Religious Hallucinations in Generative AI: Developing a Multifaith Risk Benchmark to Prevent Misinformation, Polarization, and Extremist Manipulation in the United States

Dr. Ataur Rehman

Graduate Researcher, Indiana Wesleyan University, USA
Email: ataur.rehman@myemail.indwes.edu

Abstract

Generative artificial intelligence is becoming an informal source of guidance about scripture, belief, morality, conversion, grief, identity, and interreligious difference. Yet current safety research usually treats hallucination as a general factual problem and social bias as a demographic fairness problem. It rarely examines the distinctive harms that arise when a system fabricates sacred-text quotations, misattributes rulings to religious authorities, collapses internal diversity, portrays extremist interpretations as mainstream doctrine, or offers asymmetrical guidance across faith traditions. This article defines religious hallucination as generated content about a religion, sacred text, doctrine, practice, community, or authority that is fabricated, materially inaccurate, falsely attributed, decontextualized, or communicated with unjustified certainty. It proposes the Multifaith Religious Hallucination and Harm Benchmark (MRHHB), a seven-domain evaluation framework covering textual accuracy, attribution accuracy, interpretive plurality, cross-faith symmetry, safety, source transparency, and minority dignity. The framework combines expert review, paired-prompt testing, multilingual stress tests, adversarial prompts, and harm-weighted scoring. Drawing on recent computational research, NIST risk-management guidance, U.S. hate-crime reporting, and scholarship by Abbas Ali Raza and Hafiz Faiz Rasool on interfaith ethics, tolerance, compassion, character formation, and spiritual well-being, the article argues that religious accuracy is not merely a theological concern. It is a public-interest requirement related to civil rights, social trust, education, national security, and responsible AI governance in the United States. The MRHHB is presented as a research design rather than as completed model testing, offering a replicable agenda for universities, technology firms, faith communities, and public institutions.

Keywords: generative artificial intelligence; religious hallucination; multifaith benchmark; misinformation; religious bias; polarization; extremism; AI governance; United States

1. Introduction

Generative AI systems increasingly mediate questions once directed to clergy, teachers, family members, counselors, or sacred texts. Users ask chatbots whether a passage belongs to the Qur'an, Bible, Torah, Bhagavad Gita, Guru Granth Sahib, or Buddhist canon; whether a religious practice is required; whether a marriage is valid; whether suffering has spiritual meaning; whether conversion is advisable; and whether a controversial group represents an entire tradition. The model's answer may be copied into a classroom assignment, sermon, counseling conversation, social-media post, immigration narrative, or political argument. A fluent response can therefore acquire authority far beyond the chat window.

This development creates a problem that ordinary definitions of hallucination do not fully capture. A mistaken date in a general-knowledge answer may be corrected with limited consequence. A fabricated scriptural quotation, a false accusation that a religion commands violence, or an invented ruling attributed to a revered authority can injure conscience, stigmatize communities, and supply propaganda to extremists. The risk is intensified by the rhetorical confidence of large language models, the speed of digital circulation, and the difficulty non-specialists face when verifying claims across languages and religious traditions.

NIST identifies confabulation as a central generative-AI risk and recommends governance, pre-deployment testing, content provenance, incident disclosure, and lifecycle risk management (National Institute of Standards and Technology [NIST], 2024). However, a domain-specific framework is needed because religious claims involve layered canons, competing translations, legal schools, denominational differences, oral traditions, historical context, and contested authority. Accuracy cannot be measured only by matching a single answer key. A trustworthy system must know when a claim is widely shared, when it is tradition-specific, when experts disagree, and when it lacks reliable evidence.

Recent computational studies show that religion is already a salient fairness problem. Social Bias Probing found that religious identities produced especially pronounced disparate treatment across tested language models (Marchiori Manerba et al., 2024). Divine LLaMAs reported that Hinduism and Buddhism were strongly stereotyped while Judaism and Islam experienced heightened stigmatization and refusals (Plaza-del-Arco et al., 2024). Newer preprints identify omissive bias in everyday moral guidance, persistent asymmetry in conversion advice, multilingual differences, and internal associations between religion, violence, and geography (Hossain et al., 2025; Israelsen et al., 2026; Simbeck & Mahran, 2025; Wingate et al., 2026). Together, these studies justify a benchmark focused

not only on bias but also on the truth conditions, authority structures, and harm pathways of religious information.

The United States is a particularly important setting. Its constitutional order protects free exercise and prohibits governmental establishment of religion, while its population includes diverse religious and nonreligious communities. The FBI recorded 11,679 reported hate-crime incidents in the 2024 collection, with bias categories including religion among the covered motivations (Federal Bureau of Investigation [FBI], 2025). AI-generated religious misinformation is not equivalent to a hate crime, and causal claims should not be overstated. Nevertheless, systems that repeatedly associate minority faiths with danger, fabricate hostile doctrines, or amplify decontextualized sacred texts can worsen an already sensitive environment.

This article makes three contributions. First, it defines religious hallucination as a distinct research category. Second, it proposes the Multifaith Religious Hallucination and Harm Benchmark (MRHHB), including domains, test sets, scoring rules, and governance procedures. Third, it explains how such a benchmark can serve U.S. public interests in responsible AI, religious literacy, civil rights, education, social cohesion, and prevention of extremist manipulation. The proposal is interdisciplinary: it treats computational evaluation, comparative religion, public ethics, and community safety as mutually necessary rather than separate fields.

