

Ethics and Epistemology of Generative AI in Local-Wisdom-Based Religious Moderation Research: Reflections from the JOGLO Model

Akhmad Farkhan

Universitas Islam Negeri K.H. Abdurrahman Wahid Pekalongan, Indonesia

E-mail: akhmadfarkhankua@gmail.com

Abstract

Generative artificial intelligence (AI) is increasingly used in religious studies research, yet its ethical and epistemological risks are rarely examined from users' direct experience, even though language models can produce fluent text with no grounding in real data. This study aims to document the benefits and the problematic incidents of using generative AI in developing JOGLO (Jagongan Slow), a participatory dialogue model for religious moderation in Tegal Regency, Indonesia, built on Javanese local wisdom and the song Tegal Toleransi, and to formulate an ethical protocol for AI use in Islamic studies research. The study employed an autoethnographic approach supported by document analysis of the author's interaction trail with AI assistants across the project. The results show that generative AI accelerated literature synthesis, conceptual framing, instrument design, and form automation, but also produced two patterns of problematic incident: the wholesale fabrication of findings, including statistics and quotations, for a forum that had never been held, and the presentation of social media reach as evidence of impact. Both were detected through manual cross-verification and simple conceptual questioning rather than automated checks. Read through *maqāṣid al-sharī'ah*, particularly *ḥifẓ al-‘aql* and scholarly trustworthiness as part of *ḥifẓ al-dīn*, these incidents threaten both sound reasoning and religious integrity. The study concludes with an eight-principle protocol, including mandatory human verification, a prohibition on writing findings before data exist, the separation of reach from impact, and transparent disclosure of AI involvement, offered as a reflective framework rather than a generalizable claim.

Keywords: generative artificial intelligence, research ethics, religious moderation, *maqāṣid al-sharī'ah*, academic integrity

Introduction

Generative artificial intelligence (AI), meaning systems that produce text, images, or code in response to natural-language prompts, has entered nearly every stage of academic work, from literature searching and synthesis to manuscript drafting (Lund et al., 2023; Salvagno et al., 2023). Its benefits are real, but so is its best-documented weakness: "hallucination," the tendency to generate fluent output that is not grounded in facts or data (Huang et al., 2025; Ji et al., 2023). In scholarly writing this weakness has already surfaced as fabricated references and invented

content (Alkaissi & McFarlane, 2023; Walters & Wilder, 2023). Some authors argue that “hallucination” understates the problem and that such output should be called fabrication and falsification (Emsley, 2023), or described as speech produced without regard for truth (Hicks et al., 2024).

The research-integrity response has so far been largely institutional. Publication-ethics bodies and leading journals have affirmed that AI tools cannot be authors and that human researchers remain fully accountable for AI-assisted content (Committee on Publication Ethics [COPE], 2023; Thorp, 2023; “Tools Such as ChatGPT,” 2023), while commentators have called for transparency and human verification (Hosseini et al., 2024; van Dis et al., 2023). These responses matter because fabrication has long been classed, with falsification and plagiarism, among the gravest forms of research misconduct (Fanelli, 2009). Generative AI adds a new dimension: fabrication may now occur not through a researcher’s intent but as a by-product of systems designed to produce plausible rather than verified text (Bender et al., 2021). First-person accounts of what happens when such tools go wrong inside an actual project, however, remain scarce; reflexive accounts of AI-assisted academic work have only begun to appear (Johnson & Paulus, 2024).

Islamic scholarship offers resources for this discussion that general AI-ethics guidelines, largely Western in origin, do not (Jobin et al., 2019). Recent work has proposed Islamic virtue ethics and maṣlaḥah-based benchmarking for AI (Elmahjub, 2023; Raquib et al., 2022). The maqāṣid al-sharī‘ah framework systematized by al-Shāṭibī (1997) identifies five universal objectives: the preservation of religion (ḥifẓ al-dīn), life (ḥifẓ al-nafs), intellect (ḥifẓ al-‘aql), lineage (ḥifẓ al-nasl), and property (ḥifẓ al-māl). Contemporary scholars have extended the framework beyond particular legal rulings into an instrument for analysing new problems (Attia, 2007; Auda, 2008), a position the author has also developed elsewhere (Farkhan, 2026b). Ḥifẓ al-‘aql, classically tied to the prohibition of intoxicants, can thus be extended to safeguarding reasoning that depends on accurate information. Scholarly honesty can likewise be read as part of ḥifẓ al-dīn, since the hadith tradition treats trustworthy transmission as a religious matter: Ibn Sīrīn’s dictum that “this knowledge is religion, so look to whom you take your religion from” and Ibn al-Mubārak’s statement that “the isnād is part of religion” both appear in the introduction (muqaddimah) to *Ṣaḥīḥ Muslim* (Muslim ibn al-Ḥajjāj, n.d.; see also Brown, 2009).

