Diniyyah Jurnal
https://ojs.stit-diniyyahputeri.ac.id/index.php/dj

Epistemic Authority and Religious Knowledge Verification in AI-Assisted Islamic Education: A Systematic Literature Review

Fazri Zuzano [1], Darwianis [1], Muhammad Sayuti [1], Ade Sri Madona [1], Asraf Kurnia [2]
[1] Universitas Bung Hatta, Indonesia
[2] Universitas Islam Negeri Imam Bonjol Padang, Indonesia
fazrizuzano@bunghatta.ac.id*

Article Information:
Received October 9, 2026
Revised November 12, 2026
Accepted December 19, 2026

Keywords: *Generative artificial intelligence, epistemic authority, Islamic religious education, religious knowledge verification, systematic literature review.*

Abstract

Generative artificial intelligence (GenAI) increasingly mediates Islamic learning, yet fluent responses do not guarantee authoritative or accurate religious knowledge. This review examines how epistemic authority is negotiated, how AI-generated Islamic claims are verified, and what educational and technical outcomes have been assessed. A qualitative systematic synthesis was conducted using 42 unique records from a researcher-exported Scopus dataset (20 October 2026), with PRISMA 2020-informed selection and design-sensitive appraisal. After title–abstract screening, 32 reports were sought; 10 full texts could not be retrieved through available open-access routes. Of 22 reports assessed, two were excluded, leaving 20 studies for synthesis. The evidence comprised educational interventions, stakeholder and conceptual analyses, textual audits, and technical evaluations. The HARMONY model reported significant short-term within-group gains in hadith learning competence and self-reported tolerance; a course-specific hadith chatbot showed favorable descriptive outcomes but did not establish a causal effect. Technical systems improved performance on defined retrieval and source-grounding tasks, whereas interpretive studies documented errors in translation, citation, and contextual meaning. Across studies, religious-knowledge validation depended on traceable canonical sources and qualified human judgment rather than linguistic fluency alone. A six-stage verification pathway is proposed as an evidence-informed synthesis, not a tested intervention. Findings remain limited by incomplete search documentation, access-related exclusions, and heterogeneous study designs.

INTRODUCTION

Generative artificial intelligence has become a routine means of producing explanations, study questions, translated passages and instructional material. In mainstream educational settings, this capability has been associated with opportunities for practice, formative feedback and differentiated learning, but also with erroneous claims, overreliance, unequal access and pressure on academic integrity (Kasneci et al., 2023; Miao & Holmes, 2023; Ji et al., 2023).

How to cite: Zuzano, F., Darwianis, D., Sayuti, M., Madona, A, S., Kurnia, A. (2026). Epistemic Authority and Religious Knowledge Verification in AI-Assisted Islamic Education: A Systematic Literature Review. *Diniyyah Jurnal*, 13(2), 42-57.

E-ISSN: 2810-0050
Published by: The Institute for Research and Community Service

In Islamic religious education (IRE), the consequences of inaccurate content are not confined to factual mistakes. AI outputs may appear to quote scripture, classify prophetic reports, explain juristic rules or interpret doctrinal distinctions that historically require accountable scholarly reasoning. The appearance of fluency can therefore create an unwarranted impression of religious expertise.

An Islamic claim is warranted through more than grammatical plausibility. Its status may depend on a traceable Qur'anic verse, a reliably attributed hadith, the assessment of isnād and matn, authoritative scholarly interpretation, or the context of a particular school of jurisprudence. Digital assistance can reduce the cost of accessing such materials but does not independently confer the standing associated with trained scholars, teachers or accountable institutions. Dart (2026) conceptualizes the distinction as one between instrumental epistemic trust and moral or religious authority. Related analyses in Islamic education and pesantren scholarship situate this tension within the continuing negotiation of scholarly transmission, institutional legitimacy, educational values and digitally mediated knowledge (Mustopa et al., 2026; Qulyubi et al., 2026; Anim et al., 2026).

The reliability problem itself contains several different tasks. A model can invent a source citation, reproduce the wrong verse, assign a fabricated hadith, select a superficially related passage, simplify the wording while altering meaning, or provide a doctrinal answer without specifying interpretive conditions. These errors should not be collapsed into a generic measure of "AI accuracy." Research on retrieval-augmented generation (RAG) provides a technical rationale for external evidence grounding (Lewis et al., 2020), while general language-model risk scholarship explains why confident text generation does not assure truthfulness or social accountability (Bender et al., 2021; Ji et al., 2023).

A developing technical literature addresses particular religious tasks through source-constrained retrieval, multi-agent query routing, expert-verified databases and benchmarks. For example, EMAN evaluates embedding-linked access to authenticated hadith (El Ganadi et al., 2025); HQDR/HQPS validates deterministic Qur'anic pattern search (Bakr et al., 2026); Fanar-Sadiq benchmarks task-specific Islamic question answering (Abbas et al., 2026); and QAMAR compares generated Qur'anic morphological analysis with checked annotations (Faqihi et al., 2026). Such studies are important, yet a technical improvement in a benchmark must not be described as a demonstrated improvement in students' religious understanding or critical reasoning.

Educational research supplies a different but much smaller body of evidence. Sumarna et al. (2026) introduced the HARMONY hadith-learning approach, combining verified textual tools with supervised AI-supported reflection. Nouri (2026) examined a course-constrained chatbot for undergraduate hadith sciences. Teacher focus groups (Almajdoub et al., 2026) and students' narratives (Darmu'in et al., 2025) additionally illuminate the curricular and relational implications of AI-assisted learning. Together, these works show why a review needs to distinguish classroom outcomes, expert judgment, contextual interpretation and technical evaluation rather than treating them as interchangeable evidence of effectiveness.

