SINTECHCOM: Science, Technology and Communication Journal
Vol. 6, No. 3, June 2026, pp. 339-354, DOI: 10.59190/stc.v6i3.392

Implementation of the mawaris fiqh hybrid chatbot based on retrieval-augmented generation and rule-based expert system

Irpan Afrizal Putra Eriani, Nazruddin Safaat Harahap*, Suwanto Sanjaya, Muhammad Irsyad
Department of Informatics Engineering, UIN Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia

ABSTRACT

Islamic inheritance law (mawaris fiqh) regulates the distribution of inheritance based on the Quran, Sunnah, and ijma'. However, many people still have difficulty in understanding the concept of inheritance and performing accurate inheritance calculations due to the complexity of faraidh rules and limited sources of information about faraidh. This study aims to develop a hybrid-based mawaris chatbot that integrates retrieval-augmented generation (RAG) and rule-based expert system to support both conceptual question answering and deterministic inheritance calculations. This system is implemented using the Voyage-3-Large embedding model, Qdrant vector database, semantic caching, large language models (LLM) for contextual response generation using models from GPT-4o (main) and llama3.2:3b (fallback mode) as well as semantic cache using paraphrase-multilingual-MiniLM-L12-v2. The "Ask Concept" answering mode uses semantic search, confidence router, and RAG, while the "Calculate Inheritance" answering mode uses a rule-based expert system for heir identification, validation, faraidh calculation, and division result preparation. The system performance is evaluated for conceptual questions using BERTScore and weighted scoring model (WSM) for inheritance calculation questions. Experimental results show that the conceptual question-answering mode achieves a pass rate of 91.3% on questions in that domain. For inheritance calculation, the RAG-based approach achieves an average score of 44%, while the rule-based expert system achieves 100% in all evaluation categories. These findings indicate that the proposed hybrid architecture effectively combines the contextual reasoning capabilities of RAG with the deterministic accuracy of rule-based calculation, making it a reliable solution for mawaris consultation and inheritance distribution assistance.

ARTICLE INFO

Article history:

Received Jun 23, 2026
Revised Jun 24, 2026
Accepted Jun 25, 2026

Keywords:

Hybrid Chatbot
LLM
Mawaris Fiqh
RAG
Rule-Based Expert System

* **Corresponding Author**
E-mail address: nazruddin.safaat@uin-suska.ac.id

1. INTRODUCTION

Mawaris Fiqh is a branch of Islamic law that regulates the systematic distribution of inheritance based on the Qur'an, Sunnah, and ijma'. These rules aim to ensure justice and avoid disputes between heirs [1]. However, in practice, many people still have difficulty understanding the concept and calculation of inheritance distribution according to the rules of Mawaris Fiqh. As a result, errors still occur in society in implementing it [2]. One of the causes of this problem is the lack of information sources about Mawaris Fiqh [3].

Technological developments are currently experiencing rapid progress, allowing for easy and rapid dissemination of information [4]. One rapidly developing technology is artificial intelligence, which has driven increased rapid information delivery through the implementation of chatbots.

Chatbots are one implementation of a Question Answering System (QA System) capable of quickly finding, extracting, and presenting specific answers to user questions using natural language [5]. However, chatbots that use only one approach still have limitations in providing answers.

A rule-based expert system chatbot is one approach that represents an expert's knowledge in the form of explicit logical rules (if-then rules). This system works by matching input conditions against a set of rules predetermined by the expert to generate deterministic and consistent decisions or answers [6]. But, the answers generated using this approach can sometimes make it difficult for users to obtain answers that require more in-depth explanations. On the other hand, a RAG-based chatbot is able to understand and generate natural language contextually, allowing the chatbot system to answer user questions with more in-depth reasoning and explanations [7]. However, a RAG-based chatbot system cannot guarantee absolute accuracy and precision, as the RAG approach sometimes produces inaccurate and contextually irrelevant responses [8].

Therefore, the Mawaris fiqh chatbot system implements a hybrid approach. A hybrid chatbot is a chatbot system that integrates a rule-based approach and a RAG approach, or generative model, to handle various types of user questions more effectively. This approach is very useful for addressing the shortcomings or limitations of both rule-based and RAG-based chatbots [9]. Therefore, it has three main components: a Rule-Based Expert System to perform deterministic inheritance calculations, a Large Language Model (LLM) to generate natural and contextual answers, and Retrieval-Augmented Generation (RAG) to optimize the chatbot's performance in retrieving conceptual information from documents. This helps the LLM understand and answer user questions and reduces the occurrence of hallucinations [10].

Based on the above problems, this study aims to develop a hybrid-based Mawaris fiqh chatbot system that integrates the Rule-Based Expert System and RAG approaches. The developed system is designed to be able to answer conceptual questions related to Mawaris fiqh contextually, as well as calculate inheritance distribution accurately and deterministically in accordance with the rules of faraidh science. In addition, this study also aims to improve the quality of chatbot answers by utilizing a semantic search-based retrieval mechanism, as well as optimizing system performance in reducing answer errors such as hallucinations in the LLM model.

2. RESEARCH METHODS

The research methodology consists of a series of systematically designed steps used as a guideline to achieve the research objectives. Figure 1 illustrates the research stages discussed in this chapter.

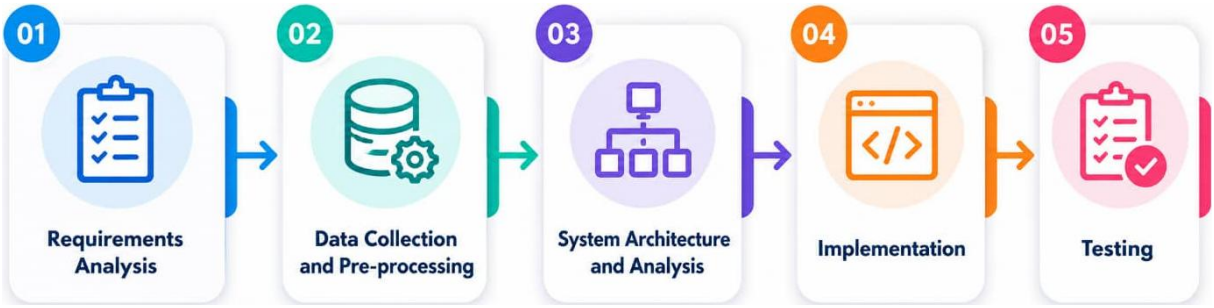

Figure 1. Research stages.

2.1. Requirements Analysis

This research methodology begins with a needs analysis to address the challenges faced by the community regarding the understanding of Mawaris fiqh. The identified challenges include limited access to authoritative sources of Mawaris fiqh, the complexity of inheritance calculations that often lead to errors, and the lack of a Mawaris fiqh question-and-answer system capable of combining deterministic inheritance calculations with the presentation of clear conceptual answers. Therefore, this methodology was developed using a hybrid approach that integrates a Rule-Based Expert System

to ensure the accuracy of asset distribution according to faraidh laws, and a RAG to ensure that each conceptual explanation is based on the fiqh literature stored in a vector database.

2.2. Data Collection and Pre-Processing

This stage is the main foundation for the RAG component of the system. The quality of answers in the "Ask Concept" mode depends heavily on how the textual data is processed and mapped into the vector database. The following is a breakdown of the data preprocessing process into a vector database, as seen in Figure 2.

