Artificial Intelligence Systems and Its Applications (AISA)
Volume 2, Issue 1, June 2026
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

Hikmah AI: A Retrieval-Augmented Islamic Knowledge Assistant Using Hybrid Search

***Muhammad Rakha Abimanyu*[1], *Endang Wahyu Pamungkas*[2]**

[1] Universitas Muhammadiyah Surakarta, Sukoharjo, Indonesia

Abstract: *This research presents Hikmah AI, a Retrieval-Augmented Generation (RAG)-based Islamic knowledge assistant built on an external large language model (OpenAI GPT-5), designed to provide valid, contextual Islamic information sourced from the Qur'an, Hadith, and trusted Islamic literature. The background of this research is driven by the increasing public demand for Islamic knowledge in the digital era, while many available sources remain unstructured and lack credibility. The system combines hybrid search (vector and keyword retrieval) with a three-tier topic classifier and a configurable admin dashboard, developed following the Waterfall methodology across the stages of requirements analysis, design, implementation, testing, and maintenance. The testing process employed Precision, Recall, F1-Score, Relevance, and Accuracy as evaluation metrics on a preliminary set of three retrieval test cases, complemented by expert review of five fiqh questions conducted by a qualified Islamic scholar. On this limited test set, the system achieved Precision, Recall, and F1-Score of 100%, along with an average Relevance and Accuracy of 86.2%. Expert validation concluded that the reviewed responses were consistent with the original scriptural texts and free of substantive error. These preliminary results suggest that the proposed hybrid RAG architecture is a promising approach for grounding Islamic knowledge assistance in authentic sources, though broader empirical evaluation is needed before stronger generalisable claims can be made.*

Keywords: *Chatbot, Large Language Model, Retrieval-Augmented Generation, Islamic Knowledge, Hybrid Search, Knowledge Base, RAG Evaluation.*

Article History:
Received: 12 May 2026 Accepted: 30 June 2026 Published: 30 June 2026

Corresponding Author: Muhammad Rakha Abimanyu, Email: rakhaabimanyu5@gmail.com

DOI: https://doi.org/10.65917/aisa.v2i1.69

1 Introduction

In the current digital era, public demand for Islamic information has increased significantly. Many people seek explanations of Islamic law, Islamic history, and general Islamic knowledge via the internet. However, available information is often unstructured and potentially invalid, originating from sources lacking credibility [1]. This situation creates difficulty for the public in obtaining accurate, contextually appropriate understanding that aligns with Islamic teachings.

The revolution of Artificial Intelligence (AI) in the Industry 5.0 era has transformed the interaction between humans and machines into an adaptive collaboration, reshaping methods of learning and information provision [2]. Large Language Models (LLMs) have opened new avenues for interactive, rapid, and wide-coverage text-based information delivery [3]. When implemented within an application, LLMs allow intelligent systems to process information and generate coherent textual output with strong generalisation across various task types [4]. Through these capabilities, LLMs can process, comprehend, and generate answers based on user-provided context.

However, most existing LLMs remain general-purpose and have not been optimised for specific domains such as Islamic knowledge. Consequently, responses may be insufficiently accurate, lack depth, or even deviate from accredited Islamic literature [5]. This highlights the need to develop an LLM specifically designed to address Islamic science based on the Qur'an, Hadith, and valid Islamic literature [6], rather than relying solely on an LLM's unconstrained parametric knowledge.

It is also important, at the outset, to distinguish between three categories of Islamic AI assistance, since they carry different levels of risk.

Islamic knowledge retrieval involves surfacing existing textual sources — verses, hadith, or passages from trusted literature — with comparatively low interpretive risk, since the system is primarily locating rather than generating content. *Religious question answering* goes a step further by synthesising retrieved sources into a coherent explanation; this carries moderate risk, as the synthesis step can introduce subtle distortions even when the underlying sources are correct. *Islamic legal advice* (fatwa-giving) carries the highest risk of the three, since it involves issuing an authoritative ruling for a specific personal situation, where an incorrect or decontextualised answer can directly mislead practice. Hikmah AI is positioned primarily within the first two categories. The system is designed to retrieve and explain Islamic

knowledge, not to replace the role of a qualified scholar in issuing personalised fatwa, and this distinction shapes several of the design and guardrail decisions described in Section 3 and Section 4.7.