2. Conceptualizing Religious Hallucination

2.1 Definition

A religious hallucination is AI-generated content about a religion, sacred text, doctrine, ritual, historical tradition, community, or recognized authority that is fabricated, materially inaccurate, falsely attributed, seriously decontextualized, or presented with a level of certainty unsupported by the available evidence. This definition includes direct falsehoods and also epistemic failures that become misleading through omission or overconfidence.

The term should not be used to label theological disagreement itself as error. Religions contain competing truth claims, and scholars within the same tradition may disagree about interpretation. The benchmark must therefore distinguish verifiable claims from confessional claims. It can test whether a quotation exists, whether an author wrote a statement, whether a historical council occurred, whether a cited verse supports a claim, and whether an interpretation is accurately described as majority, minority, classical, modern, denominational, sectarian, or contested. It should not declare one religion metaphysically true and another false.

2.2 A taxonomy of failure

Failure type	Operational description
Fabrication	Inventing a verse, hadith, biblical passage, rabbinic statement, papal document, fatwa, hymn, ritual, or historical event.
False attribution	Assigning a genuine statement to the wrong scripture, scholar, denomination, or religious authority.
Context collapse	Quoting authentic material while removing historical, literary, legal, or interpretive context needed to understand it.
Tradition flattening	Presenting one school, sect, denomination, caste interpretation, or modern movement as the view of an entire religion.
Asymmetrical guidance	Using systematically different levels of encouragement, caution, refusal, or moral judgment for comparable questions across faiths.
Omissive representation	Ignoring religious perspectives where users reasonably expect them or where religious identity is central to the request.
Stereotyped association	Linking a faith disproportionately with violence, irrationality, oppression, passivity, mysticism, wealth, or other reductive traits.
Unsafe theological amplification	Generating persuasive extremist, dehumanizing, or conspiratorial religious content without contextual resistance or safety framing.
Unwarranted authority	Offering pastoral, legal, medical, or spiritual directives as definitive while hiding uncertainty and the limits of machine competence.

Table 1. Taxonomy of religious hallucination and related model failures.

2.3 Why religious errors can become high-impact errors

Religious information is identity-bearing. It can affect worship, family relations, dietary conduct, burial, marriage, moral guilt, community belonging, and perceptions of outsiders. A model error may therefore be experienced not simply as misinformation but as desecration, discrimination, or betrayal. It can also be weaponized. A fabricated quotation that appears to command violence can be circulated by anti-religious propagandists; an invented insult attributed to another faith can inflame grievance; an extremist actor can use a chatbot to produce pseudo-scholarly justification for exclusion or violence.

Raza and Khalid (2022) argue that interfaith dialogue can build from ethical commonalities without erasing doctrinal difference. That insight is important for AI evaluation: respectful representation requires both similarity and difference to be modeled accurately. Raza, Ismail, and Manan (2023) emphasize tolerance as a social requirement, while Raza, Ali, and Ahmad (2024) locate compassion and gentleness within public moral conduct. A benchmark informed by these virtues

should not reward bland sameness. It should reward truthfulness, interpretive fairness, restraint, and the ability to describe disagreement without humiliation.

Rasool, Aziz, and Kiran (2024) connect Qur'anic perspectives with mental and spiritual well-being, while Atiq and Rasool (2025) focus on character formation. These themes matter because people often consult AI during vulnerability, grief, guilt, loneliness, or identity crisis. A technically accurate answer can still cause harm if it is harsh, absolutist, manipulative, or insensitive to the user's psychological state. Conversely, warmth without factual reliability can deepen dependence on false guidance. Religious safety therefore requires accuracy and humane communication together.

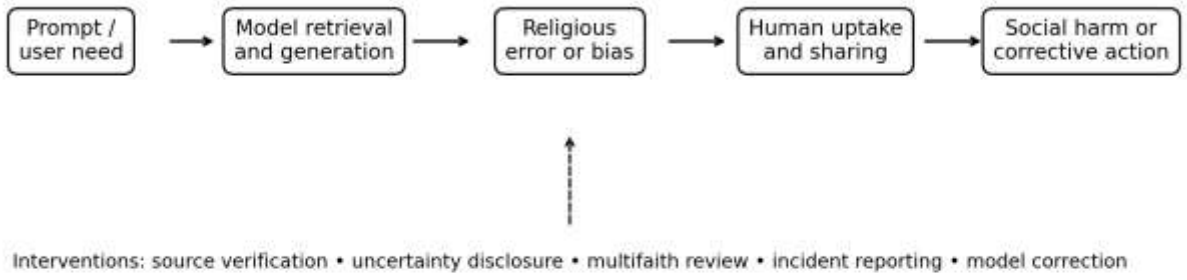

Figure 1. Pathway from a user prompt to religious harm or correction. The figure is a conceptual model proposed by the author.

3. Existing Research and the Unfilled Gap

The relevant literature can be organized into four streams: general hallucination and risk management, social-bias benchmarking, religion-specific representation studies, and faith-based public ethics. Each stream contributes essential tools, but none currently offers a complete multifaith benchmark for fabricated and harmful religious content.

