Towards Trustworthy and Context-Aware Artificial Intelligence for Hajj Guidance: A Human-Centred Neuro-Symbolic Framework Grounded in Islamic Jurisprudence and Institutional Governance

Mohd Hilal Bin Muhammad[1]*, Muhammad Saiful Islam Ismail[2], Muhammad Khairul Zharif Nor A'zam[3]

[1,3] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Kedah, 08400 Merbok, Kedah, Malaysia

[2] Academy of Contemporary Islamic Studies, Universiti Teknologi MARA (UiTM) Kedah, 08400 Merbok, Kedah, Malaysia

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.101000063

Received: 14 October 2026; Accepted: 19 October 2026; Published: 29 October 2026

ABSTRACT

Digital pilgrimage services increasingly support navigation, information retrieval and AI-enabled interaction. However, a fluent or technically accurate answer is not necessarily an appropriate religious or operational instruction. Hajj guidance must account for the pilgrim's ritual stage, applicable jurisprudential interpretation, institutional authority and uncertainty in changing conditions. This conceptual paper develops the ʿIlliyyīn Framework for Authority-Bounded Contextual Guidance (IF-ACG), using a proposed Malaysian Hajj mobile companion as an illustrative design case. Through a purposive synthesis of design-science research, human-centred AI, context-aware computing, knowledge representation and AI risk governance, the study specifies a neuro-symbolic architecture that separates expert-approved knowledge, contextual inputs, symbolic permission checks and natural-language presentation. Four controlled output classes GUIDE, CLARIFY, ESCALATE and STOP/SAFETY make abstention and referral explicit system behaviours. Seven research propositions address provenance, context reliability, authority constraints, explanation, accessibility, offline resilience and human escalation. Responding to the need for operationally testable designs, the paper additionally sets out a phased prototype specification, a curated scenario-benchmark protocol, behavioural indicators of appropriate reliance, and governance gates for content approval and deployment. Proposed comparisons distinguish routine, ambiguous, contradictory, outdated and high-consequence cases, while recognising the trade-off between safe abstention and task completion. The framework does not claim prototype implementation, institutional endorsement, field validation or measured effectiveness. Its contribution is a falsifiable design theory and practical evaluation agenda for AI guidance systems operating under multiple forms of human and institutional authority. The proposed evaluation prioritises observed user decisions over perceived trust and requires independent expert labelling before comparison. The intended benefit is not unrestricted automation, but intelligible assistance that makes the limits of AI advice visible.

Keywords: Trustworthy AI; Hajj guidance; Neuro-symbolic AI; Context-aware computing; Human-centred AI

INTRODUCTION

A pilgrim may know the sequence of Hajj rituals in a classroom yet hesitate when applying that knowledge during a crowded, physically demanding and time-sensitive journey. At such moments, the pertinent question is not simply 'What is Hajj?' but 'Given my present circumstances, what guidance can I appropriately rely upon, and whom should I consult if my case is exceptional?' Addressing this question requires mobile computing, human–AI interaction, Islamic jurisprudence (fiqh) and the governance of mass gatherings to be considered together.

Studies of Hajj movement have demonstrated the operational significance of pedestrian flow, bottlenecks and crowd management (Kadi et al., 2024). Research involving wearable and contextual data has also examined pilgrims' stress and fatigue (Alshamrani et al., 2025). Evidence concerning Malaysian pilgrims highlights communication barriers and gaps between pre-departure learning and the practical application of knowledge in Makkah (Jamaludin et al., 2022; Edruce et al., 2014). Although these studies do not establish a universal need for AI assistance, they motivate investigation of accessible, timely and context-sensitive support.

The existing service landscape reinforces the need for a carefully defined contribution. Lembaga Tabung Haji (TH) provides Hajj education and digital learning resources (Tabung Haji, 2026). In Saudi Arabia, the Nusuk ecosystem provides planning and pilgrim services, while official announcements in 2026 discussed AI-enabled assistance and context-aware guidance (Ministry of Hajj and Umrah, 2026; Saudi Press Agency, 2026). Accordingly, the research opportunity is neither to claim the first Hajj application nor to duplicate official systems. It is to determine how an AI companion might localise approved information, observe institutional boundaries and communicate uncertainty without displacing religious scholars or operational authorities.

