JNEST EISSN: 2961-8738 ● PISSN: 2961-8916 ● DOI: 10.56741/jnest.v4i02.1042
Article Received 25 May 2025, Revised 06 June 2025

Hybrid Ensemble Retrieval-Augmented Generation for Indonesian Legal Consultation with Keyword Boosting

Suharyadi*, Irwansyah Saputra

***Abstract*—This study presents the design and evaluation of a fully local, hybrid ensemble Retrieval-Augmented Generation (RAG) system tailored for Indonesian legal consultation. By integrating sparse (BM25), dense (FAISS), and keyword-aware retrieval mechanisms, the system balances lexical, semantic, and domain-specific relevance to retrieve high-quality legal context. A curated dataset of 8,450 legal consultation articles was scraped from a trusted legal platform, cleaned through multi-stage pre-processing, and indexed for efficient retrieval. Retrieved documents are formatted into structured prompts and fed into locally hosted large language models (LLMs) using Ollama, allowing for complete offline operation. Experiments comparing five retrieval configurations TF-IDF, BM25, FAISS, ensemble BM25+FAISS, and ensemble with keyword boosting demonstrate that the hybrid ensemble with keyword boosting yields the most relevant and grounded answers. Both quantitative (retrieval score analysis) and qualitative (manual relevance rating) evaluations were performed, confirming the effectiveness of the ensemble strategy in improving answer quality. Additionally, the system achieves practical response times (12–20 seconds) on consumer-grade hardware without reliance on cloud services. This work makes a novel contribution by demonstrating that a hybrid ensemble retrieval framework, specifically tuned to the linguistic characteristics and retrieval challenges of Indonesian legal texts, can significantly enhance the performance of local RAG-based legal QA systems. Future directions include real-time indexing, fine-tuning of legal-domain LLMs, and extending the system to support other legal domains such as statutory law, regulations, and court rulings.**

***Index Terms*—Retrieval-Augmented Generation, RAG, Hybrid Ensemble Retrieval, Indonesian Legal QA System, Information Retrieval, FAISS, BM25, TF-IDF**

I. Introduction

THE advancement of legal information access by incorporating large language models (LLMs) into natural language question answering (QA) systems has made considerable advancements in recent years. Retrieval-Augmented Generation (RAG) frameworks, which integrate dense or sparse retrieval along with generation by a language model, have emerged as a compelling paradigm to provide accurate contextual answers [1]. Despite their efficacy, most of the existing RAG-based legal QA systems are either developed for English corpora or they rely on paid cloud APIs, which limits their use in the legal consultation domain and language-constrained contexts, such as Indonesia.

Legal consultations in Indonesia involve highly structured but also nuanced interpretations of Laws, regulations, and precedents [2]. There are unique challenges in building QA systems for legal consultations. First, ensuring the accurate retrieval of legal knowledge from structured datasets. Second there is limited annotated corpora of Bahasa Indonesia for the legal domain. Third, inference must run in its entirety on local hardware because of privacy, regulation, or infrastructure limitations.

To address these challenges, we propose a Hybrid Ensemble RAG system specifically designed for Indonesian legal consultation. Our system has a composition of various retrievals, including Best Matching 25 (BM25) as traditional keyword matching, Facebook Artificial Intelligence Similarity Search (FAISS) as semantic similarity, and boosting keywords based on professional terminology matching weights [3]. This combination of retrievals will allow the system to account for relevancy not only in literal terms but also contextual aptitude or legal drawing to the documents recovered.

To our knowledge, there is no prior work that systematically integrates BM25, FAISS, and domain-specific keyword boosting in a RAG framework for Indonesian legal consultation. Importantly, the linguistic and structural characteristics of Indonesian legal texts — including mixed formal-informal phrasing, evolving statutory language, legal expressions influenced by Dutch and religious law, and a lack of large annotated legal corpora — create unique retrieval challenges. In this context, a hybrid ensemble approach is particularly well-suited to balance lexical accuracy, semantic generalization, and domain salience, enabling retrieval systems to better handle the nuances of Indonesian legal questions. This work demonstrates that such a design not only improves retrieval effectiveness but also enables fully local and privacy-preserving operation without reliance on external Application Programming Interfaces (APIs).

