

A hybrid multi-corpus RAG framework for faithful and source-grounded MUI fatwa question answering

Fitra Kurnia, Nazruddin Safaat Harahap*, Surya Hidayatullah Firdaus

Department of Informatics Engineering, UIN Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia

ABSTRACT

The increasing use of large language models (LLMs) for Islamic legal information retrieval presents substantial challenges related to hallucinations, source relevance, and the limited coverage of single-corpus retrieval-augmented generation (RAG) systems. This study proposes a sequential multi-corpus RAG architecture for a hybrid MUI fatwa chatbot that prioritises authoritative fatwa documents while providing controlled fallback retrieval from a curated hadith corpus. The system employs dense semantic retrieval using FAISS for the primary MUI fatwa corpus, followed by similarity threshold filtering and LLM-based relevance verification. When no sufficiently relevant fatwa context is identified, the query is redirected to sparse retrieval of over 12,370 hadith records from Sahih al-Bukhari and Sahih Muslim. The proposed system was compared with a conventional single-corpus RAG configuration using the BERTScore and Giskard's RAG evaluation toolkit (RAGET). BERTScore evaluation on ten questions showed that the proposed architecture increased the average F1-score from 0.6664 to 0.7546, representing a relative improvement of 13.23%. On 41 successfully paired evaluation records, the Giskard faithfulness score increased from 0.8537 to 0.9439, equivalent to a 10.57% relative improvement. However, both systems produced an identical abstain rate of 26.83%, indicating that the proposed architecture primarily improved answer relevance and contextual grounding rather than overall question coverage. These findings demonstrate that hierarchically organising knowledge sources according to their authority and function can enhance the reliability of RAG-based Islamic legal information systems. Nevertheless, the chatbot should be regarded as a source-grounded information-access tool and not as a substitute for fatwas or qualified Islamic scholars.

ARTICLE INFO

Article History:

Received Aug 13, 2026

Revised Sep 27, 2026

Accepted Oct 4, 2026

Published Oct 28, 2026

Keywords:

Hybrid Retrieval
Islamic Law
MUI Fatwa
Question Answering
RAG

Copyright (c) 2026:

Fitra Kurnia,
Nazruddin Safaat
Harahap, Surya
Hidayatullah Firdaus

License:

This work is licensed under a [Creative Commons Attribution 4.0 International License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).



*Corresponding Author

E-mail Address: nazruddin.safaat@uin-suska.ac.id

1. INTRODUCTION

Artificial intelligence (AI) has advanced rapidly in recent years, particularly in natural language processing (NLP). The evolution of large language models (LLMs) has fundamentally transformed the paradigm of human-machine interactions [1]. The capacity of generative AI to process complex queries enables it to function as an intelligent information retrieval assistant, commonly implemented through chatbot applications. Xu et al. [2] define a chatbot as an AI-based system capable of interacting naturally with users across various service domains. Furthermore, Ardiansyah and Widiyanto [3] demonstrated that integrating AI into instant messaging platforms can provide an interactive and accessible medium for delivering information.

Within Indonesia's predominantly Muslim population, the need for authoritative guidance on Islamic legal matters necessitates access to credible sources, particularly fatwas issued by the Indone-

sian Ulema Council (Majelis Ulama Indonesia, MUI). MUI fatwas hold considerable moral, social, and religious authority as a source of guidance for Indonesian Muslims [4]. Hasanuddin et al. [4] and Yuspiani et al. [5] further indicate that MUI fatwas strategically contribute to shaping Sharia-oriented governance and policies that respond to the evolving challenges of contemporary society. However, digital fatwa documents are predominantly stored as dispersed PDF files without a unified search system. Consequently, locating relevant rulings through manual searches is time-consuming and requires a degree of legal and contextual understanding that lay users may lack.

The application of AI to Islamic legal information raises an important scholarly concern regarding the appropriate boundaries of automated systems in this field. Dewi [6] argues that AI cannot replace Islamic scholars in performing *ijtihad* because it lacks the capacity to understand *maqasid al-Shariah* holistically. Consistent with Putra's [7] recommendation that the role of AI in religious matters should be clearly delimited, the system developed in this study positions AI strictly as a facilitator of access to textual information contained in existing fatwas rather than as an autonomous fatwa-issuing authority. This distinction is essential because generative AI systems remain vulnerable to language-model hallucinations, through which plausible but factually unsupported responses may be produced [8].

Retrieval-augmented generation (RAG) offers a promising approach for mitigating this risk. By grounding generated responses in externally retrieved documents rather than relying exclusively on the knowledge encoded within an LLM's parameters, RAG combines generative capabilities with external information retrieval and can reduce unsupported outputs [9]. More advanced approaches, such as Self-RAG, further demonstrate that adaptive retrieval and self-reflection mechanisms are instrumental in improving response accuracy and faithfulness to source documents [10]. These capabilities make retrieval-grounded generation particularly relevant to Islamic legal information systems, in which traceability and authority of supporting evidence are fundamental requirements.