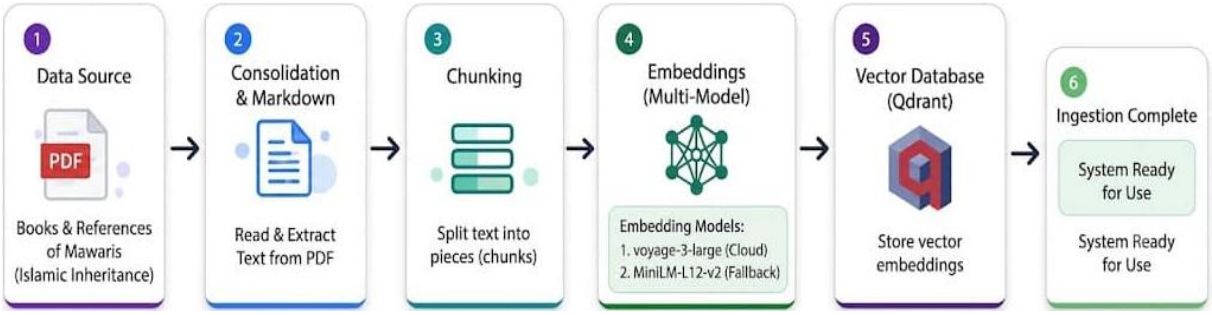

Figure 2. Data collection and pre-processing stages.

2.2.1. Data Source

The primary data used in this study are digital documents in PDF format containing comprehensive literature on the science of mawaris [11-13].

2.2.2. Consolidation and Markdown Process

In this process, the system merges and extracts all collected PDF files using the PyMuPDF4LLM library. This library is layout-aware, capable of preserving complex document semantic structures, such as chapter heading hierarchies (#,##,###), tables of heirs, and Arabic text, into a structured Markdown format. The resulting Markdown format is highly optimized for integration with LLM and RAG systems [14].

2.2.3. Chunking and Embeddings Process

The extracted clean text into Markdown files is then broken down into smaller pieces (chunks) and converted into vector representations (embedding). This embedding process uses a model from Voyage AI (voyage-3-large). This model was selected based on its capabilities as a state-of-the-art embedding model that excels in semantic similarity and information retrieval tasks [15, 16].

2.2.4. Save to Vector Database

The embedding vectors containing the original text and metadata are then saved to the Qdrant vector database. Qdrant was selected as the primary vector database based on its advantages, including its native hybrid database design, specifically designed to handle high-dimensional vector data up to 65,536 dimensions. It supports indexing methods such as Hierarchical Navigable Small World (HNSW) and Product Quantization (PQ), as well as predicated query capabilities that combine vector search with structured metadata filtering [17]. Data is uploaded to a collection named "mawaris_docs" so that the RAG component is ready to search for relevant context for each user query.

2.3. System Architecture and Analysis

The architecture and analysis of this system are designed by integrating the Rule-Based Expert System and RAG approaches to optimally handle two main service modes. Broadly speaking, the system architecture is divided into two main pipelines that work according to the type of user request, namely the RAG pipeline and cache for the "Ask Concept" mode and the Rule-Based pipeline for the "Calculate Inheritance" mode. These two pipelines are supported by shared components such as the

embedding model, the Qdrant vector database, and the semantic caching mechanism. The overall system analysis, including the data flow and its constituent components.

2.3.1. Ingest Data

This data ingestion stage has been explained in sub-chapter 2.2. Data Collection and Pre-Processing.

2.3.2. RAG Pipeline and Cache

This pipeline is designed to provide conceptual answers by combining retrieval and generative capabilities, and is equipped with a cache mechanism for efficiency. The following are the steps in the RAG and cache pipeline flow, as seen in Figure 3.

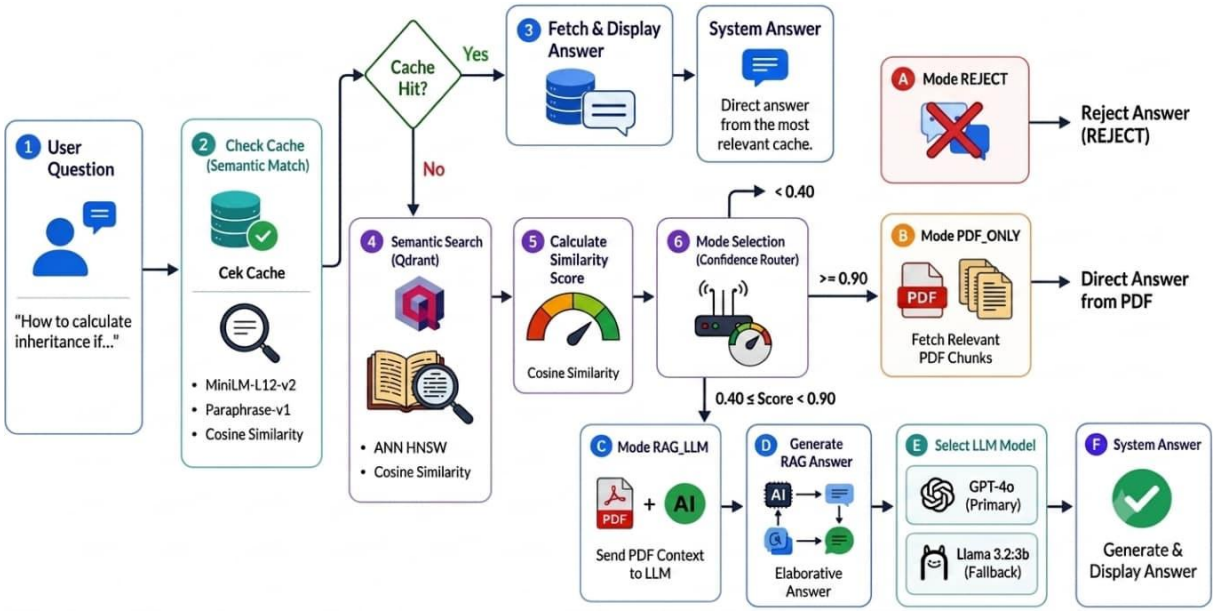

Figure 3. RAG pipeline and cache.

2.3.2.1. Cache Checking with Semantic Match

Before searching directly in the vector database, the system first checks the cache, which stores question-answer pairs from previous interactions. This mechanism functions to answer questions with previously provided answers. The purpose of using this cache is to increase the efficiency of AI model token usage and the time it takes to provide answers. This mechanism utilizes semantic matching using the MiniLM-L12-v2 multilingual paraphrase embedding model to represent questions and calculate similarity using cosine similarity [18]. The use of this model is based on its ability to handle multilingual text, including Indonesian, and is effective in capturing semantic similarity [19].

2.3.2.2. Semantic Search to Vector Database (Qdrant)

If no adequate match is found in the cache, the system proceeds to the semantic search stage in Qdrant. At this stage, the user's question is converted into a vector representation and then a similarity search is performed against the vectors of text chunks from the Mawaris fiqh book stored in Qdrant. The semantic search process in Qdrant is performed using the Approximate Nearest Neighbor Search (ANN) algorithm with the HNSW library, as described in research [20], where the cosine similarity between the query vector and the document vector is calculated to find the most relevant text snippet.

2.3.2.3. Determining the Answer Mode Based on Similarity Score

After the semantic search process in Qdrant, the system calculates the highest similarity score using cosine similarity to measure the semantic closeness between the user's query and the found text

snippet, as implemented in research [21]. Based on this score, the confidence router determines three answer modes: REJECT, PDF_ONLY, and RAG_LLM. In REJECT mode (score <0.40), the system rejects the answer because the question is deemed outside the scope of knowledge, with a threshold determined based on the threshold optimization principle by [18] to maintain a balance between precision and recall. In PDF_ONLY mode (score ≥ 0.90), the answer is provided directly from the retrieved text snippet without involving LLM. Meanwhile, in RAG_LLM Mode (score 0.40 to < 0.90), the system sends the most relevant text snippets as context to the GPT-4o model as the main LLM to generate more elaborate and contextual answers according to the RAG paradigm. The selection of GPT-4o is based on its ability to understand context and generate optimal answers, as evidenced by research [22, 23] that conducted performance tests of the GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro models. In addition, the system also uses the local Ollama model (llama3.2:3b) as a fallback to maintain service continuity when the main model is unavailable or experiencing limitations. The use of llama3.2:3b is supported by research [24] which shows that the model still has good performance in the RAG system.

