Journal of Deep Learning, Computer Vision and Digital Image Processing
E-ISSN: 2986-8939 P-ISSN: 2986-8920
Vol. 4, No. 3, September 2026
DOI. https://doi.org/10.61255/decoding.v4i3.1840

Evaluating Precision and Semantic Similarity in a Retrieval-Augmented Generation-Based Question Answering System for Shafi'i Prayer Fiqh

Thoriq Wijaya ✉
(Universitas Malikussaleh, Aceh, Indonesia)

Defry Hamdhana
(Universitas Malikussaleh, Aceh, Indonesia)

Fajriana
(Universitas Malikussaleh, Aceh, Indonesia)

Mukti Qamal
(Universitas Malikussaleh, Aceh, Indonesia)

OPEN ACCESS

ARTICLE HISTORY
Received: June 9, 2026
Revised: July 20, 2026
Accepted: August 24, 2026

KEYWORDS
Fiqh of prayer;
Large language model;
Question answering system;
Retrieval-augmented generation;
Semantic retrieval.

ABSTRACT

Purpose – This study evaluates retrieval precision and answer similarity in a closed-domain question-answering system for Shafi'i prayer fiqh using retrieval-augmented generation (RAG).

Methods – Indonesian translations of three classical fiqh texts were segmented into 251 chunks using a 1000-character chunk size and 200-character overlap. LangChain coordinated multilingual-e5-base embeddings, ChromaDB retrieval, and local Qwen2.5:3b generation. The retriever returned up to four eligible chunks. Twenty predefined queries were evaluated using Precision@4, cosine similarity to researcher-prepared reference answers, and qualitative assessment of required answer components.

Findings – Mean Precision@4 was 70.00%, and mean cosine similarity was 0.9197 (91.97%). Retrieval varied across queries, with Query 19 returning only one relevant chunk. Queries 8 and 19 produced incomplete answers despite high semantic similarity. The system returned a predefined fallback response when no retrieved passages passed the distance threshold.

Research implications – The results support preliminary prototype feasibility but do not establish expert-verified correctness of the rulings. The small query set and absence of a quantitative non-RAG baseline limit generalization and prevent estimation of any reduction in hallucination.

Originality – The study evaluates retrieval and generation separately within a constrained Indonesian fiqh corpus and illustrates the distinction between semantic similarity and answer completeness.

Correspondence Author: ✉thoriq.220170150@mhs.unimal.ac.id
To cite this article: Wijaya, T., Hamdhana, D., Fajriana, & Qamal, M. (2026). Evaluating Precision and Semantic Similarity in a Retrieval-Augmented Generation-Based Question Answering System for Shafi'i Prayer Fiqh. *Journal of Deep Learning, Computer Vision and Digital Image Processing*, *4*(3), 169–182. https://doi.org/10.61255/decoding.v4i3.1840

This is an open access article under the CC BY-SA license

INTRODUCTION

Digital transformation has changed how people access religious legal literature [1]. Independent study of Islamic jurisprudence (fiqh) can be challenging for lay readers because the rules governing the pillars, validity conditions, and invalidators of worship are complex [2]. Conventional search engines require users to locate and synthesize information from separate web documents. A question-answering system could support this process by responding to natural-language queries with structured guidance linked to relevant sources.

Large language models (LLMs) provide a foundation for interactive question-answering systems [3]. However, they can generate fluent statements that are factually incorrect, a problem commonly described as hallucination [4]. In Islamic legal consultation, generating answers without a clearly defined source corpus may produce responses that depart from the intended jurisprudential framework [5].

Retrieval-augmented generation (RAG) supplies documents from an external knowledge base as context for an LLM before answer generation [6]. Recent studies have applied this approach to Islamic information systems. Fathoni et al. developed an open-domain system drawing on sources from Nahdlatul Ulama, Muhammadiyah, and the Indonesian Ulema Council [7]. Asl and Minaei-Bidgoli developed FARSIQA to index Islamic legal documents across multiple categories [8], while Alan et al. proposed MufassirQAS for answering Islamic conceptual questions using broader knowledge sources [9].

To limit variation across jurisprudential perspectives, Ichsan and Nusantara developed a closed-domain system based exclusively on Safinatun Najah [10]. Restricting the corpus to a single source can help maintain a consistent legal perspective. Farisi et al. similarly applied instructional constraints to consultations concerning Hajj and Umrah [11].

Cross-madhhab corpora may introduce ambiguity when similar terms refer to different legal rulings. A retriever may select passages that are semantically related to a query but inconsistent with the intended school of jurisprudence. Although closed-domain systems limit this source of variation, the reviewed studies primarily emphasize the quality of generated answers. Separate quantitative evaluation of retrieval is needed to establish whether the context supplied to the language model directly addresses the query.

This study therefore evaluates a closed-domain RAG architecture using Indonesian translations of Shafi'i prayer-fiqh texts. The corpus is restricted to obligatory prayer within the Shafi'i school to limit variation arising from different legal interpretations (ikhtilaf) [12]. Evaluating retrieval and generation separately allows errors to be examined at each stage of the pipeline.