The integration of RAG into public information systems has produced encouraging results. Fauzi and Sutabri [11] developed PublicTalk and reported a text-understanding accuracy exceeding 88%, whereas Cahyanti and Raya [12] introduced Chatku AI, which effectively processed externally sourced PDF data and achieved a user acceptance score of 99.2%. Within the domain of Islamic literature, Rahayu et al. [13] developed a LangChain-based question-answering system using a PDF dataset covering the jurisprudence of four Sunni schools of law. Similarly, Helviansyah et al. [14] employed an LLM supported by a similarity search to retrieve information related to contemporary Islamic jurisprudence. The retrieval of primary Islamic evidence has also been explored by Herwanza et al. [15] for hadith extraction, while Haekal et al. [16] developed a closed-domain question-answering system for fatwa documents using a conventional RAG architecture. Moreover, comprehensive evaluations of LLMs in Islamic legal reasoning domains, including Islamic inheritance law, have revealed that standard models remain susceptible to fundamental reasoning failures and source misinterpretations [17]. These findings reinforce the importance of grounding generated responses in reliable external sources such as research papers.

Evaluating a retrieval-grounded question-answering system requires more than just measuring lexical overlap. Miyaji et al. [18] explained that BERTScore is effective for text evaluation because it employs contextual representations to capture the semantic similarity between the generated and reference texts. Dobslaw et al. [19] introduced Giskard as an evaluation framework designed specifically for LLM-based applications. The framework supports the assessment of properties, such as faithfulness to source documents, and the detection of hallucinated content. Therefore, the combined use of semantic similarity and faithfulness evaluation can provide a more comprehensive assessment of whether system-generated responses are both contextually relevant and adequately grounded in the retrieved evidence.

Despite the technical foundation established by previous studies, a fundamental conceptual limitation remains: predominant reliance on a single document corpus. MUI fatwas are inherently issue-specific and problem-oriented; consequently, they are not intended to provide exhaustive coverage of all areas of Islamic law. When a user submits a question that falls outside the scope of the available fatwa documents, a conventional single-corpus RAG system may encounter an out-of-domain query. Under such conditions, the system may retrieve weakly related passages and generate answers from inappropriate contexts, thereby undermining both response relevance and source reliability.

To address this research gap, this study develops a Hybrid MUI Fatwa Chatbot based on a sequential multi-corpus retrieval architecture. Its principal novelty lies in a hierarchical fallback mechanism that initially prioritises semantic retrieval from the MUI fatwa corpus. When the system determines that no sufficiently relevant fatwa context is available, the query is automatically redirected to a secondary corpus containing 12,370 curated authentic hadiths through lexical-retrieval. This priority-based retrieval strategy preserves MUI fatwas as the principal source while providing controlled access to supporting hadith evidence when the primary corpus cannot adequately address a query. Therefore, this study contributes theoretically by proposing a sequential multi-corpus RAG architecture for Islamic legal information retrieval. Practically, it provides an accountable digital information service intended to reduce hallucinations, improve source traceability, and facilitate public access to authoritative Islamic legal references without positioning AI as a substitute for qualified Islamic legal scholars.

2. RESEARCH METHODS

This study conducted a classification comparison of an English Qur'an translation dataset using two approaches: LSTM and RF with SMOTE. The implementation used Python with TensorFlow/Keras for LSTM and Scikit-learn along with imbalanced-learn for Random Forest and SMOTE. The general pipeline consisted of: (1) data loading and exploration, (2) text preprocessing, (3) train/validation/test splitting with a 72-8-20 ratio, (4) feature extraction, (5) SMOTE oversampling per label on training data, (6) model training, and (7) evaluation on unseen test data.

2.2. Research Design

This study adopted an engineering-oriented research design, focusing on system development and comparative performance evaluation in the context of an AI-based question-answering system. A controlled comparative experiment was conducted to assess two retrieval configurations: (1) a conventional RAG system relying exclusively on a single MUI fatwa corpus and (2) the proposed sequential multi-corpus retrieval system integrating MUI fatwa documents as the primary corpus and a curated hadith database as the fallback corpus. As this study focused on data-driven system testing and automated evaluation, it did not involve human participants.

The research hypothesis posits that the proposed sequential multi-corpus retrieval method generates higher-quality answers than a conventional single-corpus RAG system. The retrieval configuration (single-corpus RAG or sequential multi-corpus retrieval) was the independent variable. The dependent variables comprised response quality, measured using BERTScore precision, recall, and F1-score, as well as faithfulness and abstention behaviour, evaluated using the Giskard framework.

This study was motivated by the limited thematic coverage of MUI fatwas. Fatwas are generally issued in response to particular circumstances or emerging problems and are therefore not intended to cover every question concerning Islamic teachings. To address this limitation, the proposed system employs a hierarchical retrieval. It first searches the MUI fatwa corpus using semantic similarity and subsequently verifies the topical relevance of the retrieved context using an LLM-based decision-making gate. When the primary corpus does not provide adequate context, the system redirects the query to a curated hadith database serving as the secondary knowledge source.

2.2. Research Stages

This study was conducted in six stages. The first stage involved problem analysis to identify the principal limitations associated with using MUI fatwa documents in an automated question-answering system. These limitations include the unstructured nature of PDF documents and their issue-specific coverage.

