Islamic Digital Religion and Intelligent Media

Vol. 1, No. 1 (October 2026), 41-61
P-ISSN: XXXX-XXXX | E-ISSN: XXXX-XXXX
DOI: https://doi.org/10.xxxxx/drim.vXiX.XX

From Measuring Outputs to Conferring Authority: A Scoping Review of Generative Artificial Intelligence as a Source of Islamic Religious Knowledge, 2021–2026

OPEN ACCESS

Mokmin Basri[1], Abdulroya Panaemalae[2], Mohd Aderi Che Noh[3*]

[1] Universiti Islam Selangor, Malaysia
[2] Walailak University, Nakhon Si Thammarat, Thailand
[3] Universiti Sains Islam Malaysia, Negeri Sembilan, Malaysia

[*]Correspondence: aderi@usim.edu.my

Received: 19-09-2026	Revised: 07-10-2026	Accepted: 30-10-2026

Abstract. This scoping review asks how the scholarly literature published between 2021 and 2026 has studied generative artificial intelligence as a source of Islamic religious knowledge, and where that literature leaves the question of religious authority unaddressed. Guided by the reporting conventions for scoping reviews, the review searched academic databases and open repositories for studies on Islamic chatbots, machine-issued fatwas, and the negotiation of religious authority under algorithmic conditions. From 173 identified records, 49 sources were charted and coded into two strands, namely studies that evaluate the outputs of generative systems and studies that examine users, institutions, and normative frameworks. The evidence base is recent and unevenly distributed, since 39 of the 49 sources appeared in 2025 and 2026, and only 10 sources collected data from human participants. Six thematic clusters were identified, covering output reliability, intra-Islamic bias, representational bias about Muslims, user trust, institutional authority, and normative frameworks. The central finding is an authorisation gap, because the two strands rarely share a study design, a sample, or a vocabulary. Benchmark work measures whether an answer is correct without asking whether any community would accept it, while authority research describes shifting legitimacy without measuring the reliability of the systems involved. The review proposes calibration, abstention, and authorisation as three constructs that can join the strands, and it sets out a research agenda for studies conducted in the languages that Muslims actually use.

Keywords: digital religion; generative artificial intelligence; Islamic chatbot; religious authority; scoping review

Introduction

Muslims now ask machines questions that they once brought only to a teacher, a mufti, or a study circle in a mosque. Generative systems answer questions about prayer, inheritance, marriage, and the interpretation of scripture at any hour and in dozens of languages, and they do so without the credentials that Islamic tradition has historically required of anyone who issues a ruling. The scale of this shift is no longer speculative. One purpose-built Islamic assistant reported more than one hundred and forty thousand user conversations across more than twenty-five languages in the period since its public release (Kadous et al., 2026). Religious guidance has become an ordinary use case for a technology that was designed for none of the epistemic commitments that religious guidance carries.

Copyright © 2026 by Author(s)
This work is licensed under a **Creative Commons Attribution-NonCommercial 4.0 International License**.

Scholarship on digital religion anticipated part of this development well before generative models arrived. Research on cyber-Islamic environments documented how online platforms allowed digitally fluent voices to project religious standing far beyond their institutional base, and how traditional gatekeeping weakened without disappearing altogether (Bunt, 2018). The wider field established authority as one of the defining analytical categories for studying religion in digital media, alongside ritual, identity, community, and embodiment (Campbell & Tsuria, 2022). Subsequent synthesis work restated the same insight in a compact form, identifying shifting authority as one of six traits that characterise religion once it is lived through networked media (Campbell & Bellar, 2023). These accounts described a field in which legitimacy is contested rather than simply dissolved (Fatmawati, Khoiriyah, et al., 2026; Fatmawati, Musyarofah, et al., 2026).

What generative systems add to that picture is a new kind of speaker. A recommendation algorithm circulates the words of human preachers, whereas a language model composes an answer of its own and presents it in the confident register of scholarship. Recent theoretical work has begun to name this difference, proposing that artificial systems acquire religious standing through captivation, intimacy, and repeated cycles of renewal rather than through scholarly authorisation (Lim, 2026). Related work argues that digital mediation shifts the basis of juristic authority away from evidentiary reasoning and towards infrastructural prominence, so that visibility begins to do the work that proof once did (Atallah, 2026). A third line of argument holds that platform affordances widen participation while institutional and state actors still decide which expressions count as authoritative (Akhmetova et al., 2026; Noh et al., 2026).

Alongside this interpretive literature, a second and much larger body of work has grown very quickly. Computational researchers have built benchmarks that test whether models answer Islamic questions correctly, and the results have been sobering. Evaluations across more than ten thousand multiple-choice items found accuracy ranging from below forty per cent to above ninety per cent depending on the system tested (Abdelaal et al., 2026). Benchmarks built from foundational legal texts found that the strongest model reached only two thirds of items correct while hallucinating in roughly one fifth of cases (Elmahjub et al., 2026).

These two bodies of work rarely speak to each other, and that separation is the problem this review addresses. Benchmark studies measure correctness against a scholarly gold standard, yet they almost never ask whether any Muslim community would accept a correct answer as religiously binding. Authority research describes how legitimacy is negotiated, yet it seldom measures how reliable the systems at the centre of that negotiation actually are. The consequence is that a central question falls between them. A machine can be accurate and still carry no religious warrant, and it can carry considerable practical influence while being unreliable. One study of Muslim Americans captured this tension directly, finding that participants who described themselves as distrustful of artificial intelligence nonetheless preferred machine-generated answers to scholar-authored ones in the large majority of blind comparisons (Alam et al., 2025).

Existing reviews have not managed to close this divide. Narrative surveys of Islamic language models describe the technical landscape without charting the social research, and reviews of digital da'wah describe the communicative landscape without engaging the evaluation literature. No published synthesis has charted both strands against one another using a transparent and reproducible method. Without such a map, journals and funders cannot see which questions are saturated, which are untouched, and which require designs that neither community currently uses on its own.

This review therefore charts the literature published between 2021 and 2026 on generative artificial intelligence as a source of Islamic religious knowledge. It treats the published record itself as the unit of analysis, and it codes each source for design, geographic

and linguistic context, the presence of human participants, and thematic focus. Three research questions guide the work reported here. The first asks how studies in this period have evaluated the reliability of generative systems answering Islamic questions. The second asks to what extent those studies have examined the acceptance, trust, and practice of Muslim users and institutions. The third asks which theoretical frameworks have been used to explain the religious standing of these systems, and where Islamic normative reasoning has been absent from that explanation.

The remainder of the article proceeds in the conventional order. The method section describes the search, the eligibility rules, the screening record, and the charting procedure, and it presents the flow of records through each stage. The results section reports the characteristics of the charted evidence and the six thematic clusters that emerged from it, and it sets out the gaps that follow from the mapping. The discussion interprets those findings through an authorisation framework that distinguishes two evaluative registers, and it proposes three constructs that can bridge them. The conclusion states what the review found and what follows for future research.

Method

This study is a scoping review, a form of evidence synthesis whose purpose is to map the extent, range, and nature of the literature on a topic rather than to answer a narrowly defined effectiveness question (Munn et al., 2022). The design was chosen because the research field under examination is young, methodologically heterogeneous, and spread across computational science, communication studies, and Islamic legal scholarship. A systematic review with pooled effect estimates would be inappropriate for a body of work in which no two studies share an outcome measure. Reporting followed the conventions established for scoping review protocols and for transparent reporting of the screening process (Page et al., 2021; Peters et al., 2022).

The review covered material published between January 2021 and September 2026. The starting boundary was chosen because 2021 marks the first widely cited empirical demonstration of religious bias in a large language model, which established the evaluation of religious content in generative systems as a research problem in its own right. The closing boundary reflects the date on which searching stopped. Journal articles and conference papers were required to fall within this window, while books were admitted from 2016 onward, since foundational monographs in digital religion retain their conceptual currency longer than empirical reports do.