2.3.3. Rule-Based Pipeline

The inheritance calculation pipeline is a computational flow that processes user natural language input into accurate and deterministic inheritance distribution details. This pipeline implements a Rule-Based Expert System as the core algorithm, which mimics the reasoning of a faraidh expert through a series of strict if-then rules, as implemented in previous research [3]. Broadly speaking, this flow is divided into six stages, as can be seen in Figure 4.

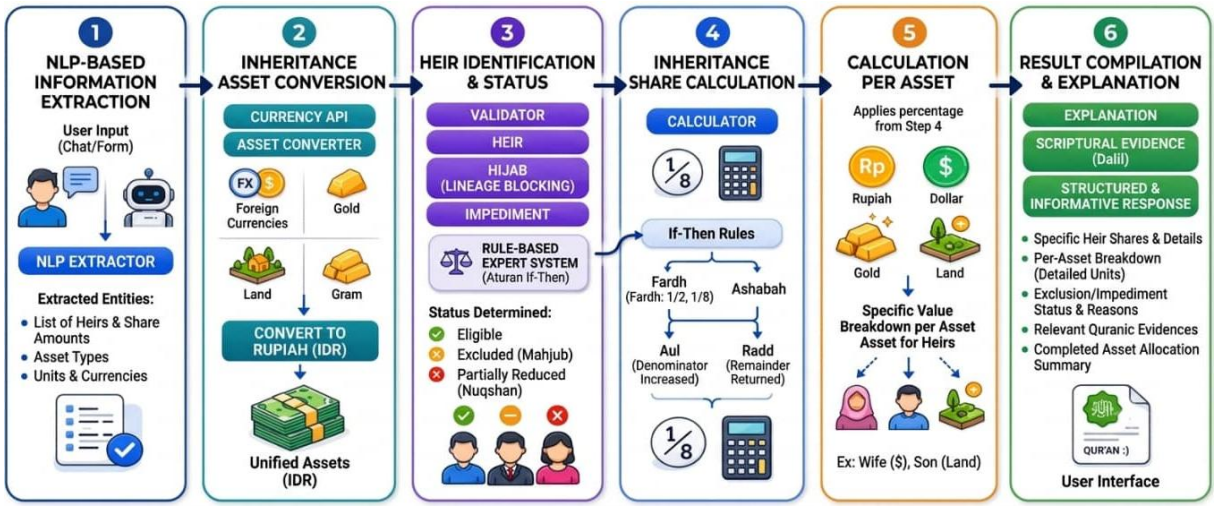

Figure 4. Rule-based pipeline.

2.3.3.1. NLP-Based Information Extraction

The inheritance calculation process begins when a user submits a query regarding an inheritance case through either the chat interface or the structured input form. To accommodate both conversational and structured inputs, the system employs an NLP Extractor module that automatically analyzes the submitted information and identifies all entities relevant to inheritance calculations. This module extracts the list of heirs along with their quantities, identifies various inheritance asset categories such as cash, gold, land, and foreign currencies, and records the associated values, units, and currency denominations. For example, when a user submits an inheritance case containing information about heirs, monetary assets, land ownership, and precious metals, the NLP Extractor transforms the natural language input into a structured representation that can be processed by subsequent modules. This structured data serves as the foundation for all subsequent inheritance calculations.

2.3.3.2. Inheritance Asset Conversion

After the required information has been extracted, the system standardizes all inheritance assets into a unified monetary representation to ensure accurate proportional calculations. This process is handled by the Asset Converter and Currency API modules. Foreign currencies are converted into Indonesian Rupiah (IDR) using the latest available exchange rates obtained through the Currency API. Meanwhile, non-cash assets such as gold and land are converted into their equivalent monetary values based on predefined valuation rules and market prices. By transforming all inheritance assets into a common valuation standard, the system establishes a unified asset pool that accurately represents the total estate value prior to inheritance distribution. This normalization process ensures consistency and prevents discrepancies during the calculation stage.

2.3.3.3. Heir Identification & Status (Rule-Based Expert System)

Once all heirs have been identified, the Rule-Based Expert System evaluates their legal status according to Islamic inheritance law. Using a collection of deterministic if-then rules derived from classical faraidh principles, the system determines whether each heir is eligible to inherit, excluded from inheritance, or subject to a reduced inheritance share. Heirs categorized under (Hijab Hirman) are completely excluded from inheritance due to the presence of closer relatives who possess a stronger inheritance claim. In contrast, heirs affected by (Hijab Nuqshan) remain eligible to inherit but receive a reduced portion because specific inheritance conditions are present. Through this evaluation process, the system establishes a complete inheritance eligibility profile for every heir before proceeding to the distribution stage.

2.3.3.4. Inheritance Share Calculation

Following the determination of heir eligibility, the Calculator module computes the inheritance shares for each qualified heir according to Islamic inheritance principles. The system first allocates the fixed Quranic shares known as (Fardh), which include prescribed fractions such as one-half, one-third, one-fourth, one-sixth, and one-eighth. After all fixed shares have been distributed, the remaining estate is allocated to residuary heirs under the concept of (Ashabah). The calculation process also incorporates advanced inheritance mechanisms such as (Aul), which adjusts the denominator when the total prescribed shares exceed the available estate, and (Radd), which redistributes any remaining assets among eligible heirs when no residuary heirs exist. These calculations ensure that the inheritance distribution remains fully compliant with established Islamic inheritance regulations.

2.3.3.5. Calculation Per Asset

Although inheritance shares are initially calculated using a unified monetary representation, the system subsequently applies each heir's calculated proportion to every individual asset category included in the estate. This process enables the system to generate detailed asset-level distributions rather than presenting only aggregate monetary values. As a result, heirs can clearly see their entitlement in terms of specific assets, including cash holdings, foreign currencies, gold ownership, and land assets. By performing inheritance calculations at the asset level, the system provides greater transparency and facilitates practical implementation of inheritance distribution in real-world scenarios.

2.3.3.6. Result Compilation & Explanation

In the final stage, the system compiles all calculated results into a comprehensive and user-friendly response. The generated output includes detailed inheritance allocations for each heir across all asset categories, explanations regarding heirs affected by (Hijab Hirman) and (Hijab Nuqshan), and relevant scriptural evidence obtained from the (Dalil) module. The system also generates a narrative summary describing how the inheritance was distributed and confirming that the entire estate has been allocated according to Islamic inheritance law. By combining numerical calculations with explanatory information and supporting religious references, the system enables users to understand not only the final inheritance distribution but also the legal reasoning behind each decision.

2.4. Implementation

The Mawaris fiqh chatbot was implemented by integrating Retrieval-Augmented Generation (RAG) and a Rule-Based Expert System into a web-based application designed to support both conceptual consultation and inheritance calculation. The user interface was developed using React, Vite, and CSS, providing two interaction modes: a chat mode for open-ended conceptual questions and a form mode for structured inheritance calculation requests.

On the backend side, the system was developed in Python and consists of two primary processing pipelines. The first pipeline is the RAG module, which is responsible for answering conceptual questions related to Islamic inheritance law. This module integrates Qdrant as the vector database, Voyage-3-Large as the primary embedding model, and Paraphrase-Multilingual-MiniLM-L12-v2 as a local embedding model for semantic representation. Retrieved knowledge is then processed by a configurable LLM, which can operate either through a locally deployed Ollama instance using Llama 3.2:3B or through the cloud-based OpenAI API using GPT-4o. To improve response efficiency and reduce redundant processing, a semantic caching mechanism is employed to store and reuse answers to semantically similar queries.

