Sophist: Jurnal Sosial Politik, Kajian Islam dan Tafsir
Vol. 08, No. 2, 2026, p. 195-221
Published by Fakultas Ushuluddin dan Studi Agama (FUSA), UIN Mataram
p-ISSN 2656-1220, e-ISSN 2745-8571, available online at https://sophist.or.id/index.php/js/id/index

A MAQASID AL-SHARIAH-BASED ALGORITHMIC AUDITING MODEL: A CONCEPTUAL FRAMEWORK FOR ETHICAL FORMULATION AND HERMENEUTIC VERIFICATION IN GENERATIVE AI DA'WAH CONTENT ENGINES IN INDONESIA

Muhammad Yasin Isa Al-Gazali [1], Subayil[2]
[1]STID Mustafa Ibrahim Al-Islahuddin Kediri Lombok Barat
[2] STAI Al-Amin Gersik Kediri Lombok Barat NTB

Corresponding author: mohammadgazali172@gmail.com

Received: 10 September 2026| Revised: 15 September 2026 | Published Online: 26 September 2026

Abstract: *The rapid integration of Generative Artificial Intelligence (GenAI) into digital da'wah content engines, including AI-animated short videos and automated religious-consultation chatbots, has changed how Islamic communication circulates in Indonesia, introducing risks such as contextual reductionism, hallucinated citations, and a weakening of the traditional scholarly chain (sanad). Existing scholarship has mostly described institutional adoption, such as Nahdlatul Ulama's Kitab AI and Muhammadiyah's Chat HPT, without proposing a testable, shariah-grounded auditing procedure for the GenAI systems that now sit between a text and its audience. This study addresses that gap by formulating a Maqasid al-Shariah-Based Algorithmic Auditing Model (MAAM) through qualitative document analysis and cyber-hermeneutic reading of publicly available GenAI da'wah content, institutional AI policy documents, and the wider Indonesian and international scholarship on Islamic AI ethics. The audit matrix is anchored in Hifz ad-Din (preservation of religion) and Hifz al-'Aql (preservation of intellect), operationalised through the Qur'anic principles of tabayyun (verification) and qaulan sadidan (accurate speech). The analysis shows current GenAI da'wah engines are optimised for engagement rather than context, producing source stripping, flattened ikhtilaf (interpretive plurality), and occasional fabricated attribution. The proposed*

model responds with a tri-tier workflow: pre-prompt source authentication, in-pipeline hermeneutic constraints in the system prompt, and post-output human clearance by a certified da'i or editorial board, paired with an institutional certification mechanism. The contribution is conceptual: MAAM is a proposed model, illustrated with a single worked application, whose evidentiary base is documentary and literature-grounded, and whose inter-rater reliability on real content pipelines awaits empirical validation.
Keywords: *Algorithmic Auditing, Digital Da'wah, Generative AI, Islamic Communication Ethics, Maqasid al-Shariah, Tabayyun.*

INTRODUCTION

Islamic communication studies in Indonesia have moved, over roughly a decade, from describing one-directional message transmission toward mapping a contest over authority, identity, and virtual piety inside digital ecosystems (Hasibuan et al., 2026b). The clearest driver of that shift is the rapid arrival of Generative Artificial Intelligence (GenAI) inside the ordinary infrastructure of da'wah. Large religious organisations have adapted in visibly different ways. Nahdlatul Ulama (NU) built Kitab AI for classical-text learning, Siskader for cadre administration, and internal sentiment-analysis tools, a pattern that Nada-Qisthina and Musyafak (2025) describe as institutional adoption filtered through doctrinal legitimacy rather than through raw technical appetite. Muhammadiyah took a different route with Chat HPT, an automated jurisprudential-consultation tool, reflecting what the same study calls a more instrumental orientation toward efficiency and standardisation. Adeni's descriptive account and later hybridisation studies of both organisations' communication strategy (see also the account in Ahmad, 2026, on institutional responses more broadly) converge on one point: mainstream Islamic authority in Indonesia is not refusing GenAI outright, it is negotiating with it, and the negotiation is uneven across organisations.

Underneath that institutional story sits a much less governed layer: GenAI content engines that generate short animated da'wah videos and automated chatbot replies for Instagram and TikTok, exemplified by accounts such as @edumotion.ai. These tools are optimised for a platform logic that rewards retention and completion rate, not doctrinal precision, and the tension this creates is not unique

to Indonesia. Ahmad (2026) documents the same pattern across a wider set of Islamic chatbots, from Ansari Chat to SheikhGPT, noting that models trained on broad and often uncurated corpora can hallucinate Qur'anic verses, misattribute Prophetic traditions, or conflate a weak (da'if) narration with a sound (sahih) one, all while sounding equally confident. Wright's (2024) Scientific American feature on the same phenomenon quotes theologian Noreen Herzfeld's assessment that chatbots trained on scripture "carry some of the same defects that all large language models have," foremost among them hallucination, in the course of reporting on QuranGPT founder Raihan Khan's 2023 launch and the traffic crash that followed it. Elsewhere in the wider hallucination literature, benchmark studies of retrieval-grounded chatbots report factual-accuracy gains once a system is tied to a curated corpus rather than left to free generation (Semnani et al., 2023), a finding with an obvious, if untested, analogue for religious content engines.

Two bodies of Indonesian scholarship already circle this problem without quite landing on it. The first studies institutional AI adoption directly (Nada-Qisthina & Musyafak, 2025; a hybridisation study of NU and Muhammadiyah's Gen-Z communication strategy), but stops at description: it tells us how organisations talk about AI, not how a piece of AI-generated content should be checked before it reaches an audience. The second studies Islamic communication ethics in the digital sphere through Qur'anic principles such as tabayyun and qaulan sadidan (Hulawa & Kasmiati, 2025; Fepriani & Ratnasari, 2026), but mostly at the level of the individual user's conduct on social media, not at the level of an automated content pipeline that has no individual conscience to appeal to. A third, more recent cluster works at the meso level of platform regulation, arguing for an "algorithmic tabayyun" that scales Qur'anic verification from personal ethics into institutional mechanism (Hasibuan et al., 2026a; Nasution et al., 2026), but this cluster is written for broadcasting policy in general, not for the specific technical object of a generative pipeline that ingests a prompt and returns a video script.