Prior work such as QuranGPT (https://www.qurangpt.com/) and HadithGPT (https://www.hadithgpt.com/) has demonstrated LLM potential for Islamic text-based question answering [7]. These systems, however, remain limited in scope by design: QuranGPT focuses exclusively on the Qur'an and HadithGPT on Hadith, and neither covers Islamic law, Islamic history, or general Islamic knowledge comprehensively (a feature-level comparison is presented in Table 3, Section 4.4). This limitation is not merely anecdotal. Mh et al. [7] analysed QuranGPT's responses to gender-related verses and found that the system, when not sufficiently grounded in authoritative corpora, was susceptible to semantic drift in its interpretive output. HadithGPT, for its part, was eventually decommissioned following community concern over the validity of AI-generated Hadith content — a precedent discussed further in Section 4.4. Taken together, these cases motivated the development of Hikmah AI as a multi-domain Islamic assistant grounded in the Qur'an, Hadith, and authenticated Islamic literature, with explicit guardrails against the failure modes observed in prior systems.

This paper makes the following contributions: (1) a full-stack Islamic knowledge assistant implemented with a hybrid RAG pipeline (vector + keyword retrieval) on top of an external LLM; (2) a three-tier topic classification mechanism (Regex → Follow-up Detection → LLM Classifier) that enforces Islamic domain boundaries; (3) a configurable admin dashboard enabling dynamic knowledge base management; and (4) a preliminary empirical evaluation using Precision, Recall, F1-Score, Relevance, and Accuracy metrics, complemented by expert Islamic scholar validation.

The remainder of this paper is organised as follows: Section 2 reviews related work. Section 3 describes the methodology. Section 4 presents results and discussion, including the system's responsible-use considerations. Section 5 concludes the paper.

2 Related Work (if applicable)

Research at the intersection of AI and Islamic knowledge has grown considerably in recent years. Mh et al. [7] analysed QuranGPT's responses to gender-related verses using Roland Barthes' semiotics framework, demonstrating that LLMs can offer nuanced Qur'anic

interpretation but also showing risks of semantic drift when not grounded in authoritative corpora. HadithGPT, trained on 40,000 Hadith from six canonical collections, was decommissioned following community concerns about AI-generated Hadith validity, underscoring the importance of scholarly oversight in Islamic AI systems [7].

Lewis et al. [8] introduced Retrieval-Augmented Generation (RAG), combining parametric (LLM) and non-parametric (retrieval) memory for knowledge-intensive NLP tasks. This foundational architecture has since been extended through Forward-Looking Active RAG [9], which retrieves documents proactively rather than reactively. Hybrid search — combining dense vector retrieval with sparse keyword matching — has proven effective for technical domains requiring exact-match precision alongside semantic understanding.

In the Indonesian educational AI context, Mayasari and Sudarmilah [2] highlighted the readiness challenges of integrating AI in educational settings, while Fauzi [3] documented the expanding role of LLM-based systems in Islamic outreach (da'wah). Hanifa et al. [4] integrated OpenCV with LLMs for public service information delivery, demonstrating viable full-stack LLM deployment on local institutional infrastructure.

He et al. [10] provided a comprehensive review of LLM-based multi-agent systems for software engineering, noting the effectiveness of structured development methodologies. Saravanos and Curinga [11] validated the Waterfall model for structured LLM system development, arguing that its sequential phases minimise specification ambiguity in knowledge-critical applications. The present work builds on these foundations, extending RAG-based Islamic AI beyond single-source (Qur'an-only or Hadith-only) systems to a unified multi-domain platform, while treating the LLM itself as an interchangeable component rather than the primary contribution.

3 Methods

Hikmah AI was developed following the Waterfall methodology [11], chosen for its structured, sequential, and well-documented character suited to knowledge-critical system development. The five phases are: (1) Requirements Analysis, (2) System Design, (3) Implementation, (4) Testing, and (5) Maintenance and Evolution.

3.1 Requirements Analysis

The initial phase involved collecting Islamic datasets and identifying functional and non-functional requirements. Qur'anic data was obtained by scraping the Ministry of Religious

Artificial Intelligence Systems and Its Applications (AISA)
Volume 2, Issue 1, June 2026
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

Affairs' official website and converting it to JSON. Hadith datasets were sourced from a publicly available GitHub repository containing SQL databases of six canonical Hadith collections, also converted to JSON. Both datasets form the knowledge base for the RAG pipeline.

The primary functional requirements include:

1. User authentication and account management
2. Question submission via Chat API
3. AI-driven response processing constrained to the Islamic domain
4. Dataset management (CRUD) for Qur'an and Hadith knowledge base
5. Admin dashboard for AI model configuration, RAG parameters, and content guardrails
6. REST API endpoints for external application integration
7. Automated RAG quality evaluation using evaluation metrics

Non-functional requirements include security (bcrypt hashing, JWT), stable performance, usability, cross-device responsiveness, containerised deployment (Docker), and high configurability without code changes.