The second pipeline is the Rule-Based Expert System, which handles deterministic inheritance calculations based on established principles of Islamic inheritance law (faraidh). This pipeline performs entity extraction from user input, validates heirs according to inheritance eligibility and hijab rules, normalizes inheritance assets into Indonesian Rupiah (IDR), calculates both fixed shares (fardh) and residuary shares (ashabah), and generates detailed inheritance distributions. The final output includes inheritance allocations for each eligible heir, explanations regarding excluded heirs, supporting Islamic legal references, and a comprehensive summary of the inheritance distribution process. The integration of these two pipelines enables the chatbot to provide both context-aware conceptual responses through RAG and highly accurate inheritance calculations through deterministic rule-based reasoning. This hybrid architecture combines the flexibility of LLM with the reliability and consistency required for Islamic inheritance computations.

2.5. Testing

2.5.1. BERTScore

This test aims to evaluate the chatbot system's ability to provide relevant and contextual answers to user questions in the "Ask Concept" mode. The evaluation focuses on the level of semantic correspondence between the system's answers and reference answers prepared based on the Mawaris fiqh literature. The method used is BERTScore, a contextual representation-based evaluation metric that utilizes the Transformer model to measure the similarity of meaning between texts [25]. In this study, each answer generated by the chatbot is compared with the reference answer (ground truth). Next, BERTScore calculates the level of similarity based on the embedding representation of each token. The evaluation results are expressed in three main metrics: Precision, Recall, and F1-Score.

2.5.1.1. Precision

Precision is used to measure the accuracy of the information generated by the chatbot compared to the reference answer. A high precision value indicates that most of the information provided by the system is relevant and appropriate to the context of the question. In the BERTScore evaluation, precision describes how well the tokens in the chatbot's answer can be matched to the reference answer based on similarity in meaning. The following is the precision calculation formula, shown in Equation (1).

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} max\, x_i{}^{\top}\, \hat{x}_j \quad (1)$$

2.5.1.2. Recall

Recall is used to measure the completeness of the information successfully conveyed by the chatbot. A high recall value indicates that most of the important information contained in the reference answer has been successfully captured and conveyed by the system. In BERTScore, recall indicates

the extent to which the tokens in the reference answer can be found in the answer generated by the chatbot. The following is the recall calculation formula, shown in Equation (2).

$$R_{BERT} = \frac{1}{|x|} \sum_{x_j \in x} max\, x_i^{\top}\, \hat{x}_j \quad (2)$$

2.5.1.3. F1-Score

F1-Score is a metric that combines precision and recall to provide a more comprehensive picture of system performance. This value is used to assess the balance between the accuracy and completeness of the generated answers. The higher the F1-Score, the better the chatbot's ability to generate answers that are not only relevant but also include the expected important information. The following is the F1-Score calculation formula, shown in Equation (3).

$$F_{BERT} = 2\, \frac{P_{BERT} . R_{BERT}}{P_{BERT} \cdot R_{BERT}} \quad (3)$$

BERTScore values range from 0 to 1. The closer to 1, the higher the semantic match between the chatbot's answers and the reference answers. Therefore, this metric is considered more representative for evaluating RAG and LLM-based chatbots because it considers not only word similarity but also the similarity of meaning in the answers provided.

2.5.2. Weighted Scoring Model (WSM)

The "Inheritance Calculation" mode test was evaluated using the Weighted Scoring Model (WSM), a multi-criteria decision-making method that combines multiple evaluation criteria according to their relative importance [26]. Unlike conventional accuracy measures that focus on a single aspect, the WSM allows for a more comprehensive assessment by considering multiple components involved in the inheritance calculation. This approach is particularly well-suited for evaluating inheritance systems because the accuracy of the final result depends not only on numerical calculations but also on the accurate identification of heirs, inheritance assets, and inheritance shares. The overall WSM score is calculated as the weighted sum of all evaluation criteria, as defined in Equation (4).

$$WSM = \sum_{i=1}^{n} w_i s_i \quad (4)$$

where, WSM represents the final weighted score, Wi denotes the weight assigned to criterion (i), Si represents the score obtained for criterion (i), and (n) is the total number of evaluation criteria. In this study, four evaluation criteria were used. The first criterion, Heir Detection, was weighted at 0.20 (20%) and evaluated the system's ability to correctly identify the type and number of heirs provided by the user. The second criterion, Asset Detection, was weighted at 0.10 (10%) and measured the system's ability to recognize inherited assets and their values. The third criterion, Portion Accuracy, was weighted at 0.35 (35%) and assessed the accuracy of the inheritance portion allocated to each heir in accordance with Islamic inheritance law. The fourth criterion, Numerical Accuracy, was also weighted at 0.35 (35%) and evaluated the accuracy of the numerical calculations generated by the system.

Table 1. Example of weighted scoring model (WSM) assessment mechanism.

Criterion	Weight	RAG	Rule-based expert system
Heir detection (h)	0.20	✓ (1)	✓ (1)
Asset detection (a)	0.10	✓ (1)	✓ (1)
Portion accuracy (p)	0.35	✗ (0)	✓ (1)
Numeric accuracy (n)	0.35	✗ (0)	✓ (1)
Total WSM score	1.00 (100%)	0.30 (30%)	1.00 (100%)

The weighting scheme is determined based on the relative importance of each criterion in the inheritance calculation process. Section Accuracy and Numerical Accuracy are given the highest weighting because they directly affect the validity of inheritance distribution. Meanwhile, Heirs

Detection and Asset Detection serve as supporting components that influence subsequent calculation stages. Each criterion is evaluated using a binary scoring scheme, where a score of 1 indicates successful fulfillment of the criterion and a score of 0 indicates failure. The final WSM score is obtained by multiplying each criterion score by its respective weight and summing the results. Consequently, the maximum achievable score is 1.00 (100%), indicating successful fulfillment of all evaluation criteria. Table 1 illustrates an example of WSM scoring using the inheritance case of "Inheritance value of IDR 300,000,000; heirs: husband, mother, and father," which falls into the Gharawain (Umariyyatain) category.

The following is a manual calculation of the scores assigned to the RAG approach using the example question "Inheritance value of Rp 300,000,000. Heirs: husband, mother, and father," as shown in Equation (5).

$$RAG = (1 \times 0.20) + (1 \times 0.10) + (0 \times 0.35) + (0 \times 0.35) = 0.30 \quad (5)$$

The resulting score of 0.30 (30%) indicates that the RAG approach successfully performed information extraction tasks, including heir and asset identification, but failed to generate valid inheritance shares and numerical calculations. This outcome demonstrates that while RAG is effective for understanding and extracting contextual information, it is less reliable for deterministic inheritance calculations that require strict compliance with faraidh rules. In contrast, the Rule-Based Expert System successfully satisfied all evaluation criteria. The system correctly identified the heirs, recognized the inheritance assets, determined the appropriate inheritance shares according to the Gharawain (Umariyyatain) rules, and produced accurate numerical calculations. Therefore, all criteria received a score of 1.00 (100%).

3. RESULTS AND DISCUSSIONS

This section contains the results and discussion of the procedural stages applied in the research methodology.

3.1. Results of the Data Collection and Pre-Processing Stages

3.1.1. Results of the Consolidation and Markdown Process

In this process, the system merges and extracts the three PDF data files using the PyMuPDF4LLM library. The goal of this process is to structure the entire text content of the PDF data and convert it into a Markdown file that combines the entire content of the Mawaris fiqh reference book into a single file, ready to be broken down into chunks in the next step. This process is shown in Figure 5.

Figure 5. Results of the consolidation and markdown process.