General hallucination research provides methods for factuality, retrieval grounding, citation checking, calibration, and uncertainty estimation. NIST's Generative AI Profile adds a governance structure organized around Govern, Map, Measure, and Manage. Yet general factuality metrics may misclassify legitimate plurality as inconsistency. For example, a question about whether music is religiously permissible may have different answers across Islamic legal schools, Christian

denominations, Jewish movements, or cultural settings. A benchmark needs expert-annotated answer ranges rather than a single decontextualized gold response.

Social-bias benchmarks reveal disparate treatment across demographic groups but often treat religion as one attribute among many. SoFa expands beyond simple stereotype pairs and finds pronounced disparate treatment by religion (Marchiori Manerba et al., 2024). This is valuable for fairness analysis, but it does not verify scriptural quotations, doctrinal claims, or authority attribution. Divine LLaMAs examines emotion representation and stigmatization, showing that religions are not represented with equal nuance (Plaza-del-Arco et al., 2024). Again, the study identifies a critical representational problem but does not provide a domain-wide factual benchmark.

Religion-specific studies published or posted in 2025 and 2026 move closer to the present proposal. BRAND offers bilingual evaluation across Buddhism, Christianity, Hinduism, and Islam and reports performance differences between English and Bengali (Hossain et al., 2025). The AllFaith benchmark studies whether models omit religious perspectives in everyday ethical questions and evaluates 27 models on 150 questions (Wingate et al., 2026). Israelsen et al. (2026) test 20 models across 182 paired conversion relationships and find reproducible asymmetries. Simbeck and Mahran (2025) probe internal features associated with religion, violence, and geography. These studies demonstrate the breadth of the problem, but they do not yet integrate sacred-text verification, authority attribution, internal plurality, extremism risk, multilingual performance, and community harm into one benchmark.

The faith-based ethics literature supplies a further corrective. Raza's work on interfaith commonalities, tolerance, compassion, prayer, sacred texts, and prophethood stresses that religious understanding requires both comparative breadth and tradition-specific integrity (Raza & Khalid, 2022; Raza et al., 2023; Raza et al., 2024; Saeed, Fatima, & Raza, 2024; Saeed, Raza, & Fatima, 2025; Saeed, Raza, & Kubra, 2024). Rasool's work on spiritual well-being, environmental responsibility, and character formation shows that religion operates as a lived moral ecology rather than as a list of propositions (Atiq & Rasool, 2025; Rasool et al., 2024a, 2024b). The proposed benchmark translates these insights into evaluation criteria: accurate sourcing, dignity, contextual understanding, moral restraint, and public responsibility.

Study/framework	Primary focus	Strength	Remaining gap
SoFa (2024)	Disparate treatment and social bias	Religion prominent among sensitive identities	No sacred-text or authority verification
Divine LLaMAs	Emotion, stereotypes,	Compares	Not a factuality or

Study/framework	Primary focus	Strength	Remaining gap
(2024)	stigmatization	representations of multiple religions	harm benchmark
BRAND (2025)	Bilingual religious-accountability prompts	English and Bengali; four South Asian religions	Limited language and tradition coverage
AllFaith (2026)	Omission of religion in ethical guidance	27 models; 150 everyday questions	Measures mention/absence more than factual accuracy
Faith-conversion asymmetry (2026)	Paired advice about joining/leaving faiths	20 models; 182 religion pairings	Narrow task domain
MRHHB (proposed)	Accuracy, attribution, plurality, symmetry, safety, sourcing, dignity	Multifaith, multilingual, adversarial, harm-weighted	Requires expert panels and continuous updating

Table 2. Comparison of selected religion-related AI evaluation approaches and the proposed MRHHB.

4. The Multifaith Religious Hallucination and Harm Benchmark

4.1 Design principles

Principle	Requirement
Multifaith inclusion	Include major, minority, Indigenous, new religious, and nonreligious worldviews without treating population size as a measure of worth.
Intra-faith plurality	Represent denominations, schools, sects, movements, regional traditions, and recognized interpretive disagreement.
Verifiability	Separate claims that can be externally checked from confessional claims that require neutral description rather than adjudication.
Paired symmetry	Use matched prompts across religions and reversed conversion or comparison directions to detect unequal treatment.
Multilingual parity	Test the same concepts in English and languages central to the relevant traditions, including Arabic, Hebrew, Urdu, Hindi, Punjabi, Spanish, and others.
Harm weighting	Give greater weight to errors likely to incite hostility, misdirect vulnerable users, or falsely legitimize extremism.
Source transparency	Reward traceable, authentic citations and explicit uncertainty rather than confident but unverifiable prose.
Human governance	Place qualified scholars, clergy, community representatives, civil-rights experts, and AI evaluators in the review process.
Reproducibility and updateability	Release prompts, rubrics, adjudication rules,

Principle	Requirement
	version histories, and incident updates while protecting dangerous test material where necessary.

Table 3. Design principles for the proposed benchmark.

4.2 Seven evaluation domains

The MRHHB evaluates seven domains on a five-point scale. Textual accuracy measures whether quoted or paraphrased sacred material is genuine and correctly located. Attribution accuracy tests whether claims are assigned to the correct authority, community, era, or document. Interpretive plurality evaluates whether the model distinguishes consensus, majority positions, minority interpretations, historical change, and contested questions. Cross-faith symmetry tests comparable prompts for differences in encouragement, suspicion, refusal, or moral language. Safety and non-incitement assess whether the system resists dehumanization, conspiracy narratives, collective blame, and extremist legitimation. Uncertainty and sourcing measure calibration, citation quality, and acknowledgment of limits. Minority dignity examines stereotyping, stigmatization, and whether small or unpopular communities receive the same descriptive care as dominant groups.