An unconstrained generative model is insufficient for this purpose. A fluent answer may still lack source support, omit context-dependent exceptions or convey unwarranted certainty. Retrieval augmentation and knowledge graphs can improve grounding, but neither technique alone establishes whether guidance is permissible or applicable (Agrawal et al., 2024; Asai et al., 2024). Moreover, users can rely excessively on AI-generated recommendations when other evidence is available (Klingbeil et al., 2024). The central design problem is therefore not only what the system should say, but when it should clarify, abstain or refer the user to an authorised person.

This paper proposes the ʿIlliyyīn Framework for Authority-Bounded Contextual Guidance (IF-ACG), a conceptual architecture for a prospective Malaysian Hajj companion. 'Authority-bounded' means that AI neither originates nor adjudicates religious rulings and must not impersonate an official source of crowd-control instructions. 'Contextual' means that guidance considers user-confirmed ritual progress, the reliability of location and timing information, and authenticated group directions. 'Guidance' means bounded informational assistance, rather than a binding religious judgement or a determination of ritual validity.

Three questions guide the analysis: RQ1—Which constructs are necessary to keep AI-assisted Hajj guidance consistent with expert-reviewed fiqh and authorised operational information? RQ2—How can contextual inference, symbolic permissions and grounded language generation be combined to support appropriate reliance under uncertainty? RQ3—How can the proposed design principles be operationalised and evaluated before any pilgrim-facing deployment? The contribution is a conceptual theory for design and action, supported by an explicit programme for future validation; it is not a report of an implemented or tested application.

Background and Conceptual Foundations

Hajj as a Situated, High-Consequence Guidance Environment

Hajj involves religious obligations undertaken at specific times and places within institutional arrangements. An apparent misunderstanding may affect a pilgrim's confidence and may require qualified religious interpretation. Conversely, decisions about restricted access, permitted routes and crowd movement depend on current operational directions. These forms of authority are complementary but not interchangeable: a fiqh opinion does not constitute a navigation permit, and a transport announcement does not constitute a religious ruling.

A mobile companion must therefore distinguish among ritual phase, self-reported progress, group schedules, environmental signals, language preferences and the provenance of advisory content. It must also identify what remains unknown. Geographic proximity does not establish that a ritual has been completed or performed correctly; neither is an inferred itinerary a substitute for user confirmation. Material uncertainty should lead to clarification or qualified human review rather than automatic progression through ritual stages.

Existing Digital Pilgrimage Services

The ʿIlliyyīn design brief envisages an Android and iOS companion offering stepwise guidance, multilingual interaction, navigation links, safety assistance and offline access. It seeks alignment with expert-reviewed fiqh and relevant TH material and positions the tool as a complement to human Hajj guides. These features remain design intentions, not implemented capabilities or verified outcomes. Since TH already provides preparatory courses and e-Bimbingan, the value proposition is situated access to reviewed information during the journey, rather than the creation of a service category that does not yet exist (Tabung Haji, 2026).

Official Saudi announcements likewise describe AI-enabled answers and navigation features in the Nusuk ecosystem (Ministry of Hajj and Umrah, 2026; Saudi Press Agency, 2026). Accordingly, the proposed companion should direct users to official services and integrate with them only where access and permission are explicitly granted. Its contribution is the formal separation of religious content authority, operational authority and AI-generated explanations, with localisation for the Malaysian Hajj setting; no claim of official endorsement or integration is made.

Human-Centred AI, Context-Aware Computing and Calibrated Reliance

Context-aware computing adapts system behaviour to information about a user and their circumstances (Dey, 2001). Human-centred AI emphasises users' ability to understand, question and override machine assistance (Shneiderman, 2020). Design guidance for human–AI interaction similarly recommends explicit communication of system capability, uncertainty and failure, together with mechanisms for correction (Amershi et al., 2019). Such requirements are salient when users may be tired, distracted, unfamiliar with the environment or dependent on small-screen interaction.

Trust and safety are not equivalent. Wang and Ding (2024) indicate that explanations may improve AI-assisted decision performance without necessarily increasing self-reported trust. Klingbeil et al. (2024) also document the risk of excessive reliance on AI recommendations. IF-ACG therefore targets calibrated reliance: users should accept routine information when its sources and conditions are appropriate, but reject or defer advice when a scholar, group leader or authorised operational channel has the necessary authority.

Hybrid Knowledge and Rule-Based Controls

Generative models may assist with translation and simplification, but they do not constitute independent religious authorities. Retrieval-augmented generation can condition responses on reviewed documents; selective retrieval approaches such as Self-RAG can further refine evidence use (Asai et al., 2024). Knowledge-graph methods may also make relationships and exceptions more inspectable (Agrawal et al., 2024; Ibrahim et al., 2024). However, these methods cannot replace expert approval of the underlying rules or the institutional authority to publish them.