3.1.2. The Results of the Chunking and Embedding Process

The text extracted into a single Markdown file is then broken down into smaller chunks and converted into vector representations (embeddings). The results of this process can be seen in Figure 6.

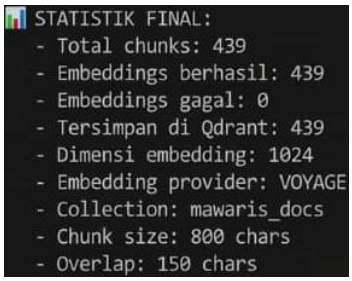

Figure 6. The results of the chunking and embedding process.

3.1.3. The Results of The Process Are Saved to The Vector Database

After pre-processing, each text chunk is converted into a vector representation using the Voyage-3-Large embedding model. The resulting embeddings are then stored in the Qdrant vector database to support semantic similarity search during the retrieval process. As seen in Figure 7, 439 vectorized document chunks have been successfully indexed and organized into a single collection named (mawaris_docs). The successful storage of all embeddings indicates that the knowledge base has been correctly constructed and is ready for retrieval operations.

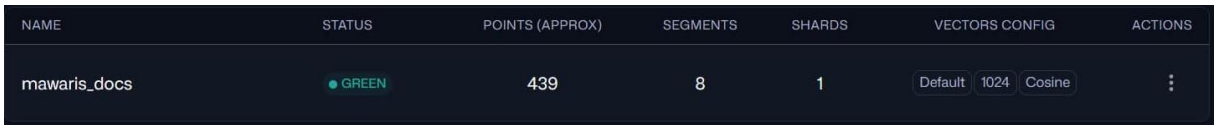

Figure 7. Results of entering data into a vector database.

3.2. Results of the System Analysis Stage

The following subsections present the implementation results of each major component and illustrate how the hybrid architecture integrates RAG and Rule-Based Expert Systems into a unified web-based application.

3.2.1. Similarity Score and Semantic Search Results

Once the system is ready to perform data ingestion or pre-processing, it is ready for use by the user. Figure 8, shows the results of determining the mode, searching the vector database, and caching the questions and answers.

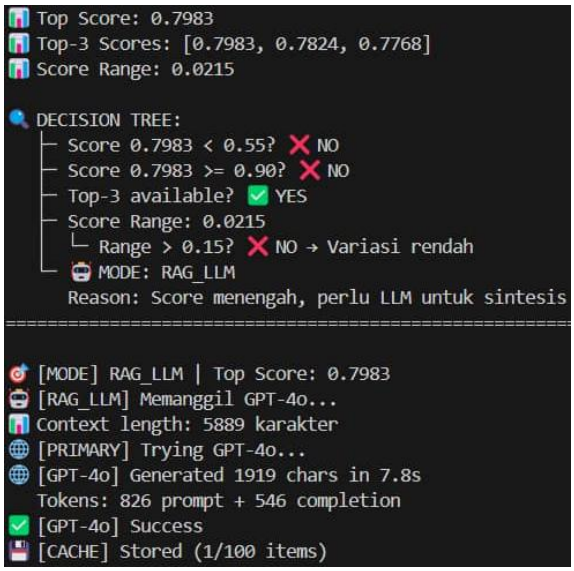

Figure 8. Similarity score and semantic search results.

3.2.2. Results of Using Semantic Cache

This semantic cache system can quickly and efficiently respond to users' previously stored questions. This process utilizes semantic matching using the MiniLM-L12-v2 multilingual paraphrase model.

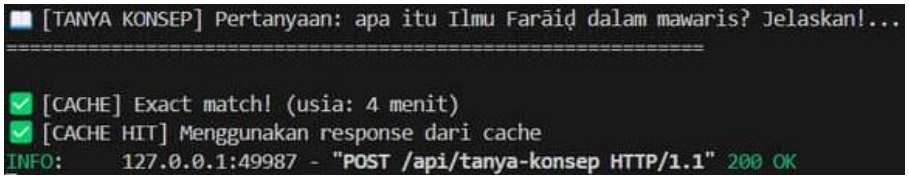

Figure 9. Results of using semantic cache.

3.2.3. Rule-Based Pipeline Results

The results of the inheritance calculation stage show how the system works in answering user questions in the "Calculate Inheritance" mode. Figure 10 shows how the system answers using the form mode in the example case "The deceased left assets in the form of cash worth $10,000 USD. Heirs: wife, 2 sons, 1 daughter".

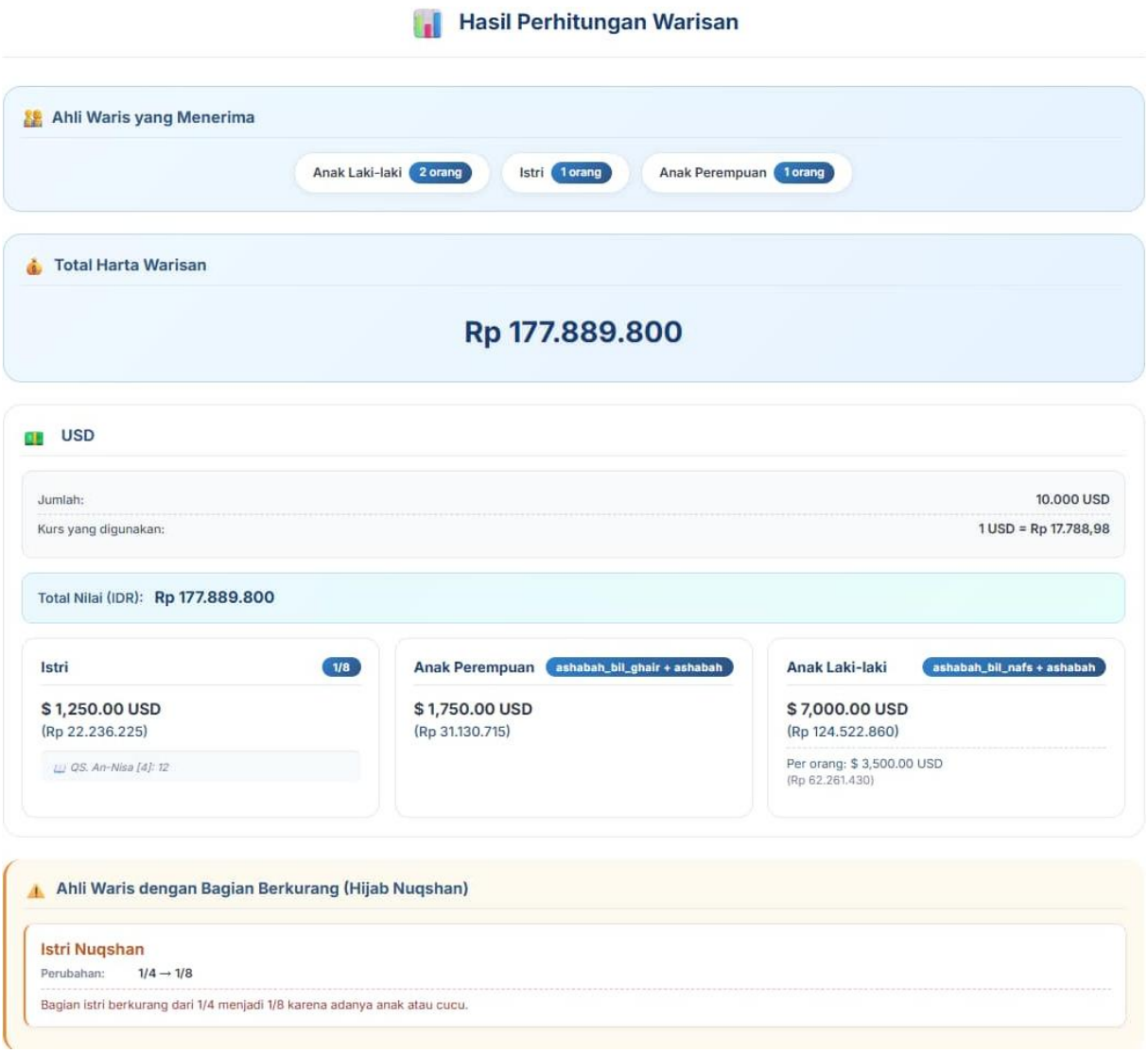

Figure 10. Rule-based pipeline results.