The study provides a preliminary evaluation of an Indonesian question-answering system for Shafi'i prayer fiqh. Its contribution is the separate assessment of retrieval precision and answer similarity within a specialized corpus containing substantial lexical repetition. The evaluation also examines whether high semantic similarity is accompanied by complete coverage of the required answer components.

The objective is to implement a closed-domain RAG system and evaluate its retrieval precision and the semantic similarity of generated answers to researcher-prepared reference answers.

METHOD

This study used a software-prototype design and evaluation approach to develop an educational question-answering system. The system was restricted to a defined corpus and evaluated using a predefined set of queries.

The knowledge base comprised Indonesian translations of three classical Shafi'i fiqh texts: Matan Safinatun Najah by Salim bin Sumair Al-Hadhrami, translated by Achmad Sunarto and published by Mutiara Ilmu; Fathul Qarib by Muhammad bin Qasim Al-Ghazi, translated by M. Faisol Fatawi and published by Lirboyo Press; and Matan Abu Syuja' (Taqrib) by Ahmad bin Al-Husain Al-Asfahani,

published by Al-Hidayah. Only chapters addressing obligatory prayer were extracted. Overlapping rulings across the texts were retained. Extraction and segmentation produced 251 document chunks for indexing in ChromaDB.

LangChain coordinated data loading, retrieval, and generation [13]. TextLoader loaded the extracted text, and RecursiveCharacterTextSplitter segmented it into chunks. The intfloat/multilingual-e5-base model generated embeddings through Hugging Face [14], and ChromaDB stored the resulting vectors [15]. Qwen2.5:3b generated responses locally through Ollama. The user interface was implemented as a progressive web app (PWA) using Laravel, with a Python FastAPI backend [16].

The experimental procedure consists of two main phases: data preparation and user interaction inference.

Data preparation: The source PDFs were converted to plain text. Cleaning removed page numbers, layout artifacts, and extraneous symbols. Case folding converted the text to lowercase.

Chunking: Document chunking is managed using static parameters. The chunk size (chunk_size) is set to 1000 characters, maintaining an overlap area (chunk_overlap) of 200 characters to preserve semantic continuity across segment borders [17]. Hierarchical separators are configured sequentially using the character strings \n[CHUNK-, \n###, \nBAB, and \n\n.

Embedding and indexing: The multilingual-e5-base model converted each text chunk into a 768-dimensional embedding vector [14]. The vectors were indexed locally in ChromaDB [15].

Inference: User queries are converted into vector representations. The retriever module executes a semantic similarity search on the ChromaDB cluster [18]. The retrieval parameter restricts the results to the top four documents (*Top-K Retrieval*, $\boldsymbol{k = 4}$). The retrieved document chunks are assembled and injected alongside the user query into the system's instruction prompt. This combined prompt is processed by the Qwen2.5:3b language model to generate the final text response [19].

Prompt template and inference constraints: LangChain combined the retrieved passages and the query in a predefined ChatPromptTemplate. The instructions required the model to answer only from the supplied context, use concise Indonesian, omit conversational filler, and avoid exposing system metadata. Predefined fallback responses were provided for out-of-domain queries and empty retrieval results. These instructions were intended to constrain unsupported generation; their presence alone does not establish factual correctness.

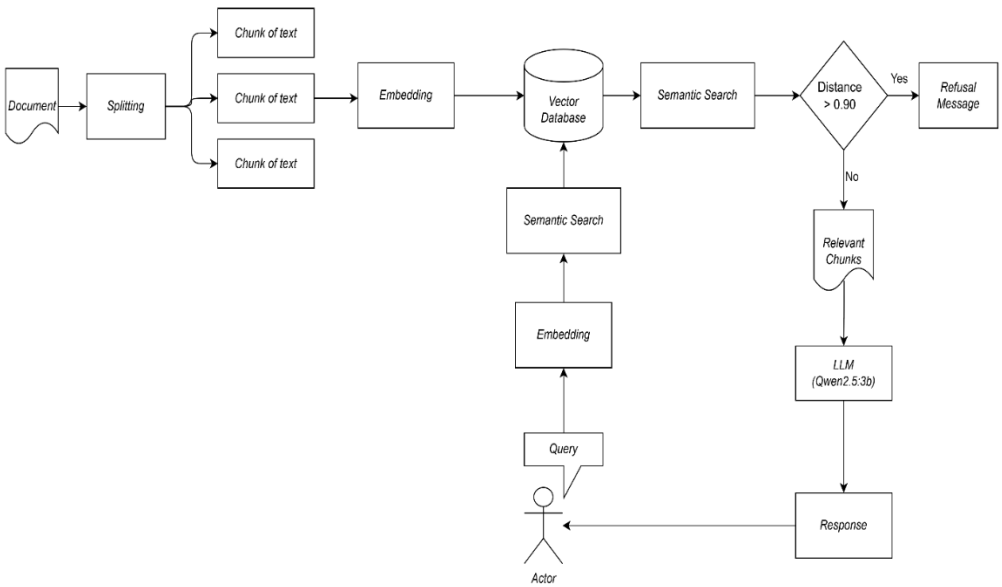


Figure 1. RAG system flow for prayer fiqh

Figure 1 shows the two stages of the RAG workflow. During data preparation, TextLoader reads the combined plain-text corpus and RecursiveCharacterTextSplitter divides it into overlapping chunks. The multilingual-e5-base model then embeds each chunk, and ChromaDB stores the vectors for subsequent retrieval.