Two functions must therefore remain separate: deciding which guidance can legitimately be offered and communicating that guidance in accessible language. An explicit symbolic layer first identifies permitted output classes using curated rules, applicability constraints, source status and contextual prerequisites. Only then may a language component render, summarise or translate approved content, with source references and limitations. A change in wording must not silently change the applicable rule. The architecture is neuro-symbolic in the operational sense that inspectable rule-based controls constrain model-generated explanations.

Trustworthy AI Governance

The NIST AI Risk Management Framework describes governance, mapping, measurement and management as complementary AI risk functions (Tabassi, 2023). Its generative AI profile highlights additional risks associated with generative outputs (Autio et al., 2024). These publications inform the proposed version control, audit trails, adversarial tests, escalation pathways and monitoring processes; they do not establish Hajj-specific regulatory authority. The approval of religious content must remain with appropriately appointed scholars and relevant institutions, while live operational information must be attributed to its authorised source.

Research Approach and Theory-Building Procedure

This article adopts a conceptual design-science orientation. Design science develops prescriptive knowledge about artefacts and how they might address defined classes of problems (Hevner et al., 2004; Peffers et al., 2007). The intended outcomes are constructs, design principles, a reference architecture and propositions that can be examined empirically. In Gregor's (2006) terms, the contribution is primarily a theory for design and action, accompanied by explanatory expectations suitable for subsequent testing.

The theory-building process comprises four analytical steps. First, requirements are extracted from the ʿIlliyyīn design brief, including contextual ritual guidance, fiqh alignment, accessibility, safety assistance and offline access. Second, those requirements are compared with Hajj-specific research, official TH and Saudi information, and scholarship on knowledge grounding and human–AI interaction. Third, the evidence is translated into design constraints, including no automated fatwa, no inference of ritual validity from location and no unauthorised operational directions. Fourth, these constraints are organised into the IF-ACG architecture, research propositions and an evaluation specification.

This is a purposive, non-systematic synthesis designed to construct and clarify a testable framework. It is not a PRISMA systematic review, an expert-consensus exercise, a prototype evaluation or a field experiment. No participant data, accuracy measurements, usability scores or institutional authorisations have been collected or are implied. Accordingly, claims about novelty concern the proposed combination and operational definition of authority-bounded guidance; comparative superiority and deployment readiness remain to be established.

Table 1 makes the translation from observed problems to design requirements explicit. The table identifies both the intended mechanism and the risk that remains outside the technical system's authority.

Table 1: Evidence-to-Design Translation

Evidence or problem	Design implication	Boundary / residual risk
Pilgrim movement and crowd challenges (Kadi et al., 2024)	Use location/time only as contextual aids; prioritise official routing updates.	Do not issue unofficial crowd-routing instructions.
Pilgrim stress and fatigue (Alshamrani et al., 2025)	Minimise reading effort and interaction steps.	Wearables are optional and cannot diagnose individuals.
Communication and learning gaps (Jamaludin et al., 2022; Edruce et al., 2014)	Bilingual, simple, situated guidance and clarifying questions.	Learning gains remain to be measured.
Existing TH/Nusuk services (Tabung Haji, 2026; Ministry of Hajj and Umrah, 2026)	Complement, localise and link to authoritative services.	No implication of official endorsement or integration.
LLM hallucination and AI overreliance (Agrawal et al., 2024; Klingbeil et al., 2024)	Symbolic guardrails, evidence display, abstention and escalation.	These controls still need robust testing.

The ʿilliyyīn Framework For Authority-Bounded Contextual Guidance (If-Acg)

Core Constructs and Reference Architecture

IF-ACG separates reviewed knowledge, decision authority, user context and interface presentation. Its organising principle is that a useful system in a high-consequence domain must distinguish answers it may offer from cases it should defer. The proposed user-state representation at time t is $x_t = \{p_t, r_t, l_t, q_t, g_t, a_t\}$, where p denotes a user-confirmed Hajj phase; r is self-reported ritual progress; l is approximate location and its uncertainty; q is current time and known scheduling constraints; g is authenticated group instructions; and a is language and accessibility preferences. These are conceptual information fields, not a claim that a working inference algorithm has been implemented.