The second stage comprised a literature review of LLM-based question-answering systems, RAG architectures, multi-corpus and hybrid retrieval strategies, and evaluation approaches based on BERTScore and Giskard software. This stage established the conceptual and technical foundations of the proposed architecture.

The third stage involved the collection of the corpus. The primary corpus consisted of MUI fatwa documents in PDF format obtained from www.fatwamui.com/data-fatwa. The secondary corpus comprised 7,008 hadiths narrated by al-Bukhari and 5,362 hadiths narrated by Muslim, obtained from

the public GitHub repository at <https://github.com/irsyadulibad>. The secondary corpus contained 12,370 hadith records.

The fourth stage involved analysing and designing the sequential multi-corpus retrieval architecture, including document segmentation, embedding generation, vector storage, semantic retrieval, lexical retrieval, relevance verification, query routing, and fallback mechanisms. The fifth stage comprised the implementation of a web-based chatbot. The sixth stage involved comparative system testing using the BERTScore and Giskard RAG evaluation toolkit (RAGET).

2.3. System Analysis and Design

The proposed architecture employs an LLM as both a response generation component and an adaptive relevance evaluator. LLaMA 3.1 8B was selected as the principal language model because it offers competitive language understanding and instruction-following capabilities while remaining sufficiently compact for local deployment. Its ability to process multilingual input is particularly relevant because the system handles Indonesian queries and Islamic source materials that contain Indonesian and Arabic expressions. Its instruction-following capabilities also support its use within an adaptive decision gate that semantically evaluates whether a retrieved fatwa passage corresponds adequately to the user's query [20].

LangChain was employed as the orchestration framework connecting the retrieval, routing, context construction, and response generation components. The architecture applies a hybrid storage strategy based on the characteristics of each corpus (see Figure 1). The MUI fatwa documents were pre-processed using the Recursive Character Text Splitter with overlapping segments. The resulting text chunks were converted into dense vector representations and stored in a vector store to facilitate semantic retrieval. Conversely, the structured hadith records were stored in a MySQL relational database equipped with a full-text index to support lexical-retrieval.

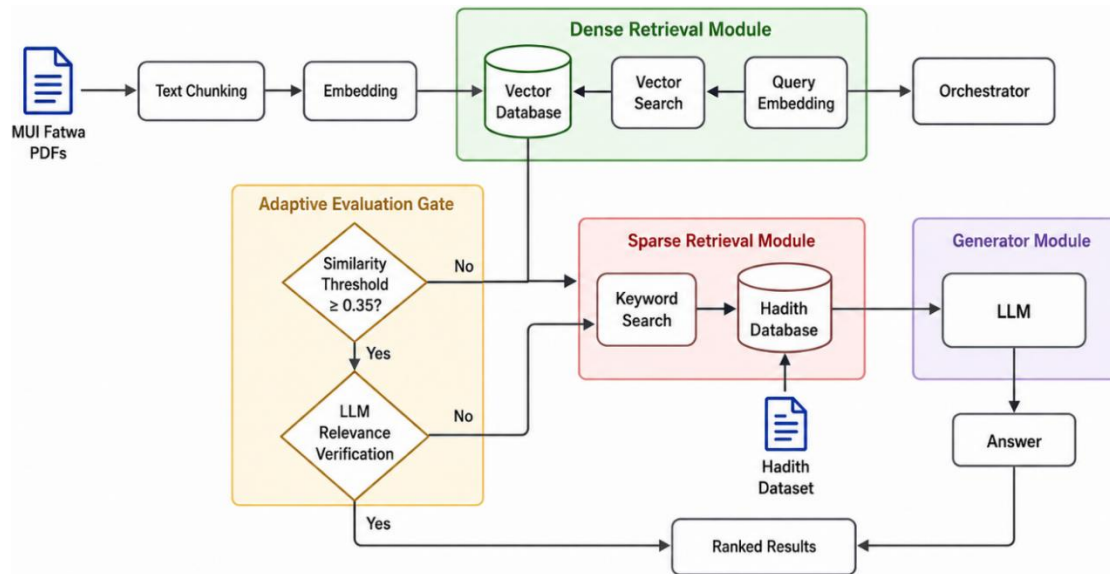


Figure 1. Architecture of the hybrid MUI fatwa Chatbot.

Unlike conventional single-corpus RAG, the proposed system implements sequential multi-corpus retrieval using three decision levels. The first level is primary semantic retrieval. At this level, the system performs a dense retrieval of the vectorised MUI fatwa collection and applies an initial similarity threshold to the retrieved passages.

The second level employs LLM-based relevance verification as an adaptive-decision gate. A fatwa passage that passes the initial similarity threshold is not immediately used to generate answers. Instead, the LLM evaluates whether the topic and meaning correspond sufficiently to the user's intent. This additional verification is intended to reduce the risk of generating responses from passages that exhibit superficial semantic similarity but do not address the actual questions.

The third level comprises dynamic routing, and fallback retrieval. When the decision gate confirms the relevance of the fatwa context, the retrieved passage is supplied to the generation component for synthesising the answer. Conversely, when the retrieved fatwa context is judged to be inadequate or irrelevant, the system redirects the query to the secondary hadith corpus. It then performs sparse retrieval using full-text keyword matching to identify the supporting hadith records. This sequential design preserves MUI fatwas as the prioritised source while extending the system’s coverage in a controlled manner.