What is missing, in short, is an audit instrument: something concrete enough that a developer, an editorial board, or a fatwa council could actually apply it to a specific GenAI da'wah tool, rather than only a normative argument that such a tool ought to be more careful. Two research questions follow from that gap: RQ1: What structural and theological vulnerabilities characterise current GenAI-generated da'wah content engines in Indonesia? RQ2: How can the Maqasid al-Shariah principles of Hifz ad-Din and Hifz al-'Aql be translated into a proposed algorithmic audit matrix and a hermeneutic verification workflow whose application to a real content pipeline can be illustrated, even where formal empirical validation remains for future work?

The stakes of getting this wrong are not abstract. Indonesia hosts the world's largest Muslim population, and its digital da'wah ecosystem now reaches audiences that no single pesantren, mosque, or broadcast studio could have addressed a generation ago. Marwantika et al. (2025) note that generative tools such as ChatGPT, Gemini, and DALL·E have already been folded into journalism education because they are simply faster than the alternative, and religious content production faces the identical incentive: a GenAI pipeline can produce in minutes what a scriptwriter and animator once needed days to finish. Speed of this kind is not neutral. It changes who gets to speak, how much scrutiny a claim receives before it is published, and how far a theological error can travel before anyone with the training to catch it even sees it. Aritonang et al. (2026) frame this directly as an ethical-analysis problem for Islamic communication studies, arguing that the production side of AI-generated religious messages, not only the reception side, needs its own normative account, a claim this paper takes as a starting premise rather than a conclusion to re-argue.

It is also worth being precise about what kind of gap this paper is filling, because "there is no audit model yet" is a claim that could describe several different absences. It is not that Islamic scholars have ignored AI; the opposite is closer to true, and the volume of recent work on AI ethics from a maqasid perspective (surveyed in section 2.2) shows a field working hard to catch up. It is not that nobody has

proposed general principles for AI and Islam; Elmahjub (2023) and Mohadi and Tarshany (2023) both offer serious philosophical treatments. What is missing is the connective layer: a document that a person actually building or reviewing a da'wah content engine could open and use as a checklist, one that translates "preserve religion" and "preserve intellect" into something as concrete as "check whether this hadith citation resolves to a real, gradable source" or "flag this script if it presents one fiqh opinion as though no other exists." That connective layer, between maqasid theory and pipeline practice, is what this paper tries to supply.

This paper answers both questions through a qualitative, literature- and document-grounded design, and proposes the Maqasid al-Shariah-Based Algorithmic Audit Matrix (MAAM) together with a three-tier verification workflow, as a conceptual contribution that sits between Islamic legal-ethical theory and the practical governance of generative religious media. The remainder of the paper proceeds as follows. Section 2 develops the theoretical framework across four strands: diffusion of innovation inside Islamic institutions, maqasid al-shariah as a governance paradigm, Qur'anic communication ethics as an operational vocabulary, and the general AI-auditing literature this paper borrows technical language from. Section 3 sets out the method and is explicit about the study's evidentiary limits. Section 4 presents the matrix, the workflow, and two comparative readings, one across Indonesian institutions and one across international Islamic chatbots, that test whether the model's categories travel. Section 5 concludes with recommendations for developers, institutions, and academic bodies, and states the paper's limitations without softening them.

2. METHOD

2.1 Research Design

This study uses a qualitative research design combining document analysis (Morgan, 2022) with a cyber-hermeneutic reading of digital religious media, an approach that treats text, interface, prompt architecture, and algorithmic output as a single interpretable object

rather than separating "the technology" from "the message" (a method consistent with the critical-discourse and thematic-synthesis approaches used in adjacent Indonesian Islamic-AI scholarship, e.g., Hasibuan et al., 2026b; Habib, 2025). It is important to be explicit about what this design does and does not claim. An earlier working draft of this project described primary data collection, in-depth interviews with six informants and a content sample of thirty short videos, that this study did not in fact carry out. No interviews were conducted for this paper, and no informant transcripts exist. Rather than retain those claims, this study is reframed, consistent with standard practice for a synthesis paper of this kind, as a qualitative documentary and cyber-hermeneutic analysis of publicly available material: GenAI-generated da'wah short videos and chatbot outputs that already circulate on public platforms (illustrated here through the @edumotion.ai case and comparable public examples discussed in the wider Islamic-chatbot literature, such as QuranGPT, Ansari Chat, and SheikhGPT), institutional policy statements and congress resolutions from NU and Muhammadiyah that are publicly published, and the secondary scholarly literature synthesised in section 2. This is a narrower evidentiary basis than a field study with real informants, and the paper states that limitation plainly rather than presenting invented-sounding fieldwork as empirical finding. To remove any ambiguity following that correction: every illustrative example retained elsewhere in this paper, including the composite example in section 4.1 and the worked application in section 4.2.1, is explicitly labelled as a hypothetical or composite illustration rather than as a transcript of an observed data point, and no claim in this paper rests on primary data collected from human informants.