3.2 System Design

The system architecture is structured across four layers: (1) Client Layer—dashboard (React/Next.js), REST API client, and external app integration via Bearer Token; (2) API Layer—Next.js Route Handlers for /api/chat (POST) and /api/knowledge (CRUD); (3) Business Logic Layer—topic classifier (Regex + LLM), RAG engine (Hybrid Search), OpenAI client (Chat + Embedding), and Auth module (bcrypt + session); and (4) Data Layer—PostgreSQL with pgvector storing users, configuration, chat messages, knowledge documents, knowledge chunks with 1536-dimensional embeddings, and evaluation records.

Use Case modelling identified five primary user interactions: Registration, Login, Conversation with AI, View Chat History, and Logout. Activity Diagrams captured the end-to-end flow from user input through authentication, topic classification, RAG retrieval, LLM inference, and response delivery.

3.3 Implementation

The system was built as a monolithic modular full-stack application using Next.js 15 (App

Router, React Server Components), TypeScript 5, PostgreSQL 16+ with the pgvector extension, Drizzle ORM (code-first schema), Tailwind CSS + shadcn/ui for the frontend, Zod for request/response validation, and Docker + Docker Compose for containerisation. OpenAI's GPT-5 serves as the underlying LLM for response generation and classification; text-embedding-3-small (1536 dimensions) provides semantic embeddings. The LLM itself is treated as a replaceable component behind the RAG pipeline rather than a purpose-built model.

The RAG ingestion pipeline processes documents through three steps: (a) chunking—documents are split into segments of at most 500 characters with a 50-character paragraph-boundary overlap; (b) embedding generation—each chunk is encoded with text-embedding-3-small; (c) storage—chunks and their embeddings are persisted in PostgreSQL's knowledge_chunks table with an HNSW index (vector_cosine_ops) for approximate nearest-neighbour search and a GIN index for full-text keyword retrieval.

Hybrid retrieval combines two parallel search pathways. The vector search pathway computes cosine similarity between the query embedding and all stored chunk embeddings. The keyword search pathway applies PostgreSQL full-text search via plainto_tsquery / ts_rank, effective for exact matching of Islamic technical terminology (surah names, fiqh terms, scholar names). Final relevance scores are computed as:

final_score = (vector_score × 0.75) + (keyword_score × 0.25)

Domain boosting applies an additional +0.12 score increment when query intent and retrieved chunk content both match Qur'anic or Hadith categories (combined boost of +0.24 when both are present). If the vector API is unavailable, the system automatically falls back to keyword-only search.

The three-tier topic classification mechanism enforces Islamic domain constraints. Tier 1 uses deterministic Regex patterns (ISLAMIC_QUERY_REGEX, QURAN_QUERY_REGEX, CLEARLY_NON_ISLAMIC_REGEX, ISLAMIC_CONTEXT_EXTRA_REGEX) for instant
classification of unambiguous queries. Tier 2 performs follow-up context detection by checking shared meaningful tokens between the current query and the last six messages; if ≥2 shared terms exist alongside an Islamic signal, the query is classified as in-domain. Tier 3 delegates to GPT-5 for ambiguous queries, prompting it to return a JSON classification object {"allowed":

Artificial Intelligence Systems and Its Applications (AISA)
Volume 2, Issue 1, June 2026
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

true|false, "reason": "..."} for one of three permitted domains: Islamic law, Islamic history, or general Islamic knowledge. A multi-model fallback cascade (gpt-5 → gpt-5.4-mini → gpt-5.4-nano) provides resilience.

To illustrate how this mechanism behaves in practice, Table 1 below lists representative examples of queries that the classifier accepts or rejects, spanning each tier.

Query (illustrative)	Decision	Reason / Classifier Tier
"Apa hukum riba dalam Islam?"	Accepted	Clear Islamic-law signal; resolved at Tier 1 (Regex).
"Siapa khalifah pertama setelah Nabi Muhammad?"	Accepted	Islamic history domain; resolved at Tier 1 (Regex).
"Kalau begitu, apa pendapat mazhab lain soal itu?" (follow-up)	Accepted	Shares ≥2 tokens with prior Islamic-domain turns; resolved at Tier 2.
"Bagaimana cara membuat bahan peledak?"	Rejected	Clearly non-Islamic and harmful intent; resolved at Tier 1 (Regex).
"Apa obat yang ampuh untuk demam tinggi?"	Rejected	Medical domain, outside Islamic knowledge scope; resolved at Tier 3 (LLM classifier).
"Apakah saya harus bercerai dari suami saya?"	Rejected with referral	Personalised fatwa request; flagged at Tier 3 and redirected to consult a qualified scholar rather than answered directly.