2.4. System Implementation

The user interface was implemented using Streamlit, which provides an interaction layer between users and the chatbot pipeline. The main application page consists of a navigation panel on the left and an interaction panel in the central region. The navigation panel presents six representative sample questions concerning online transactions, bank interest, Friday prayers during a pandemic, insurance, ablution procedures and music in Islam. It also provides access to conversation history. The main interaction panel contains a text-input field and a “Search for an Answer” button that initiates the retrieval and response generation process in real time. A minimalist green visual theme was adopted to establish a consistent interface identity appropriate for the application domain.

The system was developed and tested on a computer equipped with an Intel® Core™ i5-13450HX processor, 20 GB RAM, 512 GB solid-state drive, and Windows 11 operating system. Python was used as the primary programming language. Llama 3.1 8B was executed locally using the Ollama runtime. Local deployment eliminates the dependency on third-party API keys and external token quotas while retaining user queries within the local execution environment.

The model temperature was set to 0.8 to provide sufficient linguistic flexibility when synthesising retrieved fatwa and hadith passages into coherent responses that were contextually appropriate. The context window was limited to 2,048 tokens to balance the local memory consumption, inference speed, and capacity required to accommodate the system prompt, user query, and retrieved passages. This configuration was also intended to reduce the inclusion of excessive context that could distract the model from the most relevant evidence to be detected.

The all-MiniLM-L6-v2 sentence-transformer model was used to encode fatwa passages into 384-dimensional vector representations. This model was selected because it provides a practical balance between computational efficiency and the quality of semantic representation for local retrieval. FAISS with an IndexFlatIP index was used as a dense retriever. The document vectors were normalised so that the inner-product search could be applied as cosine similarity retrieval.

The sparse retrieval component used MySQL Full-Text Search over a secondary corpus of 7,008 al-Bukhari and 5,362 Muslim hadiths. The initial routing mechanism applied a similarity threshold of 0.35 to the data. When the highest dense retrieval score did not exceed this threshold, the query was redirected to the sparse retrieval path without proceeding to answer generation from the fatwa corpus. Queries exceeding the threshold underwent an additional LLM-based relevance verification stage. The system also incorporated a rule-based keyword bypass that directly activated hadith retrieval when the user explicitly requested a hadith or specifically referred to a hadith source.

2.5. System Testing and Evaluation

A controlled end-to-end evaluation was designed to compare the overall response quality of two system configurations: the proposed sequential multi-corpus chatbot, which combines the MUI fatwa corpus with a hadith fallback retrieval, and a conventional RAG system restricted to the MUI fatwa corpus. To isolate the effect of the retrieval strategy, all remaining system components—including the LLM, prompt configuration, temperature setting, and user interface configuration—were maintained consistently across both experimental conditions.

Two separate evaluation datasets were used because BERTScore and Giskard assess different aspects of system performance and require different evaluation inputs to do so. The first dataset comprised ten specific questions, accompanied by manually constructed reference answers. It was used to assess response relevance using the BERTScore. Each system-generated answer was treated as a candidate or hypothesis, and the corresponding human-annotated answer served as the reference. The BERTScore precision, recall, and F1-score were calculated to quantify the contextual semantic similarity between the generated and reference answers. Precision represents the degree to which the se-

semantic content of the generated response is supported by the reference answer, whereas recall indicates the extent to which the reference content is covered by the generated response. The F1-score provides a harmonic mean.

The second dataset consisted of 50 separate questions evaluated using Giskard's Retrieval-Augmented Generation Evaluation Toolkit. This evaluation focused on whether the generated answers remained faithful to the retrieved context and whether the system appropriately abstained when adequate supporting information was unavailable. The dataset covered two representative categories: questions falling within the thematic coverage of MUI fatwas and questions outside the primary corpus that required retrieval from a secondary hadith database.

The principal Giskard metrics were faithfulness and abstention rate. Faithfulness indicates the extent to which the claims contained in a generated answer can be supported by the retrieved source's context. The abstention rate represents the proportion of cases in which the system refrains from providing unsupported answers when the available context is inadequate. These measures were used as behavioural indicators of grounding reliability and hallucination risk. The combination of BERTScore and Giskard enabled the evaluation to consider both semantic correspondence with human-authored reference answers and factual grounding in the retrieved evidence.

3. RESULTS AND DISCUSSIONS

This section presents and discusses the comparative performance of the conventional single-corpus RAG system and the proposed sequential multi-corpus retrieval system. The evaluation examines complementary dimensions of response quality, including semantic correspondence with human-authored reference answers using BERTScore, and contextual grounding and abstention behaviour using Giskard RAGET. The findings are subsequently interpreted to determine the contribution of the proposed architecture compared with those of previous studies and examined in relation to their theoretical significance, practical implications, and methodological limitations.

3.1. BERTScore Evaluation Results

BERTScore was used to measure the semantic similarity between the answers generated by each system and the corresponding human-authored reference answers. The evaluation involved ten questions covering various Islamic legal topics represented in the MUI Fatwa and hadith corpora.