This study arose from the development of JOGLO (Jagongan Slow), a participatory dialogue forum for strengthening religious moderation in Tegal Regency, Central Java, rooted in Javanese local wisdom and opened symbolically by the song Tegal Toleransi (Farkhan, 2026a). The design belongs to a wider Indonesian effort to ground religious moderation in local culture (Hernawan et al., 2021; Kementerian Agama Republik Indonesia, 2019; Pajariantanto et al., 2022). One

clarification governs everything that follows. The song is real, widely heard, and has been the subject of field-based study (e.g., Allyusni et al., 2025). JOGLO, by contrast, remains a conceptual model: at the time of writing, not a single JOGLO session has been held.

Accordingly, this study asks three questions: (1) What concrete benefits did generative AI provide in a community-based religious research project? (2) How did problematic incidents arise, and how were they detected and corrected? (3) How can *maqāṣid al-sharī‘ah* serve as an ethical lens for these incidents and for a more responsible protocol? Its novelty lies in pairing a documented, first-person account of AI fabrication with an Islamic ethical framework, rather than offering policy review or speculation alone. As an interpretive study, it advances research questions rather than statistical hypotheses.

Methods

Research Design

The study used autoethnography, a qualitative method that treats the researcher’s personal experience as primary data for understanding wider social and cultural phenomena while maintaining critical reflexivity about the researcher’s dual role as subject and observer (Chang, 2008; Ellis et al., 2011). The approach fits because the research questions concern how the author himself interacted with generative AI and how the incidents were experienced and handled from a first-person perspective.

Participants

The only participant is the author, a doctoral student in Islamic law who developed the JOGLO design together with the song’s composer during 2026. No community members were involved and no field data were collected, because the forum has not been implemented. One series of interactions took place with a different AI assistant on a separate track outside the process the author directly oversaw; its outputs are treated as part of the documentary record.

Data Collection Technique

The primary data are the complete trail of conversations between the author and AI assistants across the project, including every manuscript draft, instrument, and program script, as well as early versions that were later revised. Secondary data comprise the literature on AI and research ethics and on *maqāṣid al-sharī‘ah* reviewed above.

Data Analysis Technique

Analysis proceeded in three stages: (1) chronological reconstruction of the stages of AI use, from initial literature synthesis to technical automation; (2) identification and categorization of problematic incidents from the documentary

trail; and (3) reflective reading of both through the *maqāsid* lens, which produced the protocol proposed in the Discussion. Consistent with the study's own argument, the fabricated figures, statistics, and quotations are not reproduced; the incidents are described by their pattern and structure, because reproducing fabricated content, even to criticize it, risks circulating it further. To limit the recall bias inherent in autoethnography, every statement about what happened was anchored to the documentary trail rather than to memory alone, and the author's interpretation was kept distinct from the record it interprets (Chang, 2008).

Results

Stages of AI Use

Chronological reconstruction identified six stages at which generative AI was involved (Table 1).

Table 1 Stages of generative AI use in the JOGLO project

Stage	AI-assisted task	Observed output
1	Synthesis of two earlier studies of the song Tegal Toleransi	Gap between the studies identified
2	Conceptual framework of JOGLO, including a tiered <i>maqāsid</i> structure	Coherent argumentative structure
3	Drafting manuscripts, first as a conceptual article and then in IMRAD format	Draft manuscripts
4	Survey instruments, focus group guides, and observation sheets	Instruments mapped to sub-questions
5	Automated construction of digital forms through scripts	Working digital forms
6	Separate track: interactions with another AI assistant	First pattern of incident

Observed Benefits

At Stages 1 to 5, generative AI delivered tangible benefits. It helped identify a gap between the two earlier studies of the song that had not been mapped explicitly, assembled a coherent argumentative structure in a short time, produced instrument designs with systematic coverage, including an explicit mapping between research sub-questions and instrument sections, and automated the construction of digital forms that would have taken far longer to build by hand.