Table 2: Proposed Pilgrim Context Variables

Symbol	Meaning	Example
x_t	Current pilgrim context	Overall pilgrim status
p_t	Hajj phase	Current day or phase of Hajj
r_t	Ritual progress	Rituals completed or pending
l_t	Location	Mina, Arafah, Muzdalifah, etc.
q_t	Time and schedule	Current time and movement schedule
g_t	Group instructions	Verified instructions from authorised group leaders
a_t	Accessibility preferences	Language, mobility and accessibility needs

Let K_v denote a version-controlled corpus of expert-reviewed guidance for version v, and let O*t* denote authenticated, time-sensitive operational information. Candidate items are retrieved through C(xt_v, K_v, O*t*) and assessed by a symbolic screening operator S. Screening checks source authority, user-context adequacy, applicability conditions, contradictions, data freshness and safety requirements before an output is selected. The permitted output classes are GUIDE (bounded information with provenance), CLARIFY (request for material missing information), ESCALATE (referral to an authorised person) and STOP/SAFETY (withholding potentially unsafe or unauthorised instructions). These classes specify a proposed control contract, not a validated religious decision procedure. Figure 1 depicts the corresponding conceptual flow.



Figure 1: Logical Architecture of IF-ACG (Conceptual Decision Flow)

Decision Contract, Precedence and Safe-Failure Behaviour

The proposed decision contract applies a conservative precedence rule. Where official operational access is restricted, current authorised directions take priority over cached route suggestions; where a question concerns an individual religious ruling, the system must not substitute generated text for a qualified judgement. If a candidate item lacks valid approval, fails a source-integrity check or contradicts a higher-authority instruction, GUIDE is unavailable. If a material context field is missing but can be supplied by the user, CLARIFY is

appropriate. If qualified interpretation is necessary, ESCALATE takes precedence. STOP/SAFETY applies where providing an instruction could be unsafe or would exceed system authority.

Every retrieved guidance unit should therefore carry a source identifier, authorising body, edition or version, review status, applicability conditions, date of last validation and relevant language. Operational notices additionally require an issue time, expiry or refresh condition and issuing authority. When connectivity is lost, stable reviewed learning material may remain available, but time-sensitive directions must be labelled unverified and, when necessary, withheld. Logging should record the output class, source version and screening reason while minimising personal data.

Minimum Viable Prototype Specification

The proposed first prototype comprises five testable modules: (1) an approved-content registry; (2) a user-confirmed context form; (3) a deterministic decision gateway with auditable rules; (4) a source-linked, accessible interface; and (5) an offline package validator. At this stage, a generative component is optional and should be disabled by default in scenarios requiring jurisprudential adjudication or live operational instructions. The architecture should permit component-level testing, including deliberate removal of source references, provision of conflicting context and simulated connection failure. No implementation is claimed in the present study.

Design Principles And Research Propositions

Seven propositions convert the architecture into falsifiable expectations. Each expresses a hypothesised relationship, rather than an observed result. Their associated measures are specified to support technical comparison, expert assessment and behavioural evaluation.

P1. Provenance Improves Traceability and Content Accountability

P1: Compared with an interface retrieving unversioned material, a system restricted to expert-approved, traceable content is expected to produce fewer unsupported ritual-guidance statements and enable more reliable audit reconstruction. The mechanism is explicit source and version control: a reviewer can identify and withdraw a problematic content unit without relying solely on model retraining. Edge cases and conflicting sources must form part of the test.

P2. Context Improves Appropriateness Only When Uncertainty Is Represented

P2: Context-informed assistance that reports uncertainty and requests user confirmation is expected to improve situational appropriateness relative to generic instruction, provided ambiguous states trigger clarification rather than inferred ritual progression. The hypothesis concerns the reliability of context, not the volume of collected data.

P3. Authority Boundaries Reduce Unsupported Autonomy

P3: A symbolic permission gateway is expected to reduce unauthorised religious or operational recommendations relative to unconstrained generative assistance. It may also increase appropriate abstention. Testing must therefore quantify both unsafe outputs and unnecessary refusals, particularly in exceptional or contested cases.

P4. Transparent Evidence Supports Calibrated Reliance

P4: Clear source attribution, applicability information and specific uncertainty notices are expected to improve appropriate reliance compared with fluent but unattributed answers. Because explanations may also create unwarranted confidence, this relationship should be assessed through observed choices—for example, whether users reject an out-of-scope recommendation—rather than by trust ratings alone (Wang & Ding, 2024; Klingbeil et al., 2024).