The results demonstrate that the proposed sequential multi-corpus retrieval system consistently outperformed the conventional single-corpus RAG configuration on all BERTScore metrics. As presented in Table 1, the conventional RAG system achieved an average precision of 0.6770, a recall of 0.6630, and an F1-score of 0.6664. In comparison, the proposed system achieved a precision of 0.7634, recall of 0.7506, and F1-score of 0.7546.

The largest relative improvement was observed in the F1-score, which increased by 13.23%. The recall increased by 13.21%, indicating that the proposed system captured a greater proportion of the semantic information contained in the reference answers. The precision also increased by 12.76%, suggesting that the generated content corresponded more closely to the expected answers. Collectively, these findings indicate that the proposed retrieval architecture produced answers with better semantic coverage and greater alignment with reference responses. Nevertheless, because the evaluation involved only ten questions, the observed improvements should be interpreted as descriptive rather than statistically significant improvements.

Table 1. BERTScore evaluation results for ten test questions.

Metric	Standard RAG	Proposed multi-corpus RAG	Relative improvement
Precision	0.6770	0.7634	12.76%
Recall	0.6630	0.7506	13.21%
F1-score	0.6664	0.7546	13.23%

3.2. Giskard RAGET Evaluation: Faithfulness and Abstain Rate

The Giskard RAGET was used to assess response grounding and the system's tendency to abstain when sufficient supporting evidence was unavailable. The evaluation dataset initially comprised

50 questions. However, several records could not be processed during the automated evaluation because of technical execution failures. The standard RAG configuration produced valid results for 46 questions, whereas the proposed system produced valid results for 43. To ensure a direct and fair comparison, the final paired analysis was restricted to 41 questions that were successfully evaluated under both configurations.

Faithfulness measures the extent to which the claims contained in the generated answer are supported by the retrieved context. On the 41-question intersection, the standard RAG system achieved an average faithfulness score of 0.8537 (85.37 %). The proposed sequential multi-corpus system achieved 0.9439, or 94.39%, representing a relative improvement of 10.57% (see Table 2). This result indicates that the proposed system generated answers that were more consistently grounded in retrieved source material.

Both systems recorded an identical abstain rate of 0.2683, equivalent to 26.83%. The unchanged abstain rate indicates that the proposed architecture does not increase the proportion of questions for which an answer can be generated. Its principal advantage is the grounding quality of the answers produced. In other words, incorporating a secondary hadith corpus and an adaptive relevance verification mechanism improved faithfulness without reducing the frequency of abstention.

Table 2. Giskard RAGET evaluation results.

Metric	Standard RAG	Proposed multi-corpus RAG	Change
Faithfulness	0.8537 (85.37%)	0.9439 (94.39%)	+10.57% relative
Abstain rate	0.2683 (26.83%)	0.2683 (26.83%)	0.00%
Successfully evaluated records	46	43	—
Paired intersection used for comparison	41	41	—

The faithfulness score should be interpreted as an indicator of contextual grounding rather than as direct proof that hallucinations were completely eliminated. A highly faithful answer may remain incomplete or inherit limitations from the retrieved documents. Nevertheless, this improvement provides evidence that the proposed retrieval and routing mechanism reduces the likelihood of generating claims unsupported by the supplied context.

3.3. Comparative Analysis of Standard and Proposed RAG Systems

The two evaluation procedures examine complementary dimensions of response quality. BERTScore assessed semantic correspondence with human-authored reference answers using ten questions, whereas Giskard RAGET assessed contextual faithfulness and abstention behaviour using 41-question paired intersections. Therefore, the results should not be treated as originating from a single common sample. Instead, they jointly provide descriptive evidence of response relevance and reliability.

The proposed system improved the BERTScore F1-score from 0.6664 to 0.7546. This 13.23% relative improvement indicates a better balance between semantic precision and coverage in the proposed model. The simultaneous increases in precision and recall further demonstrate that the improvement did not result merely from generating longer or more inclusive responses; the generated information also exhibited a closer semantic correspondence with the reference answers.

A similar pattern emerged for the contextual grounding. Faithfulness increased from 85.37% to 94.39%, corresponding to a relative improvement of 10.57 percentage points. This result suggests that the LLM-based relevance gate and secondary retrieval path enabled the system to construct answers from more appropriate evidence. When the primary fatwa corpus does not contain an adequate passage, the system can avoid relying on weakly related fatwa content and instead retrieve supporting material from the hadith corpus.

However, the identical abstinence rate of 26.83% reveals an important limitation. Although the proposed architecture improved the quality of the supported answers, it did not broaden the system’s answerable coverage within the paired evaluation set. Consequently, the benefit of sequential

multi-corpus retrieval was primarily qualitative rather than quantitative: the system generated more relevant and better-grounded answers, but it did not answer a greater proportion of questions.

3.4. Discussion and Comparison with Previous Studies

These findings support the research hypothesis that sequential multi-corpus retrieval can improve the response quality of an Islamic legal information chatbot compared to conventional single-corpus RAG. The improvement in BERTScore F1 and faithfulness indicates that combining prioritised corpus selection, threshold-based filtering, LLM-assisted relevance verification, and controlled fallback retrieval offers advantages over retrieving evidence exclusively from MUI fatāwā documents.