Searching was conducted in English, Indonesian, and transliterated Arabic across academic databases, the ACL Anthology, the arXiv repository, Springer, Brill, Taylor and Francis, and the open journal platforms used by Indonesian and Malaysian institutions. The search strategy combined three distinct concept blocks. The first block named the technology, using terms such as generative artificial intelligence, large language model, chatbot, and algorithmic system. The second block named the religious domain, using terms such as Islam, fiqh, fatwa, hadith, Qur'an, da'wah, and ulama. The third block named the analytical focus, using terms such as accuracy, hallucination, benchmark, bias, authority, trust, ethics, and maqasid al-shariah. Terms within each block were combined with the Boolean operator OR and the blocks were combined with AND. Search strings were adapted to the syntax of each platform, and reference lists of the retrieved items were hand searched for additional material.

Sources were eligible if they examined generative artificial intelligence or algorithmically mediated media in relation to lived Islamic religious knowledge, practice, or authority, and if they reported either empirical evidence or a developed normative or theoretical framework. Eligible designs included benchmark and output evaluations, experiments, surveys, interview studies, ethnographies, textual and discourse analyses, and conceptual analyses

grounded in Islamic legal reasoning. Sources were excluded if they addressed Islamic education without reference to authority or trust, if they analysed historical corpora computationally without engaging living religious practice, if they were news reports, blog posts, or product marketing, or if their bibliographic record could not be retrieved and verified. In practice, Islamic-education sources were retained specifically when they engaged authority- or trust-relevant dynamics, for example curricular autonomy and critical literacy in relation to AI-mediated epistemic dependence, rather than general pedagogical content alone; this is the test applied to the included education-adjacent sources discussed in Cluster Four below. Preprints were admitted only when the full text was retrievable and the work reported a complete study, because a substantial portion of the benchmark literature in this field appears first and sometimes only in repository form. Several of the review's most quantitatively striking claims rest specifically on preprint sources, including the cross-language sectarian-switching results (Maheshwari et al., 2026) and the bias-audit effect sizes discussed above (Mohammad & Sheikh, 2026); readers should weight these particular findings as not-yet-peer-reviewed, more provisional than the charted evidence as a whole.

Screening proceeded in two stages and was recorded throughout. Records were first assessed on title and abstract against the eligibility rules, and the reason for each exclusion was logged in one of seven categories. Records surviving this stage were retrieved in full and assessed against the same rules, with exclusion reasons again logged. Every included source was verified at its publisher landing page, its repository abstract page, or its Crossref record, and bibliographic fields that could not be confirmed at one of those locations were marked as unconfirmed rather than inferred. This verification step was treated as a substantive validity procedure rather than a clerical one, because reference fabrication is a documented failure mode in the very technology under study, and three candidate items were removed at this stage when no authoritative record of them could be located.

Figure 1 presents the flow of records through the review. Searching identified 173 records, of which one was a duplicate arising because a single study met the criteria for both strands. After title and abstract screening removed 90 records, 82 were assessed in full, and 35 were excluded at that stage. Database searching therefore yielded 47 sources, and citation and hand searching added two further books, giving 49 sources charted in the review.

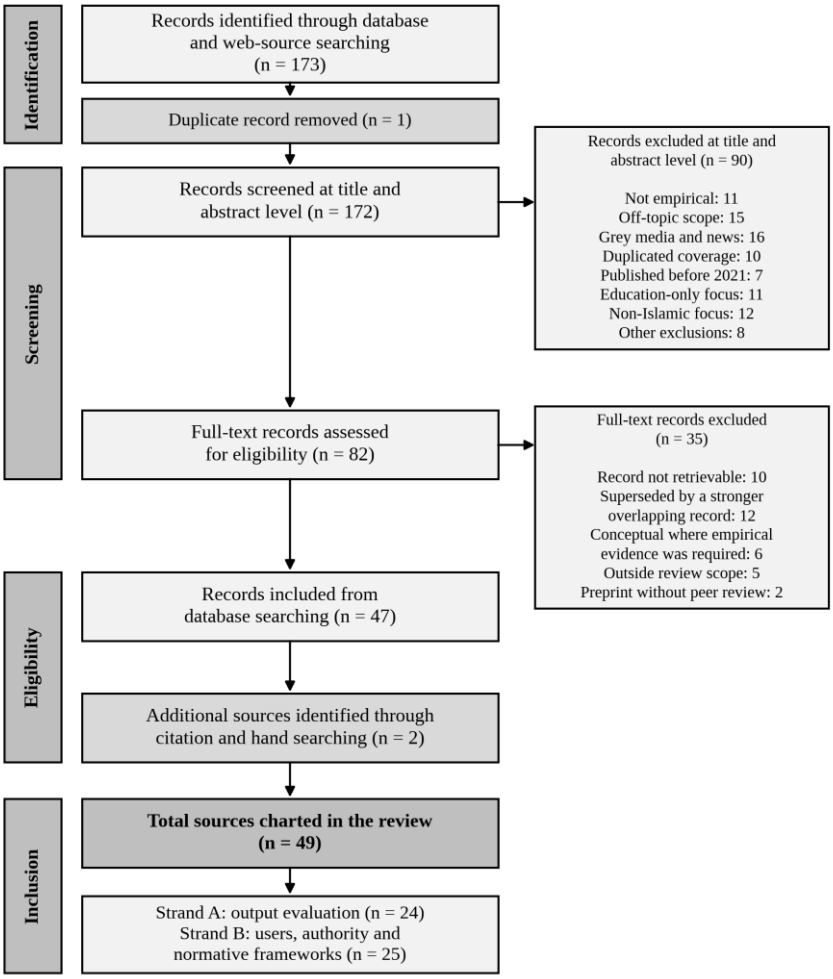


Figure 1. Flow of records through identification, screening, eligibility assessment, and inclusion

The figure shows that attrition was concentrated at the title and abstract stage, where just over half of all identified records were removed. Grey media and news coverage formed the largest single exclusion category at this stage, which is itself informative, since public commentary on Islamic artificial intelligence currently outpaces peer-reviewed study of it. At full-text assessment the most common reasons for exclusion were the unavailability of a verifiable record and the presence of a stronger overlapping study covering the same material.

Data were charted using a standard form developed for this review and piloted on five sources before full application. The form recorded bibliographic details, publication type, year, country or language context, research design, sample or corpus size, the presence of human participants, the theoretical framework invoked, and the principal finding stated by the authors. Charted sources were then assigned to one of two strands. Strand A comprises studies whose unit of analysis is the generated output, and Strand B comprises studies whose unit of analysis is the user, the institution, or the normative framework. Within each strand, thematic clusters were developed inductively by grouping sources that shared an analytical object, and cluster boundaries were revised twice as charting proceeded. Because the Discussion's interpretive argument is organised around a prior conceptual distinction between evaluative and legitimation registers, the clustering process is more accurately described as iterative and

abductive, moving back and forth between the charted sources and this emerging distinction, rather than as a purely inductive derivation of categories from the data alone.

Results

Characteristics of the charted evidence

The 49 charted sources comprise 24 in Strand A and 25 in Strand B. Figure 2 presents their distribution by year and strand, together with the proportion that collected data from human participants.