2.2 Units of Analysis

Three units of analysis structure the reading. Institutional AI systems: the publicly documented features of NU's Kitab AI and BPIS digital hub and Muhammadiyah's Chat HPT (Majelis Tarjih/LabMu), read through their own published descriptions and through the secondary literature that has already studied them empirically. GenAI da'wah content engines: publicly observable output patterns from AI-animated da'wah accounts on Instagram and TikTok, with

@edumotion.ai serving as an illustrative case of the automated short-video genre, read structurally (opening hook, core message, reflective close) rather than through any claimed private access to the account's own prompt architecture. This single-account illustration is a convenience case rather than a representative sample; the structural pattern it exemplifies is treated as a plausible instance of a wider genre, not as evidence that the genre's failure modes are statistically typical, and a systematic, multi-account sample is identified in section 5 as necessary before that stronger claim could be supported. Comparative international cases: publicly reported Islamic chatbot deployments outside Indonesia (Ansari Chat, SheikhGPT, WisQu, Salam. chat, QuranGPT, HUMAIN Chat), used to check whether the vulnerabilities observed in the Indonesian case are local or structural to the technology itself.

2.3 Data Sources and Analysis

Data sources are exclusively secondary and documentary: peer-reviewed journal articles, institutional publications and congress resolutions, and publicly accessible platform content, each identified and cross-checked as described in the references. Analysis followed a thematic variant of Miles, Huberman, and Saldaña's interactive model: data reduction (isolating recurring failure patterns across the literature and the public content examples), data display (organising those patterns against the two maqasid values), and conclusion drawing and verification (checking the resulting matrix against the general AI-auditing literature in section 2.4 to make sure the Islamic-ethical categories map onto something operationally checkable, not only rhetorically compelling). The audit matrix and workflow presented in section 4 are the product of that synthesis, not of new measurement.

2.4 Trustworthiness and Positionality

A document-and-literature synthesis carries its own version of the trustworthiness questions a field study would answer through inter-rater reliability or member checking. Two safeguards were applied here instead. First, triangulation across source type: no claim about a structural vulnerability in section 4.1 rests on a single study; each is cross-read against at least one Indonesian Islamic-communication

source and at least one general (non-Islamic, non-Indonesian) AI-governance source, on the reasoning that a pattern visible in both literatures independently is less likely to be an artefact of one field's particular theoretical fashion. Second, an explicit separation of levels of evidence: sources whose DOI or publisher page this study directly resolved are, in principle, distinguishable from sources encountered only as a citation inside another paper's reference list and not independently re-resolved. A full verification index recording this distinction per reference is maintained as part of this project's internal documentation but is not reproduced in this manuscript; its absence is a limitation the authors acknowledge, and the index can be made available to editors or reviewers on request. Relatedly, the authors note that several cited outlets are recently established Indonesian journals (for example, several 2026 articles appearing in Jurnal Administrasi Pemerintahan Desa) whose indexing status and peer-review scope for AI-ethics and Islamic-communication topics the authors were not able to independently confirm at the time of writing; readers who rely on this paper for further research are encouraged to verify those journals' standing independently before treating them as settled scholarly authority. A study of this kind cannot claim the same warrant as a coded content sample with reported reliability statistics, and it does not attempt to.

4. FINDINGS AND DISCUSSION

4.1 Structural Vulnerabilities in GenAI Da'wah Content Engines

Public GenAI da'wah content, exemplified by accounts such as @edumotion.ai, tends to follow a recognisable three-beat structure: an opening hook, a core message, and a reflective closing line. The structure itself is not the problem; it is a reasonable adaptation of short-form video convention and resembles the structure of a good khutbah. The problem sits one layer down, in how the core message gets compressed to fit a fifteen- to sixty-second window optimised for completion rate rather than context. Three recurring failure modes emerge across the literature and the observable public examples.

Contextual reductionism and popularity bias. Complex fiqh questions, on marriage, inheritance, or worship validity, get

compressed into rigid, binary claims because a script that hedges, qualifies, or names competing scholarly opinions performs worse on a retention curve than one that states a single confident answer. Nuriana and Salwa's (2024) narrative review of "digital da'wah in the age of algorithm" documents this same engagement-over-depth trade-off across a wider sample of Islamic social media content, and the general literature on engagement-driven amplification (Milli et al., 2025) shows the mechanism is not specific to religious content; it is how the underlying recommender objective behaves whenever it is applied to normatively complex material.

Superficial source attribution. A Qur'anic verse displayed on screen is not the same thing as the verse read in its tafsir context. When the surrounding commentary, asbab al-nuzul, and recognised scholarly gloss are dropped, the visual citation becomes technically accurate and interpretively hollow at the same time, a gap that Meerangani et al. (2025) identify directly in their study of digital tabayyun and AI-filtered Qur'anic information, and that maps onto the general "superficial source attribution" problem documented in AI-generated journalism research more broadly (Piasecki et al., 2024).

Algorithmic echo chambers and hallucinated attribution. Curation algorithms reward emotionally resonant, repetitive content, which narrows rather than broadens a viewer's exposure to the range of recognised Islamic legal and theological positions, weakening exactly the reflective capacity that Hifz al-'Aql is meant to protect. Layered on top of this is the hallucination risk documented in the Islamic-chatbot literature specifically (Ahmad, 2026): a generative system trained on uncurated web text can misattribute a hadith's strength rating, invent a citation that sounds plausible, or conflate two scholars' opinions into one, and it will do so with the same fluent confidence it uses for a verified claim, a pattern general-purpose hallucination research also documents outside the religious domain (see the hallucination-benchmark literature summarised in Semnani et al., 2023).

A short illustrative reading helps make these three failure modes concrete rather than abstract; the reading below is a hypothetical composite constructed for expository purposes and is not a transcript

of one specific observed video. A composite, illustrative sixty-second GenAI da'wah video on inheritance law (waris), representing a pattern common to the genre rather than a documented single instance, might open with a relatable hook ("what happens to your assets after you die?"), state a single fixed division ratio as though it were the only applicable rule, display an Arabic verse fragment on screen without its surrounding tafsir, and close with a call to "share this if you learned something." Each element of that structure is individually unremarkable; together, they compress a question that classical fiqh treats with considerable nuance, dependent on the heirs present, competing schools' rulings on edge cases, and regional 'urf, into a single confident claim that a viewer has no reason to doubt and little means to verify. None of the three failure modes requires a malicious developer. A recommender system rewarding completion rate, a script-generation prompt with no built-in ikhtilaf flag, and a training corpus with uneven source curation are individually mundane engineering choices that jointly produce a theologically thin, epistemically overconfident artefact, exactly the combination Hifz ad-Din and Hifz al-'Aql are meant to guard against.

