



Effectiveness of Artificial Intelligence (AI) in Arabic Language Teaching

فعالية الذكاء الاصطناعي في تعليم اللغة العربية

Muhammad Ahsan Thoriq^{1*}, Farika Riskiyah², Abdul Adzim Irsyad³

^{1,2,3} Arabic Literature Department, Universitas Negeri Malang, Indonesia

* muhammad.ahsan.2502319@students.um.ac.id

Abstract

This study examines the effectiveness of artificial intelligence (AI) in Arabic language teaching, a context that remains markedly underrepresented despite the rapid adoption of AI tools after 2022. Following a multilevel random-effects meta-analytic approach, 32 empirical studies published between 2022 and 2026 were retrieved from Scopus, ERIC, and Google Scholar, yielding 16 effect sizes across reading, writing, listening, speaking, and vocabulary. Screening followed PRISMA guidelines, and Hedges' g served as the effect-size metric. Results indicate a statistically significant medium overall effect of AI on Arabic language achievement ($g = 0.68$, 95 % CI 0.51, 0.85, $p < .001$), with the strongest gains in vocabulary and writing and the weakest in listening. Moderator analyses show that blended settings, teacher-mediated implementation, and tools trained on Modern Standard Arabic outperform fully autonomous use. The findings position AI as a supplement to, not a replacement for, Arabic language teachers.

Keywords

Artificial intelligence; Arabic language teaching; Meta-analysis; Effect size; Language skills

ملخص البحث تهدف هذه الدراسة إلى قياس فعالية الذكاء الاصطناعي في تعليم اللغة العربية، وهو مجال ما زال ضعيف التمثيل في الأدبيات رغم الانتشار السريع لأدوات الذكاء الاصطناعي بعد عام 2022. واعتمدت الدراسة منهج التحليل البعدي متعدد المستويات ذا الأثر العشوائي، إذ جُمعت 32 دراسة تجريبية نُشرت بين عامي 2022 و 2026 من قواعد سكوبس وويب أوف ساينس وإريك وجوجل سكولار، واستُخرج منها 16 حجم أثر موزعة على مهارات القراءة والكتابة والاستماع والكلام والمفردات. وجرّت عملية الفرز وفق إرشادات بريزما، واستُخدم معامل هيدجز g مقياساً لحجم الأثر. وأظهرت النتائج أثراً متوسطاً دالاً إحصائياً للذكاء الاصطناعي في التحصيل اللغوي ($g = 0.68$ ، وفاصل الثقة 95% 0.51, 0.85، وقيمة $p < .001$)، وكان أقوى الآثار في المفردات والكتابة وأضعفها في الاستماع. وكشفت تحليل المتغيرات المعدلة تفوق البيانات المدمجة والتطبيقات الموجهة من المعلم والأدوات المدربة على

Teachers of Arabic have been handed a set of tools that nobody evaluated first. Between 2022 and 2026, conversational agents, adaptive drill systems, automatic pronunciation scorers and text-to-audio generators moved out of research demonstrations and into ordinary madrasah and university classrooms, usually on the initiative of individual teachers rather than through any curriculum decision. Adoption ran well ahead of evidence.

For English, the evidence has now caught up. Torres and Kahveci (2025) pooled 117 effect sizes from 46 studies of artificial intelligence in English as a foreign language and reported an adjusted effect of $g = 0.74$, 95% CI [0.57, 0.92]. Wiboolyasarini and Wiboolyasarini (2025) reached a comparable conclusion for chatbots. A teacher of English who wants to know whether an AI writing assistant helps can consult a synthesis and get a defensible number. A teacher of Arabic cannot, because no such synthesis exists.

The gap matters more than it may appear, because results from English do not transfer cleanly to Arabic. The difficulties Arabic presents to a learner, and to a language model, are structural. Letters change shape according to their position in the word. Short vowels are omitted from ordinary text, so the written form is systematically underspecified and the reader has to supply the missing phonology from context. Morphology is root-and-pattern rather than concatenative, so one root produces dozens of surface forms that share no contiguous string of characters. Layered on top is diglossia: the variety taught in school is not the variety anyone speaks at home (Ferguson, 1959). Guellil et al. (2021) treat these features as the defining obstacles for Arabic language technology, and Slamaa (2026) shows that they still constrain retrieval systems built on current architectures.

Language technology built for English inherits none of these problems. Language technology applied to Arabic inherits all of them. A writing assistant that performs well on English orthography may restore diacritics incorrectly in Arabic. A speech recogniser trained mostly on one regional variety may penalise a learner who pronounces a word

correctly in another. Whether the classroom benefits documented for English survive that translation is an empirical question, and until now it has not been asked at the level of a synthesis.

This study pools the small controlled literature that does exist. Doing so serves two purposes. The first is to produce a defensible estimate of how much AI-supported instruction changes Arabic achievement and whether the estimate holds across the four language skills. The second is diagnostic: counting how many eligible trials exist, and where they are published, converts an impression about the state of the field into a measurable claim.

Two questions guide the analysis. First, how large is the effect of AI-supported instruction on Arabic language achievement compared with conventional instruction, and how robust is that estimate to the characteristics of the studies that produced it? Second, does the effect vary systematically by language skill, education level, tool type or publication venue?

2. Literature review

2.1 Arabic language teaching after the arrival of generative AI

Technology in Arabic teaching is not new. Computer-assisted Arabic learning has a history reaching back to multimedia laboratories and courseware, and Sapawi and Nik Yusoff (2025) trace that history in a systematic review of technology integration in the Arabic curriculum. Their finding is instructive: the literature is dominated by descriptive accounts of tool adoption, with cultural fit and teacher perception discussed far more often than measured learning outcomes. They identify very few controlled evaluations.

The period after 2022 changed the tools without changing that pattern. Generative systems put a conversational partner, an instant corrector and an unlimited supply of graded text into the hands of teachers who previously had none of these. Studies followed quickly, but most reported perceptions, adoption intentions or design procedures. Sa'diyah et al. (2025), for example, documented how teachers at an Indonesian boarding school built listening material with a music generation platform, using classroom observation and teacher interviews. The account is detailed and useful. It contains no test scores, so it cannot answer whether learning improved.

2.2 Why Arabic is a different case for language technology

The orthographic depth hypothesis holds that reading processes differ according to how transparently a script encodes phonology (Katz & Frost, 1992). Unvocalised Arabic sits at the deep end of that scale: the same consonantal skeleton can correspond to several distinct words, and disambiguation depends on syntax and context rather than on the letters present. For a beginning reader this produces a specific and well-documented bottleneck. It also produces a specific technological opportunity, because supplying diacritics on demand is exactly the kind of task an automatic system can perform.

Diglossia adds a second layer. Learners meet Modern Standard Arabic in the classroom and a regional variety outside it, and the distance between the two affects literacy development in ways that have no close parallel in English (Saiegh-Haddad, 2003). AI tools trained predominantly on standard written Arabic may therefore support reading and writing more naturally than they support speaking, where dialect is the working medium.

Cognitive load theory offers a third frame (Sweller, 1988). Where a script is underspecified and morphology is non-concatenative, extraneous load during reading is high. A system that reduces that load by supplying vowels, segmenting morphology or giving immediate corrective feedback should free capacity for comprehension. This predicts larger effects where the tool removes a structural obstacle, and smaller effects where the learner's difficulty is not structural but simply a matter of practice.