During retrieval and generation, a query submitted through the PWA is embedded using the same model. ChromaDB retrieves the nearest eligible chunks, which are combined with the query and passed to Qwen2.5:3b through Ollama. The generated response is then returned to the PWA. Whether the response is relevant and complete is assessed through the evaluation described below.

The system was evaluated using 20 predefined Indonesian queries about Shafi'i prayer fiqh. Two metrics were used:

Precision@4 measures the proportion of the four retrieved chunks judged relevant to a query [20]. It evaluates retrieval performance rather than answer correctness.

$$\boldsymbol{Precision@k = \frac{\text{Number of relevant documents in Top-k}}{k}}$$

Cosine similarity measures the semantic alignment between the embedding of a generated answer and that of its corresponding reference answer [18]. Higher similarity does not necessarily imply factual correctness or complete coverage of the required ruling.

$$\boldsymbol{CosineSimilarity(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|}}$$

Evaluation Metric Calculations

Query 19 illustrates how the evaluation metrics were calculated. Only one of the four retrieved chunks was judged relevant to this query.

Precision@4 Calculation (Query No. 19): For instance, in Query Number 19, the search module retrieves four top documents ($\boldsymbol{k = 4}$), but only one document is identified as contextually relevant to the user's query. Based on the *Precision@k* equation, the calculation is executed as follows:

$$\boldsymbol{Precision@4 = \frac{1}{4} = 0.25}$$

The resulting proportion was expressed as a percentage:

$$\boldsymbol{0.25 \times 100\% = 25.00\%}$$

Cosine Similarity Calculation (Query No. 19): Concurrently, the *Cosine Similarity* spatial vector computation between the LLM output text vector ($\boldsymbol{A}$) and the reference answer vector ($\boldsymbol{B}$) for Query 19 records a raw score of 0.8738.

$$\boldsymbol{CosineSimilarity(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} = 0.8738}$$

The cosine score was also multiplied by 100 for presentation in the evaluation tables:

$$\boldsymbol{Similarity = 0.8738 \times 100\% = 87.38\%}$$

Overall performance was summarized as the arithmetic mean across the 20 queries.

The sum of the 20 Precision@4 percentages was 1400%:

$$\boldsymbol{MeanPrecision = \frac{1400\%}{20} = 70.00\%}$$

The sum of the 20 semantic similarity percentages was 1839.38%:

$$\boldsymbol{MeanSimilarity = \frac{1839.38\%}{20} = 91.97\%}$$

Annotation Rubric and Answer Sufficiency

For Precision@4, a retrieved chunk was labeled relevant only if it directly contained the ruling required to answer the query. Topically related chunks addressing a different ruling were labeled irrelevant. Under this rubric, 30.00% of the retrieved chunks were judged irrelevant. Generated answers were also assessed for completeness against the required components in the source texts, including conditional rulings and requested recitations.

Scope and Limitations of the Methodology

Evaluation was limited to 20 Indonesian queries within the selected domain. Retrieval used ChromaDB's default squared Euclidean (L2) distance with normalized embeddings. Qwen2.5:3b was configured with a temperature of 0.0 and the default top-p setting. The researchers prepared the 20 reference answers by summarizing the selected source texts. No quantitative comparison with a non-RAG language-model baseline was conducted.

The multilingual-e5-base model was used both for retrieval embeddings and for answer-similarity evaluation. Retrieved candidates with distance scores above 0.90 were rejected. The research team assessed relevance by comparing the retrieved chunks with the query and its reference answer.

RESULTS AND DISCUSSION

The implemented system provided an interactive PWA interface [16]. Answers were presented as paragraphs or bullet points to improve readability while conveying information from the retrieved passages.

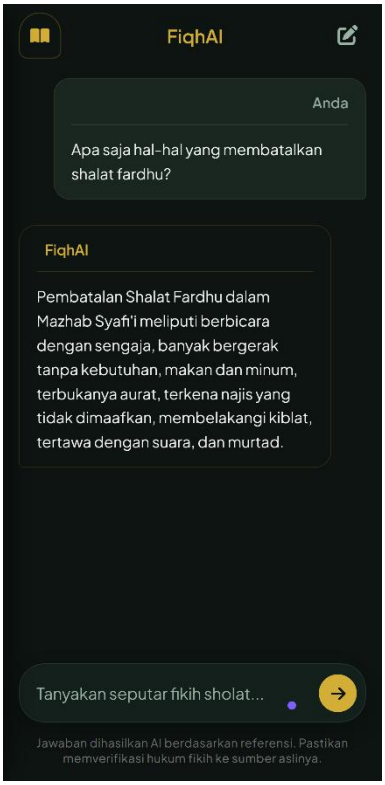


Figure 2. PWA interface displaying retrieved-context responses

Figure 2 shows an example response concerning actions that invalidate obligatory prayer (pembatal shalat). The interface displays the answer generated from the retrieved context.

For the tested out-of-domain queries, including questions about sports results, no candidate passages passed the retrieval threshold. The application therefore returned a predefined message explaining the system's scope.