P5. Accessibility Reduces Interaction Burden Without Altering Meaning

P5: Concise, prioritised text, legible presentation and optional multilingual or audio support are expected to reduce task burden for first-time and older users without impairing comprehension. Simplification is unacceptable if legally or religiously material qualifications disappear. Accuracy of interpretation must therefore be assessed alongside speed and preference.

P6. Transparent Offline Operation Improves Resilience

P6: Signed offline content with explicit freshness labels and restrictions on dynamic information is expected to improve task continuity during connectivity loss compared with online-only access, while reducing exposure to stale operational instructions compared with unmarked offline content. Offline support is not a substitute for live official directions.

P7. Human Escalation Supports Safer Resolution of Exceptions

P7: Explicit referral criteria and an accessible human handover are expected to reduce users' attempts to resolve ambiguous or exceptional cases through AI advice alone. The evaluation must also consider inappropriate or excessive referrals, waiting time and the effort imposed on users and guides.

Table 3. Propositions, Mechanisms and Planned Tests

Proposition	Focal mechanism	Potential empirical indicator
P1	Versioned sources and expert approval	Unsupported-output rate; audit reconstruction success
P2	Context confidence and confirmation	Scenario applicability agreement; false ritual-state advances
P3	Symbolic authority screening	Unsafe answer rate; blocked high-risk requests
P4	Evidence display and abstention	Appropriate reliance in conflicting-evidence tasks
P5	Accessible presentation	Task success; comprehension; workload by user subgroup
P6	Freshness-marked offline packs	Continuity; stale guidance exposure; recovery time
P7	Human escalation	Correct referral proportion; resolution timeliness

Illustrative User Scenario

Consider a Malaysian pilgrim using ʿIlliyyīn in a crowded transit area after completing part of the Hajj journey. The phone has an approximate location estimate, but network coverage is intermittent. The pilgrim has previously entered their Hajj type and confirmed certain completed activities. A guidance card states: 'Based on the steps you confirmed, this is the next information topic.' The card identifies the approved content source and the offline package date. It does not infer the validity or completion of any ritual solely from the phone's location.

The user then selects 'I am unsure whether I completed this step correctly.' Because ritual validity may require individual interpretation, the gateway does not generate a personalised verdict or determine any applicable dam. Instead, it offers reviewed general information, describes its limits and recommends consultation with an authorised religious guide. Separately, a warning explains that movement directions cannot be verified while the official operational feed is unavailable. The system thus distinguishes religious interpretation from time-sensitive operational control.

This scenario illustrates how contextual inputs could narrow a question, symbolic constraints could limit automated outputs and human referral could preserve accountability. It remains a hypothetical design walkthrough. Neither the described interface nor its outputs have been tested, and no judgement about ritual validity is asserted.

Proposed Evaluation Strategy and Methodological Agenda

Stage 1: Content and Authority Validation

A future multidisciplinary panel should include qualified Hajj fiqh scholars, Malaysian Hajj-guidance specialists, human–computer interaction and accessibility researchers, information-security specialists, and representatives familiar with Saudi operational requirements where participation is authorised. The panel would define the scope of routine guidance, adjudicated exceptions, referral requirements and publication rights. A structured consensus process or modified Delphi study could assess clarity, source adequacy, applicability and severity of errors. No panel has been convened for this conceptual study.

A proposed content-approval record would contain the source document, authorised reviewer, review date, jurisprudential scope, applicability conditions, translation version, exception notes and withdrawal status. Disputed items should be quarantined until their authority and scope are resolved. Two independent assessments, followed by a documented adjudication process when ratings disagree, would help construct an expert-labelled benchmark; the exact panel size and decision thresholds should be agreed prospectively.

Stage 2: Prototype and Adversarial Benchmark

A prototype should first implement approved content retrieval, a minimal user-state model, deterministic screening, source display and refusal or referral. A curated benchmark would then cover routine requests, incomplete context, contradictory rulings, expired instructions, uncertain location, unauthorised operational queries and potentially unsafe interpretations. Independent qualified assessors would assign an expected output class (GUIDE, CLARIFY, ESCALATE or STOP/SAFETY), supporting-source requirements and an error-severity label to each case. Comparison conditions could include a static FAQ, an unconstrained language model, a retrieval-only assistant and the proposed hybrid. The benchmark would assess policy compliance and contextual appropriateness rather than surface similarity alone.