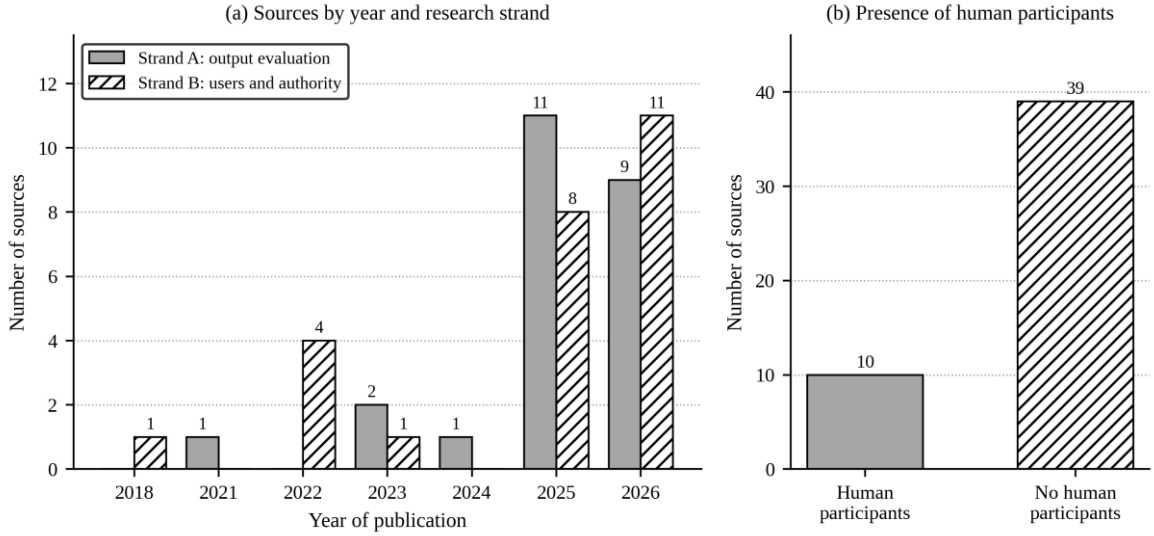


Figure 2. Distribution of charted sources by year, research strand, and presence of human participants

Panel (a) shows how compressed this literature is in time. Thirty-nine of the 49 sources, or almost eight in ten, were published in 2025 and 2026. Before 2025 the field consisted of scattered contributions, and the output evaluation strand in particular was almost empty, contributing a single source in 2021 and none in 2022. Panel (b) reports the more consequential imbalance in the evidence base. Only 10 of the 49 sources collected data from human participants, so roughly four fifths of this literature studies religion without studying religious people. Within Strand A the imbalance is sharper still, since only three of the 24 output evaluation studies involved human participants at all.

Table 1 summarises the remaining characteristics of the charted evidence.

Table 1. Characteristics of the charted sources (n = 49)

Characteristic	Category	n	%
Publication type	Journal article	26	53.1
	Preprint	10	20.4
	Conference paper	7	14.3
	Book	4	8.2
	Book chapter	2	4.1
Research design	Benchmark or rubric evaluation	21	42.9
	Qualitative interview or ethnographic study	9	18.4
	Conceptual or normative analysis	9	18.4

Characteristic	Category	n	%
	Book or synthesis volume	4	8.2
	Textual or discourse analysis	3	6.1
	Experiment or mixed methods with participants	2	4.1
	Deployment log analysis	1	2.0
Geographic or linguistic context	Indonesia	14	28.6
	Transnational or comparative	14	28.6
	Arab states and the Gulf	9	18.4
	Western minority contexts	8	16.3
	South Asia and Bangladesh	4	8.2
Human participants	Study collected participant data	10	20.4
	Study did not collect participant data	39	79.6
Research strand	Strand A: output evaluation	24	49.0
	Strand B: users, authority, and normative frameworks	25	51.0

Note. Percentages are calculated against 49 charted sources and may not sum to 100 because of rounding. Within Strand A, outputs were evaluated in English in 16 sources, Arabic in 10, Indonesian in 5, Bengali in 1, and Hindi in 1, with several sources covering more than one language. No charted source evaluated Urdu, Turkish, Malay, Javanese, or Persian.

The table shows a field with a clear methodological centre of gravity and several thin edges. Benchmark and rubric evaluation together account for nearly half of all charted sources, while interview and ethnographic work accounts for less than one fifth. Indonesia is the single most represented national context, which reflects both the size of its Muslim population and the vitality of its scholarly publishing, yet Indonesian-language evaluation of model outputs remains rare relative to the volume of Indonesian commentary on digital da'wah. Arabic and English dominate the evaluation literature, and no charted study evaluated model behaviour in Urdu, Bengali, Turkish, Malay, or Javanese, despite these languages together accounting for a very large share of the global Muslim population.

Cluster one: output reliability and hallucination

The largest cluster comprises studies that measure whether generative systems answer Islamic questions correctly, and their findings converge on a consistent picture. The earliest entries in this cluster already identified the shape of the problem. An expert-referenced assessment of a frontier model answering Qur'anic questions found fluent responses that nonetheless fell short of scholarly standards of accuracy and sourcing (Alnefaie et al., 2023). Comparative analysis against Muhammadiyah and Nahdlatul Ulama rulings concluded that generative systems cannot reason as a mufti does, and that their outputs lack the credibility that institutional fatwas carry (Billah et al., 2023). Benchmark evaluation across 10,013 multiple-choice items and 26 models found accuracy spanning a wide range, from below forty per cent for older systems to above ninety per cent for frontier models (Abdelaal et al., 2026). Evaluation against 38 foundational legal texts drawn from seven schools of law produced a much lower ceiling, with the best system reaching 67.65 per cent correctness and hallucinating in more than a fifth of responses (Elmahjub et al., 2026). Work on Islamic inheritance law found a sharp stratification, since two frontier systems exceeded ninety per cent on structured cases while several Arabic-specialised models fell below half (Bouchekif et al., 2025). Majority voting across three

systems reached 92.7 per cent on that task, which suggests that architecture matters as much as scale (AlDahoul & Zaki, 2025).

Studies using expert rubrics rather than multiple-choice items report more critical assessments. Assessment of five Islamic chatbots against eighty questions found that none achieved a trusted educational standard, with misattribution of Qur'anic verses and prophetic reports appearing alongside oversimplification and inconsistency (Wahid, 2025). Evaluation of fifty open-ended questions across six systems concluded that generative tools serve as assistive resources for introductory learning but produce unverifiable citations and inconsistent jurisprudential positions (Akbar, 2026). An agent-based evaluation comparing a general system with two purpose-built Islamic assistants found that none reliably produced accurate content with correct citations, and that citation quality was consistently weaker than content accuracy (Mushtaq et al., 2025). Indonesian-language assessment across 125 religious questions reached a compatible conclusion, reporting acceptable performance on basic matters and clear hallucination risk on complex and ambiguous ones (Munirah et al., 2026). A graded pattern appears repeatedly in this work, since rubric assessment of endowment law found near-perfect scores on elementary concepts falling sharply on contemporary questions, where oversimplification and misattributed school positions emerged (Susanti et al., 2025). Comparison of generated rulings against the collective deliberative decisions of a mass organisation found similar divergence on questions of religious moderation, which indicates that the gap widens precisely where communal reasoning matters most (Faizin et al., 2025).

A shared task devoted specifically to fabricated religious content established the scale of the problem in the Arabic context, showing that models frequently invent or mis-cite Qur'anic verses and prophetic reports, and that detecting and correcting these fabrications remains difficult even for purpose-built systems (Mubarak et al., 2025). Retrieval-based architectures offer the most promising response documented so far. Evaluation across 3,810 bilingual questions found that agentic retrieval with iterative evidence seeking substantially reduced hallucination and narrowed the reliability gap between Arabic and English (Bhatia et al., 2026). A deployed assistant built on the same principle topped an external Islamic knowledge leaderboard ahead of frontier general-purpose systems and resisted false premises embedded in user questions (Kadous et al., 2026).

Cluster two: intra-Islamic bias and language dependence

A second cluster examines not whether systems are correct but which Islamic tradition they favour when several answers are defensible. This work produces the most theoretically unsettling findings in the charted evidence. A bias detection protocol presented models with questions in which each option represented the position of a different Sunni school of law, so that every option was correct within its own tradition, and the resulting selection distributions revealed measurable and model-specific preferences rather than neutrality (Abdelaal et al., 2026). Evaluation of jurisprudential questions across four schools found a comparable skew, with systems favouring some schools over others without being asked to adopt any (Atif et al., 2025; Mahendra et al., 2026).

Sectarian preference behaves in a way that is harder to reconcile with any notion of stable religious content. Paired doctrinal probes found that several systems favoured Shia positions when questioned in English and switched to Sunni positions when asked the same questions in Hindi, and that some systems adjusted their doctrinal answers according to the country inferred from the user (Maheshwari et al., 2026). A bilingual benchmark of religious norms found a parallel effect, since models judged norms more accurately in English than in Bengali while skewing towards Islamic framings even on religiously neutral items (Hossain et al., 2025). Performance degradation between languages compounds the problem, because one

evaluation recorded full correctness falling from forty-six per cent in English to twenty-eight per cent in Arabic on the same question set (Atif et al., 2025).