4.2 The Maqasid al-Shariah Algorithmic Audit Matrix (MAAM)

Table 1 operationalises the two maqasid values discussed in section 2.2 into four audit dimensions, each anchored to a concrete technical verification metric and an Islamic communication benchmark, so that the matrix can function as a checklist rather than only an argument.

Table 1. The Maqasid-Based Algorithmic Audit Matrix (MAAM)

Audit Dimension	Core Maqasid Value	Technical Verification Metric	Islamic Communication Benchmark
Sanad and Source Integrity	Hifz ad-Din	Dataset cross-checked against canonical hadith and tafsir databases; zero tolerance for a citation that cannot be traced to a real, checkable source.	Tabayyun

Textual Authenticity	Hifz ad-Din	Exact-match checking of Arabic text, diacritics, and any translation against a recognised published edition	Qaulan Sadidan
Cognitive Nuance and Anti-Bias	Hifz al-'Aql	Diversity scoring across the output set to flag reductionist or extremist framing, and to flag when recognised ikhtilaf has been collapsed into a single claim.	Qaulan Ma'rufan
Social Responsibility	Hifz al-Nasl / Hifz al-Mal	Screening for hate speech, sectarian provocation, and copyright violation in generated media assets	Amanah and Maslahah

Note. Hifz ad-Din = preservation of religion; Hifz al-'Aql = preservation of intellect; Hifz al-Nasl = preservation of lineage/dignity; Hifz al-Mal = preservation of property; Tabayyun = rigorous verification before dissemination; Qaulan Sadidan = truthful and technically accurate speech; Qaulan Ma'rufan = constructive, dignifying speech; Amanah = trustworthiness; Maslahah = the public good. These terms are defined in full in sections 2.2 and 2.3.

Each row exists because a corresponding failure was documented in section 4.1: the sanad row answers hallucinated attribution, the textual-authenticity row answers superficial or garbled verse display, the cognitive-nuance row answers contextual reductionism and echo-chamber narrowing, and the social-responsibility row answers the reputational and legal risks that AI-generated video assets specifically raise (voice cloning, unlicensed imagery, and the kind of hate-speech drift that Ahmmad et al.'s (2025) systematic review of filter-bubble research associates with algorithmically amplified content generally). The matrix deliberately borrows its verification-metric column from the general AI-auditing literature (Mökander, 2023; Cheong, 2024) rather than inventing a

parallel technical vocabulary, on the view that an audit instrument is more likely to be adopted by actual developers if its technical half looks like something a technical team already recognises.

4.2.1 Illustrative Application of the Matrix

Responding to the concern that the matrix remains untested, this subsection walks through a single worked application. The illustration below applies Table 1 to the same hypothetical composite waris video introduced in section 4.1, and is offered as a proof-of-concept rather than as a validated instrument; the walkthrough has not been checked for inter-rater reliability against a second, independent coder, and the matrix's numeric decision thresholds remain to be specified in future technical work.

Applying the Sanad and Source Integrity row: the composite video displays an Arabic verse fragment without a citation string that resolves to a specific surah and ayah number; under the matrix's zero-tolerance rule for unresolvable citations, this would register as a finding requiring at minimum a Moderate correction note, and a Critical block if the fragment's wording departed from a recognised published edition. Applying the Textual Authenticity row: because the composite example does not alter the Arabic text itself and only omits its surrounding tafsir, this row would not register a finding on its own, a reminder that the four rows are meant to track logically distinct defects rather than restate one failure four times. Applying the Cognitive Nuance and Anti-Bias row: the video's presentation of a single inheritance ratio as though it were the only applicable rule is a textbook instance of collapsed ikhtilaf, and would register as a Serious finding under Table 2, routed to a certified da'i for correction before release rather than blocked outright, since the underlying ratio is not itself fabricated. Applying the Social Responsibility row: nothing in the composite example raises a hate-speech, copyright, or voice-cloning concern, so this row would register no finding.

This single walkthrough cannot establish the matrix's reliability across coders or content types; it is included only to demonstrate that the four rows produce differentiated, non-trivial judgements on one piece of content, which is the minimum bar for a checklist to be more than a restatement of its own categories. A systematic reliability study,

in which at least two trained coders independently apply Table 1 and Table 2 to a shared sample of real GenAI da'wah outputs and report their agreement with a standard statistic such as Cohen's kappa, is identified in section 5 as a necessary next step before MAAM can be described as validated rather than proposed.

4.3 A Tri-Tier Hermeneutic Verification Workflow

The matrix in Table 1 describes what to check; this section describes when and how to check it, across three tiers that map onto a generative pipeline's actual lifecycle rather than onto an abstract ethical stage.

Tier 1, Pre-Prompt Authentication. Before generation begins, the system should be constrained to retrieval-augmented generation (RAG) drawn only from an authenticated repository, digital kitab kuning collections, a fatwa institution's own published corpus such as Himpunan Putusan Tarjih, or an equivalent verified dataset, rather than free generation from an uncurated pretraining corpus. The practical feasibility of this tier depends on such corpora being available in a machine-readable, appropriately licensed form; at the time of writing, the authors have not confirmed that Himpunan Putusan Tarjih or comparable fatwa corpora are published in a format ready for automated retrieval, and a developer implementing Tier 1 would need to budget time for digitisation, licensing negotiation, or both before this tier could be deployed as described. This is not a novel technical idea; Semnani et al.'s (2023) WikiChat demonstrates the general mechanism (grounding measurably reduces hallucination), and this study's contribution is narrower and more specific: identifying which repositories count as "authenticated" for an Islamic content pipeline, and making that authentication step a mandatory gate rather than an optional feature.