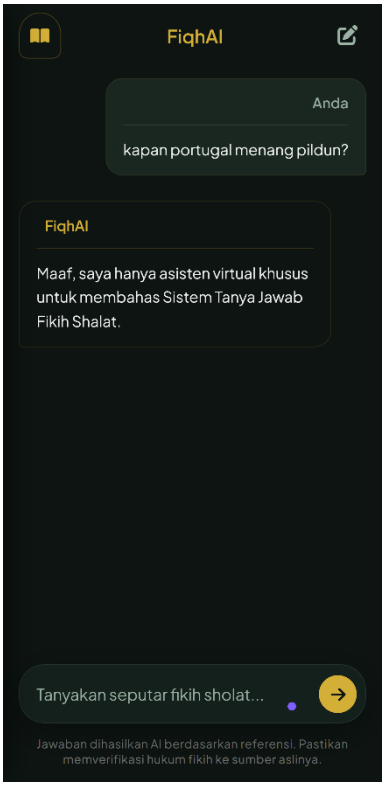


Figure 3. PWA interface for Out-of-Domain query rejection

Figure 3 illustrates this rejection behavior. When no eligible context was available, the application returned the fallback message rather than generating an unrestricted response.

Table 1 presents five representative queries from the 20-question evaluation set.

Table 1. Representative Sample of Query Scenarios

No.	Query Description
2	Are people who have lost their intellect or are insane obligated to perform prayer?
7	What is the recommended action if someone forgets the initial tashahhud and stands up?
8	Is a person's prayer valid if they forget or intentionally leave the iftitah prayer?
11	What is the ruling on closing eyes during prayer without urgent need?
19	What should be recited during sujud sahwi?

Retrieval was evaluated using Precision@4, while answer similarity was evaluated using cosine similarity. Table 2 reports the selected examples and the mean across all 20 queries. The complete query-level results are provided in the Appendix.

Table 2. Evaluation Results of Retriever and Generator Performance

Query No.	Relevant chunks (k = 4)	Precision@4 (%)	Cosine similarity	Similarity (%)
2	4	100.00%	0.9369	93.69%
7	4	100.00%	0.9251	92.51%
8	2	50.00%	0.9290	92.90%
11	3	75.00%	0.9180	91.80%
19	1	25.00%	0.8738	87.38%
Mean	-	70.00%	0.9197	91.97%

As shown in Table 2, retrieval performance varied across queries. Queries 2 and 7 each retrieved four relevant chunks, giving Precision@4 values of 100.00%, whereas Query 19 retrieved one relevant chunk, giving 25.00%. The mean cosine similarity across all 20 generated and reference answers was 0.9197. This indicates substantial semantic similarity but does not establish that every answer contains all required legal conditions. Independent expert assessment is needed to evaluate the correctness and completeness of the rulings.

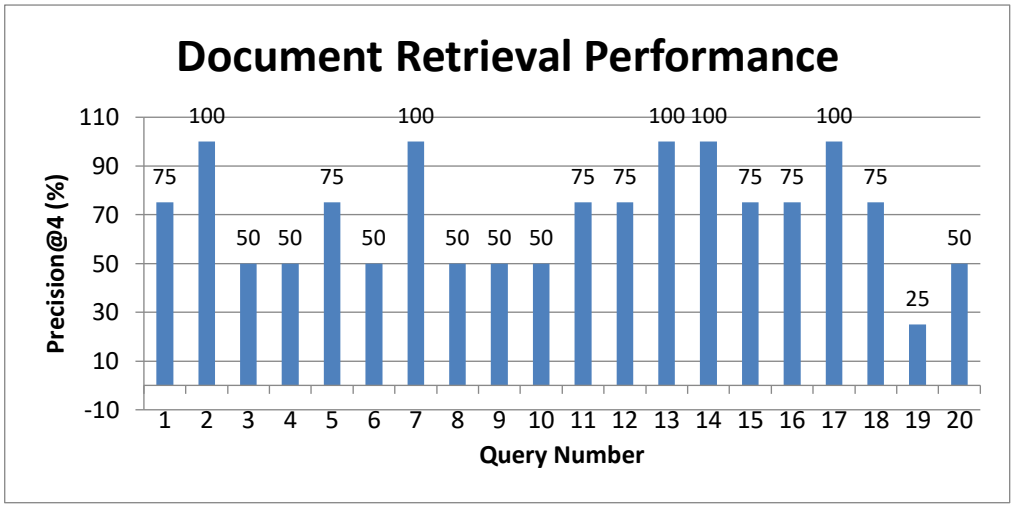


Figure 4. Precision@4 retrieval performance distribution

Figure 4 shows the variation in retrieval precision across queries. Some queries retrieved four relevant chunks, while others retrieved fewer, reducing the overall mean to 70.00%.