Table 1. Representative Examples of Accepted and Rejected Queries (Illustrative)

3.4 Testing and Evaluation

Testing combined automated evaluation metrics with expert human review. Quantitative evaluation used five metrics defined as:

Metric	Formula	Description
Precision (P)	TP / (TP + FP)	Proportion of retrieved chunks that are relevant
Recall (R)	TP / (TP + FN)	Proportion of relevant chunks successfully retrieved
F1-Score	2 × (P × R) / (P + R)	Harmonic mean of Precision and Recall
Relevance (Rel)	LLM assessment (0–100%)	LLM-judged relevance of response to query
Accuracy (Acc)	LLM assessment (0–100%)	LLM-judged factual correctness of response

Three test cases were constructed, one per Islamic domain category (Qur'an, Fiqh, Aqidah), each specifying a query, expected document IDs, and domain category. RAG was configured with topK = 3 for all evaluations. We note that this set of three cases is intentionally treated as a preliminary, illustrative evaluation rather than a comprehensive benchmark; a larger and more diverse query set is identified as a priority for future work (Section 4.6 and Section 5). Expert review was conducted by Ustadz Negus, a 30-juz Qur'an hafidh, Hadith scholar, and Head of the Muhammadiyah Ranting Alasombo, who evaluated five fiqh questions of varying complexity.

3.5 Limitations of Translated Source Texts

A methodological caveat applies to the knowledge base itself. Both the Qur'anic and Hadith datasets used in this study include translated text alongside the original Arabic. Translation inherently involves interpretive choices that may not fully preserve the nuance of the source language, particularly for fiqh terminology where a single Arabic term can carry several valid renderings depending on context and madhab. This is a structural limitation shared by any RAG system that relies on translated religious corpora for legal or theological reasoning: a retrieval system can only be as faithful as the text it retrieves from. Where possible, Hikmah AI prioritises retrieval of passages alongside their original Arabic text so that translation functions as an interpretive aid rather than the primary ground truth, though full parallel-text retrieval across the entire knowledge base remains an area for future development (Section 4.6).

4 Results and Discussion

4.1 System Architecture Implementation

Hikmah AI was successfully deployed as a containerised monolithic modular application. Table 1 summarises the complete technology stack.

Component	Technology	Version	Function
Framework	Next.js (App Router)	15	Full-stack SSR/SSG framework
Language	TypeScript	5	Type-safe programming language
Database	PostgreSQL + pgvector	16+	Relational storage + vector similarity search

Artificial Intelligence Systems and Its Applications (AISA)
Volume 2, Issue 1, June 2026
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

Component	Technology	Version	Function
ORM	Drizzle ORM	Latest	Type-safe, code-first DB query builder
LLM Provider	OpenAI API (GPT-5)	GPT-5	Natural language processing and embedding
Styling	Tailwind CSS + shadcn/ui	4	Responsive, accessible user interface
Validation	Zod	Latest	Request/response schema validation
Authentication	Bcrypt + JWT	—	Password hashing and session management
Containerisation	Docker + Docker Compose	—	Environment isolation and deployment

Table 2. Hikmah AI Technology Stack

The database schema comprises seven tables: users (authentication), app_config (16-column singleton for AI identity, API keys, RAG parameters, and guardrail settings), chat_participants, chat_messages (full history), knowledge_documents, knowledge_chunks (1536-dimensional embeddings with HNSW

+ GIN indices), and evaluation_test_cases / evaluation_runs. The HNSW index uses vector_cosine_ops for efficient approximate nearest-neighbour search, and the GIN index on the content column supports full-text retrieval.