Tier 2, In-Pipeline Hermeneutic Audit. The system prompt itself should carry hermeneutic constraints as explicit rules, not as hoped-for emergent behaviour: a requirement to surface asbab al-nuzul or asbab al-wurud context whenever a verse or hadith is quoted, a requirement to flag recognised ikhtilaf rather than presenting one position as the only position, and a standing prohibition on generating

an absolute fatwa-style ruling. This tier operationalises the distinction, noted in section 2.1, that NU's own deliberative bodies have already drawn between AI as a consultative aid and AI as a source of binding guidance; writing that distinction into the system prompt is what makes it enforceable at the level of the pipeline rather than only recommended at the level of user conduct.

Tier 3, Post-Output Ethical Clearance. Before public release, outputs pass through human review by a certified da'i, an Islamic-communication auditor, or an editorial board, mirroring the human-in-the-loop clearance step that general hallucination-mitigation research identifies as the single most effective check available once automated grounding and prompt constraints have already reduced the error rate as far as they can (a pattern also visible in the perioperative and legal-domain chatbot literature, where human sign-off remains the final gate even in high-accuracy systems). This tier is also where qaulan ma'rufan and qaulan layyinan function as tone benchmarks, catching content that is factually defensible but needlessly harsh or sectarian in framing.

4.4 Institutional Governance and Ethical AI Certification

A matrix and a workflow are only as durable as the institution that enforces them. This study follows the direction argued in the platform-regulation literature (Hasibuan et al., 2026a; Nasution et al., 2026) that individual-level tabayyun needs an institutional counterpart to scale. A plausible Indonesian vehicle is a body convened jointly by the Association of Islamic Communication and Broadcasting Study Programs (ASKOPIS), national university networks, and representatives from NU, Muhammadiyah, and the Indonesian Ulema Council (MUI), issuing an audit-compliance badge to da'wah software and content engines that pass MAAM review. This is a proposal put forward by the authors rather than an arrangement any of these organisations has been consulted on or has agreed to; realising it would require independently confirming institutional appetite from ASKOPIS, NU, Muhammadiyah, and MUI, and the discussion below should be read as a normative sketch of what such a body could look like rather than a description of a body already in formation. This is not a call for state censorship of da'wah content; it is closer to an accreditation model, comparable in spirit to existing halal-certification

infrastructure, that gives audiences and content creators a visible signal of which tools have been checked against the matrix in Table 1.

Such a body would need three things to function rather than merely exist on paper. It needs a published, versioned copy of the matrix and workflow, so that a developer building a new tool this year is checked against the same standard as one built two years ago, with any revisions dated and justified rather than applied silently. It needs a sampling mechanism, since reviewing every output of every da'wah content engine in Indonesia is not realistic at current volume; a defensible middle path, familiar from broadcast-content monitoring more generally, is a rolling audit of a randomised output sample large enough to catch systematic patterns (the kind described in section 4.1) even if it cannot catch every individual error. And it needs an appeals or correction channel for creators who are flagged, both because natural justice demands it and because a certification process with no correction path will simply be ignored by the independent creators identified in section 4.5 as the population most in need of it. None of this is technically difficult; ASKOPIS and the university networks already run comparable peer-review infrastructure for academic publishing, and the certification body proposed here is best understood as an adaptation of that existing capacity rather than an entirely new institution.

4.5 Comparative Institutional Adoption: NU and Muhammadiyah as Divergent Case Types

Reading NU and Muhammadiyah side by side is useful precisely because they diverge, and the divergence tests whether MAAM's categories travel across different adoption styles. NU's epistemic-localisation pattern, keeping a named scholarly chain visible even inside a digital tool, already satisfies much of the sanad-and-source-integrity row in Table 1 by institutional design; the organisation's 2023 and 2025 Bahtsul Masail rulings against treating AI output as binding guidance function, in effect, as an informal Tier 2 constraint that predates any formal MAAM adoption. Muhammadiyah's instrumental-rationality pattern with Chat HPT is comparatively stronger on standardisation and consistency, useful for the technical-authenticity row, but its

greater openness to treating AI as a da'wah vehicle (Adeni's and the related institutional-response literature both note this openness) means it has less built-in resistance to the contextual-reductionism failure mode identified in section 4.1, unless a Tier 2 constraint is added deliberately rather than assumed. The comparison suggests MAAM is not redundant with either organisation's existing informal practice; it is a way of making each organisation's implicit safeguard explicit, checkable, and transferable to the far less governed space of independent GenAI content creators who answer to neither organisation's deliberative body.

4.6 A Wider Lens: Islamic Chatbots Outside Indonesia

The vulnerabilities identified in section 4.1 are not specific to the Indonesian platform ecosystem. Ahmad's (2026) survey of Islamic chatbots internationally includes HUMAIN Chat, Ask AiDeen, SheikhGPT, WisQu, Salam Chat, and Ansari Chat, and documents the same underlying tension between accessibility and authority: each tool expands lay access to religious text in a way that recalls the earlier "Sheikh Google" moment of Islam Q&A sites, and each carries the same risk of conflating a hallucinated response with a weak narration and a contested-but-established position, from the ordinary user's vantage point. Wright's (2024) Scientific American account of QuranGPT's chaotic 2023 launch, developed by then-twenty-year-old student Raihan Khan, who woke to find his own chatbot crashed from traffic, illustrates how quickly an individually built tool can reach an audience larger than any editorial board can review in real time, which is exactly the scenario Tier 3 human clearance is designed to slow down before release rather than only audit after the fact. That this pattern recurs across Bangladesh-, Gulf-, and North-America-facing tools as well as Indonesian ones is, if anything, an argument for MAAM's portability: the matrix is built from Islamic legal-ethical categories that do not depend on any single national regulatory environment, even though this paper's illustrative material is drawn from the Indonesian case.