The other feature of this system is that we are able to fully implement the system and run it completely locally using open-source components (they are free), and public LLMs (Large Language Models) like Ollama, Gemma, and Mistral, thus keeping it private and resource accessible [4]. We assess the system using both quantitative and qualitative metrics. Quantitatively, we measure retrievals through given scores and with baseline results and various non-machine learning (ML) methods like Term Frequency - Inverse Document Frequency (TF-IDF) or BM25-only, FAISS-only, etc. As for qualitative metrics, we assess if the documents are legally relevant and if the generated response flows coherently enough for usability. Finally, a public-facing interface was built to demonstrate and visualize the practical, real-world application of hybrid RAG systems in consultation-specific domains.

Suharyadi and **Irwansyah Saputra** are with the Faculty of Computer Science, Universitas Nusa Mandiri, Jakarta, Indonesia.
*Correspondence: 14240001@nusamandiri.ac.id

© 2025 IISTR

II. Related Work

The use of RAG in domain-specific question answering is becoming more popular as this model can produce good and contextually accurate answers with external document corpora. In the original RAG model [1], applied dense vector retrieval combined with generative transformers and provided promising results on knowledge-intensive challenges. A plethora of different approaches and modifications have since been proposed, including hybrid retrieval approaches, ensemble scores, and relevance-based retrieval components.

In the context of the legal domain, question answering applications often rely on rule-based or sparse retrieval approaches (e.g., BM25)[5], which may still work on keyword-rich queries, but often do not yield effective results in cases with semantic variation. Research on dense retrieval models, using sentence embeddings, has explored approaches that achieve better semantic relevance in legal text corpora [6]. However, these models often assume a well-resourced language with sufficient high-quality legal datasets, leaving incomplete approaches in low-resourced languages such as Bahasa Indonesia [7].

Hybrid retrieval methods that combine sparse and dense approaches have been shown to be effective in addressing the challenges of variance in mapping retrieval techniques that can neither bridge the gap from literal matching to semantic generalization [8]. However, existing hybrid approaches often overlook the relative value of retrieval signals from domain-specific terms in both computing acceptable retrieval accuracy and how to apply these to cases [9]. Some researchers have introduced tactics to develop domain-aware ranking or term-weighting approaches (especially in biomedical or financial domains) [10]. In the legal domain and in low-resource legal languages, there are very few studies that have applied this kind of strategy.

Additionally, much of this RAG work applied in legal contexts is not fully self-managed, as a large portion of these RAG systems rely on cloud-based APIs or proprietary datasets that are developed and hosted by other companies [10]. This can be difficult for the justice sector and citizens to implement, where there are restrictive laws or policies that limit how and when sensitive data can be used, stored, and processed in cloud settings. Our current work brings a novel, fully self-contained legal consultation system using a hybrid ensemble retrieval strategy with keyword boosting technique, specifically for the linguistic structure, terminologies, and retrieval challenges common in Indonesian legal texts [7]. Our system is innovative because it expands on hybrid retrieval systems that rely upon cloud-based APIs or generalized language model (LM), whilst ensuring reliable and privacy-oriented operation through complete local deployment with open-source models and tools.

To our knowledge, there is no other work that has systematically fused BM25, FAISS, and domain-specific keyword boosting into a unified retrieval ensemble within a RAG framework for Indonesian legal consulting. Prior research seems to focus solely on either sparse or dense retrieval, or has not demonstrated sensitivity to legal-specific terminology. Moreover, implementations in Bahasa Indonesia are nearly absent, despite the distinctive linguistic and structural characteristics of Indonesian legal documents [7]. This study fills this gap by showing that fusing complementary retrieval paradigms with targeted term weighting provides stronger relevant contextual grounding for legal question answering, and in this respect presents a novel, practical contribution to retrieval research and legal natural language processing (NLP) with multilingual text [11]. Moreover, the hybrid ensemble design is particularly suited to the Indonesian legal context, where linguistic variability, limited annotated corpora, and domain-specific legal terminology present unique retrieval challenges that cannot be effectively addressed by sparse or dense methods alone.

III. Materials and Methods

A. Dataset

The dataset used for the purposes of this study is based on legal consultation articles found in Indonesia's legal information services platform, hukumonline.com. Hukumonline [12] is one of Indonesia's largest and most trusted sources of legal information for up-to-date laws, regulations, legal opinions, and answers to legal problems posed by users. Hukumonline has a section called “Klinik Hukum” [12] where users have submitted thousands of legal questions. These questions are answered by legal professionals using Indonesian law and applicable case law.