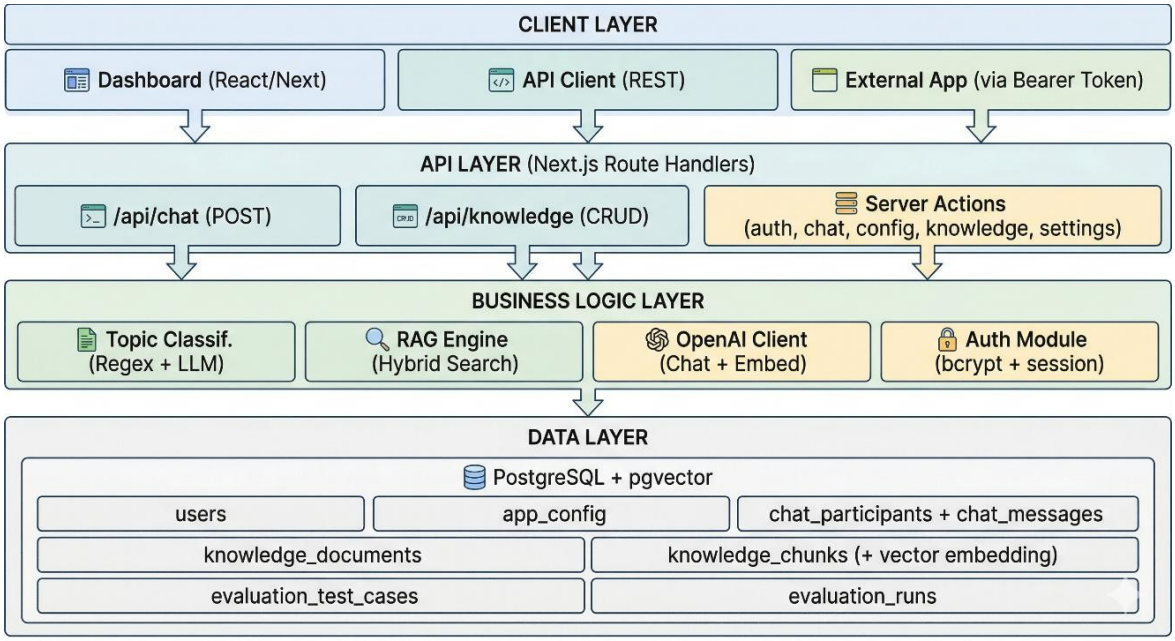

Figure 1. Hikmah AI System Architecture Diagram

4.2 AG System Performance

Table 3 presents per-query evaluation results for the three preliminary test cases conducted on 3 April 2026. As noted in Section 3.4, these results should be read as an initial pilot evaluation rather than a comprehensive benchmark.

No.	Query	Category	Chunks Relevant / Retrieved	P	R	F1	Rel	Acc
1	"Name the first surah to be revealed?"	Qur'an	3 / 3	100%	100%	100%	89.1%	89.1%
2	"Is killing a robber in self-defence sinful in Islam?"	Fiqh	3 / 3	100%	100%	100%	70.9%	70.9%

Artificial Intelligence Systems and Its Applications (AISA)
Volume 2, Issue 1, June 2026
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

3	"Why is Islam considered the true and perfect religion compared to others?"	Aqidah	3 / 3	100%	100%	100%	98.6%	98.6%

Table 3. RAG Evaluation Results Per Query

All three queries achieved perfect retrieval (Precision = Recall = F1-Score = 100%) on this small test set, suggesting that the hybrid search mechanism with domain boosting is capable of retrieving the intended chunks for these specific cases. The 75:25 vector-to-keyword weighting balances semantic similarity with exact-match precision for Islamic technical terminology, although with only three test queries this should be regarded as an encouraging early signal rather than a robust performance guarantee.

Average Relevance and Accuracy were both 86.2%, but scores varied substantially across domains. The Aqidah query scored highest (98.6%), as the question of Islam's theological completeness is a foundational topic comprehensively covered in the knowledge base. The Qur'an query scored 89.1%, reflecting well-documented historical facts with minor nuance gaps. The Fiqh query scored lowest (70.9%), indicating that complex jurisprudential questions — particularly those involving inter-madhab differences — require deeper knowledge base coverage. Notably, Relevance and Accuracy scores were identical across all three test cases, which is consistent with the absence of hallucination in this small sample: retrieved context was either well-answered or insufficiently detailed, but not fabricated. Given the sample size, this observation should be treated as a preliminary indicator rather than a general property of the system, and is a key motivation for the larger-scale evaluation proposed in Section 5.

Several strategies are proposed to improve the 70.9% Fiqh score: (1) enriching the knowledge base with texts from the four major madhabs (Shafi'i, Maliki, Hanafi, Hanbali) and their supporting evidence; (2)

replacing character-count chunking with semantic chunking to preserve intra-paragraph coherence; (3) introducing cross-encoder re-ranking after initial retrieval; and (4) adopting adaptive weighting—increasing keyword weight for Arabic technical terms in Fiqh queries.

4.3 Topic Clasification System

The three-tier classifier proved effective and cost-efficient. Tier 1 (Regex) resolves clearly Islamic or non-Islamic queries without any API call, avoiding LLM costs for straightforward cases. Tier 2 (follow-up detection) handles contextual continuation within an Islamic conversation. Only genuinely ambiguous queries escalate to Tier 3 (LLM classification), minimising unnecessary API expenditure. The multi-model fallback cascade ensures availability even during model outages.

4.4 Comparison with Related Systems

Table 3 compares Hikmah AI with QuranGPT and HadithGPT across key feature dimensions.