The proposed architecture extends the closed-domain fatwa question-answering approach developed by Haekal et al. [16], which relies on a single document corpus. In contrast, the present study preserves MUI fatwas as the primary authoritative corpus while introducing a controlled fallback to a curated hadith database when no relevant fatwa context is identified. This design addresses the issue-specific nature of fatwa documents without treating fatwas and hadith as interchangeable sources. The fallback corpus supplements the primary corpus but does not replace the interpretative and institutional authority associated with an MUI fatwa.

Rahayu et al. [13] reported an average BERTScore F1-score of approximately 0.81 for a LangChain-based question-answering system covering the jurisprudence of the four Sunni schools. Direct numerical comparison with the present study should be approached cautiously because the studies used different corpora, question sets, reference answers and experimental configurations. Nevertheless, this comparison illustrates the broader value of retrieval-grounded generation for Islamic knowledge applications. The distinct contribution of the present study lies in its sequential routing mechanism, which combines an initial similarity threshold, semantic verification by an LLM, and fallback retrieval from secondary corpus.

These findings are also consistent with previous research [16], indicating that access to multiple knowledge sources can improve retrieval performance relative to reliance on a single source. Similarly, the results correspond conceptually with the findings reported in [17], where combining sparse and dense retrieval with LLM-based re-ranking improved question-answering accuracy in the biomedical domain. Biomedical and Islamic legal information systems differ substantially in their epistemological and operational requirements. However, both domains require accurate source selection, terminology-sensitive retrieval, and robust safeguards against unsupported generation. The present results suggest that layered retrieval and relevance verification can be beneficial for accessing Islamic legal information.

The faithfulness score of 94.39% is particularly relevant because an Islamic legal information system must maintain a clear correspondence between its generated responses and the cited source material. However, this score should not be interpreted as evidence that the system is qualified to issue fatwas or conduct independent Islamic legal reasoning. The chatbot remains an information retrieval interface that summarises existing documentary sources. Questions requiring interpretation, reconciliation of conflicting evidence, contextual legal reasoning, or consideration of maqasid al-Shariah must remain within the authority of qualified Islamic scholars.

3.5. Theoretical Significance and Practical Implications

Theoretically, this study contributes to a sequential multi-corpus RAG architecture for question answering in the Islamic legal information domain. The architecture integrates dense semantic retrieval from a primary fatwa corpus, similarity-based filtering, LLM-assisted relevance verification, and sparse lexical retrieval from a secondary Hadith corpus. This design demonstrates that multi-corpus integration does not necessarily require the simultaneous retrieval of all available sources. Instead, corpora can be organised hierarchically according to their function and authority within the application context.

This study also demonstrates the importance of evaluating retrieval-grounded systems across multiple dimensions. BERTScore captures semantic correspondence with reference answers, whereas faithfulness assesses whether the generated claims are supported by the retrieved evidence. The unchanged abstention rate further shows that improvements in answer quality do not automatically translate into broader question coverage. Therefore, these dimensions should be assessed separately when evaluating RAG-based systems.

From a practical perspective, increasing faithfulness from 85.37% to 94.39% may reduce the risk of presenting users with information disconnected from the retrieved Islamic sources. This architecture can facilitate more efficient access to MUI fatwas while providing a contrrerrankingorting pathway to authentic hadith records when the primary corpus is insufficient. Therefore, it may serve as a reference architecture for Islamic institutions, pesantren-based educational organisations, and public-service platforms seeking to develop source-grounded AI applications.

However, practical implementation requires explicit safeguards. Generated responses should identify whether their supporting evidence originates from an MUI fatwa or the hadith corpus, provide traceable citations, and communicate that the retrieved hadith evidence does not automatically constitute an institutional fatwa. The interface should also direct users to qualified scholars when a question involves contextual interpretation, conflicting evidence, or circumstances that are not adequately represented in the available corpora.

3.6. Limitations and Directions for Future Research

Several limitations should be considered when interpreting these findings. First, the BERTScore evaluation was based on only ten questions. Although the results consistently favoured the proposed system, this sample is insufficient to represent the full diversity of Islamic legal questions or support claims of statistical significance. Future studies should employ a substantially larger and stratified evaluation dataset covering transactional law, worship, family law, finance, food, health, technology, and contemporary social issues.

Second, the Giskard evaluation initially involved 50 questions, but the paired comparison was restricted to 41 questions because of automated evaluation failures. Although using the intersection ensured a fair comparison, the missing evaluation records may have introduced selection bias if the failed cases systematically differed from those successfully processed. Future experiments should document the causes of evaluation failure, rerun unsuccessful records where possible and report metric uncertainty using confidence intervals or paired statistical tests.

Third, the study relied on automated evaluation and did not include usability testing or expert validation by legal scholars. BERTScore measures semantic similarity but cannot determine whether a response reflects a valid Islamic legal interpretation. Similarly, automated faithfulness measures contextual support, not the theological authority, completeness, or legal appropriateness of answers. Subsequent research should combine automated metrics with structured assessments conducted by MUI representatives, hadith specialists, Islamic jurisprudence scholars and intended users.

