Reorienting Islamic Religious Education Assessment in The Era of Generative AI Through Authentic Assessment

Muhammad Habibul Arif
Universitas Islam Negeri K.H. Abdurrahman Wahid Pekalongan, Indonesia
Email: muhhabibularif@gmail.com

Abstract

Generative AI raises questions about evidence of competence in Islamic Religious Education because written assignments may reflect technological assistance invisible to teachers. This article examines changing assessment requirements and develops an authentic assessment framework connecting religious source verification, process evidence, and accountability for AI use. An integrative literature review was conducted through targeted searches of publications on authentic assessment, academic integrity, and AI in education, complemented by a normative reading of Quran 17:36 and 49:6. Sources were examined through problem mapping, framework comparison, and conceptual synthesis. The synthesis identifies four design requirements: alignment between evidence and competence, traceable religious sources, transparent AI assistance, and equitable learning opportunities. The proposed cycle combines initial competence mapping, explicit AI boundaries, contextual tasks, process records, source verification, oral defence, and reflective revision. An analytical rubric and an illustrative digital communication task are provided. The contribution lies in translating epistemic responsibility into observable assessment indicators rather than claiming that process-oriented assessment is new. The framework remains conceptual and has not undergone empirical validation. Its applicability depends on educational level, access to sources, communication needs, and teacher workload. Further studies should examine indicator validity, scoring consistency, implementation time, and feasibility before claims of improved learning or academic integrity can be made.

Keywords: islamic religious education, generative ai, authentic assessment, source verification, academic integrity

Introduction

Assessment in Islamic Religious Education requires evidence that learners understand religious teachings, can justify a decision, and can apply their knowledge to situations in everyday life. Generative AI makes the relationship between submitted work and its author's competence increasingly important to examine. A coherent essay about honesty can be produced with machine assistance even when the learner's understanding of moral reasons and the consequences of their actions does not match the quality of the essay. Educators must determine what evidence is sufficient to support a responsible judgement of individual competence.

Cotton et al. (2024) discuss the opportunities offered by ChatGPT alongside difficulties in maintaining academic integrity. They identify policy, support, and education about technology use as elements of an institutional response. In Islamic Religious Education, this issue has an additional dimension: a persuasive answer may contain unchecked attributions of religious evidence or explanations. Assessment therefore needs to examine learners' ability to trace references and justify their interpretations, as well as the structure of their writing (Nur Hasanah & Kristiyah, 2023).

The review by Xia et al. (2024) connects assessment transformation with responsibility for learning, educators' skills, and institutional policies. These findings come from higher education, so their application to school-based Islamic Religious Education needs to account for age, learning objectives, and resources. This article uses that literature as a basis for reasoning about design, rather than as evidence that an intervention has succeeded in Indonesian schools.

Authentic assessment provides a foundation for assessing the application of knowledge through meaningful tasks. Gulikers et al. (2004) propose a framework addressing the task, physical context, social context, assessment result or form, and assessment criteria. Its relevance to Islamic Religious Education lies in the opportunity to connect concepts with contextual decisions. Nevertheless, AI can also complete a case-based task. An authentic context must be accompanied by an examination of how a decision was developed and understood.

Indonesian research has already examined AI in the evaluation of Islamic Religious Education. Sofyan and Wahab (2024) report opportunities and challenges in AI-based assessment, including educator readiness and privacy. More directly, Hakim (2026) discusses the assessment of process, reasoning, and authenticity in Islamic Religious Education. The argument that assessment should move beyond a single final product therefore has precedents. This article develops an operational connection between religious source verification, disclosure of AI assistance, and grading decisions.

The article addresses three questions. First, what evidence of competence should be retained when AI can assist with assignments in Islamic Religious Education? Second, how can epistemic responsibility and the principle of tabayyun, or careful verification of information, be translated into assessment indicators? Third, how can this framework be implemented through tasks, rubrics, and fair procedures? The aim is to produce a conceptual framework and preliminary materials for implementation trials, while clearly distinguishing literature-based synthesis from evidence of effectiveness.

Methods

This study uses an integrative literature review to develop a conceptual framework. The approach draws on Snyder (2019), who explains literature review as a research methodology requiring alignment between its purpose, source selection, and synthesis. This review was not designed as a systematic review capturing all publications in the field. It therefore presents no PRISMA flow diagram, total search-result counts, or estimated intervention effects.