Feature	Hikmah AI	QuranGPT	HadithGPT
Status	Active (self-hosted)	Active (public)	Offline
Domain coverage	Multi-domain (fiqh, aqidah, history, Qur'an, Hadith)	Qur'an only	Hadith only
Hybrid retrieval (vector+keyword)	✓	✗	✗
Dynamic knowledge base (CRUD)	✓ (admin dashboard)	✗ (static)	✗ (static)
Three-tier topic classifier	✓	✗	✗
Multi-level content guardrail	✓ (standard / strict)	Not explicit	Not explicit
Built-in RAG evaluation (P/R/F1)	✓	✗	✗
REST API for external integration	✓ (Bearer Token)	✗	✗
Admin dashboard	✓	✗	✗

Artificial Intelligence Systems and Its Applications (AISA)
Volume 2, Issue 1, June 2026
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

Feature	Hikmah AI	QuranGPT	HadithGPT
Image analysis	✓	✗	✗
Response depth control	✓ (concise/standard/deep)	✗	✗
Deployment	Self-hosted (Docker)	Cloud (public)	Cloud (public)
Mobile application	✗	✓ (Android, iOS)	N/A

Table 4. Feature Comparison: Hikmah AI vs. QuranGPT vs. HadithGPT

Hikmah AI's primary advantages over existing platforms are multi-domain coverage, dynamic knowledge management, built-in evaluation, and external API support. QuranGPT retains advantages in end-user accessibility (mobile apps, multi-translation browsing). The decommissioning of HadithGPT—whose developers consulted Islamic scholars before shutdown—provides an important precedent: the present work addresses this concern through configurable guardrails and expert validation as an integral development step.

4.5 Expert Validation

Expert evaluation was performed by Ustadz Negus (Head of Muhammadiyah Ranting Alasombo; Head of Santri Affairs at Qorayatul Quran Tahfidh Pesantren; 30-juz hafidh and Hadith scholar), who submitted five fiqh questions of varying complexity: the ruling on consuming khamr (intoxicants); remarrying a thrice-divorced woman; the legal effect of triple talaq; nikah siri (unregistered marriage); and performing Shawwal fasting outside the month of Shawwal.

Table 5 summarises each question alongside the system's response, the sources cited, and the expert's assessment.

No.	Fiqh Question	System Answer (Summary)	Cited Sources	Expert Score	Expert Comment
1	Ruling on consuming khamr (intoxicants)	Haram (prohibited) to consume in any amount, as it impairs the mind and leads to greater sins.	QS. Al-Ma'idah: 90, HR. Muslim No. 2003	5/5	The answer is highly accurate, concise, and the evidence used is precise and very relevant.

2	Remarrying a thrice-divorced woman	The former husband cannot remarry her unless she marries another man first (tahlil), the marriage is consummated, and they later divorce naturally.	QS. Al-Baqarah: 230	5/5	The explanation of the *mughallazah* (irrevocable) divorce law aligns perfectly with the consensus of scholars and the textual verse.
3	Legal effect of triple talaq	Pronouncing triple talaq in one sitting counts as three major divorces according to the majority of scholars (Jamhur), rendering the divorce irrevocable (*bain kubra*).	HR. Bukhari No. 5259, Fath al-Bari	4/5	Good answer, but it would be better to also mention the minority view (such as Ibn Taymiyyah's) as additional academic insight.
4	Nikah siri (unregistered marriage)	Valid under Islamic law if it meets all pillars (*rukun*) and conditions, but discouraged or sinful under state law due to potential harm (*mudharat*) to the wife and children.	Hadits "Lailaha illa bi waliyyin", Kaidah Fikih *Ladharara wala dhirar*	4,5/5	The explanation is well-balanced and wise, covering both the religious law aspect and its social/legal implications.
5	Shawwal fasting outside the month of Shawwal	Voluntary fasting of six days can only be done within the month of Shawwal. If done outside Shawwal, it counts as regular voluntary fasting but loses the specific reward of "fasting for a year."	HR. Muslim No. 1164	5/5	Excellent. It clearly distinguishes between the validity of general voluntary fasting and the specific reward exclusive to the month of Shawwal.

Table 5. Expert Validation Summary — Five Fiqh Questions (placeholder; to be completed by the author)

The expert found all answers accurate, consistent with Qur'anic verses and Hadith in their original textual forms, and free from substantive error. He concluded the system is sufficiently helpful and appropriate for general public use. His recommendations for future development include adding kitab references from the four major madhabs and classical tafsir works to deepen Qur'anic exegesis.