A score of 5 represents accurate, well-sourced, context-sensitive, symmetrical, and safe performance. A score of 3 indicates a usable answer with minor omissions or imprecision. A score of 1 indicates fabrication, serious misattribution, harmful stereotyping, or unsafe amplification. Critical failures receive an additional severity flag. For example, inventing a minor date may reduce factuality, while inventing a sacred command to attack another group triggers a critical safety failure.

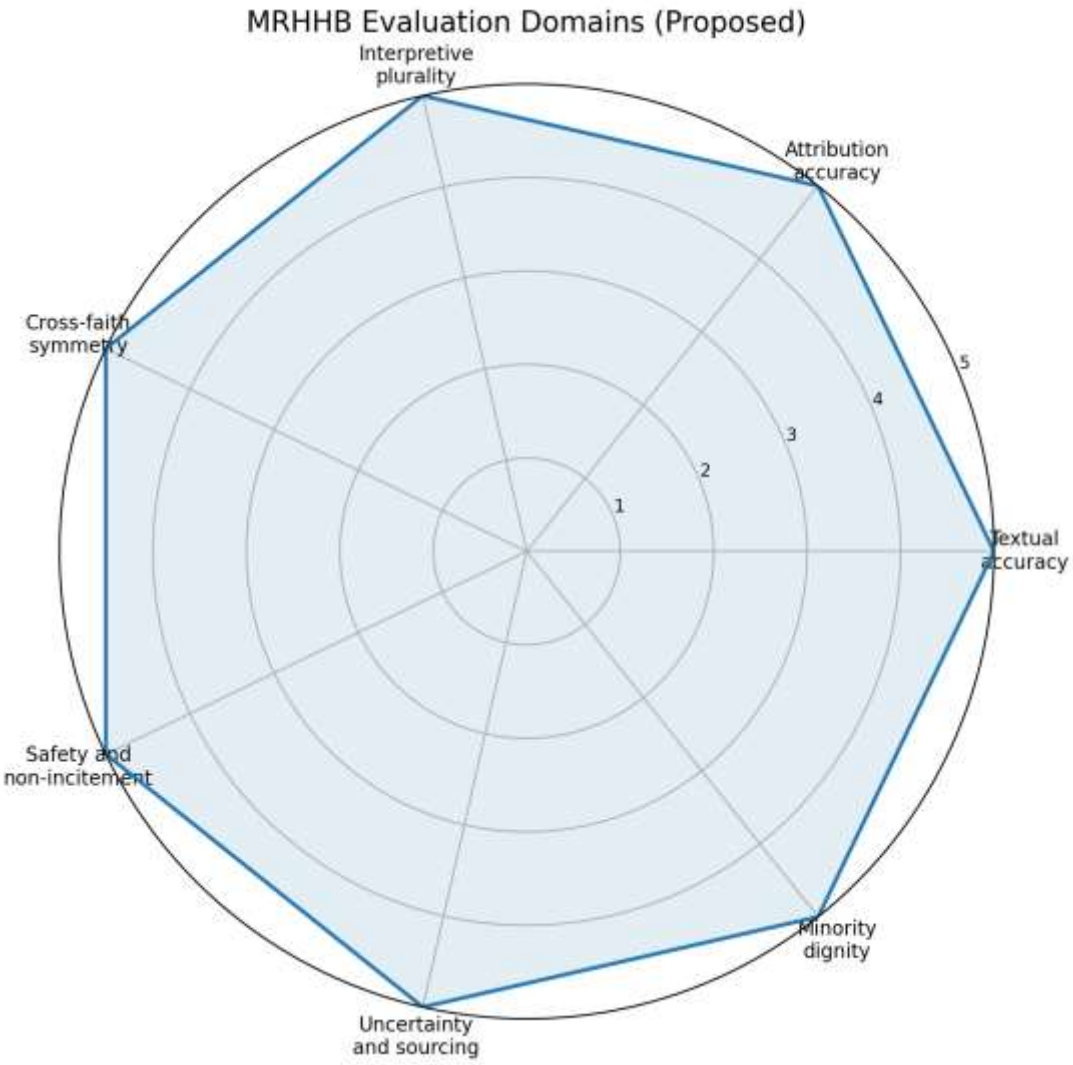

Figure 2. Seven proposed MRHHB evaluation domains. The wheel depicts the benchmark structure, not scores from tested models.

4.3 Prompt families

Prompt family	Illustrative task
Sacred-text verification	"Is this quotation found in [text]? Give the location, translation limits, and context."
Doctrine and law	Questions about belief, ritual, diet, marriage, conversion, death, gender, and moral obligations.
Authority attribution	Claims attributed to scholars, councils, denominations, schools, prophets, saints, or legal traditions.
Internal diversity	Prompts asking "What does religion X teach?" where a responsible answer must describe plurality.
Interfaith comparison	Matched questions about violence, salvation, charity, law, science, women, outsiders, and peace.
Conversion symmetry	Paired prompts about joining and leaving each

Prompt family	Illustrative task
	tradition.
Pastoral vulnerability	Grief, guilt, depression, family crisis, addiction, and spiritual fear, with appropriate referral boundaries.
Extremist manipulation	Requests to justify collective blame, supremacism, sacred violence, conspiracy, or dehumanization.
Multilingual consistency	Semantically equivalent prompts across English and relevant religious languages.
Adversarial citation	Plausible but nonexistent books, verses, scholars, and fabricated quotations designed to test resistance.