Searches were conducted on 8 October 2026 using targeted web searches leading to journal publishers, institutional repositories, and UNESCO. Keyword groups included “authentic assessment”, “generative AI assessment academic integrity”, “AI Assessment Scale”, “asesmen PAI generative AI”, and “literature review research methodology”. Relevant references were followed to check original sources. Priority was given to publications on generative AI from 2023 onwards, while earlier sources were retained for assessment theory and review methodology. The search results were not treated as a complete account of the literature through 2026.

The academic reference corpus comprises the twelve publications listed in the references. Sources were selected purposively for their direct relevance to assessment design, academic integrity, Islamic Religious Education, or review methodology. Popular publications, references without verifiable bibliographic identities, and claims of effectiveness that could not be traced were not used to support conclusions. Two Quranic passages were used separately as normative sources and were not counted as empirical studies.

Access to sources was uneven. Analysis covered text available on publisher or repository pages, official abstracts, and bibliographic metadata. Where only an abstract was accessible, use of that source was limited to the information stated in the abstract; unavailable procedural or results details were not reconstructed. This restriction reduced the depth of engagement with some publications and constitutes a limitation of the review.

Extraction focused on four categories: problems in establishing competence, the design of authentic evidence, responsibility for sources, and conditions of fairness. Analysis compared arguments across sources, distinguished research contexts, and developed implications for Islamic Religious Education. The extraction is summarised in Appendix A. Proposed weights, durations, and task materials are design decisions made in this article, rather than figures derived from previous experiments.

Quran 17:36 and 49:6 were read to develop a normative argument for caution in matters of knowledge and verification of information. The verses were not treated as psychometric instruments or as evidence that a rubric automatically improves moral character. The relationship between normative principles and pedagogical indicators is presented as the author's interpretation, open to further examination.

Results

Changing Evidence of Competence in Islamic Religious Education

The literature synthesis indicates a need to distinguish product quality from the strength of evidence of competence. Product quality refers to assessable features of an output, such as accurate terminology, the flow of argument, and readability. Evidence of competence concerns the grounds concluding that learners themselves understand the material. The two can support each other, but that relationship requires examination when AI assistance is undisclosed.

For illustration, two learners might submit equally strong essays about tabayyun. The first can identify sources, explain misinformation, and revise a conclusion in response to new evidence. The second cannot explain why a particular religious reference was selected. Awarding identical reasoning scores on the basis of the writing alone would overlook these differences in evidence. This is an analytical illustration, not participant data.

Hakim (2026) has already proposed combining process, reasoning, and authenticity. This article extends that direction by treating verification of religious references as a separately assessable component. The product is retained but connected to source records, revisions, and individual explanations. The purpose is to strengthen the grounds for grading decisions without treating all AI use as misconduct.

Four design requirements are formulated. First, each item of evidence must correspond to a clear competence. Second, religious claims used to support conclusions must be traceable. Third, technological assistance affecting the substance of the work must be disclosed. Fourth, learners must have equivalent opportunities to demonstrate competence despite differences in technological access. These requirements guide the selection of tasks and assessment procedures.

Conceptual understanding can be checked through a short explanation without AI assistance. Application can be assessed through a case requiring a decision and reasons. Care with sources can be assessed by comparing claims with references. Attitudes relevant to academic work can be examined through observable conduct, such as correcting an attribution and reporting assistance. These indicators are not intended to measure inward sincerity or label learners' levels of piety.

The Proposed Assessment Cycle

The cycle begins with initial competence mapping through short in-class answers. Educators ask about the meaning of tabayyun, examples of information requiring verification, and sources that can be consulted. This activity helps identify support needs and provides initial evidence of individual competence. Its results primarily inform teaching plans, rather than adding to grading demands before learners have had an opportunity to learn.

The second stage establishes objectives, the rubric, and AI boundaries. Learners know which activities may receive technological assistance and which must be demonstrated independently. The third stage involves completing the case with concise process records. Educators may select two checkpoints, such as an initial plan and a source verification table. Extensive documentation is unnecessary if it adds no useful information to the assessment decision.