To create the dataset, we used automated web scraping to pull the data using a custom-built Python script. This dataset was pulled from Hukumonline.com via web scraping techniques in April 26, 2025. The data contained in the web scrape adds up to 8,450 legal consultation entries obtained in a 7-hour period of automated scraping. Hukumonline [12] automated web scraper was able to successfully pull 8,450 legal consultation articles, and the articles were exported to a structured storage format as a CSV file. A dataset record contains the legal question posed by a user and the expert's written legal response.

Nonetheless, the raw data we collected was not yet ready for direct training or retrieval tasks. Many sample responses included boilerplate introductory text, such as repeated legal disclaimers or general notes regarding legal advice that was not related to the question itself. Additionally, some included promotional material, references to the platform, or advertisements for legal services embedded within the text response. Another common issue was the text merging, where terms appeared to be combined because of inconsistencies in the site’s formatting or HTML rendering.

To remedy the issue, we built a dedicated cleaning pipeline that performed regular expression-based removal of recurring disclaimers and promotional/commercial references. There were also truncation rules that removed generic opening and closing sentences that contributed to distracting textual noise without relevant legal content. We also automated whitespace correction to separate words that had inadvertently merged to their proper token boundaries. Finally, light linguistic normalization was performed to ensure clarity and consistency of grammar in the final cleaned responses.

The result is a curated dataset of Indonesian legal question-answer pairs that are not only linguistically coherent but also legally grounded, making them highly suitable for downstream retrieval and generation tasks in our RAG-based system.

The dataset used in this study was derived from publicly accessible legal consultation articles on Hukumonline.com. Prior to dataset collection, we formally requested permission from Hukumonline via email on April 21, 2025. Approval for research use of the data was granted on April 22, 2025 by representatives from Hukumonline (shavira@hukumonline.com and meisa.aulia@hukumonline.com). The dataset is used strictly for academic research purposes, with no redistribution or commercial use, and respects all applicable legal and ethical considerations regarding data use. This explicit permission ensures that the dataset collection and use in this study comply with ethical research standards and legal constraints.

B. System Architecture

The overall architecture of the proposed system follows the RAG paradigm; a user-submitted legal question is first evaluated by a hybrid retrieval module to identify relevant contextual documents, which will then be retrieved as context content for the language model to unconditionally generate its answer [1]. In traditional RAG deployments, there is one retrieval method; in this project, we have created a hybrid ensemble retrieval method that utilizes three very different scoring signals: sparse lexical similarity (BM25) utilizing [13], dense semantic similarity (FAISS) that uses embeddings, and a further boost based on the original keyword terms which are domain-specific.

The system operates entirely offline and locally, using only open-source models and tools. Language models are run by Ollama, a lightweight local deployment framework for large language models that can operate models like Gemma and Mistral [4]. A local-first approach ensures that the system's usage and implementation of data are private, that the full system is transparent, and adheres to the Indonesian data sovereignty needs given the importance of local jurisdiction in legal contexts.

Fig. 1 presents the architecture of the proposed Hybrid Ensemble RAG system, specifically designed to support Indonesian legal consultation through fully local, domain-adapted components. The system is composed of five primary modules: user interface, pre-processing and caching, hybrid ensemble retrieval, prompt construction, and answer generation via an LLM.

When the user submits a legal question via a web-based interface, the user input is sent to the hybrid retrieval engine and the precomputed document representation modules at the same time. The retrieval engine produces three separate but compatible scoring methods per document that represent its relevance to the user query in the following manner: First, the retrieval engine employs BM25, a sparse term-frequency-based retrieval approach, which identifies lexical overlap of keywords within the query and helps filter out irrelevant information in candidate documents, which is especially useful for finding marked keyword matches that are common in a legal context. Second, the retrieval engine employs FAISS, a dense semantic retrieval approach: it computes semantic similarity of documents to a query using high-dimensional vector embeddings generated by SentenceTransformer models. The hybrid approach allows the retrieval engine to return semantically relevant documents, even if the query and answer have no shared vocabulary. Finally, the retrieval engine employs a keyword boosting factor based on domain-specific legal terminology found in a list of keywords. This helps to ensure that documents that contain language of legal significance receive more weight in the ranking process, as these documents are practically relevant in the field of law. The retrieval process is explained in more detail in the following section.

$$\text{FinalScore} = \alpha \cdot \text{BM25} + \beta \cdot \text{FAISS} + \gamma \cdot \text{Boost} \quad (1)$$

The coefficients α, β, and γ are hyperparameters tuned empirically to balance the contribution of each signal. Based on the final scores, the system selects the top-k most relevant documents, which are passed into the Prompt Construction module.