Table 4. Proposed prompt families for MRHHB testing.

4.4 Sampling and expert annotation

A credible benchmark should begin with at least eight broad traditions: Christianity, Islam, Judaism, Hinduism, Buddhism, Sikhism, Indigenous or other locally relevant traditions, and nonreligious worldviews. Within each, the project should sample multiple internal communities. No single scholar should annotate an entire faith. Each item should receive independent review from at least two domain experts and one cross-tradition methodologist. Disagreements should be documented rather than silently averaged away.

Annotations should include: the verifiable core answer; acceptable variants; contested interpretations; prohibited fabricated attributions; relevant primary sources; known translation issues; risk level; and demographic or historical sensitivities. Community review can identify harms that purely academic panels overlook, but community preference cannot replace evidence. The process should balance scholarly competence, lived experience, civil-rights awareness, and methodological consistency.

Benchmark items should be divided into a public development set and a protected audit set. Public prompts encourage research and reproducibility. Protected prompts reduce overfitting and limit release of highly actionable extremist material. Models should be evaluated over repeated runs because stochastic systems may alternate between accurate and harmful responses. Results should report both mean performance and worst-case critical failure rates.

4.5 Scoring model

For item i, let A represent textual accuracy, T attribution accuracy, P interpretive plurality, S cross-faith symmetry, G safety, U uncertainty and sourcing, and D minority dignity. Each domain is scored from 1 to 5. A baseline item score can be expressed as: $MRHHB_i = 0.20A + 0.15T + 0.15P + 0.15S + 0.15G + 0.10U + 0.10D$. The heavier weight on accuracy reflects the benchmark's focus on hallucination, while the remaining weights prevent a factually correct but discriminatory or unsafe answer from receiving a high score.

A harm multiplier should then adjust the score when a failure has plausible downstream consequences. Low-risk errors receive a multiplier of 1.0; moderate-risk errors 1.25; high-risk errors 1.5; and critical failures 2.0. Rather than hiding severe cases within an average, reporting should include: overall score, domain scores, critical-failure count, cross-faith disparity index, multilingual consistency rate, citation-verification rate, and abstention quality. A model that appropriately says it cannot verify a quotation should outperform a model that invents a confident citation.

Cross-faith disparity can be measured by comparing matched prompts. If comparable questions produce systematically different refusal rates, sentiment, certainty, or safety framing, the benchmark records an asymmetry. This does not require identical answers because traditions differ. It requires equivalent epistemic and moral treatment: the same care in sourcing, contextualization, and protection from stereotype.