3.3. Results of the Implementation Stage

This implementation phase demonstrates what the Mawaris fiqh chatbot system will look like. The following are some screenshots of the chatbot system.

3.3.1. The Results of the Chat Page Display Ask About the Concept and Calculate Inheritance

This page features a chat area and a sidebar containing the main menu, chat actions (delete and download chat features), and information. The resulting page display can be seen in Figure 11.

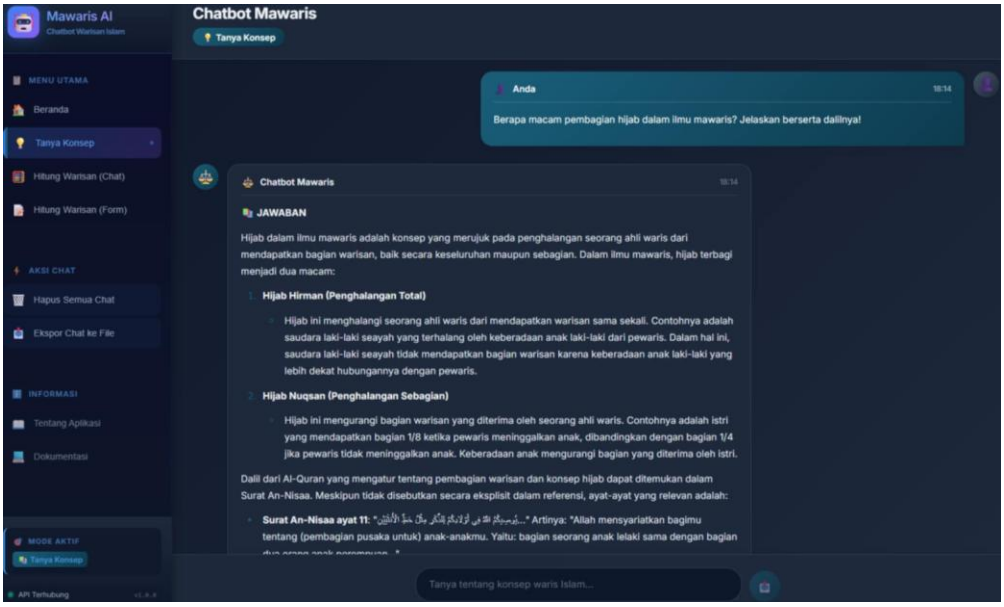

Figure 11. Chat page display results.

3.3.2. Results of the Inheritance Calculation form Page Display

This inheritance calculation form page allows you to calculate inheritances by entering the number of heirs. The resulting page display can be seen in Figure 12.

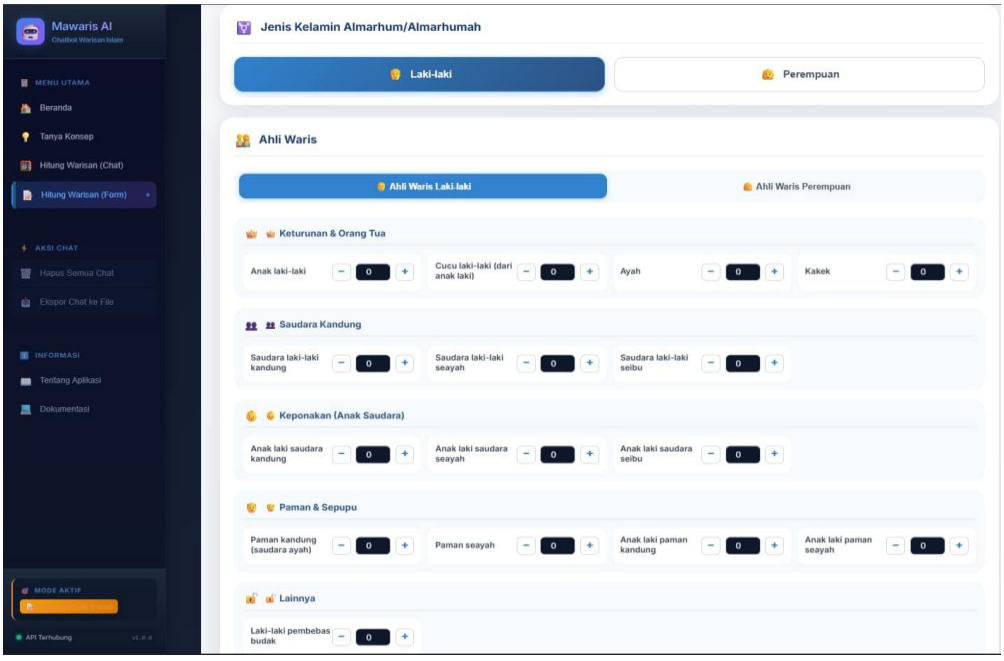

Figure 12. Inheritance calculation form display results.

3.4. Test Results

3.4.1. BERTScore

Testing was conducted to evaluate the Mawaris chatbot's ability to answer conceptual and inheritance calculation questions related to the science of faraidh. Testing in the "Ask a Concept" mode used the BERTScore method. Evaluation was carried out by comparing the system's answers to the reference answers (ground truth) using Precision, Recall, and F1-Score metrics. The test used 25 sample questions covering several material categories, namely definitions, hijab (obstructed), ashabah (remaining portion), dzawil furudh (legal heirs), pillars and conditions of inheritance, history, and questions outside the system's scope of knowledge (out-of-scope).