Adjacent Islamic education publications stress teacher mediation, ethical literacy, verification of dalil, and responsible curriculum design (Amin, 2026; Faizul et al., 2026; Ibrohim et al., 2026; Isti'ana, 2026; Muis et al., 2026; Tanjung et al., 2026). Further work examines learner acceptance, institutional preparedness and the distributed responsibility for checking AI content (Faizin et al., 2025; Purnomo et al., 2026; Shobri & Alfurqan, 2026; Hamidalloh et al., 2026). These supporting sources are relevant to framing the problem, but they do not automatically become studies within this review's pre-defined Scopus sample.

The current review is differentiated from prior discussions and syntheses by

its attempt to connect four analytical layers: (i) religious epistemic authority, (ii) distinct verification objects and procedures, (iii) evaluated educational and technical outcomes with corresponding limits of inference, and (iv) pedagogical safeguards that can be formulated without presenting a proposed pathway as an already validated model. Related existing SLRs and integrative analyses, including Mat Yusoff et al. (2026), Arif et al. (2026) and Suryanto et al. (2025), are treated as contextual literature rather than re-counted as primary studies. Other work on algorithmic daʿwah, adaptive teaching materials and ethical educational guidance further helps contextualize implementation (Marwantika & Dauda, 2025; El Yunusi et al., 2026; Suryanto et al., 2025).

This review addresses four questions: RQ1, how is epistemic authority attributed or negotiated in AI-assisted Islamic learning? RQ2, which religious-text and source-verification objects, procedures and actors are documented? RQ3, what educational outcomes, technical performance, epistemic failures and limitations are reported? RQ4, what evidence-informed educational practices follow from these findings, and what would be required to evaluate them?.

METHODS

A systematic review framework was used to map research concerning epistemic authority and AI-generated Islamic knowledge. Reporting was guided by PRISMA 2020 and PRISMA-S, with the recognition that these are reporting guidelines rather than proof of methodological rigor or independent certification of a systematic review (Page et al., 2021; Page, Moher, et al., 2021; Rethlefsen et al., 2021). The analysis was qualitative and design-stratified because outcomes ranged from university course grades to language-model accuracy, textual interpretation, interview accounts and philosophical arguments. No meta-analysis was conducted because those measures do not represent a common effect estimand (Campbell et al., 2020; Thomas & Harden, 2008; Snyder, 2019). The primary data source was a Scopus bibliographic CSV exported on 20 October 2026 by the researcher, containing 42 unique EIDs and titles, abstracts, document types and available DOIs. The collection comprised 20 journal articles, 19 conference papers, one journal review, one conference review and one book chapter; three records were dated 2023, seven 2024, twelve 2025 and twenty 2026. The exact executed Scopus Boolean string, field specifications and original interface hit count are not preserved in the working material. Therefore, the present report transparently describes an auditable exported-record set, not a independently reproducible comprehensive search across all databases. No Web of Science, ERIC or Dimensions search is claimed.

The review question and screening logic concentrated on the intersection between artificial intelligence, Islamic learning or canonical Islamic texts, religious epistemic justification, source provenance, interpretation and/or evidence of learning. Verification checks for the primary synthesis were confined to the Scopus export. Scholarly sources added later to contextualize methods and ethical implications were referenced in the introduction and discussion but were not inserted into the 42-record Scopus identification count.

Table 1. Operational inclusion and exclusion criteria and their application.

Criterion	Included in primary synthesis	Excluded or reported separately
Publication / availability	Full-length journal article or research conference paper retrievable through open-access or authorized repository routes	Book chapters, secondary reviews or reports not retrievable in full; inaccessibility is not a quality judgment

Direct topical relevance	Islamic religious learning, religious knowledge authority, or AI-generated Islamic text/answer verification	General Arabic NLP, generic AI platforms or broad commentary without review-relevant verification objects
Primary contribution	Empirical educational findings, evaluated AI outputs, textual verification studies, or original epistemic conceptual analysis	Dataset harvesting/construction only, without evaluation of AI outputs or student learning
Quality handling	Method-specific appraisal; retain limitations and distinguish conceptual from empirical claims	Do not treat absence of randomized evidence as reason to exclude relevant descriptive/qualitative studies

The restriction to full-length accessible reports was an explicit practical condition used for this evidence synthesis. Ten reports were not retrieved in full through available open-access routes after tracing candidate DOI/publisher links and repository alternatives; they are labelled "reports not retrieved," not all conclusively "closed access." This restriction introduces potential availability bias, particularly toward technical conference publications and commercial outlets. Two full-text reports Namoun et al. (2024) and Almasoud et al. (2026) were not included as primary studies because they develop source datasets without directly evaluating AI-generated output correctness or classroom outcomes. They remain usable as contextual examples of verification infrastructure (Engkizar et al., 2024; Asril et al., 2026; Saeed et al., 2025; Langputeh et al., 2026). An AI-assisted preliminary title–abstract screen of all 42 EIDs yielded 32 full-text candidates and 10 title–abstract exclusions. After duplicate-free PDF matching, 22 unique full texts were available for closer reading; one extra upload of Dart (2026) was a duplicate PDF rather than a separate study. Twenty reports met the operational scope, and two were excluded after full-text assessment. Individual records and reasons are retained in the companion screening matrix. The final 20-study inclusion set reflects a structured researcher-assisted workflow; dual independent screening, adjudicated inter-rater agreement and prospective protocol registration have not been documented and must not be inferred. Extraction recorded the bibliographic identity and year, sample or evaluation-unit count, design, religious verification object, reference standard, findings, reported numerical outcomes where appropriate, potential bias, and page-level evidence. Critical appraisal was informed by the logic of the Mixed Methods Appraisal Tool for qualitative and mixed studies and by task-appropriate scrutiny of baseline comparisons, benchmarks, expert adjudication and measurement fidelity in technical reports (Hong et al., 2018). It was not a complete independently scored MMAT exercise, and no pseudo-precise quality score was assigned across incomparable designs. Conceptual work was evaluated for argument relevance and transparency rather than forced onto a quantitative intervention checklist.