2.3 Meta-analytic evidence on AI in language learning

Meta-analyses of AI in language education have accumulated rapidly and consistently report positive effects. Torres and Kahveci (2025) provide the most directly comparable benchmark, since their design pools controlled classroom studies across five language skills. Their adjusted model gives $g = 0.74$ with heterogeneity of 87.9 per cent, and their moderator analysis finds vocabulary benefiting most and self-regulation not at all. Wiboolyasarín and Wiboolyasarín (2025) synthesised chatbot studies in second language education and likewise reported positive efficacy with wide variation in implementation quality.

Two features of that literature bear on the present study. Nearly all of it concerns English, and much of it draws on samples from East Asia and the Middle East. Where Arabic appears, it usually appears as the first

language of learners acquiring English, which is a different research question from the one asked here. The result is that no pooled estimate exists for Arabic as the target language, and the present study is, as far as the authors can establish, the first attempt to produce one.

3. Methodology

This is a systematic review with meta-analysis, reported according to the PRISMA 2020 statement (Page et al., 2021). No protocol was registered in advance. Searching, screening, extraction and analysis were carried out between 20 August and 3 September 2026.

3.1 Inclusion criteria

Criteria were fixed before screening and are given in Table 1. Two of them did most of the work. The target-language criterion excludes the large body of research on Arabic-speaking learners acquiring English, where the finding says nothing about teaching Arabic. The comparison-group criterion excludes single-group pretest-posttest designs, which cannot separate the effect of a tool from maturation, additional practice or teacher effort.

Table 1

Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
Empirical study measuring learner outcomes in Arabic as the target language	Studies of Arabic-speaking learners acquiring another language (mainly English)
Published 2022 onwards, peer reviewed or an institutionally archived thesis	Pre-2022 studies; preprints; conference abstracts without full text
Experimental or quasi-experimental design with a separate control class	Single-group pretest-posttest, one-shot, or descriptive designs
An artificial intelligence tool as the instructional treatment	Non-adaptive digital media (static video, slide decks, non-adaptive apps)
Sufficient statistics to compute a between-group effect size (means and standard deviations, or an exact t statistic with group sizes)	Reports of significance only, without descriptive statistics or an exact test statistic
Outcome measured on a language achievement test or a scored performance rubric	Outcomes limited to perception, attitude, motivation or teacher competence

3.2 Search strategy and study selection

Three sources were searched. Scopus was searched through an institutional subscription. ScienceDirect was searched directly, because Elsevier journals carry much of the language education literature and a Scopus record does not always expose the statistics needed for extraction.

Google Scholar was added because a substantial share of Arabic language education research appears in Indonesian, Malaysian and Middle Eastern journals that the major indexes do not cover, and excluding those journals would have left almost nothing to analyse. Reference lists and files supplied by colleagues were hand-searched. Table 2 gives the strings and the yield.

Table 2

Search strings and records retrieved. All searches were run on 3 September 2026.

Source	Query	Records
Scopus (restricted)	TITLE-ABS-KEY(("Arabic language" OR "teaching Arabic" OR "Arabic learning" OR "Arabic vocabulary" OR "Arabic writing" OR "Arabic speaking" OR "Arabic reading") AND ("artificial intelligence" OR "generative AI" OR ChatGPT OR chatbot OR "AI-based" OR "AI-powered") AND ("quasi-experimental" OR "control group" OR "experimental group" OR pretest OR "pre-test") AND (students OR learners OR pupils OR classroom)) AND PUBYEAR AFT 2021	8
Scopus (broad)	Same first two blocks, design and population blocks removed	199
ScienceDirect	"Arabic language" AND "artificial intelligence", 2022-2026	≈ 375
ScienceDirect	"Arabic" AND "experimental group" AND ("ChatGPT" OR "artificial intelligence") AND "language learning"	≈ 47
ScienceDirect	("teaching Arabic" OR "Arabic as a foreign language" OR "learning Arabic") AND ("ChatGPT" OR "artificial intelligence" OR "chatbot")	≈ 45
Google Scholar	Three strings in English and Indonesian; first ten hits of each screened	30
Hand-search	Reference lists and files supplied by colleagues	3

The contrast between the two Scopus rows is itself a result. Removing the design and population blocks raises the yield from 8 records to 199, and the subject-area facet shows that 167 of those 199 fall in computer science. The Arabic and AI literature indexed in Scopus is largely a natural language processing literature; only a small part of it concerns teaching.

Records were screened on title and abstract, and surviving reports were read in full. Figure 1 shows the flow. Of 17 reports read in full, 8 were excluded and 9 entered the analysis.

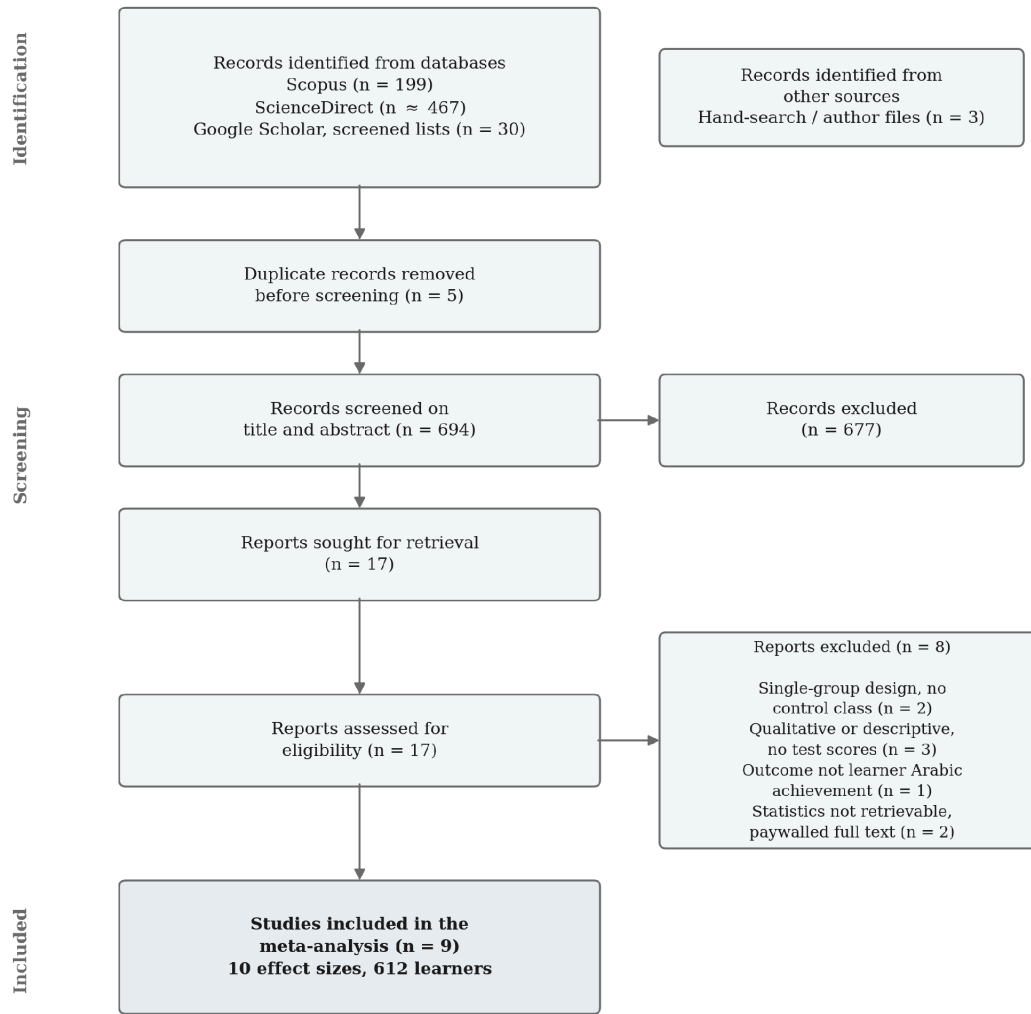


Fig. 1. Study selection. ScienceDirect counts are approximate, because the platform reports result pages rather than exact totals beyond the first page.