Fourth, the unchanged abstain rate of 26.83% indicates that the secondary hadith corpus did not sufficiently address all questions unsupported by the fatwa corpus. Future research could incorporate an additional controlled retrieval tier comprising recognised classical jurisprudential works, contemporary Islamic legal encyclopaedias and other institutionally validated sources. Such an expansion should preserve the source hierarchy and metadata so that users can distinguish between institutional fatwas, primary textual evidence, and scholarly interpretations.

Finally, retrieval performance may be constrained by the multilingual characteristics of the corpus and the queries. The all-MiniLM-L6-v2 model offers computational efficiency but may not optimally represent specialised Indonesian and Arabic terms. Future work should evaluate multilingual or domain-adapted embedding models, investigate Arabic–Indonesian cross-lingual retrieval, and explore fine-tuning using curated fatwa and hadith datasets. Further experiments could also optimise the similarity threshold, compare alternative re-ranking strategies, and assess whether lower generation temperatures improve factual consistency in source-sensitive applications.

4. CONCLUSION

This study developed and evaluated a Hybrid MUI Fatwa Chatbot based on a sequential multi-corpus retrieval architecture. The proposed system prioritises dense semantic retrieval from MUI fatwa documents and applies similarity-based filtering, followed by LLM-assisted relevance verification. When the retrieved fatwa context is insufficient or irrelevant, the query is redirected to a secondary corpus containing 12,370 curated hadith records using sparse lexical retrieval. This hierarchical mechanism preserves MUI fatwas as the principal source while extending access to supporting Islamic evidence in a controlled and traceable way.

The evaluation results indicate that the proposed architecture produces more relevant and better-grounded answers than the conventional single-corpus RAG configuration. The average BERTScore F1-score increased from 0.6664 to 0.7546, representing a relative improvement of 13.23%. Similarly, the Giskard faithfulness score increased from 0.8537 to 0.9439, equivalent to a relative improvement of 10.57%. These findings support the research hypothesis that sequential multi-corpus retrieval can improve both semantic correspondence with reference answers and faithfulness to the retrieved evidence.

However, both systems recorded an identical abstention rate of 26.83%. Therefore, the principal contribution of the proposed architecture lies in improving the quality and grounding of the generated answers rather than increasing the proportion of questions that can be answered. This finding further demonstrates that improvements in response relevance and faithfulness do not necessarily result in a broader knowledge coverage. Additional knowledge sources and more effective retrieval strategies are required to reduce abstention without compromising the reliability of the source.

Theoretically, this study contributes a hierarchical approach to multi-corpus RAG, in which knowledge sources are organised according to their function and authority rather than retrieved simultaneously. Practically, the resulting system can facilitate public access to MUI fatwas and supporting hadith evidence while reducing the risk of generating responses from weakly related contexts. Nevertheless, the chatbot should be regarded strictly as an information-access and source-summarisation tool. It is not designed to issue fatwas, perform independent Ijtihad, or replace qualified Islamic scholars.

The findings remain subject to several rankings, particularly the relatively small BERTScore dataset, the reduced number of paired records successfully evaluated using Giskard, and the absence of user testing and expert validation by Islamic legal experts. Future research should employ larger and more diverse evaluation datasets, incorporate expert assessments and usability testing, investigate domain-specific Arabic–Indonesian embedding models, and introduce additional institutionally validated knowledge tiers. Further optimisation of the retrieval threshold, reranking mechanism, and abstention strategy may also improve question coverage while maintaining the authority, transparency, and traceability required in Islamic legal information systems.

AUTHOR CONTRIBUTIONS

Fitra Kurnia: Conceptualization, methodology, supervision, and writing—review and editing. Nazruddin Safaat Harahap: Methodology, validation, supervision, and writing—review and editing. Surya Hidayatullah Firdaus: Software, investigation, data curation, formal analysis, visualization, and writing—original draft. All authors approved the final manuscript.

FUNDING STATEMENT

This research received no external funding.

DECLARATION OF AI USE

The authors declare that no AI tools were used in preparing this manuscript.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

All data supporting this study are included in the article. No separate datasets are available.

ETHICAL APPROVAL

Ethical approval was unnecessary, as no humans or animals were involved.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the institutions and individuals who supported this research. We also appreciate the providers of the publicly accessible MUI fatwa documents and curated hadith datasets used to develop and evaluate the proposed system. The authors further acknowledge the developers and research communities responsible for the open-source frameworks, language models, and evaluation tools that facilitated the implementation of this study.

