang. *INTIFA: Journal of Education and Language*, *1*(4). https://doi.org/10.62083/intifa.v1i4.259

MR, I., Saepudin, S., Herdah, H., Darmawati, D., & Ramli, K. (2024). Penggunaan media teknologi artificial intelligence dalam meningkatkan kemampuan berbahasa Arab di PPM Rahmatul Asri. *Journal on Education*, *7*(1), 3922–3933.

Mukmin, T. A., Khoirun, F., Hakim, W., & Maksum, M. S. (2025). Implementing ChatGPT in Teaching Insya'to Intermediate-Level Students: Enhancing Arabic Composition Skills. *Journal of Educational Management Research*, *4*(2), 907–920.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., & Moher, D. (2021). Updating guidance for reporting systematic reviews: Development of

the PRISMA 2020 statement. *Journal of Clinical Epidemiology*, *134*, 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003

Palomino, M., Davila, A., & Melendez, K. (2019). *Methodologies, Methods, Techniques and Tools Used on SLR Elaboration: A Mapping Study BT - Trends and Applications in Software Engineering* (J. Mejia, M. Muñoz, Á. Rocha, A. Peña, & M. Pérez-Cisneros, Eds.; pp. 14–30). Springer International Publishing.

Sa'diyah, H., & Nisa', K. (2023). Integrasi Pendekatan Konstruktivistik dalam Pembelajaran Bahasa Arab di Perguruan Tinggi. *Alsina : Journal of Arabic Studies*, *5*(2), 173–196. https://doi.org/10.21580/alsina.5.2.18934

Salam, M. Y., Fitri, S., & Zulan, H. (2024). Educational technology based on artificial intelligence in supporting distance learning in Arabic language learning in higher education. *Journal International of Lingua and Technology*, *3*(1), 134–145.

Salmona, M., & Kaczynski, D. (2024). Qualitative data analysis strategies. *In How to Conduct Qualitative Research in Finance*, *Edward Elg*, 80–96.

Sarkis-Onofre, R., Catalá-López, F., Aromataris, E., & Lockwood, C. (2021). How to properly use the PRISMA Statement. *Systematic Reviews*, *10*(1), 117. https://doi.org/10.1186/s13643-021-01671-z

Shiddiq, A. B. (2020). The Application of «Nemo» Artificial Intelegence in Arabic Language Learning in the 4.0 Revolution Era At Al-Khalifah Islamic Boarding School Cibubur. *IJLECR (International Journal of Language Education and Cultural Review)*, *6*(1), 58–61.

Solehudin, M., Hanifansyah, N., & Izzuddin, I. F. (2024). Enhancing Arabic Listening and Vocabulary Acquisition through AI-Powered Music: A Study on Suno AI. *Arabiyat: Journal of Arabic Education & Arabic Studies*, *11*(2), 175–188.

Supriyanto, S., & Toifah, N. (2024). Efektivitas penggunaan artificial intelligence (AI) dalam pembelajaran bahasa Arab di era Society 5.0: Systematic literature review. *Lugawiyyat*, *6*(2), 42–54.

Yadav, S. (2025). Reimagining education with advanced technologies: Transformative pedagogical shifts driven by artificial intelligence. In *Impacts of Generative AI on the Future of Research and Education* (pp. 1–26). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-0884-4.ch001

Zahro, I., Syakroni, A., & Mahfudh, S. (2026). The Use of ChatGPT in Enhancing Arabic Text Reading Interest among Fifth-Semester Students of the Arabic Language Education Program at Mathali'ul Falah Islamic

Boarding School Institute. *Ihtimam: Jurnal Pendidikan Bahasa Arab*, *9*(01), 42–56.

Zaimah, N. R., & Hartanto, E. B. (2026). Empathy-Driven Conversation with Chat GPT: Insights from Higher-Ed Arabic Learners. *Asalibuna*, *10*(01), 110–130.