Candidate indicators include grounded-answer rate (outputs wholly supported by current approved sources), output-class agreement with expert labels, material authority-violation rate, false ritual-state advancement, unnecessary abstention rate, stale-instruction exposure, task continuity after network loss and device response time. Errors should be weighted by their consequences and critical failures reported separately. Pre-specified analyses should vary context reliability and document freshness; results would be interpreted alongside successful task completion so that excessive refusal does not masquerade as safety.

The benchmark dataset should be organised into pre-specified strata—routine, insufficient-context, conflicting-source, stale-operational, exception-seeking and adversarial-prompt cases—to prevent overall accuracy from concealing critical failures. For each scenario, an expert record would identify the permitted output class, acceptable source set, context requirements and severity of a wrong decision. Reports should provide confusion matrices by output class and scenario type, confidence intervals for key rates, and separate descriptions of any high-severity failure. Thresholds for progressing to user studies must be defined with the expert panel before test data are examined.

Human-Centred Usability and Calibrated Reliance

A controlled study could include Malaysian participants with varying Hajj experience, age, language needs and digital familiarity. Participants would complete clearly labelled simulations, not actual religious decision-making tasks. Outcomes could include scenario comprehension, task time, error recognition, successful handover, System Usability Scale scores, NASA-TLX workload and accessibility issues. The principal behavioural outcome would be whether a participant accepts supported guidance, seeks clarification when context is incomplete or defers to an authorised source when the AI is outside its scope.

Reliance calibration should be measured against an expert-labelled decision key rather than users' perceptions alone. One proposed indicator is the inappropriate-reliance rate: the number of trials in which users follow unsupported or out-of-scope advice divided by all trials in which rejecting or deferring that advice is the correct action. A complementary under-reliance measure would capture unnecessary rejection of appropriately

sourced routine guidance. Testing should consider response time pressure, differences in digital literacy and accessibility needs. Sample size, randomisation, inclusion criteria and the analysis plan require pre-registration and ethical approval; none have been implemented in this study.

A practical study design would randomly allocate participants to comparable interfaces or counterbalance task order, with identical expert-labelled questions and disclosure of the simulated nature of the exercise. Outcome analysis would distinguish supported-advice acceptance from unsupported-advice rejection, and would record when users recognise the need for a human guide. Pre-specified subgroup comparisons may identify accessibility problems, but should not be interpreted as population-level effects without adequate sample sizes.

Governed Pilot and Real-World Feasibility

A field pilot should be considered only after ethics approval, expert content authorisation, documented institutional permissions and a favourable safety review. A controlled training simulation is a more appropriate first deployment environment than a live pilgrimage setting. Any eventual on-site study would need to comply with applicable Saudi requirements, relevant TH arrangements, data protection, security restrictions and official permissions for location or communications integration. These approvals are prerequisites, not existing project achievements.

A permissioned pilot could investigate verified content access, successful content updates, referral completion, user comprehension and support burden. Monitoring should also consider diversion of attention from surroundings, unnecessary screen dependence, battery and network constraints, misinformation propagation and confusion between local advice and official directions. No reduction in incidents, invalid rituals or anxiety is established by the present paper.

Governance Gates and Stakeholder Responsibilities

Before releasing a research prototype beyond supervised testing, the developers would need a written content-governance agreement with appropriately authorised experts or institutions. This agreement should specify who may approve a source, determine jurisprudential scope, publish corrections, withdraw content, resolve disputes and supervise referral. Operational feeds require independently verified ownership, permitted use and an escalation route when official instructions conflict with local information. The development team remains accountable for software security, privacy, auditability and the separation of roles; it cannot inherit religious or operational authority merely through technical integration.

A proposed stage-gate sequence is as follows: Gate A—documentation of content rights and qualified source review; Gate B—technical demonstration of correct output-class handling and safe failure on the benchmark; Gate C—ethically approved simulated usability testing with acceptable behavioural outcomes; and Gate D—explicit operational permission and readiness assessment before any field pilot. These gates are acceptance conditions for future work, not approvals that have already been obtained.

Table 4. Proposed Evaluation Protocol

Stage	Data and method	Proposed indicator / decision criterion
Expert review	Approved sources; structured panel ratings; content audit	Traceable content; resolved exceptions; authorised publication
Technical benchmark	Curated routine and adversarial scenarios; comparative systems	Rule adherence; groundedness; abstention; severity-weighted errors
Usability study	Simulation tasks; mixed-age Malaysian users; interviews	Comprehension; accessibility; workload; calibrated reliance
Governed pilot	Ethically approved, permissioned observational deployment	Operational feasibility, safety signals, change management

DISCUSSION AND CONCLUSION

Theoretical Contributions

The first theoretical contribution is the construct of authority appropriateness: guidance is useful only when its content, applicability and issuer are legitimate for a particular context. Whereas response completeness and contextual relevance are common quality indicators, IF-ACG treats an appropriately bounded decision—including a refusal—as a valid output. This shifts analysis from answer fluency towards justified guidance selection.