Studies were placed in one dominant analytical class for descriptive counting: two educational interventions; three stakeholder-based or conceptual accounts; six direct textual/interpretive output audits; and nine technical or human-verified output studies. The assignment is exclusive for counting but thematic interpretation crosses categories when warranted. Coding began with the four review questions and distinguished the object of verification, what counted as authoritative evidence, who was responsible for checking, the observed benefit or failure, and limits of

generalization. The themes were iteratively consolidated into an evidence-informed pathway. This is a structured qualitative thematic synthesis, rather than a pooled statistical estimate or a claim of theory saturation (Thomas & Harden, 2008; Braun & Clarke, 2006).

RESULT AND DISCUSSION

Study selection and evidence profile

The review began with 42 distinct Scopus records. Ten records were excluded during title–abstract screening: seven were judged outside the operational intersection of Islamic religious knowledge verification and AI-assisted education, two were secondary publications rather than primary studies, and one was a book chapter outside the retained publication types. The remaining 32 reports were sought for retrieval. Ten were not obtained as complete open-access texts and were therefore not assigned a negative study-quality verdict. Twenty-two reports were assessed for eligibility; two dataset-only reports were excluded for the stated outcome-scope reason. The resulting primary synthesis contains 20 studies (Figure 1). These counts describe the present accessible-evidence corpus and should be independently validated against the screening log before submission.

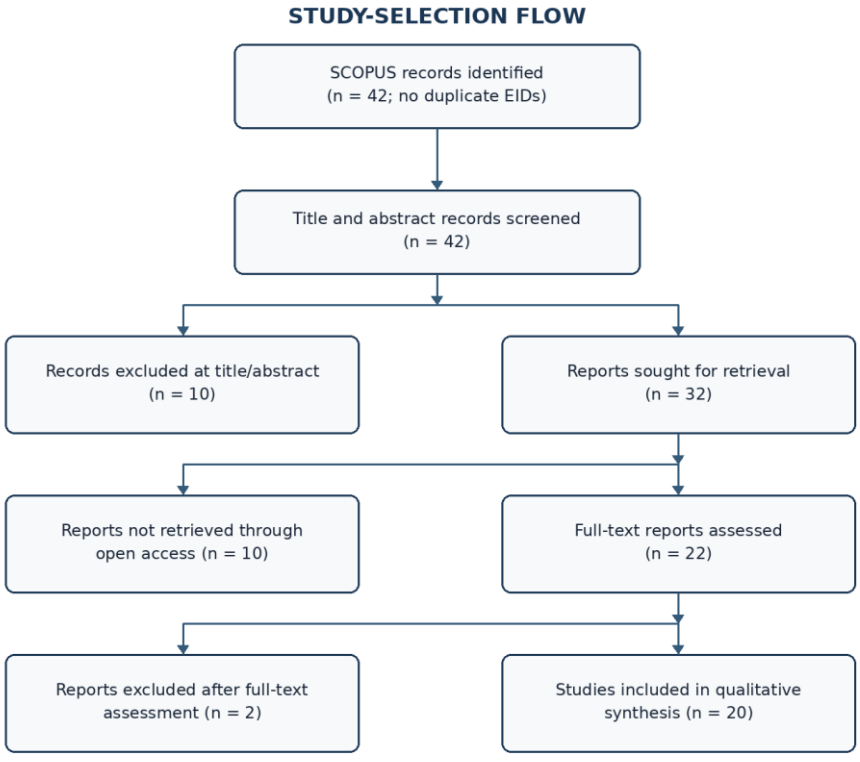


Figure 1. PRISMA 2020-informed study-selection flow for the exported Scopus corpus. "Not retrieved" indicates full-text access failure, not methodological exclusion.

The 20 included studies were published between 2023 and 2026: one in 2023, two in 2024, seven in 2025 and ten in 2026. This distribution should be read only as the time profile of the eligible accessible reports, not as a validated trend across all

global publications in AI-assisted Islamic education. The evidential profile is similarly heterogeneous: nine technical or source-verified studies, six textual interpretation or translation audits, three stakeholder or conceptual investigations, and only two classroom interventions (Figures 2 and 3).

assessment.

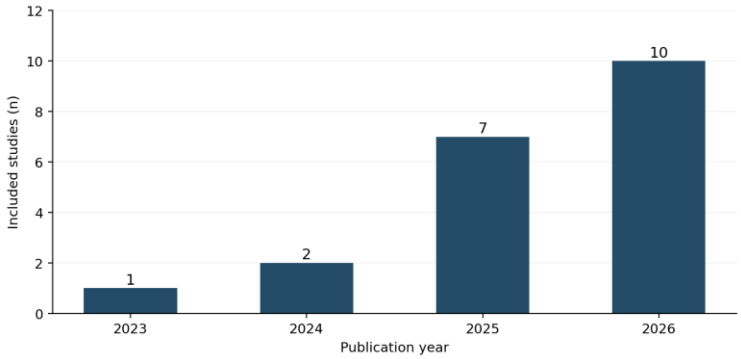


Figure 2. Publication years of the 20 studies included after accessible full-text

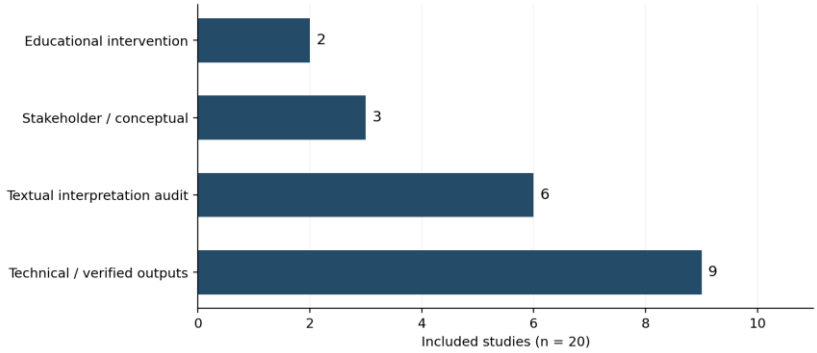


Figure 3. Primary analytic design categories of the 20 included studies; one dominant category was assigned to each study for descriptive counting