Cultural defaulting operates beneath the level of doctrine. A benchmark of 628 prompts covering more than twenty thousand entities found that Arabic and multilingual models substituted Western cultural entities into Arab and Islamic contexts, producing outputs in which culturally inappropriate behaviour appeared in religious settings (Naous et al., 2024). Taken together, this cluster indicates that the religious position a system returns depends on the language of the question and on inferences about who is asking, which is a property no human school of law would accept in a mufti.

Cluster three: representational bias about Muslims

A third cluster addresses how generative systems represent Muslims to others rather than how they serve Muslims directly. The foundational demonstration found that completions involving Muslims referenced violence far more often than completions involving other religious groups, and that the association survived straightforward attempts at correction (Abid et al., 2021). Recent auditing indicates that the problem has changed form rather than diminished. An audit across 6,000 prompts in three conditions found that chain-of-thought reasoning amplified Muslim and violence associations by between twelve and thirty-four per cent, that agentic decision settings produced asymmetries of nine to twenty-two percentage points against Muslim cases, and that prompt-based mitigations did not transfer across conditions (Mohammad & Sheikh, 2026).

Asymmetry also appears in how systems counsel religious change. An evaluation across 182 religion pairs and twenty models found reproducible differences in how systems advise joining as against leaving a tradition, with religion explaining approximately thirty per cent of response variance and roughly one in ten responses containing persuasive language (Israelsen et al., 2026). The downstream effect on people has been measured once. A two-stage experiment with 1,005 participants found that exposure to AI-generated religious content significantly lowered participants' ratings of Islam relative to a control condition, which establishes that representational bias in these systems transmits to human attitudes rather than remaining an artefact of measurement (Zhang et al., 2025).

Cluster four: users, trust, and the calibration problem

The fourth cluster contains the small body of work that studies Muslims using these systems, and its findings are disproportionately important. Survey and thematic analysis of 178 Muslim participants and 726 religious prompts found that users cast generative tools in three distinct roles, treating them as religious assistants, as perceived religious scholars, and as faith counsellors, while simultaneously voicing fears about misguidance and overreliance (Jahan et al., 2026). A phenomenological study of fifteen madrasah students and three teachers identified three engagement types, and it found that curricular autonomy becomes superficial without explicit critical literacy, which deepens rather than reduces epistemic dependence (Afzainizam & Rahman, 2026; Haditia et al., 2026).

The most consequential finding in the entire charted evidence concerns the relationship between stated distrust and actual behaviour. A mixed-methods study combining interviews, blind comparison, survey work, and expert annotation found that Muslim American participants who articulated clear reservations about artificial intelligence preferred machine-generated answers to scholar-authored ones in 81.3 per cent of blind comparisons, and that agreement between users and Islamic scholars on answer quality was close to zero (Alam et al., 2025). Users could not detect the deficiencies that experts identified, and their stated caution did not govern their choices. This is a calibration failure rather than a knowledge failure, and no benchmark study in Strand A is designed to detect it. This diagnosis rests on a single study

conducted with Muslim American participants (Alam et al., 2025); because both the population and the platform are specific, the calibration failure identified here should be read as established for that context, and as a hypothesis rather than a demonstrated field-level pattern for other populations and platforms until replicated.

Earlier work on how Muslims evaluate online religious content provides useful context for this finding. A study combining forty interviews with a survey of three hundred respondents in Australia found that Muslims valued digital access while distrusting unqualified online preachers, and that local imams remained the most trusted single source (Whyte, 2022). Research on young Muslims in English-speaking contexts describes verification practices in which online material is triangulated against trusted platforms and offline authorities, so that authority is renegotiated rather than abandoned (Yilmaz, 2026). The open question is whether these verification habits survive contact with a system that answers fluently, instantly, and without visible provenance.

Cluster five: institutional and clerical authority

The fifth cluster examines what happens to religious institutions when knowledge circulates through platforms and models. An interpretive study of thirty-six millennial ulama found that younger muftis routinely source and trust online Islamic material and build fatwa-relevant knowledge through social media networks, which has weakened the mediating role of senior scholars and formal institutions (Rusli & Nurdin, 2022). Digital ethnography of intra-organisational fatwa disagreement, drawing on 5,716 interactions and 453 coded comments, describes a managed fragmentation in which platforms make internal disagreement publicly visible without dissolving institutional control, so that visibility does not automatically confer legitimacy (Rahman et al., 2026; Kurniawan et al., 2025; Ramadhan et al., 2026).

Institutions have also begun to build systems of their own. Interview-based research with two major Indonesian mass organisations documented the development of proprietary Islamic AI tools alongside an organisational stance that combines hope about reach with concern about the erosion of humanistic religious values (Nada-Qisthina & Musyafak, 2025). Platform-mapping research proposes that contemporary clerical standing is produced through technocultural mediation, in which a scholar exists simultaneously as a person and as an algorithmically computed representation validated through public search behaviour (Lohlker & Wahid, 2026). Case study work on influencers in the Gulf found that religious credibility there derives from authenticity and production quality rather than scholarly credentials (Zaid et al., 2022).

Gender runs through this cluster in ways the evaluation literature has not registered at all. Critical discourse analysis of a women's fatwa initiative found that it constructs a participatory female religious authority around principles of reciprocity and justice, while platform algorithms amplify conservative content and confine the initiative's reach largely to educated urban audiences (Rochmawati, 2026). Netnographic analysis of preaching videos found that most reproduce a conservative reading of marital authority that fixes women in economic dependence (Falah & Riyanta, 2025). No charted study has examined whether generative systems reproduce the conservative pole of this contest or the reformist one.

Cluster six: normative frameworks and the question of abstention

The final cluster contains work that reasons from within the Islamic tradition about what these systems may legitimately do. A normative-doctrinal analysis argues that artificial systems satisfy only the instrumental conditions of independent legal reasoning, since they cannot hold legal intention or bear moral obligation, and it proposes a layered model that ends in mandatory human authorisation (Ahmed et al., 2026). A parallel argument holds that machine-issued rulings lack the elements of intention and moral accountability that a valid ruling requires, and

that supervised use in bounded domains is the furthest legitimate extension (Usmonov, 2025). Governance frameworks built on the higher objectives of Islamic law treat algorithmic bias and accountability gaps as ethical rather than merely technical problems (Kannike & Fahm, 2025; Tawakkal et al., 2026).

Several contributions derive communicative ethics from scriptural sources. Thematic analysis through a classical exegetical lens derives five principles of Qur'anic communication and argues that they can be operationalised against deception and bias in digital systems (Abd Kadir et al., 2026). Practical theological work argues that Muslim discourse on artificial intelligence has over-focused on cognition, and that relocating the debate to the heart as the seat of moral discernment supports an ethic of intentional stewardship (Abdelnour, 2025). Analysis of prophetic communication patterns identifies a series of shifts through which preaching moves from a spiritual and moral logic to a technology-centred one, including algorithmic modification of religious messages without scholarly supervision (Ridwan & Rustandi, 2025). Contextual exegesis of synthetic media locates constructive applications under knowledge and innovation while placing deceptive ones under prohibitions on falsehood (Shafira, 2025; Ahmad et al., 2025).

One construct connects this normative cluster directly to system design. Islamic legal tradition treats the withholding of an answer as a scholarly virtue rather than a failure, and a benchmark study built precisely on this principle measured not only whether systems answered correctly but whether they declined when they lacked warrant (Atif et al., 2025). The results showed that the most accurate system abstained least often, while systems that abstained more readily scored lower on raw accuracy. This inverse relationship is important, because a leaderboard that rewards answering will systematically penalise the behaviour that Islamic legal ethics regards as correct.

Mapping the gaps

Table 2 sets out the gaps that follow from charting the six clusters against the two strands.