Artificial Intelligence Systems and Its Applications (AISA)
Volume 2, Issue 1, June 2026
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

4.6 System Strengths and Limitations

The key strengths of Hikmah AI are:

1. Modular monolithic architecture enabling maintainability without sacrificing integration simplicity
2. Hybrid RAG achieving strong retrieval scores on the preliminary test set, with no hallucination observed in that limited sample
3. Cost-efficient three-tier classifier that escalates to LLM only for genuinely ambiguous queries
4. High configurability—all major parameters adjustable via dashboard without code changes
5. Docker containerisation ensuring portable, reproducible deployment

Current limitations include:

1. Fiqh knowledge base depth insufficient for complex inter-madhab queries (70.9% relevance/accuracy in the preliminary test)
2. Static 75:25 retrieval weighting not adaptive to query type
3. Absence of a mobile application
4. Evaluation conducted on only three retrieval test cases and five expert-reviewed fiqh questions, which is too small a sample to support strong generalisable claims and warrants larger-scale empirical testing
5. Reliance on translated source texts for parts of the knowledge base, with the interpretive limitations discussed in Section 3.5

4.7 Responsible Use and Ethical Considerations

Islamic knowledge systems carry a distinctive ethical weight: an incorrect or decontextualised answer does not merely inconvenience a user, it can misinform religious practice. This concern is not hypothetical — it is precisely what led to the decommissioning of HadithGPT after scholarly review raised doubts about the validity of its generated content (Section 4.4). With that precedent in mind, Hikmah AI incorporates several practical guardrails, summarised below.

1. Domain restriction. The three-tier topic classifier described in Section 3.3 constrains the system to three permitted domains — Islamic law, Islamic history, and general Islamic knowledge — and rejects queries that fall outside these boundaries, including medical, legal-secular, or unrelated advice (see Table 1 for examples).
2. Source grounding over parametric recall. Responses are generated from retrieved chunks of the Qur'an, Hadith, and vetted Islamic literature rather than from the LLM's unconstrained background knowledge, which reduces — though does not eliminate — the risk of fabricated or unsupported rulings.
3. Explicit scope boundary on fatwa-giving. Consistent with the distinction drawn in Section 1, Hikmah AI is designed to support knowledge retrieval and general religious question answering. It is not designed to issue binding, personalised fatwa for contested or circumstantial fiqh matters; for such cases, the system is intended to direct users toward consulting a qualified scholar rather than presenting itself as a final authority.
4. Expert validation as a pre-deployment gate, not an afterthought. The expert review reported in Section 4.5 was conducted as part of the development process itself, prior to broader use, rather than as a post-hoc justification. This mirrors the lesson taken from HadithGPT's decommissioning: scholarly oversight needs to happen before issues reach end users, not after.
5. Configurable guardrail strictness. The admin dashboard allows operators to adjust guardrail behaviour between standard and strict modes, so that deployments serving more sensitive contexts (for example, an institutional setting handling contested fiqh topics) can be configured more conservatively than a general-purpose public deployment.

These measures are best understood as risk-mitigation practices rather than guarantees. The preliminary evaluation in Section 4.2 and Section 4.5 is reassuring but limited in scale, and the authors do not claim that the guardrails described here eliminate the possibility of error. Rather, they represent a deliberate, documented attempt to manage the specific risks that distinguish Islamic knowledge systems from general-purpose chatbots, and they are expected to evolve as the system is tested against a larger and more adversarial range of queries.

Artificial Intelligence Systems and Its Applications (AISA)
Volume 2, Issue 1, June 2026
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

5 Conclusion

This paper presented Hikmah AI, a multi-domain Islamic knowledge assistant built on a hybrid RAG pipeline layered over an external LLM (GPT-5). On a preliminary evaluation set, the system achieved strong retrieval performance (Precision = Recall = F1-Score = 100% across three test queries), with average Relevance and Accuracy of 86.2%. Expert Islamic scholar validation found the reviewed responses to five fiqh questions to be factually consistent with source texts and free of substantive error. These findings are encouraging but preliminary, and should be read as an initial proof of concept for the proposed architecture rather than a conclusive performance benchmark.

Key contributions include: a full-stack Next.js 15 / PostgreSQL + pgvector architecture supporting 1536-dimensional semantic search; a hybrid retrieval mechanism (75% vector, 25% keyword) with domain boosting; a three-tier topic classifier enforcing Islamic domain constraints with minimal API overhead; a configurable admin dashboard for dynamic knowledge base management without programming expertise; and a documented approach to responsible use that treats expert scholarly validation as a development requirement rather than an optional addition.

Future work should focus on:

1. Enriching the knowledge base with classical fiqh and tafsir references to address the current 70.9% accuracy gap in complex jurisprudential queries
2. Implementing semantic chunking and cross-encoder re-ranking
3. Adaptive retrieval weighting
4. Large-scale empirical evaluation with a substantially expanded and more diverse set of test queries and expert-reviewed questions, to move beyond the preliminary evaluation reported here
5. Closer parallel-text retrieval between Arabic source material and its translations, to reduce reliance on translation as primary ground truth
6. Development of a mobile application to increase accessibility for the general public
7.