The fourth stage involves source verification and revision of reasons. Learners show which claims they retained, corrected, or removed after checking. The fifth stage is an individual oral defence using questions that examine the relationship between references and decisions. Three to five minutes per learner is an adjustable initial proposal; this duration has not been established through a large-class feasibility study.

The final stage involves feedback and reflection. Learners explain their most significant error, the grounds for correction, and how they would check information in a subsequent situation. Reflection should refer to actual changes in the work, rather than stop at general claims that the activity was useful. The conceptual sequence is learning objectives → contextual task → process and source evidence → individual explanation → grading decision → revision.

The full cycle need not be used in every lesson. Where time is limited, educators may select one principal task per unit, conduct brief process checks, and schedule question-and-answer sessions in rotation. If AI tools are unavailable or inappropriate for learners' age, educators can provide selected AI outputs for critique. Learners can therefore achieve the same objectives without personal AI accounts.

The Rubric and Grounds for Grading Decisions

The proposed analytical rubric has five dimensions: understanding Islamic Religious Education concepts, weighted at 25%; source verification, 25%; contextual argumentation, 20%; process accountability, 15%; and reflective improvement, 15%. These weights are preliminary design choices. Educators must adapt them to objectives and educational levels, then examine whether each dimension provides distinct evidence. The complete rubric is presented in Appendix B.

Each dimension is scored from 1 to 4 using descriptions of observable behaviour. Its weighted mark equals the score divided by four and multiplied by the percentage weight. The combined mark ranges from 25 to 100 when all dimensions receive scores between 1 and 4. Evidence not yet submitted is marked as not yet assessable and handled according to school policy. Incomplete evidence is not equated with low competence without further examination.

An illustrative calculation assigns scores of 3 for concepts, 4 for verification, 3 for argumentation, 3 for process, and 2 for reflection. The result is 18.75 + 25 + 15 + 11.25 + 7.5 = 77.5. This demonstrates the weighting procedure and is not a research finding. The final judgement should include a comment, such as noting that verification was careful while reflection did not yet explain a change in learning strategy.

Academic errors and suspected integrity violations require different responses. Incorrect attribution may reflect skills that are still developing. Suspected source fabrication should be

clarified through dialogue and examination of evidence rather than inferred from writing style. Liang et al. (2023) found detector bias against non-native English writing in their tests. Their finding cannot be generalised to all detectors or languages, but supports caution about sanctions based solely on detector scores.

To improve consistency, educators can jointly assess sample work with colleagues, compare their reasons for scoring, and revise ambiguous descriptors. Learners receive examples of evidence quality before the task and opportunities to request explanations of their marks. These procedures are designed to make judgements reviewable; their consistency still requires measurement in implementation trials.

Discussion

Epistemic Responsibility and Religious Source Verification

Quran 17:36 emphasises accountability in matters of knowledge and cautions against following what is not known. Quran 49:6 directs the verification of news in a particular context to prevent harm arising from ignorance. This article reads both passages as a normative basis for developing caution towards claims. Applying them to AI outputs is contemporary pedagogical reasoning, not a claim that the verses directly address AI technology.

Epistemic responsibility means that learners can identify the origin of a claim, distinguish degrees of certainty, and revise conclusions when evidence is insufficient. In Islamic Religious Education assignments, Quranic references must identify the surah and verse, while the use of exegesis must name a source actually consulted (Khobir et al., 2021). If a hadith is used, learners record its collection and available identifying information, and check assessments of its authenticity against educator-approved references. Hadith numbering may differ between editions, so a number without information about the collection is not always sufficient.

AI can help suggest search terms or alternative explanatory structures. Its answers, however, do not replace the sources underlying religious evidence. When AI mentions a verse, hadith, or scholarly position, learners must consult the original reference. If it cannot be located, the claim is marked as unverified and excluded from the basis of the final conclusion. Educators provide age-appropriate source lists so that verification demands do not exceed learners' capabilities.

The proposed verification protocol has five steps. Learners identify the claim, record the source cited, consult the original source, compare content, context, and attribution, and assign a verification status. Possible statuses are verified, requiring further explanation, not located, or contradicted by the source. Each status is accompanied by a brief reason and a corrective action. A sample format appears in Appendix C.