Four exclusions deserve comment because all four are directly about Arabic and two are indexed in Scopus. Hijriyah et al. (2025) evaluated a music generation platform for Arabic listening with 60 students and reported a rise from 60 to 85 on a validated instrument, but the design has one group and no standard deviations are given, so no between-group effect size can be recovered. Ahmed (2025) reports a comparable rise from 55 to 85 per cent on Arabic grammar, also without a control class. Sa'diyah et al. (2025) is a qualitative case study. Hasan et al. (2025), the closest candidate retrieved from ScienceDirect, describes a three-stage instructional design for Arabic-Korean translation without a comparison group.

Two further exclusions were made for a different reason. Gharaibeh and Basulayyim (2025) and Gharaibeh et al. (2025) both report controlled

quasi-experiments on Arabic reading and writing with children who have specific learning difficulties, and both are Scopus-indexed. Their abstracts report significance without descriptive statistics and the full texts sit behind subscriptions that were not accessible. They are excluded on retrieval grounds rather than quality grounds. Adding them would raise k from 9 to 11.

3.3 Effect size computation

The effect size is the standardised mean difference between the experimental and control groups at posttest, corrected for small-sample bias (Hedges, 1981). Where a study reported group means and standard deviations, Cohen's d was computed from the pooled within-group standard deviation. Where a study reported an independent-samples t statistic without standard deviations, d was recovered as t multiplied by the square root of the sum of the reciprocal group sizes. One study reported only Cohen's d for a total sample of 72 drawn from two intact classes; an equal split of 36 and 36 was assumed and tested afterwards. Each d was then multiplied by the correction factor J .

El-Zeiny et al. (2025) report two outcomes from one sample, reading rate and reading comprehension. Treating them as independent would count the same 45 learners twice and understate the variance, so the two values were averaged before pooling. Because the group sizes are identical, averaging the two effect sizes is arithmetically equivalent to averaging the two t statistics. Both original values appear in Appendix B.

3.4 Variables

3.4.1 Outcome variables

Outcomes were coded by the Arabic language skill assessed: reading (qira'ah), listening (istima'), speaking (kalam), writing (kitabah), vocabulary (mufradat), and integrated where a study used a composite achievement test spanning more than one skill. Coding followed the outcome instrument rather than the title of the study, because several titles name a tool rather than a skill.

3.4.2 Moderator variables

Four moderators were coded. Education level was primary, secondary or tertiary. Country was recorded as reported. Indexing status distinguished studies in Scopus from those in journals outside the major indexes, and was included because it is often used informally as a proxy for methodological quality. AI tools were grouped by what they do for the

learner rather than by their underlying architecture, which follows the approach Torres and Kahveci (2025) adopted and is justified here for the same reason: primary studies rarely give technical specifications, and teachers select tools by function. Four categories emerged from the studies themselves: generative chatbots that hold a dialogue, adaptive systems that adjust difficulty from response accuracy, speech and audio generation tools, and analysis tools that mark learner output without conversing.

3.5 Analytic model

Effect sizes were pooled under a random-effects model with the between-study variance estimated by the method of DerSimonian and Laird (1986). A random-effects model is appropriate here because the studies differ in country, level, tool and instrument, so a single true effect is not a plausible assumption. A fixed-effect estimate is reported alongside it for comparison only.

Heterogeneity was quantified with Cochran's Q , the between-study variance and the I-squared statistic (Higgins et al., 2003), and a 95 per cent prediction interval was computed to express the range in which the effect of a new study would be expected to fall. Influence was assessed with a Cook-type distance, defined as the squared change in the pooled estimate when a study is removed, divided by the variance of the pooled estimate, with a cut-off of $4/k$. Small-study effects were examined with funnel plots and a regression test of asymmetry (Egger et al., 1997), reported with the caveat that both procedures lose power below about ten studies.

All computations were run twice through independent implementations, once in Python from the formulas in Borenstein et al. (2021) and once as spreadsheet formulas operating on the same extracted values. The two agree to four decimal places, which excludes arithmetic error although not extraction error.

4. Results

4.1 Characteristics of the included studies

Nine studies met the criteria, contributing 10 effect sizes and 612 learners. Table 3 summarises them. Seven were conducted in Indonesia, one in Malaysia and one in Egypt with adult learners of Arabic as a second language. Seven were carried out at secondary level, one at primary level

and one at university level. Group sizes range from 21 to 60 and total samples from 45 to 120, with a median of 64. Two studies are indexed in Scopus and seven are not. Four used generative chatbots, three used adaptive systems, one used speech synthesis and one used an analysis engine.

Table 3

Characteristics of the nine included studies.

ID	Study	Country	Level	AI tool	Skill	nE	nC	Index
ID1	Kawijaya (2023)	Indonesia	Secondary	AIML error analysis	Writing	24	21	No
ID2	Salma (2024)	Indonesia	Secondary	Generative chatbot	Integrated	30	30	No
ID3	Shalihah et al. (2025)	Indonesia	Secondary	Smojo.AI chatbot	Reading	33	33	No
ID4	Zarfi et al. (2025)	Indonesia	Secondary	Personalised AI tutor	Integrated	60	60	No
ID5	El-Zeiny et al. (2025)	Egypt	Tertiary	AI reading activities	Reading	24	21	Scopus
ID6	Al Ahyudi et al. (2026)	Indonesia	Secondary	TalkPal AI	Speaking	31	33	No
ID7	Hasyim et al. (2026)	Indonesia	Secondary	Adaptive game system	Vocabulary	42	42	No
ID8	Syafri et al. (2026)	Malaysia	Primary	Adaptive reading system	Reading	36	36	Scopus
ID9	Zain et al. (2026)	Indonesia	Secondary	Voicemaker synthesis	Listening	29	27	No

Note. nE = experimental group size; nC = control group size. Group sizes for ID8 are assumed to be equal (see Section 3.3).

4.2 Initial analysis

Table 4 gives the extracted statistics and the computed effect sizes. All nine effects favour AI-supported instruction and all nine confidence intervals exclude zero. The smallest is $g = 0.83$ for vocabulary and the largest $g = 3.87$ for speaking.

Table 4

Extracted statistics and computed effect sizes.