First Pattern: Fabricated Findings for a Forum That Never Took Place

A supplementary document meant to provide additional material for the questionnaire and focus group guide turned out to contain a complete "Results and Discussion" section, with tables of quantitative scores and direct quotations attributed to categories of respondents (students, community figures, residents), although the JOGLO forum had never been held and no data had been collected. The output had three features. First, the figures and quotations were as precise as genuine research data rather than rough estimates, so they were hard to distinguish from real data at a glance. Second, the narrative followed a convincing argumentative sequence, including an abstract written as if the forum had been

implemented and evaluated. Third, the same document contained biographical and institutional claims that required independent verification.

Second Pattern: Reach Presented as Evidence of Impact

At a separate moment, the author asked another generative AI whether the “effectiveness” of the song Tegal Toleransi could be the topic of a paper for this conference. Instead of noting that an effectiveness claim requires impact data, the AI suggested presenting social media reach (view counts, reposts, derivative interactions) as “empirical evidence” of effectiveness. In the same exchange it proposed replicating the song in other regions through voice cloning and automatic lyric generation, without mentioning the composer’s consent to the use of his voice or the risk of shallow cultural fit if lyrics were generated without local stakeholders.

Detection and Correction

No automated system flagged the first incident. It was detected through human vigilance and manual cross-verification in four steps: (1) re-reading the original texts of the two earlier studies to confirm that no claim had been exaggerated or misquoted; (2) confirming that the reported figures had no basis, since the activity they described had never occurred; (3) independently tracing the biographical and institutional claims through web searches and visual and video documents; and (4) applying a precautionary rule: no figure or quotation from the document entered any later manuscript, even where the broader context it described proved credible. The second incident was detected through a single conceptual check, namely asking whether the proposed metric actually measured what it claimed to measure. In both cases correction was granular: claims that could be verified independently were retained, while the fictitious findings and the reach–impact conflation were rejected outright, however convincing their presentation.

Discussion

The first pattern is consistent with the hallucination literature (Huang et al., 2025; Ji et al., 2023), but it is more severe than the errors usually reported. Studies of AI-assisted writing have documented fabricated citations, for instance in 55% of references produced by GPT-3.5 and 18% of those produced by GPT-4 in one analysis (Walters & Wilder, 2023), and invented scientific content (Alkaissi & McFarlane, 2023). Here the model did not misreport an event that occurred; it created complete “results” for an event that never happened. This supports the argument that such output is better named fabrication than hallucination (Emsley, 2023). It also illustrates why plausibility itself is the danger: a system that generates text without regard for truth will produce findings as fluently as it produces anything else (Bender et al., 2021; Hicks et al., 2024).

The second pattern is subtler and therefore potentially more dangerous, because nothing is plainly invented. Reach measures how widely content has been exposed; impact measures change in attitude or behaviour. The bibliometrics literature has made the same point about altmetrics, whose counts register attention rather than demonstrated impact (Bornmann, 2014). The accompanying suggestion to clone the composer’s voice shows a related tendency: the AI offered scale and technology as added value without weighing consent, an issue that synthetic-media

scholarship treats as central (Chesney & Citron, 2019). In both cases the safeguard was human judgement, consistent with the emphasis on human oversight and accountability in COPE (2023) and UNESCO (2021).

Taken together, the results suggest a gradient of verifiability. The benefits clustered in tasks whose output the author could check against something that already existed: the texts of earlier studies, the logic of an argument, an instrument mapped to stated sub-questions, or a script that either runs or fails. The incidents arose where no such anchor existed, because there were no data against which invented findings could be compared, and no impact study against which reach could be weighed. This pattern is consistent with reports that AI assistance is most productive in the early, checkable stages of research (Lund et al., 2023; Salvagno et al., 2023), and it implies that the risk of a task depends less on its difficulty than on whether its output can be verified independently.

Read through *ḥifẓ al-‘aql*, the first incident threatens the integrity of the researcher’s reasoning: conclusions drawn from fabricated data would rest on an epistemologically corrupted foundation, a form of harm to the intellect broader than its classical meaning. Read through *ḥifẓ al-dīn* as the safeguarding of scholarly trustworthiness, the incident touches religion directly: research on religious moderation built on invented data would violate the honesty (*ṣidq*) and trust it seeks to promote. That research on honesty and harmony was nearly built on dishonest data is perhaps the most important lesson of this experience. On this basis the study proposes eight principles for AI use in Islamic studies research (Table 2).