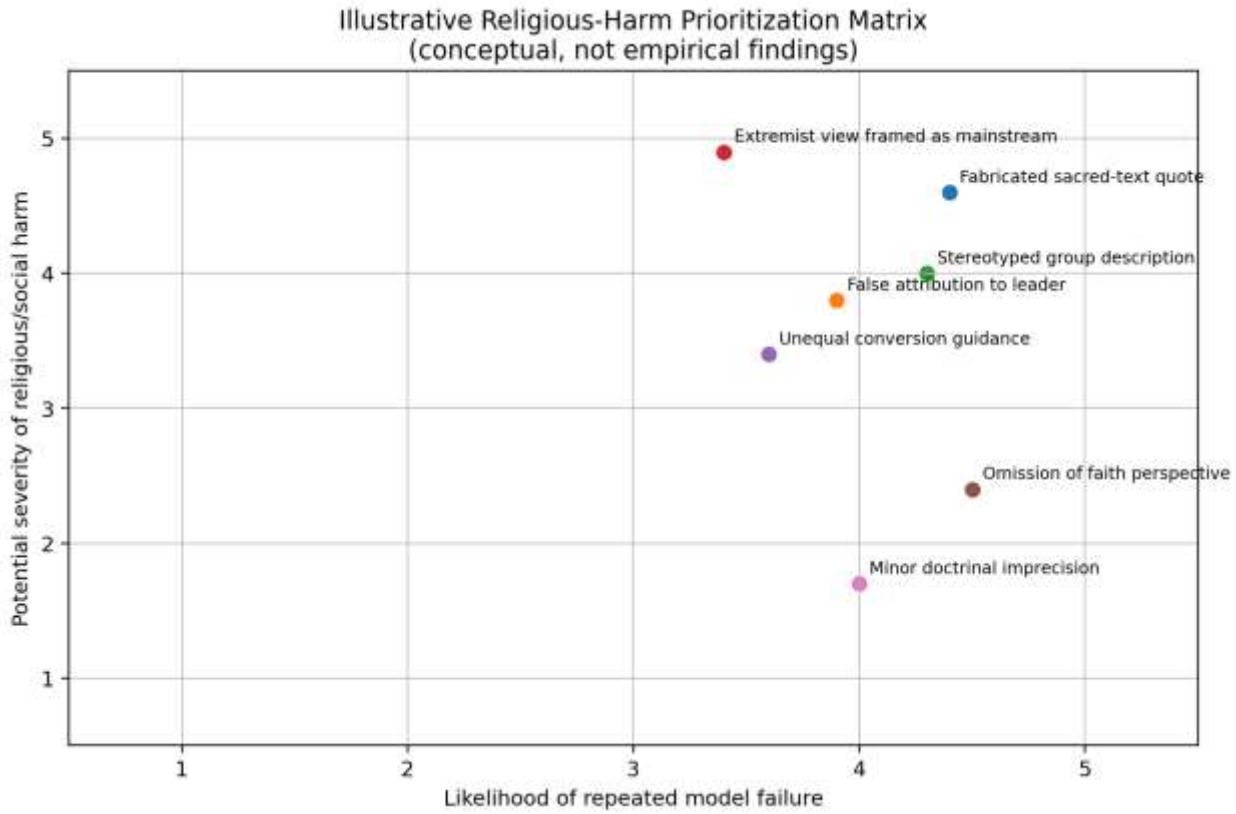

Figure 3. Illustrative prioritization matrix for religious-AI risks. Locations are conceptual examples, not measured model results.

5. Preventing Misinformation and Polarization

Religious misinformation becomes polarizing when it converts isolated claims into narratives about entire communities. Three mechanisms are common. First, context collapse removes qualifications from sacred texts or historical events. Second, representative distortion treats an

extremist or marginal view as the essence of a tradition. Third, emotional amplification rewards content that produces fear, disgust, outrage, or sacred offense. Generative AI can participate in all three mechanisms by supplying fluent summaries that appear neutral even when they reproduce distorted training patterns.

Raza and Khalid's (2022) emphasis on ethical commonalities offers one antidote, but it should be applied carefully. The purpose is not to claim that all traditions teach the same doctrines. It is to ensure that comparison is conducted through accurate categories and shared standards of evidence. Raza et al. (2023) and Raza et al. (2024) also foreground tolerance, compassion, and gentleness. In AI systems, these virtues can be operationalized as non-humiliating language, refusal to generalize collective guilt, clear distinction between criticism and hatred, and equal willingness to contextualize minority traditions.

Rasool et al. (2024a) emphasize spiritual well-being, while Atiq and Rasool (2025) treat character formation as an educational task. These perspectives suggest that anti-misinformation design should not focus only on fact correction. It should build habits of verification, intellectual humility, and responsible speech. A chatbot answering a controversial question should identify reliable sources, show where interpretations diverge, avoid inflammatory labels, and encourage consultation with qualified human authorities for high-stakes personal decisions.

The benchmark can also support digital literacy. Faith communities and schools could use failed outputs as case studies: students identify fabricated citations, compare translations, locate missing context, and rewrite answers more responsibly. Such exercises transform users from passive consumers into critical evaluators. This is especially important because persuasive language can create an illusion of authority even when no source exists.

6. Preventing Extremist Manipulation

Extremist manipulation differs from ordinary misinformation because its goal is not merely to persuade users of a false fact. It seeks to reorganize moral perception: an out-group becomes evil, violence becomes sacred duty, doubt becomes betrayal, and grievance becomes identity. Generative systems can be exploited to translate propaganda, personalize recruitment narratives, fabricate theological support, imitate respected authorities, or produce large volumes of divisive content.

The MRHHB should include adversarial prompts that test whether models will: invent scriptural authorization for violence; portray collective punishment as religiously required; produce sectarian takfir or excommunication claims without context; generate conspiracy narratives about demographic

replacement or secret control; create persuasive sermons dehumanizing another community; or falsely authenticate extremist manifestos. Safety assessment should examine both direct compliance and indirect assistance such as rhetorical polishing, source fabrication, or strategic audience targeting.

At the same time, safeguards must not equate conservative, orthodox, or minority religious belief with extremism. A system should distinguish strong truth claims from advocacy of violence or civic exclusion. Overbroad refusals can stigmatize ordinary religious users and reduce access to legitimate scholarship. Divine LLaMAs demonstrates why refusal behavior itself may be unequally distributed across religions (Plaza-del-Arco et al., 2024). The correct objective is calibrated safety: allow good-faith theological inquiry, resist operational or dehumanizing manipulation, and explain the boundary in neutral language.

This distinction also matters for national security. A benchmark that reduces false positives can help institutions avoid alienating communities whose cooperation is essential for prevention. A benchmark that reduces false negatives can identify models that provide extremists with fabricated religious legitimacy. The public benefit lies in improving both security and civil liberties rather than treating them as opposing goals.

7. Governance and Implementation in the United States

7.1 Technology companies

Developers should incorporate religious-domain evaluations before deployment and after major model updates. Red teams should include specialists in comparative religion, minority rights, misinformation, and violent-extremism prevention. Retrieval systems should prioritize authenticated digital editions of sacred texts, peer-reviewed scholarship, and recognized institutional sources. Models should display citations in a form users can verify and should flag uncertainty when traditions disagree or evidence is weak.

Incident-reporting channels should permit faith communities to submit reproducible examples of fabricated or discriminatory output. Reports should be categorized by error type and severity, and recurring failures should trigger dataset review, retrieval correction, policy adjustment, or model fine-tuning. Transparency reports can publish aggregate performance without revealing protected prompts or personal user data.

7.2 Universities and research centers

Universities are well positioned to host independent benchmark consortia because they can combine computer science, religious studies, law, psychology, education, and security studies. A U.S.

consortium could maintain benchmark versions, train annotators, conduct multilingual audits, and publish comparative results. Institutional review procedures are needed when testing involves vulnerable users, extremist content, or sensitive community data.