Decision point	Number	Interpretation
Unique Scopus records	42	No duplicate EIDs in the export
Title/abstract excluded	10	7 out-of-scope; 2 secondary; 1 book chapter
Reports sought	32	Potential full-text reports
Reports not retrieved	10	No accessible complete text established
Full texts assessed	22	Distinct papers; author-upload duplicate removed
Full-text excluded	2	Data resource construction only without eligible evaluated outcome
Studies synthesized	20	Design-specific qualitative synthesis

Table 2. Study-selection and appraisal counts

Table 3 provides a study-level overview of the evidence actually reviewed in full. It separates educational interventions from theological/linguistic critiques and technical system evaluations, and records a claim-specific boundary for each study.

Study	Design / evidence unit	Verification focus	Findings and inference boundary
Abbas et al. (2026)	Fanar-Sadiq: five QA benchmarks	Quotation validation, RAG, task routing, rule tools	Mean accuracy 76.5%; partly LLM-as-judge
Adel et al. (2023)	AraQA religious QA model benchmark	Canonical verse/hadith retrieval and text generation	Current perplexity 5.221; religious experts not included
Ahmad & Musa (2025)	Two simulations, 50 questions each	Qur'anic text/juz' and recitation markers	31/50 and 33/50 fully valid; no learner study
Al-Janabi (2024)	Critical comparative tafsir study	AI output versus classical scholarly interpretation	Contexts omitted; Arabic coding needs second reviewer
Albalawi & Abdul-Ghafour (2026)	Eight Qur'anic euphemistic verses	AI translation against culturally nuanced expert benchmark	Qualitative meaning losses; not population error rate
Almajdoub et al. (2026)	20 teacher focus-group participants	Tarteel and ChatGPT checked by teachers	Qualified teacher oversight; no student impact test
Altammami (2025)	250 hadith simplification pairs	Human checking of AI-generated hadith paraphrase	Expert-verified text corpus; no learning experiment
Bakr et al. (2026)	Search system and Tajweed benchmarks	Verified Qur'anic corpus, patterns, expert review	10-query exact-match success; narrow technical task
Basem et al. (2025)	Extractive Qur'anic QA benchmark	Span extraction with Arabic instruction tuning	Best pAP@10 .637; domain-specific extraction
Darmu'in et al. (2025)	Six interview participants	Students' sanad and contextual dalil concerns	Perceived risks; not directly tested accuracy
Dart (2026)	Normative conceptual analysis	Epistemic tool vs moral/religious authority	AI cannot supply scholarly accountability; not an experiment
Dzulraidi et al. (2025)	Four short surahs, two GPT versions	Translation versus authoritative Malay reference	GPT-4.0 closer; nuanced meaning still uncertain
El Ganadi et al. (2025)	EMAN hadith-retrieval comparisons	Authenticated hadith embeddings and source citation	F1 .67; do not conflate with a separate 100% result
Faqihi et al. (2026)	QAMAR: 10,000-word model evaluation	Human-verified Qur'anic morphology labels	LLM means 39.47% and 43.93%; not religious ruling accuracy
Hadi et al. (2025)	Qualitative analysis of al-	AI semantic output checked against	Illustrative exegetical fit; no measured

	maraḍ	classical tafsir	general effect
Nouri (2026)	Course chatbot, cohorts n=40 and n=40	Instructor-approved hadith curriculum knowledge	A grades 22.5% vs 42.5%; confounded cohort comparison
Obaid & Bawany (2024)	SeerahGPT: 100-question comparison	RAG source grounding and expert answer checks	F1 .312 vs .299; low absolute lexical metrics
Saeed et al. (2026)	QurSim + 98 human-rated explanations	Ontology/embeddings + generated explanations	Human metric labels inconsistent across sections
Shehab (2026)	Six Qur'anic near-synonym pairs	AI English lexical rendering against tafsir	Semantic flattening; small purposive lexical set
Sumarna et al. (2026)	HARMONY, n=140 one-group study	Textbooks + authenticated hadith; supervised AI reflection	Within-group hadith gain (d=1.56); no control group

Table 3. Characteristics of the 20 included full-text studies and limits of the reported evidence.

RQ1: Epistemic authority is not conferred by generated fluency

The primary conceptual study by Dart (2026) argues that a technology may be reliable as a limited instrument for navigating vetted information without becoming a legitimate source of moral or religious authority. This distinction is not an intervention effect but an epistemic criterion: a justified answer must preserve traceable sources, professional accountability and the interpretive conditions under which guidance is offered. The argument is especially salient when training-data opacity and engagement-driven digital systems encourage users to confuse assertive presentation with warranted judgment.

Two empirical accounts put practical content behind the distinction. Interview narratives from six Malaysian students studying Islamic subjects in Indonesia identified three recurrent barriers to uncritical adoption: perceived lack of scholarly authority and sanad, difficulties obtaining accurate contextual dalil, and anxiety about distortion of Qur'anic or hadith interpretation (Darmu'in et al., 2025). The participants' statements support the existence and structure of these concerns in that purposive sample; they do not establish their prevalence among all Muslim students. Focus groups with 20 teachers similarly favoured specialized recitation support over unverified generic explanations and maintained the importance of teacher correction and spiritual guidance (Almajdoub et al., 2026).