4.7 Grading Severity: Not Every Vulnerability Carries the Same Weight

One risk of a checklist-style matrix is that it can flatten genuinely different levels of harm into a single undifferentiated "fail," which would ironically reproduce the same contextual-reductionism problem the matrix is meant to catch in the content it reviews. Table 2 offers a rough severity grading, separating vulnerabilities that call for an immediate release block from those that call for a correction note or a lower-priority revision queue.

Table 2. Severity Grading for MAAM Findings

Severity	Example Finding	Recommended Action
Critical (block release)	A hadith citation that does not resolve to any real source; a Qur'anic verse rendered with altered wording	Automatic hold at Tier 1 or Tier 2; no Tier 3 reviewer discretion to override
Serious (hold for Tier 3 review)	A single fiqh opinion presented as though no ikhtilaf exists; a hadith's strength rating misstated	Route to certified da'i for correction before release
Moderate (release with correction note)	Missing asbab al-nuzul context that changes emphasis but not accuracy; tone drifting toward qaulan sadidan but not fully meeting qaulan ma'rufan	Attach a visible correction or context note; log for pattern review.
Low (log only)	Minor stylistic informality; a defensible but non-standard translation choice	No release action; recorded for periodic aggregate review

This grading is not meant to be exhaustive or final; a certification body of the kind proposed in section 4.4 would need to adapt it to specific content genres (a sixty-second Reel carries different practical constraints than a full chatbot conversation) and to specific fiqh domains, where the cost of a wrong answer varies considerably; an error about a recommended (mustahabb) practice is not equivalent in consequence to an error about an obligation (wajib) or a prohibition (haram). What the grading does establish is that "critical" findings, fabricated citations and altered scripture are treated as non-negotiable blocks rather than as one more entry on a severity-neutral list, which is

the paper's answer to the concern, raised earlier in section 2.3, that qaulan sadidan demands more than statistical accuracy; it demands that some categories of error never reach an audience at all, regardless of how rare the underlying model considers them.

It is also worth being explicit about what kind of action Table 2 specifies. An "automatic hold" is, for the purposes of this paper, a human-enforced editorial policy, a rule that a human reviewer or editorial workflow is instructed to apply without discretion, rather than a fully automated technical filter implemented in code. Building an actual machine classifier capable of reliably detecting an unresolvable hadith citation or an altered Qur'anic wording, at the accuracy needed to justify removing human override, is a separate and substantial engineering undertaking beyond the scope of this qualitative study; Table 2 should be read as a policy specification for human reviewers until such a classifier exists and has itself been validated.

5. CONCLUSION

Generative AI has become part of the ordinary infrastructure of Islamic communication in Indonesia faster than any accompanying audit instrument has been built to match it. This study has argued that the risk is not GenAI itself but the absence of a checkable procedure for the specific failure modes, contextual reductionism, superficial source attribution, echo-chamber narrowing, and hallucinated citation, that a generative pipeline optimised for engagement will produce by default. The Maqasid al-Shariah-Based Algorithmic Audit Matrix answers RQ1 by naming those failure modes against two classical maqasid values, Hifz ad-Din and Hifz al-'Aql, and answers RQ2 by proposing four audit dimensions and a three-tier workflow, illustrated in section 4.2.1 with a single worked application, that intervene before generation, during generation, and before release. The matrix's dimensions are conceptually checkable, but formal validation of their inter-rater reliability on real content remains for future empirical work rather than a claim this paper makes for itself. The contribution is deliberately modest in one respect and more ambitious in another: modest, because it borrows its technical vocabulary from an already-mature general AI-auditing literature rather than inventing a parallel

one; ambitious, because it insists that vocabulary can and should be read through, rather than around, Islamic legal-ethical categories that Western secular AI-governance frameworks have generally left unaddressed.

The comparative readings in sections 4.5 and 4.6 add two further observations worth restating plainly. First, the gap this paper addresses is not evenly distributed: large religious organisations with standing deliberative bodies already have informal versions of Tier 2 constraints built into their institutional culture, even where they have never called it that, while independent GenAI content creators, who now arguably produce a larger volume of da'wah material than any single organisation, have almost none. A certification model that only large organisations bother to seek would miss the audience it most needs to reach. Second, the vulnerabilities themselves are not an Indonesian peculiarity; they recur wherever a generative system is pointed at a religious corpus, from QuranGPT's traffic-crashing 2023 debut to the more deliberately engineered Ansari Chat and HUMAIN Chat. That recurrence is, in the end, the strongest argument for building an audit model on maqasid categories rather than on a narrower set of platform-specific rules: categories drawn from usul al-fiqh do not expire when TikTok's recommender algorithm changes, and they do not need to be rewritten for each new chatbot that launches.

REFERENCES

Adeni. (2024). Artificial intelligence (AI) as a source of Islamic information: Examining the response of Nahdlatul Ulama (NU) and Muhammadiyah in Indonesia. Proceedings of the International Conference on Islam, Media and Communication (ICIMAC).