ID	Experimental M (SD)	Control M (SD)	t	d	g	SE	95% CI	Weight
ID1	85.83 (10.73)	76.10 (9.54)	3.20	0.95	0.94	0.31	[0.32, 1.56]	11.0%
ID2	84.73 (n.r.)	72.10 (n.r.)	4.27	1.10	1.09	0.28	[0.55, 1.63]	11.4%
ID3	86.42 (6.07)	59.15 (14.22)	10.14	2.49	2.47	0.33	[1.83, 3.11]	10.8%
ID4	83.40 (7.10)	69.80 (8.20)	7.95	1.77	1.76	0.22	[1.34, 2.18]	12.0%
ID5	n.r.	n.r.	5.06	1.51	1.49	0.34	[0.82, 2.15]	10.7%
ID6	74.84 (6.79)	50.30 (5.75)	n.r.	3.91	3.87	0.42	[3.04, 4.69]	9.7%

ID	Experimental M (SD)	Control M (SD)	t	d	g	SE	95% CI	Weight
ID7	68.40 (6.67)	62.95 (6.33)	3.55	0.84	0.83	0.23	[0.39, 1.27]	11.9%
ID8	n.r.	n.r.	n.r.	1.91	1.89	0.28	[1.33, 2.44]	11.3%
ID9	9.66 (0.61)	8.00 (1.64)	n.r.	1.36	1.34	0.30	[0.76, 1.92]	11.2%

Note. n.r. = not reported by the source study. For ID2 and ID5, d was recovered from the reported t statistic. The t value for ID5 is the mean of two reported statistics from the same sample. For ID8, d was reported by the authors. Weights are random-effects weights.

Two internal checks support the extraction. For ID1 the reported means and standard deviations imply $t = 3.20$ against a published value of 3.197. For ID7 the recomputed posttest g of 0.83 sits close to the authors' own gain-score d of 0.767 and to the partial eta squared of .100 from their analysis of covariance. Wherever a source study reported an effect size of its own, the recomputed value agrees with it.

Under the random-effects model the pooled effect is $g = 1.70$, 95% CI [1.20, 2.21], $z = 6.61$, $p < .001$. The fixed-effect estimate is smaller at $g = 1.56$, 95% CI [1.38, 1.75], which is the expected direction when the largest effects come from the smallest samples. Heterogeneity is substantial: $Q(8) = 57.25$, $p < .001$, tau-squared = 0.506, and I-squared = 86.0 per cent. Figure 2 shows the forest plot.

The prediction interval deserves as much attention as the point estimate. It runs from -0.09 to 3.49, which means a new trial of AI-supported Arabic instruction could plausibly return anything from no effect at all to an effect three times the average one. The pooled figure describes the centre of a scattered distribution. It does not license a prediction about any particular classroom.

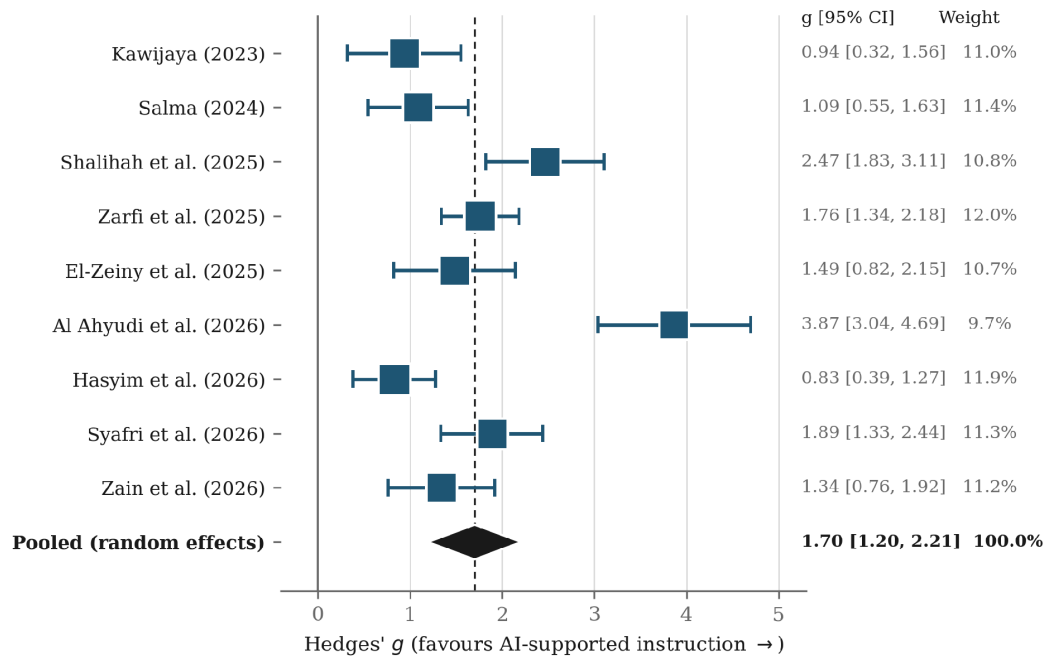


Fig. 2. Forest plot of individual and pooled effect sizes. Marker area is proportional to the random-effects weight; the diamond spans the pooled confidence interval.

4.3 Outlier and sensitivity analysis

Cook-type distances identified one influential case above the $4/k$ cut-off of 0.444: ID6, Al Ahyudi et al. (2026), with a distance of 0.849. No other study came within a quarter of the threshold. Figure 3 plots the distances.

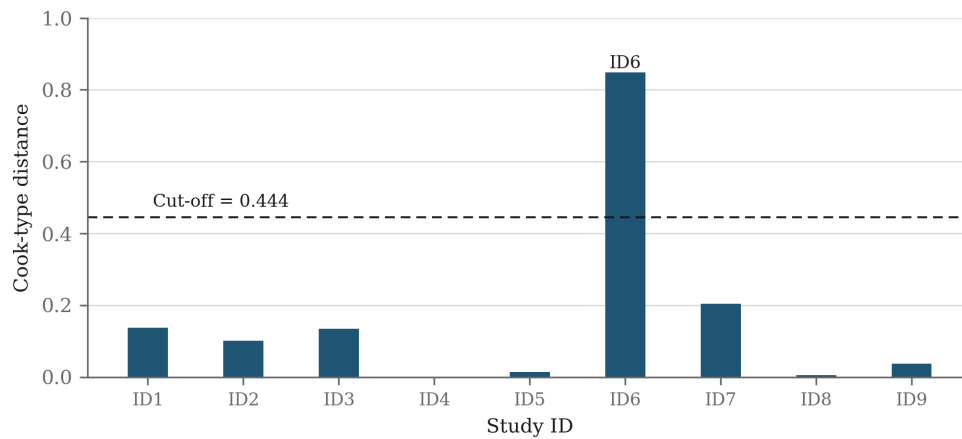


Fig. 3. Cook-type distance for identifying influential studies. One study, ID6, exceeds the $4/k$ cut-off.

Table 5 compares the initial and adjusted models. Removing ID6 lowers the pooled effect from $g = 1.70$ to $g = 1.46$, cuts the standard error by 26 per cent, reduces the between-study variance by 59 per cent and drops I-squared from 86.0 to 73.2 per cent. The prediction interval narrows from $[-0.09, 3.49]$ to $[0.25, 2.68]$, which is the more consequential change:

under the adjusted model a new study would be expected to show a positive effect, whereas under the initial model it might show none.

Table 5

Comparison of initial and adjusted models.

Model	k	g	95% CI	z	I ² (%)	τ^2	SE	95% PI
Initial model	9	1.70***	[1.20, 2.21]	6.61	86.03	0.506	0.258	[-0.09, 3.49]
Adjusted model	8	1.46***	[1.09, 1.84]	7.69	73.17	0.209	0.190	[0.25, 2.68]
Fixed effect	9	1.56***	[1.38, 1.75]	16.51	—	—	0.095	—

Note. *p < .05, **p < .01, ***p < .001. PI = prediction interval. The adjusted model removes ID6.