Taken together, these accounts distinguish at least three loci of authority: textual authority (what the primary source actually says), interpretive authority (which claims follow from the source and under what scholarly standards), and pedagogical authority (who judges the suitability of the explanation for learners). An AI system can help find, organize or rehearse material across these levels, but neither an internally generated citation nor a user-friendly presentation by itself certifies interpretive or pedagogical legitimacy. Contextual publications on Islamic AI literacy, educational governance and institutional verification reinforce the practical relevance of supervised checking (Shobri & Alfurqan, 2026; Ibrohim et al., 2026; Hamidalloh et al., 2026; Mustopa et al., 2026).

RQ2: Verification objects range from text identity to contextual interpretation

The 20 papers reveal that the meaning of "verification" depends on the type of claim. For verbatim Qur'anic or hadith citation, evaluators must verify a reference exists and that the transmitted wording, attribution, punctuation and orthographic

detail match an authorized text. For translations, the criterion widens to semantic faithfulness, register, rhetorical force and exegetical fit. For fiqh answers, apparently correct arithmetic or retrieval must be distinguished from school-specific rules and accountable explanation. For educational outputs, a teacher also needs to determine whether an otherwise accurate passage facilitates defensible learning rather than mere answer consumption.

Studies of Qur'anic translation show why textual alignment and interpretive adequacy should not be equated. Shehab (2026) compared generated translations of six near-synonym contrasts against tafsir-grounded criteria and identified flattening of connotation and semantic distinctions. Albalawi and Abdul-Ghafour (2026) examined eight euphemistic Qur'anic verses; the fluency of generated English did not preserve every culturally and theologically sensitive nuance. Dzulraidi et al. (2025), comparing two ChatGPT versions with Abdullah Basmeih's translation of four short surahs, found that an improved surface rendering could still omit interpretive precision. Hadi et al. (2025) and Al-Janabi (2024) likewise underline the importance of expert tafsir comparison for evaluating contextual claims. These are purposive textual examinations, not probability-based samples from the whole Qur'an.

A particularly concrete example comes from Ahmad and Musa (2025), who examined two sets of 50 AI-generated Qur'anic murājaʿah items. In the first set, 31 of 50 items (62%) reproduced fully valid verse text; in the second, after adding a mushaf resource as contextual input, 33 of 50 (66%) met the authors' criterion. The paper also identifies other errors in verse references and recitation markings. These results concern the accuracy of generated practice material, not gains in students' memorization or retention. Altammami (2025) documented a different human-verification workflow, in which 250 AI-simplified hadith pairs were checked by an Islamic-studies reader to preserve intended meanings. Together, the studies suggest that output checking requires explicit source standards and genre-sensitive human inspection.

The technical studies extend these checks to narrower reproducible tasks. HQDR/HQPS applies deterministic search over verified Qur'anic characters and rule patterns; its 10-query keyword test achieved 100% exact matching in that bounded evaluation, while five tajwīd-rule extraction families were checked against specialist references (Bakr et al., 2026). QAMAR created a manually checked morphological resource and tested model responses for stem, lemma, root and part-of-speech tagging on a 10,000-word subset; reported model means were approximately 39.47% for GPT-4o and 43.93% for LLaMA-4 on the paper's feature-based measure (Faqihi et al., 2026). Those metrics are not estimates of the correctness of whole tafsir or fatwas.

RQ3: What outcomes have actually been evaluated?

Only two included publications directly evaluated an educational intervention. HARMONY was developed through a sequential exploratory design with 16 qualitative participants and implemented with 140 students at four Indonesian universities. It integrated authoritative hadith resources, lecturer supervision and AI-supported discussion. In a one-group pretest–posttest comparison, hadith learning competence increased ($t(139) = 18.42$, $p < .001$, $d = 1.56$, N-gain = .71) and self-reported tolerance scores also increased ($t(139) = 17.16$, $p < .001$, $d = 1.45$, N-gain = .68); expert ratings averaged 3.80/4.00 (Sumarna et al., 2026). The size of these pre/post changes warrants attention but does not demonstrate that the model caused the improvements; history, testing and maturation are plausible alternative explanations in the absence of a comparison group.

Nouri (2026) compared naturally occurring university hadith-sciences cohorts, with a traditional spring group ($n = 40$) and a summer group ($n = 40$) additionally

using a course-constrained chatbot for six weeks. A grades were reported for 22.5% of the spring group and 42.5% of the chatbot group; 37 of 40 chatbot users (92.5%) recommended continued integration. However, grade distributions were evaluated descriptively across different terms, and no adjusted or randomized between-group learning-effect estimate was reported. The data support feasibility, acceptance and a favorable descriptive achievement pattern—not strong proof of causal superiority.

Technical model performance offers separate evidence of verifiable operations. EMAN compared prompt-based strategies with embedding-grounded hadith retrieval and reported F1 = .67 for its principal embedding-based condition versus F1 = .47 for Bukhari GPT and .43 for GPT-4 on the specified comparison (El Ganadi et al., 2025). The same paper uses a different measure in another accuracy table; its headline percentage must not be substituted for the F1 test. Fanar-Sadiq reported mean accuracy of 76.5% across five heterogeneous Islamic QA benchmarks, compared with 62.4% for a Fanar baseline and 79.6% for one Gemini-3 Pro comparator in the authors' table (Abbas et al., 2026). Some tasks rely on exact matching and others on LLM-as-judge, so the aggregate is benchmark-defined rather than an objective universal measure of religious truth.