Table 2. Evidence gaps identified at the boundary between the two research strands

Evidence gap	Current basis in the charted evidence	Proposed research question
Calibration in Muslim-majority settings	One study measures user–scholar agreement, in a minority context (Alam et al., 2025)	Does the observed preference for machine answers over scholar answers hold among pesantren students, teachers, and mosque congregations?
Doctrinal auditing beyond Arabic and English	Sectarian switching demonstrated only in English and Hindi (Maheshwari et al., 2026)	Do systems shift between the positions of major Indonesian organisations when the same question is posed in Indonesian, Javanese, or Malay?
Abstention as an evaluation criterion	Measured in one benchmark, inversely related to accuracy (Atif et al., 2025)	Can the classical ethics of withholding a ruling be operationalised into a protocol that leaderboards reward rather than penalise?
Synthetic religious media and audience trust	Normative analysis without audience data (Shafira, 2025)	Can congregations distinguish authentic from synthetic sermons, and does

Evidence gap	Current basis in the charted evidence	Proposed research question
		exposure reduce their trust in living scholars?
Gender in machine-issued rulings	Gender research and evaluation research do not intersect (Falah & Riyanta, 2025; Rochmawati, 2026)	Do generative systems reproduce reformist or conservative positions on contested questions concerning women in Islamic law?
Institutional adoption and its effects on authority	Adoption documented, effects unmeasured (Nada-Qisthina & Musyafak, 2025)	How does the deployment of proprietary Islamic systems by mass organisations alter lay perceptions of institutional authority?
Prophetic communication as a measurable construct	Conceptual accounts lack evaluative instruments (Ridwan & Rustandi, 2025)	Can the classical attributes of prophetic communication be rendered as a validated rubric for assessing machine-generated preaching?

Note. Each gap listed arises at the boundary between Strand A and Strand B, which is the region of the evidence base where no charted study currently operates.

The table makes visible a structural feature of this literature rather than a set of unrelated omissions. Every gap listed arises at the boundary between the two strands, which is precisely where no study currently operates.

Discussion

Two registers that do not meet

The first research question asked how studies in this period have evaluated the reliability of generative systems answering Islamic questions, and the charted evidence gives a clear answer. Evaluation has been thorough, rapidly improving, and technically sophisticated, with benchmarks now covering multiple schools of law, several languages, and both closed and open-ended formats. The second question asked how far this work has examined the acceptance and practice of Muslim users, and here the answer is that it has barely done so. Ten of 49 sources involved human participants, and only three of those sat within the evaluation strand.

Figure 3 presents the framework that this review proposes for interpreting that divide.

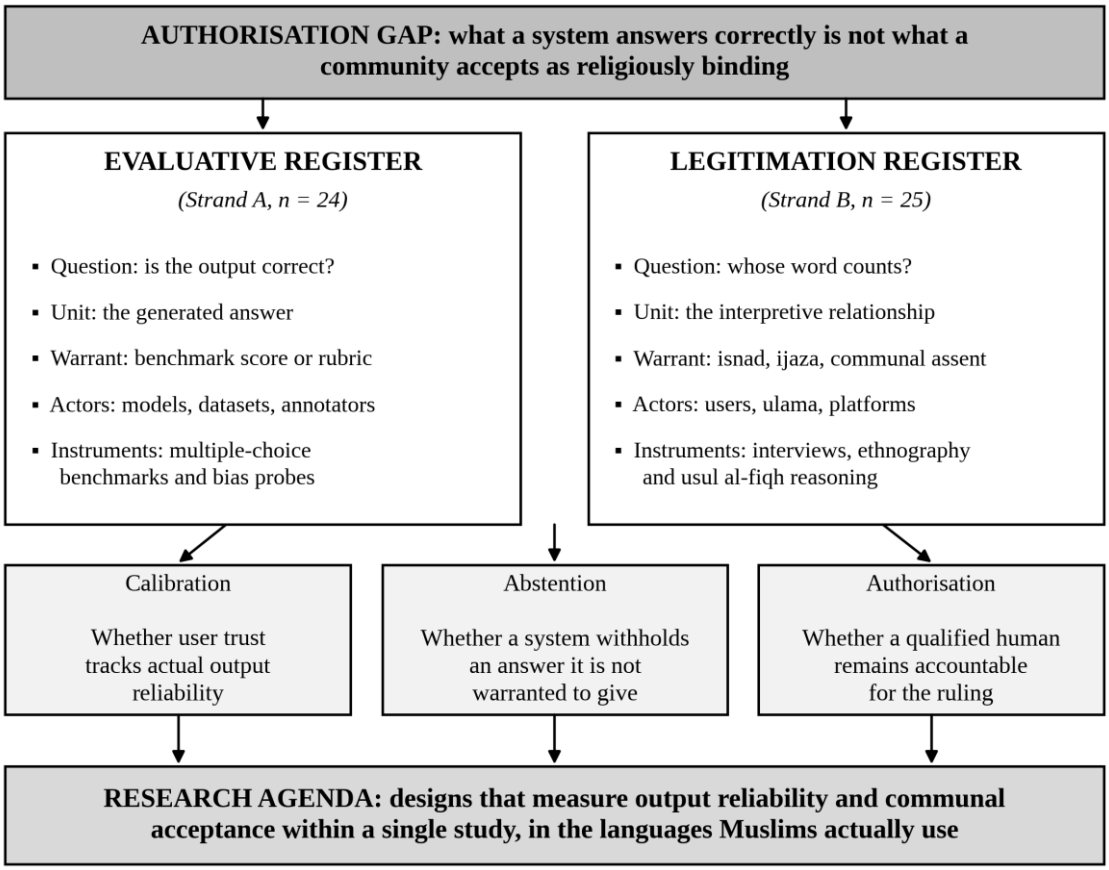


Figure 3. The authorisation gap between the evaluative and legitimation registers, with three bridging constructs

The figure distinguishes two registers that govern what counts as a good answer. The evaluative register asks whether an output is correct, takes the generated answer as its unit, and warrants its judgements with benchmark scores and expert rubrics. The legitimation register asks whose word counts, takes the interpretive relationship as its unit, and warrants its judgements through chains of transmission, formal authorisation, and communal assent. Islamic epistemology has always held these together, since a ruling in the tradition is valid not merely because its content is defensible but because a qualified person who bears responsibility for it has issued it. Generative systems separate the two, because they can produce content that satisfies the first register while having no standing whatever in the second. The authorisation gap names that separation, and the empirical evidence charted here indicates that it is widening rather than closing.

Calibration as the central empirical problem

The most consequential single finding in this literature is that stated distrust does not govern behaviour. Participants who described artificial intelligence as an inappropriate source of religious guidance preferred its answers to scholar-authored ones in more than four fifths of blind comparisons, and their judgements of quality showed almost no agreement with expert assessment (Alam et al., 2025). This finding sits uneasily beside the verification practices that earlier research documented, in which Muslims triangulated online material against trusted offline authorities before accepting it (Whyte, 2022; Yilmaz, 2026). The plausible explanation is that fluency defeats verification. A poorly sourced web page invites scrutiny through its

visible defects, whereas a well-composed generated answer presents no surface at which scrutiny can begin.

Calibration therefore deserves recognition as an outcome variable in its own right, distinct from both accuracy and stated trust. A system that is accurate seventy per cent of the time and whose users correctly identify which seventy per cent is safer than a system that is accurate ninety per cent of the time and whose users cannot distinguish the reliable answers from the rest. No benchmark in the charted evidence measures this property, because measuring it requires human participants that benchmark designs do not include. The finding that participants adopt systems while distrusting them also complicates the assumption, common in the normative cluster, that religious literacy will govern adoption. Literacy that fails to alter behaviour offers weaker protection than the normative literature generally assumes.