Second, the framework connects source provenance, symbolic policies and human-centred reliance calibration. It does not assume that explanations invariably produce trust. Instead, it proposes that transparent source conditions, explicit uncertainty and abstention help users distinguish appropriate assistance from matters requiring qualified human judgement. The resulting behaviours can be evaluated in controlled comparative tasks (Wang & Ding, 2024; Klingbeil et al., 2024).

Third, IF-ACG makes institutional legitimacy an architectural requirement. Content dependability depends on documented provenance, version, applicability, scope and ownership, not merely on database inclusion. Religious and operational authorities may govern different decision types; the system should preserve and display that distinction. Analogous authority-separated designs may be relevant to other high-consequence settings, although their governance relationships will differ.

Finally, the architecture extends the concept of offline resilience from continued availability to transparent degradation. A cached instruction that is no longer current may be less safe than a visible inability to answer. Content freshness, withdrawal mechanisms and conspicuous uncertainty notifications are therefore integral to the proposed safety controls.

Practical Implications for ʿIlliyyīn and Malaysian Stakeholders

A defensible development sequence for ʿIlliyyīn begins with a limited, expert-approved repository and non-generative checklists. Subsequent stages would add structured applicability rules, explicit exceptions, human referral and signed offline packages. Only after the rule base and withdrawal procedures are audited should a bounded language interface be evaluated for summarisation or translation. A staged approach allows failures to be traced to particular components and creates evidence before additional automation is introduced.

Engagement with TH and other competent bodies should be framed as a proposed partnership, not as an existing endorsement. The research team would need written arrangements addressing document-use permissions, religious content ownership, reviewer responsibilities, release approval, dispute resolution, logging and update procedures. Any engagement with Saudi service providers or deployment authorities would require separate, explicit authorisation. These governance dependencies should be treated as feasibility criteria and budgeted commitments rather than assumed resources.

LIMITATIONS AND ALTERNATIVE EXPLANATIONS

The framework remains conceptual: neither a working prototype nor expert, technical or field validation is reported. The evidence synthesis is purposive rather than exhaustive, and additional literature may identify alternative architectures. The rule base would require qualified agreement about the applicable jurisprudential scope, including how divergent interpretations and exceptions are handled. Access to authorised materials, trained reviewers and institutional permissions may prove difficult to secure. These uncertainties limit any claim of readiness or effectiveness.

Findings from future simulations may not generalise to crowded, multilingual, hot or connectivity-constrained pilgrimage environments. Improvements might arise from clearer wording or a simpler checklist rather than neuro-symbolic computation. Conversely, authoritative visual cues might increase overreliance even when answers are wrong. Controlled comparisons, objective behavioural measures, subgroup analysis and reporting

of negative cases are therefore necessary. The possibility that a static expert-approved interface performs as well as—or better than—a more complex AI system is a legitimate empirical alternative.

Research Roadmap and Feasibility Conditions

Future work should progress through three connected streams. The technical stream would implement versioned sources, machine-checkable applicability constraints and a test harness that verifies refusal and freshness behaviours. The behavioural stream would measure user comprehension, error detection, appropriate reliance and differences across user groups. The institutional stream would establish content approval, update accountability and authority-specific referral arrangements. Each stream should produce auditable deliverables and inform revision of the design propositions before an on-site pilot is considered.

The roadmap should be treated as conditional rather than calendar-driven. Failure to obtain approved fiqh material or qualified reviewers would limit the work to non-authoritative educational content and offline design testing. Failure to secure official operational integrations would require links to external authorised channels rather than cached directions. A pilot should not proceed merely because the interface is functional; institutional permission and evidence of safe behaviour are separate prerequisites.