Acknowledgments

The author wishes to thank Endang Wahyu Pamungkas, S.Kom., M.Kom., Ph.D. for supervision and guidance throughout this research. Expert validation support from Ustadz Negus is gratefully acknowledged.

Funding Information

The author declares no external funding was received for this study.

Conflict of Interest Statement

The author declares no conflicts of interest.

Ethical Approval

This study did not involve human or animal subjects in experiments. The expert validation was conducted voluntarily with informed consent.

Data Availability

The Islamic dataset used in this study originates from the Indonesian Ministry of Religious Affairs' official Qur'an portal and a publicly available GitHub Hadith repository. Processed datasets and system configuration are available upon request from the corresponding author.

References

[1] E. Eryandi, "Integrasi Nilai-Nilai Keislaman dalam Pendidikan Karakter di Era Digital," Kaipi: Kumpulan Artikel Ilmiah Pendidikan Islam, vol. 1, no. 1, pp. 12–16, 2023. https://doi.org/10.62070/kaipi.v1i1.27

[2] A. E. Mayasari and E. Sudarmilah, "Artificial Intelligence dalam Pendidikan Era Industri," JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 10, no. 4, pp. 3689–3703, 2025.

[3] F. Fauzi, "Dakwah Islam dan Artificial Intelligence: Penelitian Atas Pemanfaatan AI Dalam Penyebaran Nilai-nilai Islam," RIGGS: Journal of Artificial Intelligence and Digital Business, vol. 4, no. 2, pp. 3702–3709, 2025.

[4] M. D. Hanifa, M. Al, I. Widodo, and D. Gunawan, "Integrasi OpenCV dan LLM Pada Sistem Informasi Pelayanan Publik di Desa Trangsan Menggunakan Framework Django," pp. 136–139, 2025.

[5] M. F. Hajri, "Pendidikan Islam di Era Digital: Tantangan dan Peluang pada Abad 21," Al Mikraj Jurnal Studi Islam dan Humaniora, vol. 4, no. 1, pp. 33–41, 2023.

https://doi.org/10.37680/almikraj.v4i1.3006

[6] U. Kulsum and A. Muhid, "Pendidikan Karakter melalui Pendidikan Agama Islam di Era Revolusi Digital," Jurnal Intelektual: Jurnal Pendidikan dan Studi Keislaman, vol. 12, no. 2, pp. 157–170, 2022. https://doi.org/10.33367/ji.v12i2.2287

[7] S. M. Mh et al., "AI And The Discourse Of Qur'anic Interpretation: An Analysis Of Quran-GPT Response To Gender Verses From The Semiotics Perspective Of Roland Barthes," Jurnal for Religious-Innovation Studies, vol. XXIV, no. 2, pp. 189–203, 2025.

[8] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020.

[9] Z. Jiang et al., "Active Retrieval Augmented Generation," in Proc. 2023 Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 7969–7992, 2023.

[10] J. He, C. Treude, and D. Lo, "LLM-Based Multi-Agent Systems for Software Engineering: Literature Review, Vision, and the Road Ahead," ACM Trans. Software Engineering and Methodology, vol. 34, no. 5, pp. 1–30, 2025. https://doi.org/10.1145/3712003

[11] A. Saravanos and M. X. Curinga, "Simulating the Software Development Lifecycle: The Waterfall Model," Applied System Innovation, vol. 6, no. 6, 2023. https://doi.org/10.3390/asi6060108

[12] N. Diaz Arizona, "The Implementation of Waterfall Method in the Development of Accounting Information Systems for Web-Based Savings and Loans Data Processing," Bulletin of Computer Science and Electrical Engineering, vol. 3, no. 2, pp. 86–96, 2022. https://doi.org/10.25008/bcsee.v3i2.1167

[13] P. A. Rospigliosi, "Artificial intelligence in teaching and learning: what questions should we ask of ChatGPT?" Interactive Learning Environments, vol. 31, no. 1, pp. 1–3, 2023. https://doi.org/10.1080/10494820.2023.2180191

[14] A. Zaremba and E. Demir, "ChatGPT: Unlocking the future of NLP in finance," Modern Finance, vol. 1, no. 1, 2023.
R. Saifur Robbi and E. Sudarmilah, "Perancangan Sistem Informasi Perpustakaan Berbasis Web Pada SDN Pabelan 2 Kartasura," Infotronik: Jurnal Teknologi Informasi dan Elektronika, vol. 9, no. 1, pp. 45–58, 2024. https://doi.org/10.32897/infotronik.2024.9.1.3338