SeerahGPT reported a comparative F1 score of .312 against .299 for its Llama-2 baseline and higher ROUGE-L (.230 versus .149) on a 100-question evaluation, alongside expert-based considerations of religious answer adequacy (Obaid & Bawany, 2024). AraQA reported perplexity of 5.221 for the current experiment despite referencing 2.3 for an earlier configuration; perplexity itself cannot validate cited religious sources without a separate expert-grounded accuracy test (Adel et al., 2023). A hybrid verse-retrieval study by Saeed et al. (2026) reported a relative classification improvement (66.6% versus 64.1% for the cited Word2Vec comparison), yet the labels for human-assessed explanation completeness and correctness differ across the paper's sections. Those percentages are therefore not assigned to a definitive outcome dimension here. In extractive Qur'anic QA, Basem et al. (2025) reported pAP@10 = .637 for the best configuration, a task-specific retrieval result rather than evidence of religious reasoning competence.

Study	Verified measured outcome	Interpretive boundary
Sumarna et al. (2026)	n=140; d=1.56 (hadith competence), d=1.45 (self-reported tolerance); one-group pre/post	Evidence of within-group change; no causal comparison
Nouri (2026)	n=40+40; A grades 22.5% vs 42.5%; user endorsement 37/40	Term/cohort confounding; descriptive comparison only
El Ganadi et al. (2025)	EMAN F1=.67 vs Bukhari GPT .47 and GPT-4 .43	Distinct metrics and limited hadith prompt sample
Abbas et al. (2026)	Fanar-Sadiq mean benchmark accuracy 76.5%	Heterogeneous tasks; some LLM-judged responses
Bakr et al. (2026)	10-query keyword	Not general religious

	benchmark 100% exact match	conversational accuracy
Faqihi et al. (2026)	GPT-4o mean 39.47% and LLaMA-4 mean 43.93% on selected morphological features	No inference to exegesis/education
Obaid & Bawany (2024)	F1 .312 vs .299; ROUGE-L .230 vs .149	Lexical similarity is not expert certification
Saeed et al. (2026)	Classification 66.6% vs 64.1% Word2Vec	Human metric-label discrepancy not resolved
Adel et al. (2023)	Current perplexity 5.221 (different from prior 2.3)	Perplexity does not validate religious claims
Ahmad & Musa (2025)	Valid verse text 31/50, then 33/50 prompts	No student learning evaluation

Table 4. Selected measurable outcomes and limits of interpretation

Methodological certainty: positive signals without uniform effectiveness

The two educational reports do not provide randomized controlled effects, and the numerical outcomes from technical systems answer different questions. Human-evaluated religious-text studies tend to use small, purposive verse or hadith samples, while tool comparisons often evaluate closed task sets selected by the system developers. There is a recurring mismatch between a model's ability to reproduce a reference string and the broader authority of an interpretation. A second recurring weakness is the incompleteness of evaluator description—especially assessor training, blinded scoring, and inter-rater agreement. A third is internal reporting ambiguity, visible in the metric inconsistencies noted in Saeed et al. (2026) and Adel et al. (2023).

The review therefore assigns conclusion strength by claim type: the evidence is promising but preliminary for short-term educational outcomes; moderate within specified technical retrieval or matching tasks; moderate to low for broader correctness of generated religious interpretation; and contextually useful but non-generalizable for teacher/student experiences. Conceptual statements about religious authority are theoretical contributions, not estimates of an intervention effect. A vote count of positive/negative findings or a pooled "AI effectiveness" percentage would conceal these differences rather than synthesize them (Hong et al., 2018; Campbell et al., 2020).

CONCLUSION

The accessible research corpus indicates that AI-assisted Islamic education is most defensible when the technology is treated as a support for access, rehearsal and source search—not as a self-authorizing religious interpreter. Across 20 full-text primary studies, the recurring problem is not simply whether a fluent answer sounds plausible, but whether the cited text exists, remains faithfully reproduced, is interpreted within recognized scholarly standards and is pedagogically endorsed by an accountable educator. The studies document progress in constrained retrieval and source-fidelity checking, alongside significant vulnerabilities in scriptural translation, verse attribution, and context-sensitive exegesis. The two educational interventions offer preliminary encouraging signals, but their designs do not permit strong causal statements about student learning improvement attributable specifically to generative AI. The review consequently recommends a six-stage verification pathway based on claim classification, retrieval, authentication, contextual interpretation, human review and documented decision-making. This is an evidence-informed synthesis rather than an effectiveness-validated model. Its practical value now lies in defining testable competencies and transparent institutional responsibilities. Future research should evaluate the pathway with comparison groups, domain-expert audit of outputs,

delayed learner outcomes, and well-described adjudication rules. The 20-study evidence base should be interpreted with particular caution because the search was limited to a researcher-exported Scopus corpus and ten otherwise eligible reports were not retrieved in full through open-access routes.

REFERENCES

Abbas, U., Ouzzani, M., Eltabakh, M. Y., Sinan, O., Bhatia, G., Mubarak, H., Hawasly, M., Hashim, M. Q., Darwish, K., & Alam, F. (2026). Fanar-Sadiq: A multi-agent architecture for grounded Islamic QA. Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Industry Track), 1567–1588. https://doi.org/10.18653/v1/2026.acl-industry.109

Adel, Y., Dorrah, M., Ashraf, A., Elsaadany, A., Mohamed, M., Wael, M., & Khoriba, G. (2023). AraQA: An Arabic Generative Question-Answering Model for Authentic Religious Text. Proceedings of the 11th International Japan-Africa Conference on Electronics, Communications, and Computations, JAC-ECC 2023. https://doi.org/10.1109/JAC-ECC61002.2023.10479645

Ahmad, A.M.B., & Musa, M.A.B. (2025). Artificial Intelligence (AI) in the Field of Tahfiz: A Simulation Study of ChatGPT Interaction in Qur'anic Murājaʿah; [Kecerdasan Buatan (Ai) Dalam Bidang Tahfiz: Kajian Simulasi Interaksi Chatgpt Dalam Murajaah Al Quran]. Quranica. https://www.scopus.com/pages/publications/105018180292?origin=resultslist