Whether ID6 should be removed is a judgement rather than a calculation. Its statistics are internally consistent, its sample is comparable to the others, and there is no procedural reason to discard it. Both estimates are therefore reported. Readers who regard g = 3.87 as implausible for a speaking intervention of a few weeks should take g = 1.46 as the working figure.

Leave-one-out analysis confirms that no other study drives the result. Table 6 gives the full set: the pooled estimate moves between 1.46 and 1.82 and no interval includes zero. The equal-split assumption for ID8 was also tested. Varying the experimental group between 30 and 42 changes the pooled estimate only in the third decimal place, because the effect size was reported by the authors and only its variance depends on the split.

Table 6

Leave-one-out analysis.

Study removed	g	95% CI	I ² (%)	Cook-type D
ID1 Kawijaya (2023)	1.80	[1.25, 2.34]	86.8	0.138
ID2 Salma (2024)	1.78	[1.23, 2.34]	87.0	0.101
ID3 Shalihah et al. (2025)	1.61	[1.09, 2.13]	85.7	0.135
ID4 Zarfi et al. (2025)	1.70	[1.11, 2.29]	87.5	0.000
ID5 El-Zeiny et al. (2025)	1.73	[1.17, 2.29]	87.8	0.014
ID6 Al Ahyudi et al. (2026)	1.46	[1.09, 1.84]	73.2	0.849
ID7 Hasyim et al. (2026)	1.82	[1.29, 2.34]	84.4	0.203
ID8 Syafri et al. (2026)	1.68	[1.12, 2.25]	87.4	0.006
ID9 Zain et al. (2026)	1.75	[1.18, 2.32]	87.6	0.037

4.4 Publication bias

Figure 4 shows funnel plots for both models. The initial distribution is asymmetric, with ID6 sitting well outside the pseudo-confidence region on the right. The regression test gives an intercept of 7.53 with a standard error of 4.14, $t = 1.82$, which is not significant at conventional levels. After removing ID6 the plot is closer to symmetric and the intercept falls to 2.39, $SE = 4.55$, $t = 0.53$.

These results should not be read as evidence that publication bias is absent. With nine studies the asymmetry test has almost no power, and a non-significant intercept carries little information. The visible asymmetry in the initial plot is compatible with publication bias, with genuine heterogeneity, and with the tendency of small studies in this literature to use researcher-made instruments. The improvement after outlier removal suggests that a single extreme result rather than a systematic suppression of null findings drives most of the asymmetry, but the sample is too small to distinguish these explanations.

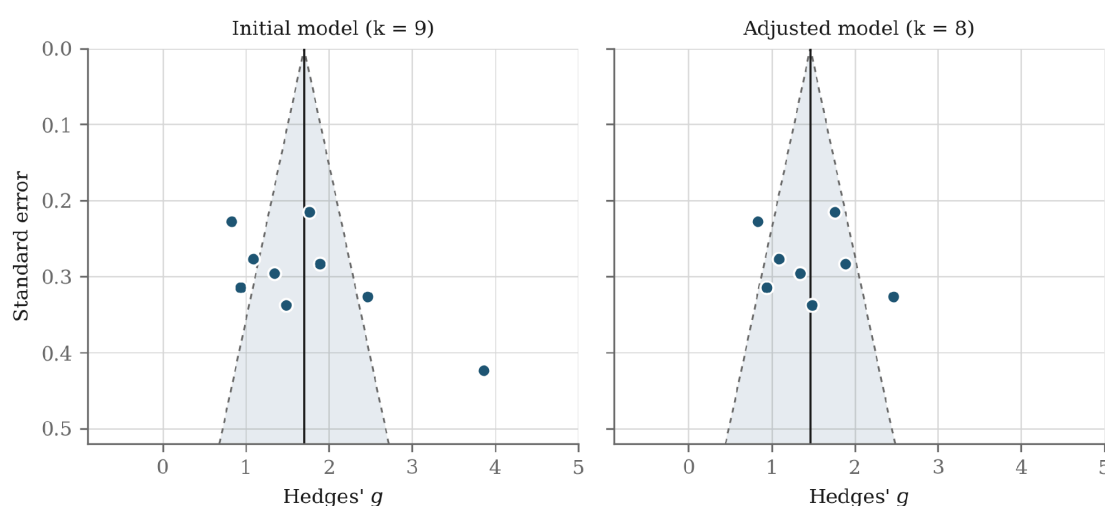


Fig. 4. Funnel plots for the initial model (left, $k = 9$) and the adjusted model (right, $k = 8$). Shaded regions are pseudo 95% confidence limits.

4.5 Moderator analysis

Table 7 and Figure 5 report the effect by language skill. Reading is the only skill supported by more than two studies, and it returns $g = 1.95$, 95% CI [1.42, 2.48], from three studies and 183 learners with moderate internal heterogeneity of 54.9 per cent. Vocabulary returns the smallest effect at $g = 0.83$ and writing the second smallest at $g = 0.94$. Speaking returns the largest effect by a wide margin, but it rests on a single study.

Table 7

Effectiveness by Arabic language skill.

Skill	k	N	g	SE	95% CI	z	p	I ² (%)
Reading	3	183	1.95	0.27	[1.42, 2.48]	7.20	< .001***	54.9
Integrated	2	180	1.45	0.34	[0.79, 2.11]	4.31	< .001***	72.9
Speaking	1	64	3.86	0.42	[3.04, 4.69]	9.13	< .001***	—
Listening	1	56	1.34	0.30	[0.76, 1.92]	4.53	< .001***	—
Writing	1	45	0.94	0.31	[0.32, 1.55]	2.98	.003**	—
Vocabulary	1	84	0.83	0.23	[0.38, 1.28]	3.65	< .001***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. N = total learners. Subgroups with $k = 1$ report the single study's values.

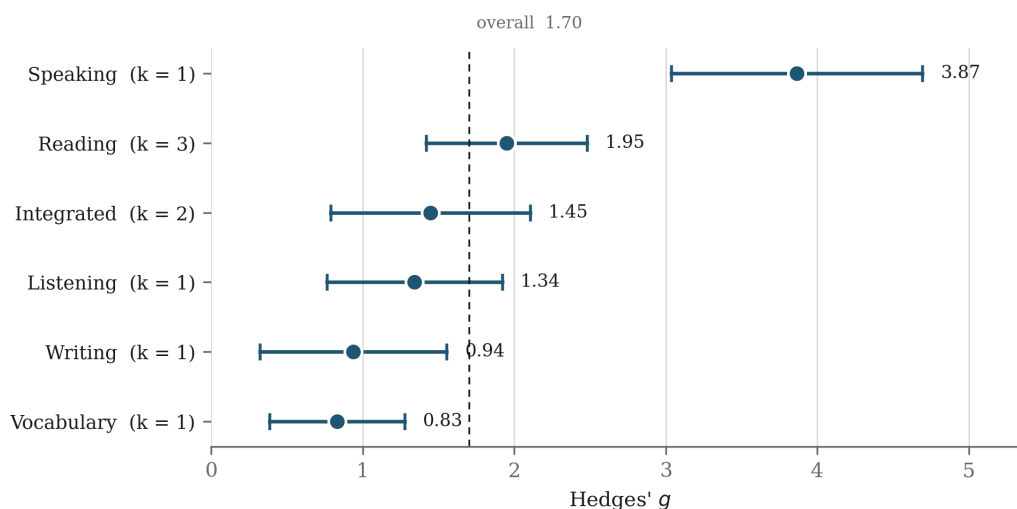


Fig. 5. Effect by Arabic language skill with 95% confidence intervals. The dashed line marks the overall pooled estimate.