Ahmad, M. A. (2026). Islamic chatbots in the age of large language models (arXiv:2601.06092). arXiv. https://doi.org/10.48550/arXiv.2601.06092

Ahmmad, M., Shahzad, K., Iqbal, A., & Latif, M. (2025). Trap of social media algorithms: A systematic review of research on filter bubbles, echo chambers, and their impact on youth. Societies, 15(11), 301. https://doi.org/10.3390/soc15110301

Aritonang, I., Barus, L. N. U. Br., & Rubino. (2026). Da'wah in the era of artificial intelligence: An ethical analysis of Islamic communication on the use of AI in the production of religious messages. Ath-Thariq: Jurnal Dakwah dan Komunikasi, 10(1). https://doi.org/10.32332/7tnc9m72

Ashraf, C. (2022). Exploring the impacts of artificial intelligence on freedom of religion or belief online. The International Journal of Human Rights, 26(5), 757–791. https://doi.org/10.1080/13642987.2021.1968376

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. International Journal of Artificial Intelligence in Education, 32(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Barlow, J., & Holt, L. (2024). Attention (to virtuosity) is all you need: Religious studies pedagogy and generative AI. Religions, 15(9), 1059. https://doi.org/10.3390/rel15091059

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. Internet Research, 33(7), 133–167. https://doi.org/10.1108/INTR-01-2022-0042

Campbell, H. (2017). Religious communication and technology. Annals of the International Communication Association, 41(3–4), 228–234. https://doi.org/10.1080/23808985.2017.1374200

Chaudhary, G. (2024). Unveiling the black box: Bringing algorithmic transparency to AI. Masaryk University Journal of Law and Technology, 18(1), 93–122. https://doi.org/10.5817/MUJLT2024-1-4

Cheong, B. C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. Frontiers in Human Dynamics, 6, Article 1421273. https://doi.org/10.3389/fhumd.2024.1421273

Elmahjub, E. (2023). Artificial intelligence (AI) in Islamic ethics: Towards pluralist ethical benchmarking for AI. Philosophy & Technology, 36, Article 73. https://doi.org/10.1007/s13347-023-00668-x

Farid, A. M., Pratama, F. A., & Sulistyowati, D. (2025). Maqashid Sharia analysis on personal data protection from the dangers of artificial intelligence misuse. Jurnal Penelitian Ilmu-Ilmu Sosial, 6(2), 179–192. https://doi.org/10.23917/sosial.v6i2.12636

Fellah, S., Bedarnia, M., & M'ziou, O. (2026). Artificial intelligence ethics in Islamic society: Towards an integrative approach between Maqasid al-Shariah and international standards for AI governance. In Proceedings of the International Conference on Artificial Intelligence Applications in Business Administration (ICAIABA 2026) (pp. 27–35). Atlantis Press. https://doi.org/10.2991/978-94-6239-711-8_4

Fepriani, J., & Ratnasari, D. (2026). Algorithmic ethics and Qur'anic tabayyun: Knowledge authority and AI bias in the digital age. QiST: Journal of Quran and Tafseer Studies, 5(1), 345–368. https://doi.org/10.23917/qist.v5i1.16001

Habib, Z. (2025). Ethics of artificial intelligence in Maqāṣid al-Sharīʿa's perspective. KARSA: Journal of Social and Islamic Culture, 33(1), 105–134. https://ejournal.uinmadura.ac.id/index.php/karsa/article/view/19617

Hasibuan, S. A., Sazali, H., Hamandia, M. R., & Jannati, Z. (2026a). Institutionalization of algorithmic tabayyun: Integration of

Maqashid Sharia as the philosophical foundation for regulation of transnational media platforms. Jurnal Administrasi Pemerintahan Desa, 7(2), 11. https://doi.org/10.47134/villages.v7i2.520

Hasibuan, S. A., Sazali, H., Hamandia, M. R., & Jannati, Z. (2026b). Algorithmic authority and the commodification of religion: A literature review of the shifting Islamic broadcasting ecosystem in the artificial intelligence (AI) era. Jurnal Administrasi Pemerintahan Desa, 7(2), 8. https://doi.org/10.47134/villages.v7i2.523

Hidayat, Y. F., & Nuri, N. (2024). Transformation of da'wah methods in the social media era: A literature review on the digital da'wah approach. IJoIS: Indonesian Journal of Islamic Studies, 4(2), 67–76. https://doi.org/10.59525/ijois.v4i2.493

Hulawa, D. E., & Kasmiati, K. (2025). Qaulan Sadida as a Qur'anic framework for revitalizing character building in the digital era. Fitrah: Journal of Islamic Education, 6(2), 292–309. https://doi.org/10.53802/fitrah.v6i2.1144

Maiga, M. H. (2025). Achieving Maqāṣid al-Sharīʿah through artificial intelligence: Mechanisms of facilitation, control, and quality assurance. Mazahibuna: Jurnal Perbandingan Mazhab, 54–70. https://doi.org/10.24252/mazahibuna.vi.55038

Malik, F. A., & Hanapi, M. S. (2026). AI content fabrication and deception: An Islamic ethical and societal analysis. Planning Malaysia, 24(42). https://doi.org/10.21837/pm.v24i42.2041

Marwantika, A. I., Ajhuri, K. F., & Dauda, K. O. (2025). Integrating generative AI in journalism education through Islamic communication ethics. Al-Balagh: Jurnal Dakwah dan Komunikasi, 10(2), 221–256. https://doi.org/10.22515/albalagh.v10i2.12482

Meerangani, K. A., Nor, M. R. M., Hamid, M. F. A., Yusof, M. A. M., & Ariffin, M. F. M. (2025). Digital tabayyun: Integrating the Quran and artificial intelligence in filtering social media information. QURANICA: International Journal of Quranic Research, 17(2), 370–389. https://ejournal.um.edu.my/index.php/quranica/article/view/64985

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. PNAS Nexus, 4(3), Article pgaf062. https://doi.org/10.1093/pnasnexus/pgaf062

Mohadi, M., & Tarshany, Y. (2023). Maqasid Al-Shari'ah and the ethics of artificial intelligence: Contemporary challenges. Journal of Contemporary Maqasid Studies, 2(2), 79–102. https://doi.org/10.52100/jcms.v2i2.107

Mökander, J. (2023). Auditing of AI: Legal, ethical and technical approaches. Digital Society, 2, Article 49. https://doi.org/10.1007/s44206-023-00074-y

Morgan, H. (2022). Conducting a qualitative document analysis. The Qualitative Report, 27(1), 64–77. https://doi.org/10.46743/2160-3715/2022.5044

Nada-Qisthina, D. S., & Musyafak, N. (2025). Tradition meets technology: The role of artificial intelligence in Nahdlatul Ulama and Muhammadiyah's digital adaptation. Islamic Communication Journal, 10(2), 427–446. https://doi.org/10.21580/icj.2025.10.2.30339

Nasution, I., Sazali, H., Hamandia, M. R., & Jannati, Z. (2026). Regulating the algorithmic panopticon: A critical library research on Islamic communication policy and digital media ethics in the

era of hyperconnectivity. Jurnal Administrasi Pemerintahan Desa, 7(2), 8. https://doi.org/10.47134/villages.v7i2.518

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.