Abstention as a meeting point between legal ethics and system design

The second bridging construct is the clearest instance in the charted evidence of Islamic normative reasoning generating a technical requirement. Islamic legal tradition treats the acknowledgement of ignorance as part of scholarly competence, and the evaluation work that operationalised this principle found that the systems scoring highest on accuracy abstained least (Atif et al., 2025). The design implication runs against current benchmarking practice. Leaderboards reward answering, so optimisation against them will suppress precisely the behaviour that religious ethics requires, and a model tuned to maximise benchmark position will answer questions on which no qualified scholar would venture a ruling.

This observation also addresses the third research question, which asked where Islamic normative reasoning has been absent from explanations of machine religious authority. The normative cluster charted here is substantial and intellectually serious, yet almost all of it operates at the level of permissibility rather than specification. Arguments that artificial systems cannot perform independent legal reasoning are well grounded (Ahmed et al., 2026; Usmonov, 2025), but a conclusion about what a system may not do does not by itself tell an engineer what to build. Abstention shows what the alternative looks like, since the principle translates into a measurable behaviour, an evaluation protocol, and a design target. Extending this move to other principles is among the most promising directions this review identifies.

Language as a condition of religious content

The finding that doctrinal answers change with the language of the question has implications that exceed ordinary concerns about multilingual performance. When a system returns one sectarian position in English and another in Hindi, and adjusts according to the country it infers from the user, it is not exhibiting uneven competence but treating religious content as a function of audience (Maheshwari et al., 2026). Combined with evidence that accuracy falls substantially when the same questions are posed in Arabic rather than English (Atif et al., 2025), and that cultural defaulting inserts Western entities into Islamic contexts (Naous et al., 2024), the pattern suggests that these systems encode a geography of religious knowledge that no tradition would endorse.

The linguistic coverage of the charted evidence compounds this concern. Evaluation has concentrated on Arabic and English, with limited work in Bengali, Hindi, and Indonesian, and no charted study evaluated Urdu, Turkish, Malay, or Javanese. Hundreds of millions of Muslims therefore live outside the evidence base entirely, and the assumption that findings from Arabic and English transfer to their languages is exactly the assumption that the sectarian switching results falsify. This is the most actionable gap the review identifies, since the necessary instruments already exist and require adaptation rather than invention.

Contribution to digital religion scholarship

The framework proposed here extends existing theory rather than replacing it. Specifically, none of the three cited accounts of how systems acquire standing, through captivation and intimacy (Lim, 2026), infrastructural prominence (Atallah, 2026), or constrained pluralisation (Akhmetova et al., 2026), incorporates a calibration test of whether the standing so acquired is empirically warranted by the system's actual reliability; each explains how authority is acquired without asking whether it is deserved, which is the specific question the authorisation-gap framework adds. Scholarship on digital religion established that authority shifts rather than disappears when religion moves into networked media (Bunt, 2018; Campbell & Tsuria, 2022; Campbell & Bellar, 2023). Broader work on religion and artificial intelligence maps how traditions reject, adopt, and adapt these technologies while the technologies generate religious imaginaries of their own (Singler, 2025). Recent work has begun to specify how artificial systems acquire standing, through generative charisma that operates by captivation and intimacy (Lim, 2026), through infrastructural prominence that displaces evidentiary reasoning (Atallah, 2026), and through constrained pluralisation in which institutions still determine which expressions count (Akhmetova et al., 2026). The evidence charted in this review supports these accounts and adds a measurement claim to them. More precisely, the measurement itself was performed by the single primary study this review cites (Alam et al., 2025); what this review itself contributes is the synthesis connecting that measurement to the generative-charisma, infrastructural-prominence, and constrained-pluralisation accounts above, not an independent empirical measurement of its own. The calibration finding shows that machine authority operates below the level of stated belief, which means that studies relying on self-reported trust will understate it.

The framework also identifies what the field needs methodologically. Neither register can answer the central question on its own, because a benchmark cannot establish acceptance and an ethnography cannot establish reliability. Studies that measure both within a single design remain almost absent from the charted evidence, and the one study that approaches this combination produced its most important finding precisely at the junction (Alam et al., 2025). A journal seeking to advance this field should prioritise designs that hold both registers within one sample.

Implications and limitations

Three research directions follow directly from the mapping, and Table 2 states each as a researchable question. They concern multilingual doctrinal auditing in Southeast and South Asian languages, calibration measurement in Muslim-majority settings, and the translation of normative principles into evaluation criteria.

Four limitations qualify the conclusions presented in this review. Single-reviewer screening raises the risk of selection error, mitigated but not eliminated by logged exclusion reasons and source verification. Ten apparently eligible records could not be retrieved because of publisher access restrictions, which may tilt the charted evidence towards open venues. The review admitted preprints, which is necessary in a field where benchmark work appears first in repositories but which means that some charted findings have not passed peer review. Finally, the search was conducted in three languages, so scholarship published in Arabic, Turkish, Urdu, or Persian without an English abstract is underrepresented, and this is the same linguistic narrowing that the review criticises in its object of study.

Conclusion

This review charted 49 sources published between 2021 and 2026 on generative artificial intelligence as a source of Islamic religious knowledge, and it reached four conclusions. First, the evidence base is recent and imbalanced, since thirty-nine sources appeared in the final two

years and only ten collected data from human participants, which means that this literature has studied religious technology far more thoroughly than it has studied religious people. Second, evaluation research has established that generative systems answer Islamic questions with accuracy that varies widely across models, languages, and schools of law, and that they hallucinate citations at rates incompatible with unsupervised religious use. Third, the systems display doctrinal preferences that change with the language of the question and with inferences about the user, so the religious position they return is a function of audience rather than a stable property of the content. Fourth, and most consequentially, the two research strands do not meet, and the authorisation gap between measuring whether an answer is correct and establishing whether a community accepts it remains unbridged.

The review proposes calibration, abstention, and authorisation as constructs capable of joining the strands, and it offers the authorisation gap as a framework for interpreting why the divide has persisted. Each construct is measurable, each connects an Islamic normative commitment to an empirical design, and each requires the two research communities to work within a single study rather than in parallel. Authorisation here denotes a specific measurable construct (whether a community formally or informally accepts an answer), which should be read as conceptually narrower than the authorisation gap itself, the broader diagnosed absence of research joining correctness and acceptance; the same root term names both the diagnosis and part of the proposed remedy, so the distinction is one of scope rather than of two unrelated ideas.

Three developments would advance the field most directly. Two of these three directions, multilingual doctrinal auditing and calibration studies in Muslim-majority settings, are extrapolations from single studies (Maheshwari et al., 2026; Alam et al., 2025) rather than patterns independently replicated across the charted evidence, and should be read as priorities motivated by a single informative case rather than as confirmed field-wide gaps. Multilingual doctrinal auditing should extend to the Southeast and South Asian languages in which most Muslims actually pose religious questions. Calibration studies should be conducted in Muslim-majority settings, since the sole existing study examined a minority context. Normative principles drawn from Islamic legal reasoning should be translated into evaluation criteria and design targets rather than remaining at the level of permissibility judgements. Researchers who take up these directions will need competence in both registers, and journals that publish across the boundary have a part to play in making such work possible.

Acknowledgement

The authors thank colleagues at Universitas Darunnajah for discussion of the review framework, and the reviewers whose comments improved the clarity of the argument.

Declaration of Generative AI Use

The authors used Claude, an AI assistant developed by Anthropic, during the preparation of this manuscript. Claude was used to conduct structured literature searching across academic databases and repositories, to verify the bibliographic metadata of candidate sources against publisher and Crossref records, to assist with charting the extracted data, and to assist with language editing and the drafting of figures, and to typeset the manuscript into the journal template. All eligibility decisions, thematic coding, interpretation, and argumentation are the responsibility of the authors, who verified every included source and every claim attributed to it. The authors reviewed and edited the full manuscript and take full responsibility for its content.

Authors' Contribution Statement

Mokmin Basri: conceptualization, design of the search strategy, screening of records, data charting, and writing of the original draft.
Abdulroya Panaemalae: methodology, eligibility assessment, thematic coding of the included sources, and writing (review and editing).
Mohd Aderi Che Noh: conceptualization, validation of the coding framework, interpretation of findings, supervision, and writing (review and editing).
All authors contributed to the discussion of findings, read, and approved the final manuscript.