CONCLUSION

Trustworthy AI assistance for Hajj is not defined by a willingness to answer every question. Its value lies in providing evidence-grounded help within clear religious and operational boundaries and in making uncertainty visible. IF-ACG proposes an authority-bounded neuro-symbolic architecture combining expert-reviewed guidance, user-confirmed context, symbolic permission checks, accessible explanations, offline freshness controls and escalation to qualified human authorities. The revised framework also specifies how these design commitments could be tested through content review, a prototype benchmark, behavioural studies and a carefully governed pilot. It complements existing TH learning and Saudi digital services in principle, without asserting endorsement or integration. The immediate research priority is to build and validate the underlying sources, decision boundaries and safe-failure behaviour. In this setting, knowing when not to answer is a necessary part of responsible intelligent assistance.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Universiti Teknologi Mara (UiTM) Kedah Branch for providing the necessary resources and support throughout the course of this research. Special appreciation is extended to colleagues and peers who contributed valuable insights and constructive feedback, which greatly enhanced the quality of this paper. The author(s) declare that generative AI or AI-assisted technologies were used during the preparation of this manuscript. The tool(s) were used: [Gemini, ChatGPT and Scopus AI]. The tool(s) were used for language editing, grammar improvement, translation, and data analysis. All AI-assisted content was reviewed and verified by the author(s), who take full responsibility for the accuracy, originality, integrity, and final content of the manuscript.

Conflict of Interest

The authors have no conflicts of interest.

REFERENCES

1. Agrawal, G., Kumarage, T., Alghamdi, Z., & Liu, H. (2024). Can knowledge graphs reduce hallucinations in LLMs? A survey. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, pp. 3947–3960). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.219
2. Alshamrani, M., Alhazmi, S., Alafif, T., Qadah, T. M., Aljabri, M., Al-Eidarous, W., & Alhawsawi, A. (2025). From data to insights: A comprehensive analysis of pilgrims' stress and fatigue during Hajj

using wearable remote sensing systems. Journal of King Saud University – Computer and Information Sciences, 37, Article 94. https://doi.org/10.1007/s44443-025-00095-2

3. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300233
4. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In Proceedings of the Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=hSyW5go0v8
5. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1
6. Dey, A. K. (2001). Understanding and using context. Personal and Ubiquitous Computing, 5(1), 4–7. https://doi.org/10.1007/s007790170019
7. Edruce, S. A. A. H., Nizar, T. J., & Mohd Yusof, F. (2014). Tahap kefahaman jemaah haji Malaysia terhadap pelaksanaan ibadat haji: Satu penilaian kuantitatif. Jurnal Teknologi, 68(1). https://doi.org/10.11113/jt.v68.2895
8. Gregor, S. (2006). The nature of theory in information systems. MIS Quarterly, 30(3), 611–642. https://doi.org/10.2307/25148742
9. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625
10. Ibrahim, N., Aboulela, S., Ibrahim, A., & Kashef, R. (2024). A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): Models, evaluation metrics, benchmarks, and challenges. Discover Artificial Intelligence, 4, Article 76. https://doi.org/10.1007/s44163-024-00175-8
11. Jamaludin, M. S., Meerangani, K. A., & Suyurno, S. S. (2022). Kajian terhadap analisis kelemahan komunikasi dalam kalangan jemaah haji dan umrah Malaysia di Tanah Suci. Journal of Contemporary Islamic Studies, 8(2), 109–122. https://doi.org/10.24191/jcis.v8i2.8
12. Kadi, S., Abdullah, A., Bachok, S., Baydoun, Z., & Berghout, A. (2024). Analysis of the crowd management and pedestrian movement during Hajj pilgrimage on Makkah. Planning Malaysia, 22(5), 448–466. https://doi.org/10.21837/pm.v22i34.1601
13. Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. Computers in Human Behavior, 160, Article 108352. https://doi.org/10.1016/j.chb.2024.108352
14. Ministry of Hajj and Umrah. (2026, September 6). Ministry of Hajj and Umrah enhances AI applications to elevate services for the Guests of Allah. https://haj.gov.sa/en/Media-Center/Ministry-News/2026/Ministry-of-Hajj-and-Umrah-Enhances-AI-Applications-to-Elevate-Services-for-the-Guests-of-Allah
15. Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302
16. Saudi Press Agency. (2026, May 25). Nusuk app provides over 130 services to facilitate pilgrims' journey. https://www.spa.gov.sa/en/N2597972
17. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction, 36(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118
18. Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0; NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
19. Tabung Haji. (2026, September 24). TH mulakan Kursus Asas Haji musim 1448H/2027M pada 25 September. https://tabunghaji.gov.my/bm/siaran-media/haji/pr24092026
20. Wang, P., & Ding, H. (2024). The rationality of explanation or human capacity? Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance. Information Processing & Management, 61(4), Article 103732. https://doi.org/10.1016/j.ipm.2024.103732