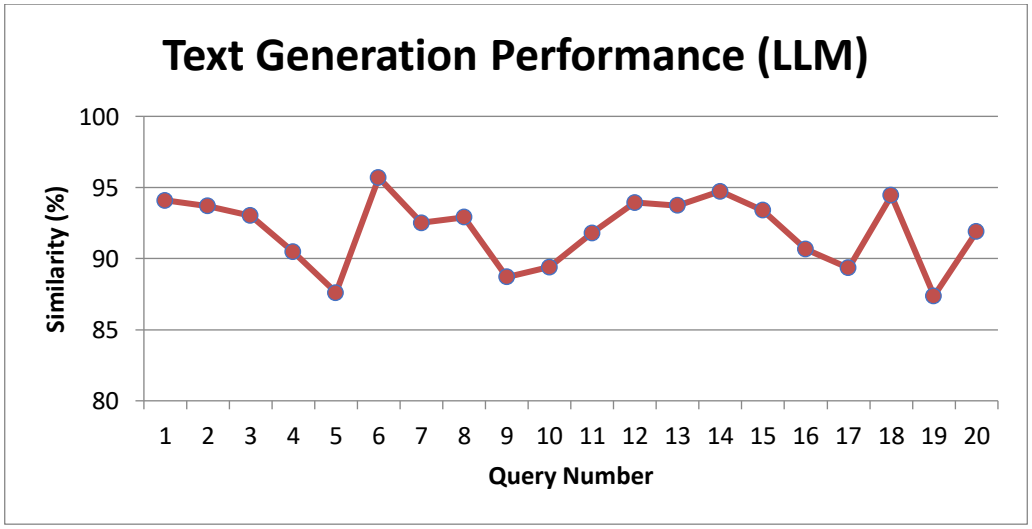


Figure 5. Semantic similarity between model outputs and reference answers

Figure 5 shows that all cosine similarity scores exceeded 0.87 despite variation in retrieval precision. This pattern indicates semantic resemblance to the reference answers, but similarity alone does not establish answer completeness.

Table 3 compares selected reference answers with the corresponding generated responses.

Table 3. Observational Comparison of Semantic Alignment

Query No.	Reference Answer	System Generated Answer
2	Insane persons are not burdened with the obligation of prayer as long as the loss of intellect persists.	People who lose their intellect, such as insane people, are not obligated to perform prayers as long as that condition persists. The obligation applies to those who have the ability to think.
8	The prayer remains valid because reading the iftitah prayer is considered sunnah hai'at, and it is not recommended to perform sujud sahwi if left out.	A person's prayer remains valid even if they forget or intentionally leave the iftitah prayer. It is included in sunnah hai'at, thus it does not invalidate the prayer.
11	Closing eyes during prayer without a need is basically considered a makruh (disliked) act.	In the Shafi'i school, closing the eyes during prayer without reason is ruled as makruh. This occurs if there is no necessity requiring it.
19	As for the prostration reading in prayer that is recommended: "Subhaana Rabbiyal A'laa" (Glory be to my Lord, the Most High). This reading is recommended to be read three times.	The sujud sahwi reading is basically the same as the prostration reading in prayer.

The mean Precision@4 of 70.00% indicates that the configuration retrieved relevant context for many queries, although 30.00% of the returned chunks were judged irrelevant. Static chunking used 1000 characters per chunk with a 200-character overlap [17]. Repeated terms such as sujud (prostration), niat (intention), and bacaan (recitation) occur across different rulings in the selected fiqh texts [10]. This repetition may contribute to the retrieval of passages that share vocabulary but address a different question.

For Query 19, which asks about the recitation for sujud sahwi, the retriever returned general prostration passages alongside the relevant material, yielding Precision@4 of 25.00%. This example suggests difficulty distinguishing closely related subtopics within the corpus. A comparison with alternative retrievers would be needed to determine whether lexical repetition caused the observed error and whether another retrieval method would reduce it [20].

The mean cosine similarity of 91.97% reflects semantic alignment between generated and reference answers [18]. However, Query 19 retained a score of 87.38% despite incomplete coverage of the requested recitation. The result illustrates why embedding similarity should not be treated as a measure of religious or legal correctness. It also does not, by itself, establish that Qwen2.5:3b faithfully used every relevant element of the supplied context [19].

Answer Sufficiency and Component Evaluation

The researchers assessed all 20 generated answers against the required components derived from the source texts. An answer was classified as complete if it retained the required elements and partial if it omitted a supported element. Table 4 identifies Queries 8 and 19 as partial: the former omitted a detail concerning sujud sahwi, and the latter omitted the requested recitation. These examples show that high semantic similarity can coexist with incomplete answers.

Table 4. Qualitative Assessment of Answer Sufficiency

No.	Original Query (Indonesian)	Required Fiqh Component(s)	Adjudicated Judgment
1	Apa tujuan utama dari pelaksanaan ibadah shalat bagi seorang muslim?	Worship, fulfillment of obligation	Complete
2	Apakah orang yang hilang akal atau gila dibebani kewajiban untuk melaksanakan shalat?	No, obligation is lifted	Complete