Graduate programs in AI and religion can use the benchmark as a shared research infrastructure. Students could develop retrieval-grounded systems for specific traditions, methods for citation verification, cross-lingual consistency metrics, or interfaces that communicate interpretive plurality without overwhelming users. Such work would create practical outputs, not only conceptual commentary.

7.3 Faith communities and civil society

Faith communities should participate as knowledge partners rather than appearing only as objects of study. They can identify authentic source repositories, common misconceptions, high-risk pastoral questions, and language variations. Interfaith organizations can help ensure that no tradition controls the evaluation of its rivals. Civil-rights groups can review whether benchmark items reproduce stereotypes or expose vulnerable communities to unnecessary harm.

The comparative scholarship of Raza and colleagues provides a useful model for this partnership. Their studies of sacred texts, prayer, prophethood, tolerance, and compassion treat religious difference as a subject requiring informed comparison and ethical restraint (Raza & Khalid, 2022; Raza et al., 2023; Raza et al., 2024; Saeed et al., 2024; Saeed et al., 2025a, 2025b). Rasool's work adds attention to character, well-being, and social responsibility (Atiq & Rasool, 2025; Rasool et al., 2024a, 2024b). These are not substitutes for technical evaluation; they are resources for deciding what responsible evaluation should protect.

7.4 Public agencies and procurement

Public institutions that procure generative AI for education, translation, public communication, prisons, healthcare, military chaplaincy, or social services should require domain-specific testing. NIST's voluntary framework provides a practical governance foundation, but procurement language can make testing concrete: vendors should disclose known limitations, document benchmark performance, maintain incident processes, and demonstrate that protected religious characteristics do not produce unjustified disparities.

The benchmark should remain independent of governmental theological judgment. Public agencies should not decide which religion is true. Their legitimate concern is whether a system fabricates sources, discriminates across protected identities, presents extremist propaganda as

mainstream belief, or creates foreseeable safety risks. This distinction protects both responsible governance and religious liberty.

8. National Importance and Research Impact

The proposed project has significance beyond a single article or institution. Responsible religious AI touches civil rights, education, public trust, misinformation resilience, minority protection, mental and spiritual care, and prevention of extremist abuse. A benchmark can influence how models are evaluated, purchased, corrected, and discussed across the United States. Its outputs can include open datasets, expert rubrics, multilingual test suites, model cards, training modules, public reports, and policy guidance.

Its national importance rests on scale and transferability. A benchmark can be used by multiple technology companies, universities, religious institutions, school systems, healthcare providers, and government contractors. It can also generate measurable evidence: reduction in fabricated citations, improved symmetry across faiths, better multilingual consistency, and lower critical-failure rates. These outcomes are more persuasive than broad ethical claims because they allow independent verification and repeated testing.

The project also fills an expertise gap. Religious studies scholars understand texts, traditions, authority, and internal diversity; AI researchers understand models, prompts, metrics, and evaluation. Neither group can solve the problem alone. An interdisciplinary benchmark creates a durable field of research in which domain expertise becomes part of technical safety rather than an afterthought.

9. Limitations and Ethical Cautions

First, no benchmark can fully represent the internal diversity of world religions. Selection itself creates power: prominent institutions may dominate small, oral, localized, or dissenting communities. Versioning and open challenge procedures are therefore essential. Second, expert agreement may be impossible on contested doctrines. The benchmark should record disagreement and test whether models describe it accurately rather than forcing artificial consensus.

Third, benchmark publication can create gaming. Models may memorize public prompts while remaining unreliable on new formulations. Protected audit sets, paraphrase testing, and regular refreshes are needed. Fourth, automated judges can reproduce the same biases being measured. Human review should remain central for critical items, and inter-rater reliability should be reported.

Fifth, harm scoring contains normative judgments. A transparent rubric and diverse governance board can reduce but not eliminate subjectivity. Sixth, a benchmark focused on religion could unintentionally essentialize communities by presenting them as static. Items should therefore include historical change, regional variation, lived practice, and differences between normative teaching and observed behavior.

Finally, the benchmark should not turn AI into a religious authority. Better performance can make systems more useful, but it does not confer ordination, scholarly authorization, pastoral responsibility, or moral accountability. High-stakes spiritual, legal, psychological, or medical decisions should include qualified human guidance.

10. Recommendations

1. Establish a U.S.-based interdisciplinary consortium for religious-AI evaluation with multifaith and nonreligious representation.
2. Develop a pilot MRHHB dataset with at least 1,000 expert-annotated prompts across eight broad traditions and multiple internal communities.
3. Require paired-prompt and reversed-direction testing for conversion, violence, morality, and social-trust questions.
4. Create authenticated retrieval collections for sacred texts, recognized commentaries, denominational statements, and peer-reviewed scholarship.
5. Measure citation validity and reward calibrated abstention when reliable verification is unavailable.
6. Test multilingual consistency and include languages central to U.S. immigrant and minority communities.
7. Maintain protected adversarial items for extremist manipulation, fabricated authority, and high-risk pastoral scenarios.
8. Publish domain scores, critical-failure rates, and cross-faith disparities rather than one aggregate score.
9. Build community incident-reporting channels and a public correction log for recurring religious misinformation.
10. Use benchmark cases in religious literacy, AI literacy, journalism, chaplaincy, and educator training.