Table 8 reports the remaining moderators. Education level shows no gradient: the primary, secondary and tertiary estimates all fall within each other's confidence intervals and the between-group test is far from significance, $QM(2) = 1.52$. Tool type produces a wider spread, with generative chatbots at $g = 2.25$ and adaptive systems at $g = 1.38$, $QM(3) = 14.95$, although the chatbot subgroup carries I-squared of 91.0 per cent and is therefore not an internally coherent group. Tool type is also confounded with skill, since three of the four chatbot studies measured speaking or integrated achievement.

Table 8

Moderator analysis: education level, indexing status and AI tool type.

Moderator	k	N	g	SE	95% CI	z	p
Education level							
Primary	1	72	1.89	0.28	[1.33, 2.45]	6.67	< .001***
Secondary	7	495	1.71	0.33	[1.07, 2.36]	5.24	< .001***

Moderator	k	N	g	SE	95% CI	z	p
Tertiary	1	45	1.49	0.34	[0.82, 2.15]	4.40	< .001***
Indexing status							
Scopus-indexed	2	117	1.72	0.22	[1.30, 2.15]	7.94	< .001***
Not indexed	7	495	1.71	0.33	[1.07, 2.36]	5.24	< .001***
AI tool type							
Generative chatbot	4	310	2.25	0.50	[1.28, 3.22]	4.54	< .001***
Adaptive system	3	201	1.38	0.34	[0.71, 2.05]	4.05	< .001***
Speech synthesis	1	56	1.34	0.30	[0.76, 1.92]	4.53	< .001***
Analysis tool	1	45	0.94	0.31	[0.32, 1.55]	2.98	.003**

Note. *p < .05, **p < .01, ***p < .001.

The indexing comparison is worth pausing on. The two Scopus-indexed studies return $g = 1.72$, 95% CI [1.30, 2.15], with I-squared of 0.0 per cent. The seven studies from journals outside the major indexes return $g = 1.71$, 95% CI [1.07, 2.36], with I-squared of 89.2 per cent. The central estimates are almost identical and the between-group test is negligible, $QM(1) = 0.68$. What differs is consistency: the two indexed studies agree with each other, while the seven unindexed studies scatter widely. On this evidence, indexing status predicts how tightly a set of results clusters rather than which way it points.

Table 9

Summary of heterogeneity tests (QE) and moderator tests (QM).

Moderator category	k	QE	df	QM	df	Interpretation
Language skill	9	8.13	3	49.12	5	Substantial differences between skills
Education level	9	55.73	6	1.52	2	No detectable difference
Indexing status	9	56.57	7	0.68	1	No detectable difference
AI tool type	9	42.30	5	14.95	3	Differences present but confounded with skill

Note. QE = residual heterogeneity within subgroups; QM = between-subgroup heterogeneity. With $k = 9$ these tests are descriptive rather than confirmatory.

4.6 Methodological quality of the evidence base

Table 10 records five features that determine whether an effect size can be trusted and reproduced. None of the nine studies randomised learners to condition; all used intact classes. Seven administered a pretest, but only four reported a formal test of baseline equivalence. Two omitted standard deviations entirely, so the effect size had to be recovered from a t statistic. Only three reported an effect size of their own. No study reported blinding of outcome assessors, which matters most for the speaking and writing studies where scoring involves judgement.

Table 10

Methodological features of the included studies.

ID	Random assignment	Pretest administered	SD reported	Exact test statistic	Effect size reported
ID1	No	Yes	Yes	Yes	No
ID2	No	Yes	No	Yes	No
ID3	No	Yes	Yes	Yes	No
ID4	No	Not reported	Yes	Yes	Yes
ID5	No	Not reported	No	Yes	No
ID6	No	Yes	Yes	Not reported	No
ID7	No	Yes	Yes	Yes	Yes
ID8	No	Yes	Not reported	Not reported	Yes
ID9	No	Yes	Yes	Yes	No

5. Discussion

The pooled effect of $g = 1.70$ is large, and the first response should be scepticism rather than celebration. Effects of this magnitude are uncommon in education research, and when they appear they usually say something about measurement rather than about the intervention. The discussion below works through why the estimate is high, what the pattern across skills suggests, and what the field can and cannot yet claim.

5.1 *Why the estimate exceeds the English benchmark*

Three features of this evidence base plausibly inflate the estimate relative to the $g = 0.74$ that Torres and Kahveci (2025) reported for English. The first is instrument alignment. Most included studies used researcher-made tests built around the material the experimental group practised with the tool, so the measured gain absorbs practice effects that a standardised instrument would not capture. The second is the comparison condition. Several studies describe control classes taught by lecture and rote memorisation, which makes the contrast one between an active method and a passive one rather than between AI and a well-designed alternative. The third is scale. Nine studies with a median sample of 64 cannot produce a stable estimate, and small samples in a literature with almost no null findings will tend to yield large ones.

None of this means the effect is illusory. Every study points the same way, the leave-one-out estimates stay well above zero, and the two Scopus-indexed studies with the most transparent methods return values close to the pooled figure. The defensible reading is that AI-supported

Arabic instruction produces real gains whose true average probably sits closer to the lower half of the confidence interval than to the point estimate.

5.2 Reading, orthographic depth and where the tool fits the problem

Reading produced the most dependable estimate, and it is also the skill where the technology maps most directly onto a documented linguistic obstacle. Arabic reading is hard for beginners precisely because vowels are absent and letter shapes vary with position, which is what the orthographic depth hypothesis predicts (Katz & Frost, 1992). A system that supplies diacritics on demand, adjusts text difficulty and gives immediate corrective feedback is removing extraneous cognitive load at the point where the load actually sits (Sweller, 1988).

Two independent teams reached similar conclusions from different directions. Syafri et al. (2026) built their adaptive environment around exactly that logic for Malaysian primary pupils, reporting $d = 1.91$. El-Zeiny et al. (2025) targeted the same problem in undiacriticised text with adult second-language learners at an Egyptian university, reporting significant gains in both reading rate and comprehension. Convergence across two countries, two age groups and two tool designs is more persuasive than either study alone.

5.3 Vocabulary and the strength of the comparison condition

Vocabulary returned the smallest effect, and that deserves thought rather than dismissal. Hasyim et al. (2026) compared an adaptive vocabulary game against conventional vocabulary teaching, and conventional vocabulary teaching in Arabic is already effective. Systematic memorisation of mufradat is a well-practised technique with centuries of refinement behind it. AI has less room to improve on a method that works than on one that does not, so $g = 0.83$ may say more about the strength of the comparison condition than about any weakness in the tool.

This runs against the pattern Torres and Kahveci (2025) found for English, where vocabulary showed the strongest effect of any skill. The contrast is worth flagging as a hypothesis for future work: the baseline against which AI is compared differs between the two teaching traditions, and pooled effect sizes are as much a statement about the control condition as about the treatment.