No.	Original Query (Indonesian)	Required Fiqh Component(s)	Adjudicated Judgment
3	Apa saja syarat sah shalat yang wajib dipastikan sebelum memulai ibadah?	Purity, qibla, time entry, awrah	Complete
4	Kapan batas waktu berakhirnya pelaksanaan shalat Asar?	Sunset (tenggelam matahari)	Complete
5	Apakah niat shalat wajib dilafalkan dengan lisan, atau cukup di dalam hati saja?	Mandatory in heart, sunnah on lips	Complete
6	Sebutkan 7 anggota tubuh yang wajib menyentuh alas ketika melakukan sujud!	Forehead, 2 palms, 2 knees, 2 toes	Complete
7	Apa tindakan yang disunnahkan jika seseorang lupa melaksanakan tasyahud awal lalu terlanjur berdiri?	Perform Sujud Sahwi before salam	Complete
8	Apakah shalat seseorang tetap sah jika ia lupa atau sengaja meninggalkan doa iftitah?	Valid; explanation of Sahwi applicability	Partial (Omitted Sahwi detail)
9	Apakah sekadar tersenyum tanpa mengeluarkan suara termasuk perkara yang membatalkan shalat?	No, unless producing two letters	Complete
10	Bagaimana hukumnya menelan sisa makanan yang ada di mulut dengan sengaja ketika shalat sedang berlangsung?	Invalidates the prayer (Batal)	Complete
11	Apa hukum memejamkan mata ketika sedang shalat jika tidak ada kebutuhan mendesak?	Makruh (Disliked)	Complete
12	Siapa kriteria orang yang paling berhak atau diutamakan untuk ditunjuk menjadi imam shalat?	Faqih, Qari', mature, etc.	Complete
13	Pada momen gerakan apa seorang makmum masbuk dianggap telah berhasil mendapatkan satu rakaat bersama imam?	Reaching perfect ruku' with Imam	Complete
14	Apa yang harus langsung dilakukan oleh makmum masbuk sesaat setelah imam mengucapkan salam?	Stand up to complete missed rak'ahs	Complete
15	Shalat fardhu apa saja yang diperbolehkan untuk diringkas (diqashar) menjadi dua rakaat?	Dhuhr, Asr, Isha	Complete
16	Berapa batas jarak tempuh minimal sebuah perjalanan agar seseorang diperbolehkan mengqashar shalatnya?	2 Marhalah (~82 km)	Complete
17	Apa bacaan tasbih yang disunnahkan untuk dibaca ketika sedang posisi ruku'?	Subhana rabbiyal adzim...	Complete
18	Kapan letak waktu yang tepat untuk melaksanakan dua kali sujud sahwi?	Before the final salam	Complete
19	Bagaimana bacaan yang dibaca ketika melakukan sujud sahwi?	Explicit recitation text provided	Partial (Text omitted)
20	Sebutkan ke-17 rukun shalat secara lengkap sesuai urutan dalam Mazhab Syafi'i!	17 sequential pillars (Niat, dll)	Complete

An exploratory qualitative comparison was also made by prompting Qwen2.5:3b without retrieved context. The researchers observed broader responses that sometimes mixed jurisprudential perspectives or omitted source-specific terminology. Restricting retrieval to the selected Shafi'i texts

was intended to reduce such variation [7]. However, no matched quantitative baseline evaluation was conducted, so the observations do not establish a measured reduction in hallucination or cross-madhhab inconsistency.

The out-of-domain mechanism used a distance threshold: candidates with scores above 0.90 were rejected. When no eligible passages remained, the application returned its fallback response. This filtering behavior constrained generation in the tested cases, although its effectiveness across a broader range of in-domain and out-of-domain queries remains to be evaluated [21], [4].

The findings suggest that retrieval in a specialized corpus may require methods that distinguish semantically related but contextually different rulings. Future comparisons could combine sparse keyword retrieval with dense retrieval and evaluate whether the combination improves relevance and answer completeness [22], [23].

The study contributes a preliminary evaluation of static character-based segmentation and multilingual embeddings for a closed-domain Indonesian fiqh question-answering system.

Several limitations constrain interpretation of the results.

The evaluation used only 20 predefined queries and does not support broad generalization to other questions, users, or jurisprudential sources.

No quantitative non-RAG baseline comparison was performed. Consequently, the effect of retrieval on hallucination frequency cannot be estimated from these results.

Further work should use a larger and more diverse query set, independent expert assessment, and matched retrieval and generation baselines. Comparing dense retrieval with BM25 and hybrid approaches would help establish whether the errors observed in this corpus can be reduced [24], [25].

CONCLUSION

This study implemented and evaluated a closed-domain RAG question-answering prototype for Shafi'i prayer fiqh. Across 20 queries, mean Precision@4 was 70.00%, and mean cosine similarity between generated and reference answers was 0.9197 (91.97%). The system returned a fallback response when no retrieved passages passed the distance threshold. However, incomplete answers to Queries 8 and 19 demonstrate that semantic similarity does not establish complete coverage of the requested ruling. The small query set, lack of an independently validated answer assessment, and absence of a quantitative non-RAG baseline limit generalization and prevent estimation of any reduction in hallucination. Larger evaluations should assess retrieval relevance, answer completeness, and expert-verified correctness separately.

APPENDIX

Table 5 presents all 20 evaluation queries, their retrieval precision, and the similarity of their generated answers to the reference answers.