Albalawi, A.F., & Abdul-Ghafour A.-Q.K. (2026). An investigation into human vs AI English translations of Qur'anic euphemisms. Social Sciences and Humanities Open. https://doi.org/10.1016/j.ssaho.2025.102249

Al-Janabi, M.K.A. (2024). Artificial Intelligence in Quranic Exegesis: A Critical Analytical Study of ChatGPT Technology. Quranica. https://www.scopus.com/pages/publications/85217026451?origin=resultslist

Almajdoub, Z., Fidalgo, P., Alramamneh, Y., AlHakmani, R., & Kasasbeh, N. (2026). Artificial intelligence in Qur'anic education: pedagogical promise, practical challenges, and ethical concerns. Frontiers in Education. https://doi.org/10.3389/feduc.2026.1876380

Almasoud, A., Al-Ghamdi, S., Alqifari, R., Alfear, N., & Al-Khalifa, H. (2026). MirathQA: A dataset for evaluating large language models on Hanbali Islamic inheritance reasoning tasks. Data in Brief. https://doi.org/10.1016/j.dib.2026.112589

Amin, M. (2026). Ethical Use of Generative Artificial Intelligence in Islamic Religious Education: Perspectives of Madrasah Aliyah Teachers and Students in Banjarmasin. Al-Munawwarah: Journal of Islamic Education. https://doi.org/10.38073/almunawwarah.v2i2.5654

Anim, S., Khairiyah, N., Muchtar, I., & Jakaria, J. (2026). Artificial Intelligence and the Crisis of Religious Authority: A Conceptual Review of Islamic Discourse in the Age of Machine Mediation. RSF Conference Series. https://doi.org/10.31098/bmss.v6i1.1095

Arif, M., Aziz, M. K. N. A., Hassan, H., Fikri, Y., Zekkari, M., & Rismawati, B. V. (2026). Artificial Intelligence in Islamic Religious Education: A PRISMA-Based Review and the SEI-CHAT Framework. Online Learning In Educational Research (OLER). https://doi.org/10.58524/oler.v6i1.993

Bakr, A., Medhat, W., & Arafa, T. (2026). HQDR and HQPS: A Holy Quran digital repository with advanced pattern search and AI-enhanced natural language interface. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1007/s44443-026-00934-w

Basem, M., Oshallah, I., Hamdi, A., & Mohamed, A. (2025). Few-Shot Prompting for Extractive Quranic QA with Instruction-Tuned LLMs. 3rd International Conference on Intelligent Methods, Systems and Applications, IMSA 2025.

https://doi.org/10.1109/IMSA65733.2025.11166890
Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922
Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa
Campbell, M., McKenzie, J. E., Sowden, A. J., Katikireddi, S. V., Brennan, S., Ellis, S., Hartmann-Boyce, J., Ryan, R. E., Shepperd, S., Thomas, J., Welch, V. A., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. BMJ, 368, l6890. https://doi.org/10.1136/bmj.l6890
Darmu'in, Nasikhin, Darnoto, Maarif, M.A., Sofanudin, A., & Haryanto, J.T. (2025). Perceived Barriers to ChatGPT Integration in Islamic Studies Education: Insights from Malaysian International Students in Indonesia. Journal of International Students. https://doi.org/10.32674/h99a4p06
Dart, B.C. (2026). Epistemic Trust and the Limits of AI Authority in Islam. Open Theology. https://doi.org/10.1515/opth-2025-0069
Dzulraidi, D.H., Jalil, A.A., & Sin, M.L.M. (2025). The Accuracy of Chatgpt In Translating The Qur'an: A Comparative Analysis With Authoritative Translation; [Ketepatan Chatgpt Dalam Terjemahan Al-Quran: Analisis Perbandingan Dengan Terjemahan Autoritatif]. Quranica. https://www.scopus.com/pages/publications/105018322493?origin=resultslist
Engkizar, E., Jaafar, A., Sarianto, D., Ayad, N., Rahman, A., Febriani, A., ... & Rahman, I. (2024). Analysis of Quran Education Problems in Majority Muslim Countries. International Journal of Islamic Studies Higher Education, 3(1), 65-80. https://doi.org/10.24036/insight.v3i1.209
El Ganadi, A., Aftar, S., Gagliardelli, L., & Ruozzi, F. (2025). Generative AI for Islamic Texts: The EMAN Framework for Mitigating GPT Hallucinations. International Conference on Agents and Artificial Intelligence. https://doi.org/10.5220/0013312800003890
El Yunusi, N. S., Salmany, M. R., Fikry, A. J. N., Mashuda, M., Syafi'i, I., & Munawir, M. (2026). Generative AI and the Development of Adaptive Islamic Education Teaching Materials for 21st Century Learning. An-Nuha. https://doi.org/10.64810/annuha.v13i1.1000
Faizin, N., Alfan, M., Basid, A., Ramadhan, M. R., Panatik, S. A., & Kawakip, A. N. (2025). Muslim students' acceptance of artificial intelligence in Islamic religious education: an extended TAM approach. Discover Education. https://doi.org/10.1007/s44217-025-00767-1
Faizul, M., Mukmin, T., Fauzi, M., & Isnaini, M. (2026). Construction of Islamic AI Literacy Dimensions for The Development of Islamic Religious Education Curriculum. Berkala Ilmiah Pendidikan. https://doi.org/10.51214/bip.v6i2.2225
Faqihi, S., Bouzoubaa, K., Tajmout, R., & Namly, D. (2026). QAMAR: A New Fully Verified and Accurate Quranic Arabic Morphological Analysis Resource. EACL 2026 - 19th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script, AbjadNLP 2026. https://doi.org/10.18653/v1/2026.abjadnlp-1.38
Hamidalloh, A. S., Thobroni, A. Y., & Suyudi, M. (2026). ChatGPT as an Algorithmic Epistemic Mediator in Contextual Islamic Religious Education and Religious Moderation. Jurnal Ilmiah Guru Madrasah.