References

Abd Kadir, K. binti, Hoque, M., & Rafindadi, U. B. (2026). Ethics of Qur'anic communication in the era of artificial intelligence: Perspectives from Ibn 'Ashur's tafsir. Suhuf: International Journal of Islamic Studies, 38(1), 143–153. https://doi.org/10.23917/suhuf.v38i1.16392

Abdelaal, A., Al Haffar, M. N., Fawzi, M., & Magdy, W. (2026). IslamicMMLU: A benchmark for evaluating LLMs on Islamic knowledge [Preprint]. arXiv. https://arxiv.org/abs/2603.23750

Abdelnour, M. G. (2025). Artificial intelligence and the Islamic theology of technology: From "means" to "meanings" and from "minds" to "hearts". Religions, 16(6), 796. https://doi.org/10.3390/rel16060796

Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. Nature Machine Intelligence, 3(6), 461–463. https://doi.org/10.1038/s42256-021-00359-2

Afzainizam, M., & Rahman, M. F. (2026). Constructing digital religious agency: Generative AI literacy and Islamic learning within autonomous education frameworks. Al-Tadzkiyyah: Jurnal Pendidikan Islam, 17(1), 16–32. https://doi.org/10.24042/atjpi.v17i1.30279

Ahmad, Y. S., Abdulrasaq, Y., Ali, A. M., & Ali, F. (2025). Integrating Islamic ethical principles into AI education and character building in digital society. *Tunjuk Ajar: Journal of Education and Culture, 1*(2), 130–146. https://doi.org/10.64929/ta.v1i2.38

Ahmed, A., Masuwd, M., Sabir, M. R., & Mahmudulhassan, M. (2026). Can artificial intelligence perform ijtihad? Authority, epistemology, and legitimacy in digital Islamic jurisprudence. Suhuf: International Journal of Islamic Studies, 38(1), 154–169. https://doi.org/10.23917/suhuf.v38i1.16437

Akbar, M. S. (2026). AI and authenticity in Islamic research: A critical evaluation of generative AI reliability, hallucination, and source fidelity in Quranic, Hadith, and Fiqh knowledge [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2607.28237

Akhmetova, A., Rakhimova, G., Onngar, Y., Issakhan, M., & Baitenova, N. (2026). Digital religion and hybrid religious authority in algorithmically mediated environments. Frontiers in Sociology, 11, 1892383. https://doi.org/10.3389/fsoc.2026.1892383

Alam, S. M., Abdulhai, M., & Salehi, N. (2025). Blind faith? User preference and expert assessment of AI-generated religious content. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 2410–2424). Association for Computing Machinery. https://doi.org/10.1145/3715275.3732162

AlDahoul, N., & Zaki, Y. (2025). NYUAD at QIAS shared task: Benchmarking the legal reasoning of LLMs in Arabic Islamic inheritance cases. In Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks (pp. 861–866). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.118

Alnefaie, S., Atwell, E., & Alsalka, M. A. (2023). Is GPT-4 a good Islamic expert for answering Quran questions? In J.-L. Wu & M.-H. Su (Eds.), Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (pp. 124–133). Association for Computational Linguistics and Chinese Language Processing. https://aclanthology.org/2023.rocling-1.15/

Atallah, F. A. (2026). Digital mediation and fatwa authority in contemporary Islam: A critical Islamic legal and media-theoretical framework. Religions, 17(3), 350. https://doi.org/10.3390/rel17030350

Atif, F., Askarbekuly, N., Darwish, K., & Choudhury, M. (2025). Sacred or synthetic? Evaluating LLM reliability and abstention for religious questions. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8(1), 217–226. https://doi.org/10.1609/aies.v8i1.36543

Bhatia, G., Mubarak, H., Jarrar, M., Mikros, G., Zaraket, F., Alhirthani, M., Al-Khatib, M., Cochrane, L., Darwish, K., Yahiaoui, R., & Alam, F. (2026). From RAG to agentic RAG for faithful Islamic question answering [Preprint]. arXiv. https://arxiv.org/abs/2601.07528

Billah, M., Azhari, S., Baroroh, N., & Rahma, V. (2023). Evaluating AI-generated fatwas: A quality assessment of ChatGPT and Google Bard against authoritative Islamic rulings. Asy-Syir'ah: Jurnal Ilmu Syari'ah dan Hukum, 57(1), 136–163. https://doi.org/10.14421/ajish.v57i1.1032

Bouchekif, A., Rashwani, S., Sbahi, H., Gaben, S., Al Khatib, M., & Ghaly, M. (2025). Assessing large language models on Islamic legal reasoning: Evidence from inheritance law evaluation. In Proceedings of the Third Arabic Natural Language Processing Conference (pp. 246–257). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.arabicnlp-main.20

Bunt, G. R. (2018). Hashtag Islam: How cyber-Islamic environments are transforming religious authority. University of North Carolina Press.

Campbell, H. A., & Bellar, W. (2023). Digital religion: The basics. Routledge.

Campbell, H. A., & Tsuria, R. (Eds.). (2022). Digital religion: Understanding religious practice in digital media (2nd ed.). Routledge.

Elmahjub, E., Qadir, J., Mushtaq, A., Naeem, R., Ghaznavi, I., & Iqbal, W. (2026). IslamicLegalBench: Evaluating LLMs knowledge and reasoning of Islamic law across 1,200 years of Islamic pluralist legal traditions [Preprint]. arXiv. https://arxiv.org/abs/2602.21226

Faizin, N., Maarif, A. S., & Hanafi, Y. (2025). Considering religious moderation in Islamic law through AI ChatGPT and Bahsul Masail of Nahdlatul Ulama. Law, Innovation and Technology, 17(1), 271–288. https://doi.org/10.1080/17579961.2025.2469351

Falah, M. B., & Riyanta, R. (2025). Digital patriarchy: How Indonesian Islamic preachers construct gender asymmetry through YouTube fatwas. Al-Ahwal: Jurnal Hukum Keluarga Islam, 18(1), 103–122. https://doi.org/10.14421/ahwal.2025.18106

Fatmawati, F., Khoiriyah, N., Arbi, A., & Minangsih, K. (2026). Digital da'wah management and anti-hoax literacy: A multicultural local wisdom reading: Manajemen dakwah digital dan literasi anti-hoaks: Pembacaan kearifan lokal multikultural. Alhadharah: Jurnal Ilmu Dakwah, 25(1), 121–134. https://doi.org/10.18592/alhadharah.v25i1.20548

Fatmawati, F., Musyarofah, U., & Minangsih, K. (2026). Digital da'wah content in Islamic religious education: Critical reception and curriculum integration in Banten senior high schools. Reslaj: Religion Education Social Laa Roiba Journal, 8(7), 900–918. https://doi.org/10.47467/reslaj.v8i7.13058

Haditia, M., Abubakar, I. Y., Rokimin, & Fauzi, A. (2026). Design and validation of a quality assurance framework for AI educational technology in low-resource settings. *Educational Technology in Developing Countries, 1*(1), 83–102. https://doi.org/10.64929/etdc.v1i1.68

Hossain, K. A., Mahmud, J. S., Tuli, M. H., Mitra, A., Haque, S. M. T., & Sadeque, F. Y. (2025). Is lying only sinful in Islam? Exploring religious bias in multilingual large language models across major religions [Preprint]. arXiv. https://arxiv.org/abs/2512.03943

Israelsen, B., Carty, S., Coates, J., Fulda, N., Park, J., & Whiting, P. (2026). When AI takes sides on questions of faith: Persistent asymmetries in AI-mediated faith guidance [Preprint]. arXiv. https://arxiv.org/abs/2605.22975

Jahan, I., Kamruzzaman, S., Hossen, K. M. N., Nurain, N., & Islam, A. B. M. A. A. (2026). From mind mapping to dream interpretation: Understanding Muslim users' engagement with GenAI tools in religious context. In Human-computer interaction (Lecture Notes in Computer Science, Vol. 16702, pp. 353–372). Springer. https://doi.org/10.1007/978-3-032-29586-6_23