Table 2 Proposed ethical protocol for generative AI use in Islamic studies research

Principle	Description
Mandatory human verification	Every factual claim, figure, or quotation produced by AI is traced to its original source before use.
No “results” before data	AI is never asked or allowed to write results or findings before primary data exist, even as an “example” or “draft.”
Transparency of AI involvement	Manuscripts state the extent and stages of AI involvement.
Separating facts from synthesis	Factual content (data, quotations, biographical claims) faces stricter verification than argumentative synthesis.
Granular caution	Problematic AI output is corrected claim by claim rather than accepted or rejected wholesale.
Reach is not impact	Engagement metrics are never treated as evidence of attitudinal or behavioural change without impact measurement.
Ethical review of proposed expansion	AI proposals for added technology (e.g., voice cloning) are reviewed for consent and cultural risk first.
Continuous reflection	Collaboration with AI is reviewed periodically during the project, not only at the final manuscript stage.

The protocol is not meant to replace existing guidance. COPE (2023) and UNESCO (2021) already establish human accountability and oversight at the level of principle, and reflexive workflows for AI-assisted writing have been proposed (Johnson & Paulus, 2024). What the protocol adds are operational rules that emerged from concrete failures, notably the prohibition on drafting results before data exist and the separation of reach from impact, which general guidelines do not spell out. It also changes their normative basis: grounded in *ḥifẓ al-‘aql* and *ḥifẓ al-*

dīn, verification becomes an expression of religious responsibility rather than mere compliance.

At a broader level, generative AI does not change the basic structure of Islamic scholarly epistemology, which places the sanad, the chain of transmission and accountability, at the centre of any knowledge claim (Brown, 2009). It demands a renewed application of that principle: every AI-generated claim needs its own “sanad” of verification before it is trusted or shared. The analogy is epistemological, not a claim that AI occupies the place of a hadith transmitter. It converges with the Islamic AI-ethics proposals noted above (Elmahjub, 2023; Raquib et al., 2022) and with international calls for human oversight (UNESCO, 2021).

The study has three limitations. First, as autoethnography, it reflects one researcher’s experience in one project, so generalization must be cautious. Second, the decision not to reproduce fabricated content limits readers’ ability to assess the incidents independently. Third, the protocol is preliminary and has not been tested in other Islamic studies projects; because JOGLO has not been implemented, the benefits reported concern the research process, not the forum’s outcomes. Future research could apply the protocol across several projects and compare incident rates with and without it.

Conclusion

Generative AI delivered real benefits in accelerating the development of the JOGLO model, but it also produced two serious patterns of incident: the wholesale fabrication of findings for a forum that had never been held, and the presentation of social media reach as evidence of impact. Both were caught by human verification and simple critical questioning, not by any automated safeguard. Read through maqāṣid al-sharī‘ah, particularly ḥifẓ al-‘aql and ḥifẓ al-dīn understood as scholarly trustworthiness, these incidents show that research integrity is not merely an administrative norm but part of the religious responsibility of the Muslim researcher. The eight-principle protocol proposed here is offered as an initial reference for Islamic studies researchers who use generative AI, with the conviction that technological progress should strengthen, not erode, the commitment to scholarly honesty at the heart of the Islamic scholarly tradition.

Acknowledgments

Acknowledgements, funding information and declaration of competing interests have been removed for blind review.

Declaration of generative AI use: Generative AI assistants were used during the JOGLO project as described in this article, and in preparing this manuscript for language editing, restructuring into IMRAD format, and reference checking. All content and references were verified by the author, who takes full responsibility for the article.

Conflict of interests

The author declares that there is no conflict of interest.