https://doi.org/10.69548/jigm.v5i1.116

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

Ibrohim, W. M., Aziz, A., & Sahrodi, J. (2026). Islamic Ethical Pedagogy for Artificial Intelligence Use: Insights from Islamic Religious Education Teachers in Indonesia. Journal of Islamic Education Research. https://doi.org/10.35719/jier.v7i4.589

Isti'ana, A. (2026). Budaya Literasi Digital-AI dalam Pendidikan Agama Islam: Penguatan Tabayyun Digital dan Integritas Akademik. SAKALIMA. https://doi.org/10.70211/sakalima.v3i1.644

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474. https://arxiv.org/abs/2005.11401

Marwantika, A. I., & Dauda, K. O. (2025). Da'wah in the Algorithmic Era: Investigating Bias and Validity of Islamic Artificial Intelligence Applications. MUHARRIK. https://doi.org/10.37680/muharrik.v8i2.7730

Mat Yusoff, S., Mohamad Marzaini, A.F., Zynuddin, S.N., Mat Jusoh, A.M., Basal, M.H., Alias, N.R., & Saad, N. (2026). Ethical integration of generative artificial intelligence in Islamic religious education involves human mediation and source verification. Discover Artificial Intelligence. https://doi.org/10.1007/s44163-026-02251-7

Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Muis, M. A., Chanifudin, C., Nuriyati, T., & Basir, A. (2026). Ethical Islamic AI Literacy: Model Literasi Artificial Intelligence Berbasis Paradigma Pendidikan Agama Islam. Attractive: Innovative Education Journal. https://doi.org/10.51278/aj.v8i1.2653

Mustopa, S., Basri, H., & Suhartini, A. (2026). ChatGPT and the Reconfiguration of Religious Authority: An Islamic Epistemological Perspective on Artificial Intelligence in Education. IJIE International Journal of Islamic Education. https://doi.org/10.35719/ijie.v5i1.2659

Namoun, A., Humayun, M.A., & Nawaz, W. (2024). A Multimodal Data Scraping Tool for Collecting Authentic Islamic Text Datasets. International Journal of Advanced Computer Science and Applications. https://doi.org/10.14569/IJACSA.2024.0151224

Nouri, R. (2026). Effectiveness of a Generative AI-Powered Conversational Bot in Teaching Hadith Sciences: Evidence from Al Qasimia University. International Journal of Information and Education Technology.

https://doi.org/10.18178/ijiet.2026.16.9.2699
Obaid, S., & Bawany, N.Z. (2024). SeerahGPT: Retrieval Augmented Generation based Large Language Model. 2024 18th International Conference on Open Source Systems and Technologies, ICOSST 2024 - Proceedings. https://doi.org/10.1109/ICOSST64562.2024.10871159
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., ... Welch, V. A. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
Page, M. J., Moher, D., 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., ... Welch, V. A. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. BMJ, 372, n160. https://doi.org/10.1136/bmj.n160
Purnomo, D. D. A. D., Fariyati, S., Arif, A., & Hasan, B. M. (2026). Islamic Education: Examining AI Use, Reliance, and Perceived Validity among Tarbiyah Students. Paedagogia: Jurnal Pendidikan. https://doi.org/10.24239/pdg.vol15.iss1.1436
Qulyubi, M., Bahri, S., Hayaturrohman, H., & Yasik, F. (2026). Negotiating Artificial Intelligence in Contemporary Pesantren: Knowledge, Religious Authority, and Islamic Educational Traditions. Munaddhomah. https://doi.org/10.31538/munaddhomah.v7i3.3291
Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. Journal of the Medical Library Association, 109(2), 174–200. https://doi.org/10.5195/jmla.2021.962
Saeed, S., Rajput, Q., & Haider, S. (2026). On Combining Embeddings, Ontology and LLM to Retrieve Semantically Similar Quranic Verses and Generate Their Explanations. Communications in Computer and Information Science. https://doi.org/10.1007/978-3-032-08333-3_14
Shehab, E. (2026). A Shallow Echo: Artificial Intelligence and the Semantic Flattening of the Qur'an. Learned Publishing. https://doi.org/10.1002/leap.2063
Shobri, F., & Alfurqan, A. (2026). From Individual Competence to Distributed Capacity: How Islamic Religious Education Teachers Verify Generative-AI Content in Padang, Indonesia. FIKROTUNA. https://doi.org/10.32806/jf.v15i02.2227
Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039
Sumarna, E., Supriyadi, T., Abdullah, M., Rahminawati, N., Avkharovna, K.Z., & Al-Sharoot, D. (2026). Developing and Validating the HARMONY Model: A Prophetic Values-Based Hadith Learning Approach Integrating Digital Tools to Strengthen University Students' Religious Tolerance. International Journal of Learning, Teaching and Educational Research. https://doi.org/10.26803/ijlter.25.8.28
Suryanto, D., Sukmawati, E., Humaira, K., Yasmin, A., Khoirunnisa, S., & Nisa, U. (2025). Pemanfaatan Kecerdasan Artifisial Generatif dalam Pembelajaran Pendidikan Agama Islam di SMA: Telaah Integratif. SAKALIMA. https://doi.org/10.70211/sakalima.v2i2.722

Tanjung, E. F., Ginting, N., & Nasution, N. (2026). Repositioning Artificial Intelligence in Value-Based Islamic Religious Education. Urwatul Wutsqo. https://doi.org/10.54437/urwatulwutsqo.v15i01.2928

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology, 8, 45. https://doi.org/10.1186/1471-2288-8-45

Copyright holder :
© Zuzano, F., Darwianis, D., Sayuti, M., Madona, A, S., Kurnia, A.

First publication right:
Diniyyah Jurnal

This article is licensed under:

CC-BY-SA