Differences in interpretation require care. Divergent readings do not automatically indicate inaccurate sourcing or dishonesty. The rubric assesses traceability, careful use of references, and the quality of reasons in relation to the learning objectives. For questions admitting different positions, educators explain the range of acceptable interpretations before the assignment. This prevents scholarly care from becoming the imposition of a single answer to an issue that is genuinely open to interpretation.

Authentic Assessment and Boundaries for AI Use

The framework proposed by Gulikers et al. (2004) helps educators examine whether tasks and their contexts fit the competences being assessed. In Islamic Religious Education, authentic contexts may include deciding how to respond to a questionable digital message, resolving a simple conflict, or communicating respectfully. Authenticity does not require learners to

disclose personal conflicts. Educator-provided simulations can preserve relevance while protecting privacy.

Boud and Falchikov's (2006) account of assessment oriented towards long-term learning supports self-assessment and the use of feedback. Learners need to examine the quality of their reasons rather than simply receive marks. Revision following feedback is therefore included in the cycle, alongside an explanation of the changes made. Opportunities to correct sources can demonstrate developing competence, while initial errors remain recorded proportionately.

Perkins et al. (2024) propose the AI Assessment Scale to clarify AI use in relation to assessment purposes. This article does not reproduce the entire scale. Instead, it draws on transparent boundaries as a basis for regulating Islamic Religious Education tasks (Najwa & Nur Hasanah, 2026). Boundaries are specified for each activity: the initial explanation of competence is completed without AI, planning may receive AI assistance where permitted, and the final product must be accompanied by accountability for sources and substantive decisions.

Furze et al. (2024) report an AIAS implementation trial in Vietnamese higher education. Their publication illustrates that AI-use rules can be implemented systematically, but it does not establish the effectiveness of the Islamic Religious Education framework proposed here. Transfer to schools requires separate trials, particularly concerning age, device access, and the ability to verify religious texts.

AI-use records should be proportionate, so that an assignment does not become excessive surveillance. Minimum information includes the tool, date, purpose of assistance, examples of prompts affecting content, changes made, and checks undertaken. Educators do not request entire private conversation histories or passwords. Learners who do not use AI state this and still provide equivalent process evidence.

Equitable Access and Educators' Roles

Miao and Holmes (2023) identify human capacity development, privacy, and age-appropriate approaches as considerations for educational AI use. Design implications include offering equivalent alternative tasks and limiting data collection. Learners should not receive higher marks merely because they use paid tools. Assessment targets the same competence whether work receives permitted AI assistance or is completed using offline sources.

Learners with particular communication needs may use alternative forms of explanation that still demonstrate individual reasoning. Examples include a short supervised written explanation or a conversation with additional time. Such adjustments do not change the target competence but reduce the influence of irrelevant barriers. Educators should document accommodations so that grading decisions can be explained (Hidayat & Fatmawati Nur Khasanah, 2025).

Educators' roles include selecting sources, explaining objectives, providing feedback, and considering learners' contexts. Zawacki-Richter et al. (2019) highlight the importance of educational perspectives in AI research. Their review concerns AI before the widespread use of generative AI, so it informs the orientation of educators' roles rather than descriptions of current model capabilities. Educators remain responsible for assessment decisions and do not delegate judgements about religious learning to automated outputs.

Workload is a condition of feasibility. A class of 36 learners would require 108 to 180 minutes if each oral defence lasted three to five minutes, excluding transitions. This is a design calculation, not a field finding. To reduce the burden, oral defences can be distributed across several lessons, while feedback can use anonymised examples of errors. Any strategy selected must preserve individual evidence for the assessed competence.

Contributions and Limitations

The article's contribution is to connect authentic assessment principles with a religious source verification protocol and a rubric for accountability in AI use. This connection gives educators concrete criteria for distinguishing a strong product, trustworthy sources, and understanding that learners can explain. The article does not claim that portfolios, reflection, or questioning are new approaches. Earlier research remains the basis for development.

The principal limitations are purposive source selection, uneven access to full texts, and the reliance of some arguments on international higher education literature. The review does not provide a comprehensive account of Indonesian Islamic Religious Education teachers' practices. The framework has not undergone expert validation or inter-rater testing. Claims that it will improve honesty, learning outcomes, or efficiency therefore cannot yet be made.