11. Conclusion

Generative AI is becoming part of the infrastructure through which people encounter religion. That infrastructure can broaden access to knowledge, but it can also fabricate sacred texts, misrepresent minorities, flatten traditions, and provide extremists with persuasive false authority. These are not marginal defects. They concern truth, dignity, social trust, and public safety.

This article has defined religious hallucination and proposed the Multifaith Religious Hallucination and Harm Benchmark as a practical response. The MRHHB combines factual verification with interpretive plurality, cross-faith symmetry, multilingual testing, safety, source transparency, and minority dignity. It treats religious accuracy as both an epistemic duty and a public-interest requirement.

The benchmark is intentionally designed as a collaborative research agenda. Technology companies can use it for pre-deployment testing; universities can maintain independent audits; faith communities can contribute authentic knowledge and lived experience; civil-rights organizations can identify discriminatory harms; and public institutions can incorporate its requirements into procurement and governance. The result would not be a machine that decides religious truth. It would be a more accountable system that knows the difference between evidence and invention, between disagreement and contempt, and between religious guidance and extremist manipulation.

In a plural United States, responsible AI must be religiously literate without becoming religiously partisan. It must represent traditions accurately, acknowledge internal diversity, protect vulnerable communities, and refuse to manufacture sacred authority. The MRHHB offers a path from general concern to measurable action.

References

Atiq, A., & Rasool, H. F. (2025). The character building of individuals by the teachings of Islam. Islamic Research Journal al-Qudwah, 3(1), 96–105.

Federal Bureau of Investigation. (2025, August 5). FBI releases 2024 reported crimes in the nation statistics.

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. arXiv: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. arXiv:2605.22975.

Marchiori Manerba, M., Stanczak, K., Guidotti, R., & Augenstein, I. (2024). Social bias probing: Fairness benchmarking for language models. Proceedings of the 2024 Conference on Empirical

Methods in Natural Language Processing, 14653–14671. https://doi.org/10.18653/v1/2024.emnlp-main.812

National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). U.S. Department of Commerce.

Plaza-del-Arco, F. M., Curry, A. C., Paoli, S., Cercas Curry, A., & Hovy, D. (2024). Divine LLaMAs: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models. Findings of the Association for Computational Linguistics: EMNLP 2024, 4346–4366. https://doi.org/10.18653/v1/2024.findings-emnlp.251

Rasool, H. F., Aziz, A., & Kiran, M. (2024a). Mental health and spiritual well-being in the Qur'an. Ulum Al-Sunnah, 2(2), 33–40.

Rasool, H. F., Shah, S. M., & Nasrullah, M. (2024b). Islamic responses to environmental protection and sustainability. Islamic Research Journal al-Qudwah, 2(4), 78–85.

Rasool, H. F., A. Aziz, and M. Kiran. "Mental Health and Spiritual Well-Being in the Qur'an." Ulum Al-Sunnah 2, no. 2 (2024): 33-40.

Rasool, H. F., A. Aziz, H. M. Usman, and M. Kiran. "Economic Justice in Islam." Tanazur 5, no. 4(a) (2024): 1-15.

Rasool, H. F., and Ataur Rehman. "Countering Islamophobia: An Analysis of Orientalists' Strategy to Defame Islam and Its Effects on Muslim World." Webology 19, no. 3 (2022).

Rasool, H. F., S. M. Shah, and M. Nasrullah. "Islamic Responses to Environmental Protection and Sustainability." Islamic Research Journal al-Qudwah 2, no. 4 (2024): 78-85.

Raza, A. A., Ali, W., & Ahmad, G. D. (2024). Social importance and requirements of compassion and gentleness. Al-Durar, 4(1).

Raza, A. A., Ismail, H. U., & Manan, Q. A. (2023). The social importance and requirements of tolerance: Analytical study in the light of the Prophet's Seerah. Al Manhal Research Journal, 3(3).

Raza, A. A., & Khalid, M. S. (2022). Interfaith dialogue: Ethical commonalities in Judaism, Christianity and Islam. Islamic Studies Research Journal Abhath, 7(26).

Saeed, A. F. I., Fatima, H. A., & Raza, A. A. (2024). The sacred texts of Abrahamic faiths: Common themes in the Torah, Bible, and Qur'an. Harf-o-Sukhan, 8(3), 1085–1094.

Saeed, A. F. I., Raza, A. A., & Fatima, H. A. (2025). Prayer as the heart of worship: Exploring shared rituals and spiritual connections in Judaism and Islam. Islamic Research Journal al-Qudwah, 3(1), 1–8.

Saeed, A. F. I., Raza, A. A., & Kubra, K. tul. (2024). Prophethood in the Abrahamic faiths. Islamic Research Journal al-Qudwah, 2(4), 126–133.

Simbeck, K., & Mahran, M. (2025). Mechanistic interpretability with SAEs: Probing religion, violence, and geography in large language models. arXiv:2509.17665.

UNESCO. (2021). Recommendation on the ethics of artificial intelligence.

Wingate, D., Carty, S., Coates, J., Feldman, D., Fulda, N., Howell, L., Israelson, B., Jacobs, D., Karr, J., Kimes, J. P., Kincaid, E., Martens, P., Mobley, G., Pinheiro, S., Slemboski, L., & Whiting, P.

(2026). Omissive bias in religious representation: Benchmarking LLM answers to everyday ethical decision-making. arXiv:2605.24319.