REFERENCES

- [1] Zhao, W. X., Zhou, K., Li, J., Tang, T., Dong, Z., Hou, Y., Zhang, B., Min, Y., Zhang, J., Liu, P., Wang, X., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Hu, Y., Nie, J.-Y., & Wen, J. R. (2026). A survey of large language models. *Frontiers of Computer Science*, **20**(12), 2012627.
DOI: <https://doi.org/10.1007/s11704-026-60308-3>
- [2] Xu, L., Sanders, L., Li, K., & Chow, J. C. (2021). Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR Cancer*, **7**(4), e27850.
DOI: <https://doi.org/10.2196/27850>
- [3] Ardiansyah, M. I. & Widiyanto, M. H. (2021). Development of online learning media based on Telegram Chatbot (Case studies: Programming courses). *Journal of Physics: Conference Series*, **1987**(1), 012006.
DOI: <https://doi.org/10.1088/1742-6596/1987/1/012006>
- [4] Hasanuddin, M., Sharuddin, M. B., & Mahri, J. W. (2023). Pemodelan fatwa ekonomi syariah dan karakteristiknya di Indonesia. *Asy-Syari'ah*, **25**(1), 1-16.
DOI: <https://doi.org/10.15575/as.v25i1.25373>
- [5] Yuspiani, Y., Muhsin, M., Mahfudz, M., Shuhufi, M., & Muhammad, F. (2026). MUI fatwas and sharia-based policy governance in Islamic education management in Indonesia. *Samarah: Jurnal Hukum Keluarga dan Hukum Islam*, **10**(1), 184-200.
DOI: <https://doi.org/10.22373/sjkh.v10i1.28762>
- [6] Dewi, R. R. (2024). Problematika artificial intelligence sebagai pemberi fatwa dalam perspektif hukum Islam. *Jurnal Analisis Hukum*, **7**(2), 209-223.
DOI: <https://doi.org/10.38043/jah.v7i2.5137>
- [7] Putra, S. A. (2024). Mengkaji penerapan tabungan perumahan rakyat dari sudut pandang prinsip keadilan. *Jurnal Ad Dustur*, **1**(1), 1-17.
DOI: <https://doi.org/10.58326/jad.v1i1.195>
- [8] Tonmoy, S. M., Zaman, S. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.
DOI: <https://doi.org/10.48550/arXiv.2401.01313>
- [9] Gao, Y. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
DOI: <https://doi.org/10.48550/arXiv.2312.10997>
- [10] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-rag: Learning to retrieve, generate, and critique through self-reflection. *International Conference on Learning Representations*, **2024**, pp. 9112-9141.
URL: <https://proceedings.iclr.cc>
- [11] Fauzi, M. E. & Sutabri, T. (2025). PublicTalk: Sistem chatbot pintar berbasis natural language processing untuk layanan pemerintahan digital. *Journal Sains Student Research*, **3**(2), 426-433.
DOI: <https://doi.org/10.61722/jssr.v3i2.4325>
- [12] Cahyanti, F. L. D. & Raya, R. D. A. (2025). Perancangan sistem informasi Chatbot retrieval augmented generation berbasis website pada PT. Revolusi Cita Edukasi. *Indonesian Journal Computer Science*, **4**(1), 15-21.
DOI: <https://doi.org/10.31294/m75d4782>

- [13] Rahayu, S., Harahap, N. S., Agustian, S., & Pizaini, P. (2024). Penerapan teknologi LangChain pada question answering system fikih empat Madzhab. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, **4**(3), 974-983.
DOI: <https://doi.org/10.57152/malcom.v4i3.1397>
- [14] Helviansyah, T., Harahap, N. S., Irsyad, M., & Negara, B. S. (2025). Sistem tanya jawab berbasis Chatbot website menggunakan Gemini AI pada data fiqih kontemporer. *Journal of Information System Management (JOISM)*, **7**(1), 38-47.
DOI: <https://doi.org/10.24076/joism.2025v7i1.2082>
- [15] Herwanza, N. A. M., Harahap, N. S., Yanto, F., & Insani, F. (2024). Penerapan Langchain retriever dengan model chat openai dalam pengembangan sistem Chatbot hadis berbasis Telegram. *Jurnal Teknologi Informasi dan Multimedia*, **6**(1), 70-83.
DOI: <https://doi.org/10.35746/jtim.v6i1.514>
- [16] Haekal, F. I., Perdana, R. S., & Adikara, P. P. (2025). Sistem Tanya jawab closed-domain terhadap dokumen fatwa menggunakan retrieval augmented generation dan large language model. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, **9**(5).
URL: <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/14854>
- [17] Boucekif, A., Rashwani, S., Sbahi, H., Gaben, S., Al Khatib, M., & Ghaly, M. (2025). Assessing large language models on islamic legal reasoning: Evidence from inheritance law evaluation. *Proceedings of The Third Arabic Natural Language Processing Conference*, 246-257.
DOI: <https://doi.org/10.18653/v1/2025.arabicnlp-main.20>
- [18] Miyaji, R., Moulin, R., Monção, S., & Machado, L. (2025). Evaluating rag-based qa systems: a comparative analysis of llm as a judge, traditional metrics, and human alignment. *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 247-258.
DOI: <https://doi.org/10.5753/stil.2025.37829>
- [19] Dobslaw, F., Feldt, R., Yoon, J., & Yoo, S. (2026). Challenges in testing large language model-based software: A faceted taxonomy. *ACM Transactions on Software Engineering and Methodology*, **35**(9), 1-24.
DOI: <https://doi.org/10.1145/3806396>
- [20] Team, L. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, **2**(3).
DOI: <https://doi.org/10.48550/arXiv.2407.21783>