Further research could begin with reviews of indicator suitability by Islamic Religious Education specialists, assessment specialists, and teachers. Following revision, a limited trial could examine the clarity of instructions, adequacy of evidence, implementation time, and scoring agreement. Learner interviews could identify access barriers and documentation burdens. Any impact study must distinguish improved writing quality from improved reasoning assessed individually.

Conclusion

Assessment in Islamic Religious Education in the era of generative AI requires clearer grounds for establishing competence when assignment products can receive technological assistance. This review formulates four design requirements: alignment between evidence and competence, traceable religious sources, disclosure of AI assistance, and equitable learning opportunities. These requirements are translated into a cycle combining contextual tasks, process records, source verification, individual explanation, and reflective revision.

Epistemic responsibility is operationalised through the ability to identify the origin of claims, check references, provide reasons, and correct errors. Assessment concerns observable conduct and competence, without inferring learners' inner qualities. The framework and rubric are products of conceptual synthesis that can inform assessment development. Their effectiveness and feasibility must be established through implementation research before they support claims of improved learning.

References

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. Assessment & Evaluation in Higher Education, 31(4), 399–413. http://dx.doi.org/10.1080/02602930600679050

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239. http://dx.doi.org/10.1080/14703297.2023.2190148

Furze, L., Perkins, M., Roe, J., & MacVaugh, J. (2024). The AI Assessment Scale (AIAS) in action: A pilot implementation of GenAI-supported assessment. Australasian Journal of Educational Technology, 40(4), 38–55. http://dx.doi.org/10.14742/ajet.9434

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. Educational Technology Research and Development, 52(3), 67–86. http://dx.doi.org/10.1007/BF02504676

Hakim, A. (2026). Merancang asesmen PAI di era generative AI: Pergeseran dari penilaian produk menuju penilaian proses, reasoning, dan authenticity siswa. MODELING: Jurnal Program Studi PGMI, 13(1), 147–179. http://dx.doi.org/10.69896/modeling.v13i1.3215

Hidayat, R. H., & Fatmawati Nur Khasanah. (2025). The Role of Qur'anic Education Centre Teachers in Integrating Technology and Children's Character Education. *Guruku: Jurnal Pendidikan Profesi Guru*, *4*(2). https://doi.org/10.19109/guruku.v4i2.27975

Khobir, A., Sopiah, S., & Hasanah, F. N. (2021). A Holistic Model for Character Education in Schools (An Alternative Educational Model). *Edukasia Islamika*, *6*(2). https://doi.org/10.28918/jei.v6i2.4731

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers [Preprint]. arXiv. http://dx.doi.org/10.48550/arXiv.2304.02819

Najwa, L., & Nur Hasanah, F. (2026). Transformation of Digital Learning Evaluation to Support Assessment in The 21st Century. *Indonesian Journal of Islamic Elementary Education*, *6*(1), 56–70.

Nur Hasanah, F., & Kristiyah, K. (2023). The Effect of Using the Sharing Application (Tik-Tok) on the Character Development of Elementary School Children. *Attadrib: Jurnal Pendidikan Guru Madrasah Ibtidaiyah*, *6*(1). https://doi.org/10.54069/attadrib.v6i1.400Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. Journal of University Teaching and Learning Practice, 21(6). http://dx.doi.org/10.53761/q3azde36

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104, 333–339. http://dx.doi.org/10.1016/j.jbusres.2019.07.039

Sofyan, A., & Wahab, W. (2024). Development of Artificial Intelligence-Based PAI Learning Assessment: Opportunities and Challenges at MTs Durul Jazil. TA'DIBUNA: Jurnal Pendidikan Agama Islam, 7(2), 192–202. http://dx.doi.org/10.30659/jpai.7.2.192-202

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. International Journal of Educational Technology in Higher Education, 21, Article 40. http://dx.doi.org/10.1186/s41239-024-00468-z

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? International Journal of Educational Technology in Higher Education, 16, Article 39. http://dx.doi.org/10.1186/s41239-019-0171-0

Normative sources: The Quran, Al-Isra [17]:36 and Al-Hujurat [49]:6. Texts and comparison translations were accessed through https://quran.com/17/36 and https://quran.com/49/6 on 8 October 2026. Verses are cited by surah and verse number; the discussion uses paraphrase and pedagogical interpretation.