ID	Kategori	Pertanyaan	Precision (P)	Recall (R)	F1-Score (F1)	Mode	Status	Pass (F1 ≥ 0.5)	Response Time (detik)
Q001	definisi	Apa pengertian ilmu faraidh atau ilmu mawaris?	0.467	0.661	0.547	RAG_LLM	SUCCESS	✓	24.15
Q002	definisi	Apa saja sumber hukum ilmu faraidh?	0.406	0.585	0.479	RAG_LLM	SUCCESS	✗	9.07
Q003	hijab	Siapa saja ahli waris yang tidak pernah terhalang secara hijab hirman?	0.483	0.582	0.528	RAG_LLM	SUCCESS	✓	10.36
Q004	hijab	Apa perbedaan antara hijab nuqshan dan hijab hirman?	0.569	0.690	0.623	RAG_LLM	SUCCESS	✓	8.54
Q005	definisi	Apa pengertian ashabah secara bahasa dan istilah?	0.555	0.697	0.618	RAG_LLM	SUCCESS	✓	8.85
Q006	ashabah	Sebutkan tiga macam ashabah nasab!	0.422	0.642	0.509	RAG_LLM	SUCCESS	✓	7.92
Q007	ashabah	Berapa jumlah penerima ashabah nasab?	0.544	0.728	0.623	RAG_LLM	SUCCESS	✓	7.63
Q008	definisi	Apa yang dimaksud dengan aul dalam ilmu faraidh?	0.430	0.601	0.501	RAG_LLM	SUCCESS	✓	7.48
Q009	definisi	Apa pengertian rad dalam warisan?	0.533	0.636	0.580	RAG_LLM	SUCCESS	✓	7.71
Q010	definisi	Apa itu masalah gharrawain atau masykul?	0.492	0.601	0.541	RAG_LLM	SUCCESS	✓	6.98
Q011	sejarah	Siapa yang pertama kali menyelesaikan ilmu faraidh?	0.512	0.694	0.589	RAG_LLM	SUCCESS	✓	8.96
Q012	rukun_syarat	Apa saja tiga rukun waris?	0.520	0.656	0.580	RAG_LLM	SUCCESS	✓	6.02
Q013	tirkah	Apa saja hak-hak yang terkait dengan tirkah?	0.483	0.622	0.544	RAG_LLM	SUCCESS	✓	8.50
Q014	definisi	Apa pengertian hijab secara bahasa dan istilah?	0.499	0.616	0.552	RAG_LLM	SUCCESS	✓	7.60
Q015	dzawil_furud	Siapa saja ahli waris penerima bagian pasti (dzawil furud)?	0.437	0.659	0.525	RAG_LLM	SUCCESS	✓	7.37
Q016	dzawil_furud	Kapan suami mendapatkan bagian warisan?	0.438	0.596	0.505	RAG_LLM	SUCCESS	✓	6.69
Q017	dzawil_furud	Kapan istri mendapatkan bagian warisan?	0.441	0.584	0.502	RAG_LLM	SUCCESS	✓	7.08
Q018	dzawil_furud	Apa saja syarat anak perempuan mendapatkan warisan?	0.403	0.605	0.483	RAG_LLM	SUCCESS	✗	6.88
Q019	definisi	Jelaskan pengertian tirkah dalam ilmu faraidh!	0.531	0.636	0.579	RAG_LLM	SUCCESS	✓	7.73
Q020	definisi	Apa perbedaan antara kakek shar'i dan kakek 'asabah?	0.605	0.748	0.669	RAG_LLM	SUCCESS	✓	9.24
Q021	out_of_scope	Siapa presiden pertama Indonesia?	1.000	1.000	1.000	REJECT	REJECT_CORRECT	-	3.21
Q022	out_of_scope	Bagaimana cara membuat kue bolu?	1.000	1.000	1.000	REJECT	REJECT_CORRECT	-	2.87
Q023	definisi	Apa yang dimaksud dengan dzawil arham?	0.461	0.627	0.531	RAG_LLM	SUCCESS	✓	7.29
Q024	rukun_syarat	Sebutkan sebab-sebab mendapatkan warisan!	0.442	0.619	0.516	RAG_LLM	SUCCESS	✓	6.65
Q025	definisi	Apa itu asal masalah dalam ilmu faraidh?	0.446	0.614	0.517	RAG_LLM	SUCCESS	✓	6.53

Keterangan: Pass jika F1-Score ≥ 0.5

2. KESIMPULAN

F1-Score Rata-rata	Pass Rate (F1 ≥ 0.5)	Total Pertanyaan	SUCCESS	REJECT_CORRECT	Rata-rata Response Time
0.5496	91.3% (21/23)	25	23 (92.0%)	2 (8.0%)	8.16 detik

Figure 13. Conceptual test results of the mawaris fiqh chatbot system.

The system evaluation results are shown in Figure 13. Based on the test results, the chatbot obtained an average F1-Score of 0.5496, with a pass rate of 91.3%, or 21 of the 23 questions within the system's knowledge scope successfully reached the set pass threshold (F1-Score ≥ 0.5). This value indicates that the system is able to produce answers that have a good level of semantic fit with the reference answer. In addition, the system successfully answered all relevant questions with a SUCCESS status and was able to reject two questions outside the Mawaris fiqh domain with a REJECT_CORRECT status. This indicates that the implemented confidence router mechanism is able to distinguish between questions that are appropriate and inappropriate to the system's knowledge base. In terms of performance, the average response time obtained by the system was 8.16 seconds per question. Overall, the test results show that the Mawaris chatbot is able to provide relevant answers and has a good ability to detect questions outside the scope of its knowledge.

3.4.2. Weighted Scoring Model (WSM)

Testing related to inheritance calculation using the WSM method by testing 15 questions with question categories consisting of 5 types of categories.

Table 2. Summary of inheritance calculation method test results.

Question category	Total questions	Average RAG score	Average rule-based score
Fardh, Hijab Hirman, Nuqshan, Ashabah	3	76.67%	100%
Gharawain / Umariyyatain	3	30%	100%
'Aul	3	30%	100%
Radd	3	30%	100%
Multi asset	3	53.33%	100%
Overall average	15	44%	100%

Based on Table 2, the Rule-Based Expert System approach obtained a score of 100% in all test categories. These results indicate that the system is able to consistently perform heir detection, asset identification, inheritance share determination, and numerical calculations in accordance with the rules of faraidh science, including in special cases such as Gharawain (Umariyyatain), 'Aul, Radd, and Multi-Asset scenarios involving several types of assets and currency conversion. The consistency of these results indicates that the rule-based approach is able to produce decisions that are deterministic, stable, and in accordance with the established rules. In contrast, the RAG approach obtained an average overall score of 44%, with varying performance in each test category. In the Fardh, Hijab Hirman, Nuqshan, and Ashabah categories, the RAG approach still showed relatively good performance with an average score of 76.67%, and was able to achieve a perfect score in several specific scenarios. However, the system's performance experienced a significant decline in the Gharawain, 'Aul, and Radd categories, with an average score of only 30%. This shows that although the model is able to detect heirs and assets quite well, the LLM-based approach still has difficulty in determining inheritance shares that involve more complex faraidh rules and require high consistency of calculation logic. In the Multi-Asset category, the system obtained an average score of 53.33%, which indicates that managing various types of assets and converting property values can still cause inaccuracies in the aspects of determining shares and numerical calculations.

The results of this test demonstrate that the use of the RAG approach as the sole method in the Mawaris chatbot is not yet able to guarantee the required level of accuracy in the inheritance calculation process. On the other hand, the use of a Rule-Based Expert System alone also has limitations in producing flexible, natural, and contextual conceptual answers for users. Therefore, the results of this test further reinforce the importance of implementing a hybrid architecture, namely by utilizing RAG to handle conceptual questions regarding Mawaris fiqh, as well as a Rule-Based Expert System for the inheritance calculation process to be carried out accurately, consistently, and in accordance with the principles of faraidh science.

4. CONCLUSION

Based on the research conducted, a hybrid implementation of the Mawaris fiqh chatbot based on Retrieval-Augmented Generation (RAG) and a Rule-Based Expert System was successfully developed to address two primary user needs: answering conceptual questions related to the science of faraidh and calculating inheritance distribution accurately and deterministically. The hybrid approach was implemented by utilizing RAG and a Large Language Model (LLM) in the "Ask Concept" mode, and a Rule-Based Expert System in the "Calculate Inheritance" mode.

In the "Ask Concept" mode, the system successfully implemented semantic search, semantic cache, and confidence router mechanisms to determine answer modes: REJECT, PDF_ONLY, and RAG_LLM based on similarity scores. Evaluation results using the BERTScore method showed that the system was able to generate answers with a good level of semantic match to reference answers, with a pass rate of 91.3% for questions within the system's knowledge base. Furthermore, the system was also able to accurately reject questions outside the Mawaris fiqh domain, thereby helping to reduce the potential for irrelevant answers and hallucinations.

In the "Calculate Inheritance" mode, testing using the Weighted Scoring Model (WSM) method demonstrated a significant performance difference between the RAG approach and the Rule-Based Expert System. The test results showed that the RAG approach achieved an average score of 44%, indicating that the RAG-based approach still faces limitations in handling inheritance calculations that require logical consistency, faraidh portion determination, and high numerical

precision, particularly in the Gharawain, 'Aul, Radd, and Multi-Asset scenarios. Conversely, the Rule-Based Expert System approach achieved the best score of 100% in all test categories, demonstrating its ability to detect heirs, identify assets, determine inheritance portions, and perform deterministic numerical calculations in accordance with the principles of faraidh science.