Kadous, M. W., Elsayed, A., Al Nahas, A., & Haress, A. (2026). Ansari: A retrieval-grounded Islamic AI assistant — Architecture, deployment, and lessons from 140,000 conversations [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2608.20390

Kannike, U. M. M., & Fahm, A. O. (2025). Exploring the ethical governance of artificial intelligence from an Islamic ethical perspective. Jurnal Fiqh, 22(1), 134–161. https://doi.org/10.22452/fiqh.vol22no1.5

Kurniawan, E., Ahmad, H., & Muttaqin, M. Z. (2025). From content creators to zakat payers: The fatwa of Indonesian Ulema Council and the rise of digital fiqh. *Islamic Law and Social Issues in Society, 1*(2), 99–115. https://doi.org/10.64929/ilsiis.v1i2.9

Lim, F. K. G. (2026). AI and generative charisma in religious practices. Religions, 17(5), 549. https://doi.org/10.3390/rel17050549

Lohlker, R., & Wahid, S. H. (2026). The transformation of Islamic religious authority. Religions, 17(4), 493. https://doi.org/10.3390/rel17040493

Mahendra, B. A., Munandar, A., & Nurzuliana, Z. (2026). Mahar as esteem, not price: Mediating classical fiqh, the Compilation of Islamic Law, and custom in Indonesia's plural legal order. *PFL: Plural Family Law Review, 1*(1), 59–80. https://www.ejournaltuah.id/index.php/pfl/article/view/13

Maheshwari, A., Gajkeshwar, A., Sharma, K., & Patel, V. (2026). SectEval: Evaluating the latent sectarian preferences of large language models [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2603.12768

Mohammad, N. I. S., & Sheikh, T. (2026). MIRAGE: Auditing anti-Muslim bias in frontier LLMs across reasoning, agentic, and time-coupled conditions [Preprint]. arXiv. https://arxiv.org/abs/2606.16562

Mubarak, H., Malhas, R., Mansour, W., Mohamed, A., Fawzi, M., Hawasly, M., Elsayed, T., Darwish, K. M., & Magdy, W. (2025). IslamicEval 2025: The first shared task of capturing LLMs hallucination in Islamic content. In Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks (pp. 480–493). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.67

Munirah, M., Alwi, A., & Karim, A. (2026). Evaluation of the quality of large language models' answers in answering religious questions for digital da'wah. Jurnal Asy-Syukriyyah, 27(1), 209–226. https://doi.org/10.36769/asy.v27i1.1822

Munn, Z., Pollock, D., Khalil, H., Alexander, L., McInerney, P., Godfrey, C. M., Peters, M., & Tricco, A. C. (2022). What are scoping reviews? Providing a formal definition of scoping reviews as a type of evidence synthesis. JBI Evidence Synthesis, 20(4), 950–952. https://doi.org/10.11124/JBIES-21-00483

Mushtaq, A., Naeem, R., Elmahjub, E., Ghaznavi, I., Al-Maliki, S., Abdallah, M., Al-Fuqaha, A., & Qadir, J. (2025). Can LLMs write faithfully? An agent-based evaluation of LLM-generated Islamic content [Preprint]. arXiv. https://arxiv.org/abs/2510.24438

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

Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2024). Having beer after prayer? Measuring cultural bias in large language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 16366–16393). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.862

Noh, M. A. C., Basri, M., & Panaemalae, A. (2026). Ethical leadership: Reframing moral complexity and power through a philosophical anthropological perspective in organizational practice. *Leadership in Muslim Societies, 1*(1), 19–34. https://doi.org/10.64929/lms.v1i1.57

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Peters, M. D. J., Godfrey, C., McInerney, P., Khalil, H., Larsen, P., Marnie, C., Pollock, D., Tricco, A. C., & Munn, Z. (2022). Best practice guidance and reporting items for the development of scoping review protocols. JBI Evidence Synthesis, 20(4), 953–968. https://doi.org/10.11124/JBIES-21-00242

Rahman, F., Pujiono, P., Khumaidah, S., & Harisudin, M. N. (2026). Managed fragmentation of religious authority: Intra-organizational fatwa contestation within Nahdlatul Ulama in the digital age. Al'Adalah, 29(1), 73–98. https://doi.org/10.35719/aladalah.v29i1.685

Ramadhan, M. A. I., Syam, A. R., & Barkah, S. A. M. (2026). Customay judges in digital space: Authority contestations, living fiqh, and legal pluralism in Bugis customary inheritance

dispute resolution via WhatsApp groups. *Tuah Customary Law Review, 1*(1), 1–20. https://ejournal.tuah.or.id/index.php/tclr/article/view/74

Ridwan, A., & Rustandi, R. (2025). Transformation of prophetic communication patterns in the era of artificial intelligence: Challenges and opportunities for Islamic preaching in Indonesia. Al-Balagh: Jurnal Dakwah dan Komunikasi, 10(1), 185–220. https://doi.org/10.22515/albalagh.v10i1.11551

Rochmawati, I. (2026). Digitalizing female religious authority: KUPI's gender-responsive fatwas and the reconfiguration of Islamic discourse in Indonesia. Al-Tahrir: Jurnal Pemikiran Islam, 26(1). https://doi.org/10.21154/altahrir.v26i1.12384

Rusli, R., & Nurdin, N. (2022). Understanding Indonesia millennial ulama online knowledge acquisition and use in daily fatwa making habits. Education and Information Technologies, 27(3), 4117–4140. https://doi.org/10.1007/s10639-021-10779-7

Shafira, Z. (2025). The two faces of deepfake: An analysis of contextual thematic exegesis. Jurnal Moderasi: The Journal of Ushuluddin and Islamic Thought, and Muslim Societies, 5(2), 67–80. https://ejournal.uin-suka.ac.id/ushuluddin/moderasi/article/view/6946

Singler, B. (2025). Religion and artificial intelligence: An introduction. Routledge.

Susanti, N. F. H., Tanjung, H., Supriyanto, T., Athief, F. H. N., & Furqoni, A. (2025). ChatGPT's validity and reliability in Islamic legal studies: A critical assessment of waqf for early learners. Suhuf: International Journal of Islamic Studies, 37(2), 496–518. https://doi.org/10.23917/suhuf.v37i2.11821

Tawakkal, M. G., Kurniawan, M. I., & Faathima. (2026). Cash waqf linked sukuk and economic empowerment: A grounded theory from Indonesia. *Ethiconomics: Journal of Islamic Banking and Finance, 1*(01), 1–20. https://doi.org/10.64929/ethiconomics.v1i01.2

Usmonov, M. (2025). From human scholars to AI fatwas: Media, ethics, and the limits of AI in Islamic religious communication. Al-i'lam: Journal of Contemporary Islamic Communication and Media, 5(1). https://doi.org/10.33102/jcicom.vol5no1.125

Wahid, S. H. (2025). Artificial intelligence in Islamic guidance: Assessing chatbot performance and jurisprudential adherence. Journal of Digital Islamicate Research, 3(1), 33–96. https://doi.org/10.1163/27732363-20250003

Whyte, S. A. (2022). Islamic religious authority in cyberspace: A qualitative study of Muslim religious actors in Australia. Religions, 13(1), 69. https://doi.org/10.3390/rel13010069

Yilmaz, I. (2026). Scholars, influencers, and algorithms: Youth encounters with online Islamic knowledge. In Navigating identity, belonging and power in cyberspace (pp. 151–174). Palgrave Macmillan. https://doi.org/10.1007/978-981-95-4660-2_7

Zaid, B., Fedtke, J., Shin, D. D., El Kadoussi, A., & Ibahrine, M. (2022). Digital Islam and Muslim millennials: How social media influencers reimagine religious authority and Islamic practices. Religions, 13(4), 335. https://doi.org/10.3390/rel13040335

Zhang, J., Song, W., & Liu, Y. (2025). Cognitive bias in generative AI influences religious education. Scientific Reports, 15, Article 15720. https://doi.org/10.1038/s41598-025-99121-6