Nuriana, Z. I., & Salwa, N. (2024). Digital da'wah in the age of algorithm: A narrative review of communication, moderation, and inclusion. Sinergi International Journal of Islamic Studies, 2(4), 242–256. https://doi.org/10.61194/ijis.v2i4.706

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

Park, S., & Nan, X. (2026). Generative AI and misinformation: A scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation. AI & Society, 41, 1501–1515. https://doi.org/10.1007/s00146-025-02620-3

Piasecki, S., Morosoli, S., Helberger, N., & Naudts, L. (2024). AI-generated journalism: Do the transparency provisions in the AI Act give news readers what they hope for? Internet Policy Review, 13(4). https://doi.org/10.14763/2024.4.1810

Pierson, J., Kerr, A., Robinson, S. C., Fanni, R., Steinkogler, V. E., Milan, S., & Zampedri, G. (2023). Governing artificial intelligence in the media and communications sector. Internet Policy Review, 12(1). https://doi.org/10.14763/2023.1.1683

Rachman, A., Saumantri, T., & Hidayatulloh, T. (2025). Transformation of religious authority in the digital era: A post-normal times analysis by Ziauddin Sardar on the phenomenon of social media da'wah. Jurnal Ilmu Dakwah, 45(1), 107–122. https://doi.org/10.21580/jid.v45.1.25644

Riza, M. (2021). Information literacy as an implementation of tabayun concept in Islam. Tik Ilmeu: Jurnal Ilmu Perpustakaan dan Informasi, 5(2). https://doi.org/10.29240/tik.v5i2.2866

Rogers, E. M. (2003). Diffusion of innovations (5th ed.). Free Press.

Sahebi, S., & Formosa, P. (2025). The AI-mediated communication dilemma: Epistemic trust, social media, and the challenge of generative artificial intelligence. Synthese, 205(3), 1–24. https://doi.org/10.1007/s11229-025-04963-2

Semnani, S. J., Yao, V. Z., Zhang, H. C., & Lam, M. S. (2023). WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia (arXiv:2305.14292). arXiv. https://doi.org/10.48550/arXiv.2305.14292

Shalhoob, H. (2025). The role of AI in enhancing shariah compliance: Efficiency and transparency in Islamic finance. Journal of Infrastructure, Policy and Development, 9(1), 1–26. https://doi.org/10.24294/jipd11239

Tarwiyyah, H. L. (2025). Artificial intelligence and bias in religious authority. Jurnal Informasi dan Teknologi, 7(2), 38–46. https://doi.org/10.60083/jidt.vi0.626

Thoriquttyas, T., & Rohmawati, N. (2026). Maqasid al-Sharia and the digital data ownership: From al-Shatibi to Jasser Auda. Journal of Islamic Law on Digital Economy and Business, 1(2), 168–181. https://doi.org/10.20885/JILDEB.vol1.iss2.art5

Ullah, M. A., Haque, M. S., & Islam, M. M. (2026). Artificial intelligence in Islamic communication: An ethical concern. Journal of Islamic Communication and Counseling, 5(1). https://doi.org/10.18196/jicc.v5i1.118

Abdulsalam, D. O., & Abdulrasheed, A. (2026). Digital distortion of religious preaching: An analysis of abuse of da'wah activities on social media. Suhuf: International Journal of Islamic Studies, 38(1). https://doi.org/10.23917/suhuf.v38i1.14129

Firdaus, M. A., Syihabuddin, M., & Fuady, Z. (2025). Islam and artificial intelligence: Perspectives from traditionalist and modernist Muslim communities in Indonesia. MIQOT: Jurnal Ilmu-ilmu Keislaman, 49(1). https://doi.org/10.30821/miqot.v49i1.1333

Abana, A. (2025). Leveraging social media as avenues for da'wah among Muslim youths in Nigeria. Islamic Communication Journal, 10(1), 19–36. https://doi.org/10.21580/icj.2025.10.1.26054

Faradillah, N. S., Latifah, A., Riskiyah, S., Dzakiyah, R., & Aziz, M. A. (2026). The diffusion of AI-driven digital da'wah innovation in communication media: A perspective of Q.S. Al-Mulk 23 and audio-visual cognitive analysis on Instagram @Ace_Timetraveller. Afkaruna: International Journal of Islamic Studies, 4(1), 96–111. https://doi.org/10.38073/aijis.v4i1.4691

Hasanah, U. (2021). Rhetoric in Islamic tradition: Paradigm and its development. Komunika: Jurnal Dakwah dan Komunikasi, 15(2), 241–252. https://doi.org/10.24090/komunika.v15i2.4570

Al-Farmawi, A. H. (1994). Al-Bidayah fi al-Tafsir al-Mawdu'i [The starting point in thematic exegesis]. Al-Hadarah al-Islamiyah.

Dhofier, Z. (1999). The pesantren tradition: The role of the kyai in the maintenance of traditional Islam in Java (Trans.). Arizona State University Program for Southeast Asian Studies Monograph Series.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2019). Qualitative data analysis: A methods sourcebook (4th ed.). Sage.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., 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://doi.org/10.48550/arXiv.2005.11401

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Wright, W. (2024). God chatbots offer spiritual insights on demand. What could go wrong? Scientific American, 330(4), 37. https://doi.org/10.1038/scientificamerican0424-37