References

- Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2), e35179. <http://dx.doi.org/10.7759/cureus.35179>
- Allyusni, O., Basukiyatno, B., & Maufur, M. (2025). The impact of listening to the Tegalana song “Tolerance” on the attitude of religious moderation. *Journal of English Language and Education*, 10(1), 274–281. <http://dx.doi.org/10.31004/jele.v10i1.630>
- al-Shāṭibī, A. I. (1997). *Al-muwāfaqāt fī uṣūl al-sharī‘ah* (Vol. 2). Dār al-Kutub al-‘Ilmiyyah.
- Attia, G. E. (2007). *Towards realization of the higher intents of Islamic law: Maqasid al-Shariah, a functional approach* (N. Roberts, Trans.). International Institute of Islamic Thought.
- Auda, J. (2008). *Maqasid al-Shariah as philosophy of Islamic law: A systems approach*. International Institute of Islamic Thought.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. <http://dx.doi.org/10.1145/3442188.3445922>
- Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, 8(4), 895–903. <http://dx.doi.org/10.1016/j.joi.2014.09.005>
- Brown, J. A. C. (2009). *Hadith: Muhammad’s legacy in the medieval and modern world*. Oneworld.
- Chang, H. (2008). *Autoethnography as method*. Left Coast Press.
- Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1820. <http://dx.doi.org/10.15779/Z38RV0D15S>
- Committee on Publication Ethics. (2023, February 13). *Authorship and AI tools* [Position statement]. <https://publicationethics.org/node/56046>
- Ellis, C., Adams, T. E., & Bochner, A. P. (2011). Autoethnography: An overview. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 12(1), Article 10. <http://dx.doi.org/10.17169/fqs-12.1.1589>
- Elmahjub, E. (2023). Artificial intelligence (AI) in Islamic ethics: Towards pluralist ethical benchmarking for AI. *Philosophy & Technology*, 36, Article 73. <http://dx.doi.org/10.1007/s13347-023-00668-x>
- Emsley, R. (2023). ChatGPT: These are not hallucinations – they’re fabrications and falsifications. *Schizophrenia*, 9, Article 52. <http://dx.doi.org/10.1038/s41537-023-00379-4>

- Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, 4(5), e5738. <http://dx.doi.org/10.1371/journal.pone.0005738>
- Farkhan, A. (2026a). *JOGLO (Jagongan Slow): Model dialog partisipatif berbasis kesenian lokal untuk penguatan moderasi beragama di Kabupaten Tegal* [JOGLO (Jagongan Slow): A participatory dialogue model based on local arts for strengthening religious moderation in Tegal Regency] [Unpublished manuscript].
- Farkhan, A. (2026b). *Maqāṣid 'āmmah dan khāṣṣah* [General and specific maqāṣid] [Unpublished manuscript]. Universitas Islam Negeri K.H. Abdurrahman Wahid Pekalongan.
- Hernawan, W., Riyani, I., & Busro, B. (2021). Religious moderation in Naskah Wawacan Babad Walangsungsang: A Sundanese religious diversity wisdom. *HTS Teologiese Studies/Theological Studies*, 77(4), Article a6773. <http://dx.doi.org/10.4102/hts.v77i4.6773>
- Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, 26, Article 38. <http://dx.doi.org/10.1007/s10676-024-09775-5>
- Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (2024). Using AI to write scholarly publications. *Accountability in Research*, 31(7), 715–723. <http://dx.doi.org/10.1080/08989621.2023.2168535>
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55. <http://dx.doi.org/10.1145/3703155>
- 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. <http://dx.doi.org/10.1145/3571730>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <http://dx.doi.org/10.1038/s42256-019-0088-2>
- Johnson, C. W., & Paulus, T. M. (2024). Generating a reflexive AI-assisted workflow for academic writing. *The Qualitative Report*, 29(10), 2772–2792. <http://dx.doi.org/10.46743/2160-3715/2024.7634>
- Kementerian Agama Republik Indonesia. (2019). *Moderasi beragama* [Religious moderation]. Badan Litbang dan Diklat Kementerian Agama RI.
- Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial intelligence-written

- research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5), 570–581. <http://dx.doi.org/10.1002/asi.24750>
- Muslim ibn al-Ḥajjāj. (n.d.). *Ṣaḥīḥ Muslim* (M. F. ‘Abd al-Bāqī, Ed.; Vol. 1, Muqaddimah). Dār Iḥyā’ al-Turāth al-‘Arabī.
- Pajarianto, H., Pribadi, I., & Sari, P. (2022). Tolerance between religions through the role of local wisdom and religious moderation. *HTS Teologiese Studies/Theological Studies*, 78(4), Article a7043. <http://dx.doi.org/10.4102/hts.v78i4.7043>
- Raquib, A., Channa, B., Zubair, T., & Qadir, J. (2022). Islamic virtue-based ethics for artificial intelligence. *Discover Artificial Intelligence*, 2(1), 1–16. <http://dx.doi.org/10.1007/s44163-022-00028-2>
- Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care*, 27, Article 75. <http://dx.doi.org/10.1186/s13054-023-04380-2>
- Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, 379(6630), 313. <http://dx.doi.org/10.1126/science.adg7879>
- Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. (2023). *Nature*, 613(7945), 612. <http://dx.doi.org/10.1038/d41586-023-00191-1>
- UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
- van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, 614(7947), 224–226. <http://dx.doi.org/10.1038/d41586-023-00288-7>
- Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13, Article 14045. <http://dx.doi.org/10.1038/s41598-023-41032-5>