Table 5. Complete Dataset of System Evaluation

No.	Original Query (Indonesian)	Query Description	Relevant chunks (k = 4)	Precision @4 (%)	Cosine Similarity	Similarity (%)
1	Apa tujuan utama dari pelaksanaan ibadah shalat bagi seorang muslim?	What is the main purpose of performing prayer for a Muslim?	3	75.00%	0.9409	94.09%

No.	Original Query (Indonesian)	Query Description	Relevant chunks (k = 4)	Precision @4 (%)	Cosine Similarity	Similarity (%)
2	Apakah orang yang hilang akal atau gila dibebani kewajiban untuk melaksanakan shalat?	Are people who have lost their intellect or are insane obligated to perform prayer?	4	100.00%	0.9369	93.69%
3	Apa saja syarat sah shalat yang wajib dipastikan sebelum memulai ibadah?	What are the validity conditions of prayer that must be ensured before starting the worship?	2	50.00%	0.9301	93.01%
4	Kapan batas waktu berakhirnya pelaksanaan shalat Asar?	When is the time limit for the Asr prayer?	2	50.00%	0.9047	90.47%
5	Apakah niat shalat wajib dilafalkan dengan lisan, atau cukup di dalam hati saja?	Must the prayer intention be pronounced verbally, or is it sufficient in the heart?	3	75.00%	0.8760	87.60%
6	Sebutkan 7 anggota tubuh yang wajib menyentuh alas ketika melakukan sujud!	Mention the 7 body parts that must touch the ground during prostration!	2	50.00%	0.9569	95.69%
7	Apa tindakan yang disunnahkan jika seseorang lupa melaksanakan tasyahud awal lalu terlanjur berdiri?	What is the recommended action if someone forgets the initial tashahhud and stands up?	4	100.00%	0.9251	92.51%
8	Apakah shalat seseorang tetap sah jika ia lupa atau sengaja meninggalkan doa iftitah?	Is a person's prayer valid if they forget or intentionally leave the iftitah prayer?	2	50.00%	0.9290	92.90%
9	Apakah sekadar tersenyum tanpa mengeluarkan suara termasuk perkara yang membatalkan shalat?	Does merely smiling without making a sound invalidate the prayer?	2	50.00%	0.8870	88.70%
10	Bagaimana hukumnya menelan sisa makanan yang ada di mulut dengan sengaja ketika shalat sedang berlangsung?	What is the ruling on intentionally swallowing food residue in the mouth while the prayer is ongoing?	2	50.00%	0.8940	89.40%
11	Apa hukum memejamkan mata ketika sedang shalat jika tidak ada kebutuhan mendesak?	What is the ruling on closing eyes during prayer without urgent need?	3	75.00%	0.9180	91.80%

No.	Original Query (Indonesian)	Query Description	Relevant chunks (k = 4)	Precision @4 (%)	Cosine Similarity	Similarity (%)
12	Siapa kriteria orang yang paling berhak atau diutamakan untuk ditunjuk menjadi imam shalat?	Who meets the criteria to be the most entitled or prioritized to be appointed as the prayer imam?	3	75.00%	0.9394	93.94%
13	Pada momen gerakan apa seorang makmum masbuk dianggap telah berhasil mendapatkan satu rakaat bersama imam?	At what exact movement is a late follower (masbuk) considered to have successfully caught one rak'ah with the imam?	4	100.00%	0.9374	93.74%
14	Apa yang harus langsung dilakukan oleh makmum masbuk sesaat setelah imam mengucapkan salam?	What should a late follower (masbuk) immediately do right after the imam says the salam?	4	100.00%	0.9472	94.72%
15	Shalat fardhu apa saja yang diperbolehkan untuk diringkas (diqashar) menjadi dua rakaat?	Which obligatory prayers are allowed to be shortened (qashar) to two rak'ahs?	3	75.00%	0.9340	93.40%
16	Berapa batas jarak tempuh minimal sebuah perjalanan agar seseorang diperbolehkan mengqashar shalatnya?	What is the minimum travel distance required for a person to be allowed to shorten their prayer?	3	75.00%	0.9066	90.66%
17	Apa bacaan tasbih yang disunnahkan untuk dibaca ketika sedang posisi ruku'?	What is the recommended tasbih recitation during bowing (ruku')?	4	100.00%	0.8934	89.34%
18	Kapan letak waktu yang tepat untuk melaksanakan dua kali sujud sahwi?	When is the correct time to perform the two prostrations of forgetfulness (sujud sahwi)?	3	75.00%	0.9444	94.44%
19	Bagaimana bacaan yang dibaca ketika melakukan sujud sahwi?	What should be recited during sujud sahwi?	1	25.00%	0.8738	87.38%
20	Sebutkan ke-17 rukun shalat secara lengkap sesuai urutan dalam Mazhab Syafi'i!	Mention all 17 pillars of prayer completely in order according to the Shafi'i school!	2	50.00%	0.9190	91.90%
-		Overall Mean (n=20)	-	70.00%	0.9197	91.97%

ACKNOWLEDGMENT

The authors express their gratitude to the leadership, academic advisors, and the entire academic community of Universitas Malikussaleh for the facilities and support provided during the completion of this research.

CONFLICT OF INTEREST

The authors declare no conflict of interest regarding the publication of this manuscript.

AUTHOR CONTRIBUTION STATEMENT

TW: Conceptualization, methodology, software development, investigation, data collection, formal analysis, and writing—original draft. DH: Supervision, methodological guidance, system architecture review, and writing—review and editing. FA: Co-supervision, validation of evaluation methods, and writing—review and editing. MQ: Technical review, system evaluation, and review of the data analysis.