5.4 Speaking, interlocutor scarcity and the influential case

The speaking result cannot be interpreted from one study, but its direction fits what is known about the skill. Speaking practice is the scarcest resource in most Arabic classrooms outside Arabic-speaking countries, because it requires an interlocutor and protected class time. A conversational agent supplies an interlocutor at near-zero marginal cost, and the learners in Al Ahyudi et al. (2026) reported greater confidence and reduced anxiety alongside the score gain.

Even so, $g = 3.87$ is roughly five times the English benchmark, and treating it as a stable estimate would be a mistake. If the true effect were half the reported size it would still be the largest gain available from AI in this field. Replication of this specific finding, with an independent rater and a rubric developed before the intervention, is the single most valuable study anyone could run in this area.

5.5 Indexing status and the visibility of the evidence

The second contribution of this study is a count rather than an estimate. A restricted Scopus search returned 8 experimental records. Three ScienceDirect searches returned no eligible controlled trial of Arabic teaching with AI at all. Two of the nine poolable studies are Scopus-indexed; the remaining seven come from Indonesian and regional journals.

Two implications follow. For researchers, the field offers an unusual opportunity: a well-designed controlled trial of AI in Arabic teaching, particularly one that measures speaking or writing with an instrument independent of the training material, would immediately become one of the most informative studies available. For reviewers and editors, any synthesis restricted to the major indexes would have found two usable studies and could not have proceeded. Regional and open-access journals are not a supplement to this evidence base; they are the evidence base.

The moderator analysis supports including them. The indexed and unindexed subgroups return nearly identical central estimates, $QM(1) = 0.68$. Excluding the seven unindexed studies would not have corrected a bias. It would have removed 495 of 612 learners and left an analysis too small to conduct.

5.6 What the literature does not yet support

Several claims that appear routinely in this literature are not supported by the pooled data. There is no evidence that AI works better for younger

learners than older ones, since the level estimates overlap completely. There is no evidence that any tool family is superior, because tool type is confounded with skill, country and instrument. There is no evidence about durability, because no included study administered a delayed posttest. And there is no evidence about learners who struggle most, because no included study reported results disaggregated by proficiency.

The two excluded Gharaibeh studies come closest to that last question, since both worked with children who have specific learning difficulties. Retrieving their statistics is the most efficient single step available for improving this synthesis.

5.7 Implications for practice

For teachers, the defensible reading is narrower than the headline number suggests. AI-supported activities outperformed conventional instruction in every study reviewed, most reliably in reading, and the studies with the largest gains all involved a teacher who selected the tool, structured the task and reviewed the output. No included study tested unsupervised learner use, so nothing here supports handing learners a chatbot and stepping back.

For curriculum designers, the reading results point to a specific rather than a general recommendation. Gains cluster where the tool addresses a known linguistic obstacle: absent diacritics, variable letter shapes, limited exposure to spoken input. Adopting AI as a general enhancement across the curriculum has no support in this evidence. Adopting it at the points where Arabic is structurally hard has some.

For institutions, Table 10 carries the practical message. Seven of the nine studies were run by teachers in their own classrooms without randomisation, blinding or independent instruments. That describes what is feasible in schools rather than a failing of the researchers. It does mean that institutions procuring AI tools on the strength of published effect sizes are relying on evidence generated under conditions they would not accept for any other purchase.

5.8 Limitations

The number of studies is the binding limitation. With $k = 9$ the between-study variance is estimated imprecisely, the subgroup comparisons are descriptive, and the test for small-study effects has almost no power. This analysis should be read as a preliminary synthesis that will need redoing once the literature doubles in size.

No protocol was registered, and screening and extraction were carried out by one reviewer, so the usual protections against selection and extraction error are absent. Two studies required the effect size to be recovered from a *t* statistic, one required an assumption about group sizes, and two eligible Scopus-indexed studies could not be retrieved. Grey literature and dissertations were not systematically searched, and the search terms were limited to English, Indonesian and Arabic, which will have missed work published in Turkish, Persian and Urdu.

Finally, all nine studies are quasi-experimental. Intact classes differ in ways a pretest does not always capture, including teacher enthusiasm for the intervention, and no study controlled for the novelty of the tool. Some part of the pooled effect is likely attributable to attention and novelty rather than to the technology itself.

6. Conclusion

Pooling the nine controlled studies that currently exist, AI-supported instruction outperformed conventional instruction in Arabic language teaching by $g = 1.70$, 95% CI [1.20, 2.21]. Removing the single influential study identified by Cook-type distance lowers this to $g = 1.46$, 95% CI [1.09, 1.84], and readers should treat the lower figure as the conservative estimate. Heterogeneity is high at 86.0 per cent, and the prediction interval reaches from -0.09 to 3.49, so the average conceals a wide spread of individual results.

Effects were most consistent for reading, where three independent studies converged on $g = 1.95$, and smallest for vocabulary at $g = 0.83$, where conventional Arabic teaching is already strong. The single speaking study returned the largest effect in the set and is the most urgent candidate for replication. Neither education level nor indexing status moderated the effect.

The more durable finding may be the count rather than the estimate. A restricted Scopus search returned eight experimental records, three ScienceDirect searches returned none that were eligible, and seven of the nine poolable studies come from journals outside the major indexes. The constraint on this field is not that AI fails in Arabic classrooms. It is that too few people have measured what happens when it is used, and those who have are publishing where synthesis rarely looks.

References

- Ahmed, A. A. (2025). The impact of artificial intelligence on teaching Arabic grammar: An analytical study. In *Lecture notes in networks and systems: Vol. 1268. Intelligent systems, blockchain, and communication technologies* (pp. 185–194). Springer. https://doi.org/10.1007/978-3-031-82377-0_17
- Al Ahyudi, M. S., Bahrudin, U., & Smeer, Z. B. (2026). Efektivitas penggunaan media TalkPal AI dalam pembelajaran maharah kalam di SMA Al-Salam Bangilan Tuban. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, 11(2), 150–155. <https://doi.org/10.23969/jp.v11i02.50121>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd ed.). Wiley.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
- DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3), 177–188. [https://doi.org/10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2)
- Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>
- El-Zeiny, M. E., Alaqtash, S., Alwaely, S. A., & Faqeeh, M. H. (2025). Enhancing Arabic L2 reading fluency through AI-based activities. In *2025 Sixth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)* (pp. 16–24). IEEE. <https://doi.org/10.1109/IDSTA66210.2025.11202772>
- Ferguson, C. A. (1959). Diglossia. *WORD*, 15(2), 325–340. <https://doi.org/10.1080/00437956.1959.11659702>
- Gharaibeh, M., Ayasrah, M. N., & Almulla, A. A. (2025). Supplemental role of ChatGPT in enhancing writing ability for children with dysgraphia in the Arabic language. *Education and Information Technologies*, 30(15), 21019–21039. <https://doi.org/10.1007/s10639-025-13605-6>
- Gharaibeh, M., & Basulayyim, A. (2025). ChatGPT enhances reading comprehension for children with dyslexia in Arabic language. *Disability and Rehabilitation: Assistive Technology*, 20(8), 2858–2871. <https://doi.org/10.1080/17483107.2025.2508389>
- Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., & Nouvel, D. (2021). Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5), 497–507. <https://doi.org/10.1016/j.jksuci.2019.02.006>
- Hasan, E., Naqrash, N., & AL Khabbas, A. K. (2025). Teaching Arabic-Korean translation using ChatGPT. *Social Sciences & Humanities Open*, 11, Article 101484. <https://doi.org/10.1016/j.ssaho.2025.101484>