These research results demonstrate that building a Mawaris chatbot system capable of answering conceptual questions and accurately calculating inheritance assets is not sufficient to rely solely on a single chatbot approach. The use of RAG alone has not been able to guarantee the accuracy of calculations required in the faraidh domain, while the use of Rule-Based Expert System alone has limitations in producing natural, contextual, and flexible conceptual answers for users. Therefore, a hybrid chatbot approach that combines RAG and Rule-Based Expert System has proven to be a more effective solution, where each approach complements each other according to the characteristics of the task being handled. Thus, this study shows that the implementation of a hybrid chatbot is able to produce a relevant, accurate, stable, and appropriate fiqh mawaris chatbot system to help the community in understanding the concept and calculating inheritance distribution based on the rules of faraidh science.

ACKNOWLEDGMENTS

The authors would like to thank Sultan Syarif Kasim State Islamic University, Riau, Indonesia, for their academic support during the implementation of this research. Appreciation is also extended to the supervisors, examiners, co-researchers, anonymous reviewers, and the editorial team for their constructive input and suggestions that helped improve the quality of this article.

REFERENCES

[1] Aldia, & Ghafurb, A. (2025). Analisis Fiqh Mawaris Islam: Status Janin sebagai Ahli Waris dan Kasus Kematian Massal (al-Halāk al-'Ām). *Journal of Religion and Social Community*, **02**(2), 2025.

[2] Bahrah, M. (2022). Ulumuddin: Jurnal Ilmu-Ilmu Keislaman Urgensi Ilmu Mawaris Dan Hukum Penerapannya Dalam Praktik Kewarisan Islam. *Ulumuddin: Jurnal Ilmu-Ilmu Keislaman*, **12**, 79–94.

[3] Irfan, M., Lustinasari, K., & Zulfikar, W. B. (2025). Optimalisasi Layanan Konsultasi Fiqih Mawaris Berbasis Chatbot dengan Pendekatan Rule-Based. *TELKA - Telekomunikasi Elektronika Komputasi Dan Kontrol*, **11**(3), 284–292.

[4] Afriani, E., Safaat H, N., Fikry, M., & Affandes, M. (2024). Aplikasi Tanya Jawab Tentang Fiqih Bersuci Berbasis Web. *ZONAsi: Jurnal Sistem Informasi*, **6**(2).

[5] Nurhapiza, N., Harahap, N. S., Fikry, M., & Affandes, M. (2024). Penerapan Chatbot pada Aplikasi Web Tanya Jawab Tentang Fiqih Jual Beli Islam Menggunakan LangChain. *Journal of Computer System and Informatics (JoSYC)*, **5**(3), 548–557.

[6] Solomon, E., & Tilahun, S. L. (2024). Rule based chatbot design methods: A review. *Journal of Computational Science & Data Analytics © AASTU Press JCSDA*, **1**(1), 75–84.

[7] Alan, A. Y., Karaarslan, E., & Aydin, O. (2025). Improving LLM Reliability with RAG in Religious Question-Answering: MufassirQAS. *Turkish Journal of Engineering*, **9**(3), 544–559.

[8] Maeng, W., & Lee, J. (2021). Designing a Chatbot for Survivors of Sexual Violence: Exploratory Study for Hybrid Approach Combining Rule-based Chatbot and ML-based Chatbot. *5th Asian CHI Symposium 2021*, 160–166.

[9] Mikael, K., Oz, C., Rashid, T. A., & Nariman, G. S. (2025). A Hybrid Chatbot Model for Enhancing Administrative Support in Education: Comparative Analysis, Integration, and Optimization. *IEEE Access*, **13**, 50741–50760.

[10] Mohammed, M. Y., Ali, S. A., Ali, S. K., Majeed, A. A., & Mohamed, E. H. (2025). Aftina: enhancing stability and preventing hallucination in AI-based Islamic fatwa generation using LLMs and RAG. *Neural Computing and Applications*, **37**(25), 20957–20982.

[11] Dewi, D., Abdul, K., Alhamdani, K., Swasti, I., Bhakti, G., Nurul, O., Mardhatillah, B., Kasiani, A., Sholihah, H., Abdul, H., Farhan, H., Marlisa, A., Yudi, E., Mega, W., Mukhammad, A. N., Hadi, N., Solihah, C., Sahala, H., Sinaga, R., & Rifai, A. (2024). Hukum Kewarisan Islam.

[12] Muhibbussabry, L. M. (2020). Fikih Mawaris.

[13] Dr. Nofiardi, M. A. (2023). Hukum Kewarisan Islam Antara Teori & Praktek.
[14] Cirillo, L., Gotelli, M., Massei, M., Sina, X., & Solina, V. (2025). A Synergistic Multi-Agent Framework for Resilient and Traceable Operational Scheduling from Unstructured Knowledge. *AI (Switzerland)*, ***6***(12).
[15] Chen, E., Luo, L., Gunturkun, F., Sambara, S., Arora, R., Tom Jin, B., Rajpurkar, P., & Kim, D. A. (2026). Evaluation of Large Language Models as Emergency Department Revisit Predictors. *Biocomputing*, 130–143.
[16] Butler, U., Butler, A.-R., & Malec, A. L. (2025). *The Massive Legal Embedding Benchmark (MLEB)*.
[17] Wang, S., Zhao, Y., Xie, Y., Liu, Z., Hou, X., Zou, Q., & Wang, H. (2025). *Towards Reliable Vector Database Management Systems: A Software Testing Roadmap for 2030*.
[18] Holis, R. M., Utomo, P. E. P., & Hutabarat, B. F. (2025). Semantic FAQ Chatbot Using SBERT (Sentence-BERT) and Cosine Similarity for Academic Services. *Brilliance: Research of Artificial Intelligence*, **5**(2), 915–922.
[19] Suharyadi, & Saputra, I. (2025). Hybrid Ensemble Retrieval-Augmented Generation for Indonesian Legal Consultation with Keyword Boosting. *Journal of Novel Engineering Science and Technology*, **4**(02), 71–85.
[20] 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. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, **6**(1), 70–83.
[21] Setiawan, G. H., & I Made, A. B. (2023). Improving Helpdesk Chatbot Performance with Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity Models. *Journal of Applied Informatics and Computing (JAIC)*, **7**(2), 252.
[22] Liu, M., Okuhara, T., Dai, Z., Huang, W., Gu, L., Okada, H., Furukawa, E., & Kiuchi, T. (2025). Evaluating the Effectiveness of advanced large language models in medical Knowledge: A Comparative study using Japanese national medical examination. *International Journal of Medical Informatics*, **193**.
[23] Sonoda, Y., Kurokawa, R., Nakamura, Y., Kanzawa, J., Kurokawa, M., Ohizumi, Y., Gonoi, W., & Abe, O. (2024). Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. *Japanese Journal of Radiology*, **42**(11), 1231–1235.
[24] Khalila, Z., Nasution, A. H., Monika, W., Onan, A., Murakami, Y., Radi, Y. B. I., & Osmani, N. M. (2025). Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models. *(IJACSA) International Journal of Advanced Computer Science and Applications*, **16**.
[25] Maulana, M. R., Harahap, N. S., Okfalisa, O., & Yusra, Y. (2025). Implementasi Chatbot Tafsir Al-Qur'an Menggunakan Chainlit dengan Pendekatan Groq. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, **5**(3), 920–929.
[26] Putri, L. U., Hutahaean, J., Hutagalung, J. E., Amin, M., & Azhar, Z. (2025). Metode Weighted Scoring Model Dalam Pemilihan Karyawan Terbaik di Central Busana Kisaran. In *Prosiding Seminar Nasional Teknologi Komputer dan Sains*, **3**(1).