AI DISCLOSURE STATEMENT

Gemini was used during manuscript preparation for organization, language editing, and citation formatting. ChatGPT (OpenAI) was used during copyediting for language, formatting, and consistency checks. The authors remain responsible for the scientific content and the final manuscript.

REFERENCES

[1] S. Suaidi, F. Faridi, and S. Sunarto, "Technology-Based Digitalization of Islamic Religious Education," *Halaqa: Islamic Education Journal*, vol. 9, no. 2, 2025.

[2] Khairuddin, "Sistematika Penulisan Fiqh Syafi'i (Analisa Buku Fiqh Al-Manhaji 'Ala Mazhab Imam Al'syafi'i) [Systematics of Shafi'i Fiqh Writing]," *Sangaji: Jurnal Pemikiran Syariah dan Hukum*, vol. 8, no. 2, pp. 288-298, 2024.

[3] D. Suhartono, M. R. N. Majiid, and R. Fredyan, "Towards automatic question generation using pre-trained model in academic field for Bahasa Indonesia," *Education and Information Technologies*, vol. 29, no. 16, 2024.

[4] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval Augmentation Reduces Hallucination in Conversation," *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784-3803, 2021.

[5] O. Ram, Y. Levine, I. Dalmedigo, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham, "In-Context Retrieval-Augmented Language Models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1316-1331, 2023.

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459-9474, 2020.

[7] K. Fathoni, A. Sudaryanto, I. Syarif, Z. M. E. Darmawan, A. A. Yunanto, Prasetyo, R. S. Aji, A. I. Fauzy, and R. Y. Hakkun, "EMasjid, Islamic Application with RAG (Retrieval-Augmented Generation)," *Informatics, Electrical and Electronics Engineering* (Infotron), vol. 5, no. 1, pp. 49–58, 2025.

[8] M. A. Asl and B. Minaei-Bidgoli, "FARSIQA: Faithful and Advanced RAG System for Islamic Question Answering," arXiv preprint arXiv:2510.25621, 2025.

[9] AA. Y. Alan, E. Karaarslan, and O. Aydin, "A RAG-based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM," *TechRxiv*, 2024. DOI: 10.36227/techrxiv.170723304.41988020/v2.

[10] M. Ichsan and A. G. Nusantara, "Perancangan Chatbot Fiqih Berbasis Retrieval-Augmented Generation Kitab Safinatun Najah [Designing a Fiqh Chatbot Based on Retrieval-Augmented Generation of the Safinatun Najah Book]," *Seminar Nasional Teknologi & Sains (STAINS)*, vol. 5, no. 1, pp. 604-610, 2026.

[11] A. Farisi, Chriestina, D. Dafid, and A. Jansa, "Penerapan Model AI Gemini pada Pengembangan Chatbot Tanya Fiqih Haji dan Umrah [Implementation of the Gemini AI Model in the Development of a Hajj and Umrah Fiqh QA Chatbot]," *INSOLOGI: Jurnal Sains Dan Teknologi*, vol. 4, no. 6, pp. 1778–1787, 2025.

[12] S. Anwar, F. Bawazir, R. Sakina, M. Lukita, N. Hernata, M. Miranda, and M. Ridwan, "Mazhab Syafi'i sebagai Paradigma dalam Pemikiran dan Penetapan Hukum Islam Di Indonesia," Varia Hukum, vol. 5, no. 2, pp. 101–123, 2023.

[13] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," *18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024)*, pp. 1-16, 2024.

[14] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 Text Embeddings: A Technical Report," *arXiv preprint arXiv:2402*.05672, 2024.

[15] Chroma Core Team, "Chroma: The AI-native Open-source Embedding Database," Official Documentation, 2024. [Online]. Available: https://docs.trychroma.com.

[16] Mozilla Developer Network, "Progressive Web Apps (PWA)," MDN Web Docs, 2024. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/Progressive_web_apps.

[17] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven Failure Points When Engineering A Retrieval Augmented Generation System" *in Proceedings - 2024 IEEE/ACM 3rd International Conference on AI Engineering*, 2024.

[18] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," *35th Conference on Neural Information Processing Systems (NeurIPS 2021),* 2021.

[19] Qwen Team, "Qwen2.5 Technical Report," *arXiv preprint arXiv:2412.*15221, 2024.

[20] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, "When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 9802-9822, 2023.

[21] P. Xu, W. Ping, X. Wu, L. McAfee, C. Zhu, Z. Liu, S. Subramanian, E. Bakhturina, M. Shoeybi, and B. Catanzaro, "Retrieval Meets Long Context Large Language Models," *12th International Conference on Learning Representations*, ICLR 2024, 1–16, 2024.

[22] W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, and W. Yih, "REPLUG: Retrieval-Augmented Black-Box Language Models," *arXiv preprint arXiv:2301*.12652, 2023.

[23] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," *arXiv preprint arXiv:2310*.11511, 2023.

[24] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Few-shot Learning with Retrieval Augmented Language Models," *Journal of Machine Learning Research*, vol. 24, pp. 1-43, 2023.

[25] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Bhowmik, and S. Nanayakkara, "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1-17, 2023.