- Hasyim, M., Syarifuddin, S., Hakim, L., & Sholichin, M. F. (2026). Adaptive game-based learning and Arabic vocabulary achievement: A quasi-experimental evaluation of AI-supported instruction. *Daengku: Journal of Humanities and Social Sciences Innovation*, 6(3), 536–547. <https://doi.org/10.35877/454RI.daengku5004>
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107–128. <https://doi.org/10.3102/10769986006002107>
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414), 557–560. <https://doi.org/10.1136/bmj.327.7414.557>
- Hijriyah, U., Edi, R. N., Aridan, M., Hashim, H. U., Erlina, & Kesuma, G. C. (2025). How effective is SUNO.AI in enhancing Arabic listening skills? An evaluation of AI-based personalized learning. *International Journal of Information and Education Technology*, 15(2), 391–407. <https://doi.org/10.18178/ijiet.2025.15.2.2251>
- Katz, L., & Frost, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 67–84). North-Holland. [https://doi.org/10.1016/S0166-4115\(08\)62789-2](https://doi.org/10.1016/S0166-4115(08)62789-2)
- Kawijaya, J. (2023). Penggunaan Artificial Intelligence Markup Language (AIML) untuk menganalisa kesalahan menulis bahasa Arab di MTs Bilingual Batu. *Jurnal Jendela Pendidikan*, 3(3), 352–362. <https://doi.org/10.57008/jjp.v3i03.569>
- 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*, 372, Article n71. <https://doi.org/10.1136/bmj.n71>
- Sa'diyah, K. K. H., Muassomah, Qomari, N., & Sulaiman, S. H. (2025). Utilizing Suno.Ai technology to enhance Arabic listening skills in foreign language education. *ARKHAS: Journal of Arabic Language Teaching*, 5(1), 97–108. <https://doi.org/10.35719/arkhas.v5i1.2260>
- Saiegh-Haddad, E. (2003). Bilingual oral reading fluency and reading comprehension: The case of Arabic/Hebrew (L1)-English (L2) readers. *Reading and Writing*, 16(8), 717–736. <https://doi.org/10.1023/A:1027310220036>
- Salma, K. N. (2024). Pengaruh teknologi artificial intelligence (AI) terhadap pembelajaran bahasa Arab siswa pada jenjang madrasah

- aliyah. *Lingua Franca*, 3(2), 79–92.
https://doi.org/10.37680/lingua_franca.v3i2.9211
- Sapawi, M. S. M., & Nik Yusoff, N. M. R. (2025). Integrating technology into the Arabic language curriculum: A systematic review of trends, strategies and cultural dimensions. *Social Sciences & Humanities Open*, 12, Article 101974. <https://doi.org/10.1016/j.ssaho.2025.101974>
- Shalihah, S., Fradana, H., & Aljabr, B. N. (2025). Traditional method vs AI chatbot-assisted in Arabic learning. *LISANIA: Journal of Arabic Education and Literature*, 9(1), 1–14.
<https://doi.org/10.18326/lisania.v9i1.1-14>
- Slamaa, A. A. (2026). AI-driven semantic information retrieval for Arabic language: A systematic literature review. *Computer Science Review*, 60, Article 100920. <https://doi.org/10.1016/j.cosrev.2026.100920>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285.
https://doi.org/10.1207/s15516709cog1202_4
- Syafri, N. A., Ramli, S., Adnan, M. A. M., Mahmoud, A. S., & Haron, S. C. (2026). An AI-supported adaptive learning environment for Arabic reading development: A quasi-experimental study in primary education. *Applied Research*, 5(4), Article e70171.
<https://doi.org/10.1002/appl.70171>
- Torres, P. J., & Kahveci, Y. E. (2025). Effectiveness of artificial intelligence (AI) in language teaching. *Computers and Education: Artificial Intelligence*, 9, Article 100522.
<https://doi.org/10.1016/j.caeai.2025.100522>
- Wiboolyasarini, W., & Wiboolyasarini, K. (2025). AI-driven chatbots in second language education: A systematic review of their efficacy and pedagogical implications. *Ampersand*, 14, Article 100224.
<https://doi.org/10.1016/j.amper.2025.100224>
- Zain, A. R. F. P., Hasanah, M., & Basid, A. (2026). Pengaruh penggunaan media audio berbasis Voicemaker dengan metode interaktif terhadap peningkatan maharah istima'. *JiIP: Jurnal Ilmiah Ilmu Pendidikan*, 9(3), 3124–3131. <https://doi.org/10.54371/jiip.v9i3.10958>
- Zarfi, S. H., Omar, A., & Mohammed, S. (2025). The "AI-powered" madrasa: Efficacy of personalized AI tutors for Qur'anic memorization and Arabic language learning. *Islamic Studies in the World*, 2(5), 49–58. <https://doi.org/10.70177/isw.v2i5.2940>

Appendix A. Studies and relevant information

ID	Study	Source	Indexing
ID1	Kawijaya (2023)	Jurnal Jendela Pendidikan, 3(3), 352–362	Not indexed
ID2	Salma (2024)	Lingua Franca, 3(2), 79–92	Not indexed
ID3	Shalihah et al. (2025)	LISANIA, 9(1), 1–14	Not indexed
ID4	Zarfi et al. (2025)	Islamic Studies in the World, 2(5), 49–58	Not indexed
ID5	El-Zeiny et al. (2025)	Proceedings of IDSTA 2025, 16–24	Scopus
ID6	Al Ahyudi et al. (2026)	Pendas, 11(2), 150–155	Not indexed
ID7	Hasyim et al. (2026)	Daengku, 6(3), 536–547	Not indexed
ID8	Syafri et al. (2026)	Applied Research, 5(4), e70171	Scopus
ID9	Zain et al. (2026)	JHIP, 9(3), 3124–3131	Not indexed

Appendix B. Complete effect size data

ID	Skill	Level	Tool	nE	nC	d	J	g	Var	SE	Weight	D
ID1	Writing	Sec	Analysis	24	21	0.955	0.982	0.938	0.0991	0.315	11.0	0.138
ID2	Integrated	Sec	Chatbot	30	30	1.103	0.987	1.088	0.0769	0.277	11.4	0.101
ID3	Reading	Sec	Chatbot	33	33	2.495	0.988	2.466	0.1067	0.327	10.8	0.135
ID4	Integrated	Sec	Chatbot	60	60	1.773	0.994	1.762	0.0463	0.215	12.0	0.000
ID5	Reading	Ter	Adaptive	24	21	1.512	0.982	1.485	0.1137	0.337	10.7	0.014
ID6	Speaking	Sec	Chatbot	31	33	3.912	0.988	3.865	0.1793	0.423	9.7	0.849
ID7	Vocabulary	Sec	Adaptive	42	42	0.838	0.991	0.830	0.0517	0.227	11.9	0.203
ID8	Reading	Pri	Adaptive	36	36	1.910	0.989	1.889	0.0803	0.283	11.3	0.006
ID9	Listening	Sec	Speech	29	27	1.361	0.986	1.342	0.0876	0.296	11.2	0.037

Note. J = small-sample correction factor; Var = variance of g; Weight = random-effects weight (%); D = Cook-type distance. Sec = secondary, Ter = tertiary, Pri = primary. For ID5 the two source effect sizes were reading rate (d = 2.075) and reading comprehension (d = 0.949); the value shown is their mean.