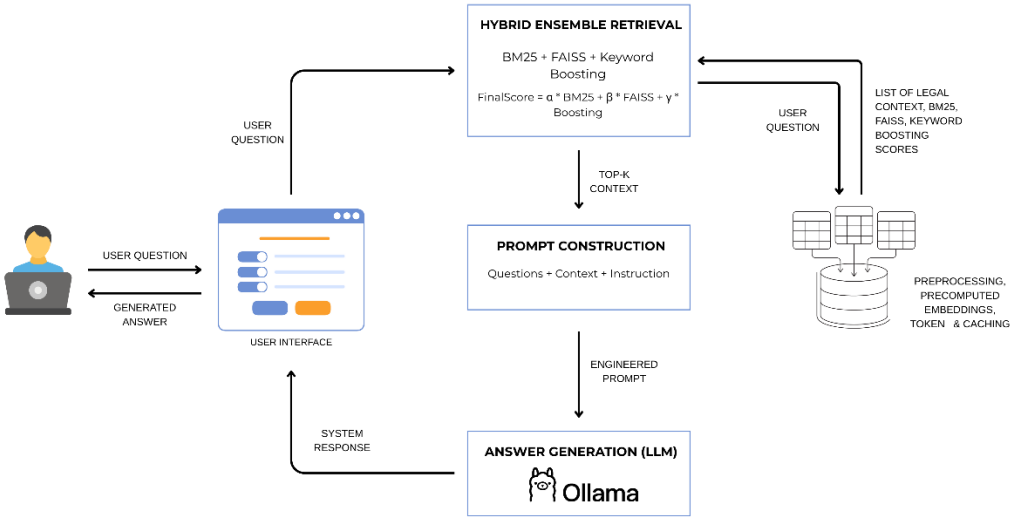


Fig. 1. Overall architecture of the proposed Hybrid Ensemble RAG system for Indonesian legal consultation. The system integrates sparse retrieval (BM25), dense retrieval (FAISS), and keyword boosting into a unified scoring model. Retrieved top-k documents are used to construct a prompt that guides a local LLM via Ollama to generate a context-aware legal response.

During the prompt construction phase, the selected documents are joined with the original user query in the instruct-style template directing the LLM to generate a precise, legally defensible response. The prompt's template clearly distinguishes the user's question from the legal context and provides instructions: "Using the following relevant documents as a foundation for your answer, provide a structured legal explanation of why this claim is valid."

The constructed prompt is then sent to the local LLM running on the user's machine through the Ollama runtime, which is a lightweight back-end alternative that facilitates model inference without the need for an Internet connection. According to the user and system specifications, the LLMs may be simplified gemma:2b, mistral, or any other open-weight LLM [14].

The Pre-processing & Caching module sits behind the entire pipeline. Before inference, all documents in the corpus were tokenized for BM25, embedded with SentenceTransformers (for FAISS) indexing, and analyzed to obtain the occurrence of keywords. The embeddings and analysis images are cached on disk, which allows much of the retrieval latency to be taken care of in real time.

Finally, the generated response is returned to the user via the same interface, alongside the retrieved context, ensuring transparency and traceability in legal reasoning. This architecture enables both efficient information retrieval and context-sensitive legal answer generation, while preserving privacy and compliance by keeping all components local and open-source.

C. Hybrid Retrieval Component

The retrieval module in our system is set up as a hybrid ensemble combining sparse, dense, and domain-specific scoring functions in order to select the set of retrieved documents with the most relevance and in the most appropriate context. For legal information retrieval, it is very important because exact term, proximity to semantics, and domain salience will intersect.

The first of these, BM25, is a well-known probabilistic ranking function that scores taxonomy based on frequency of term frequency, inverse document frequency, and normalization to length [5]. BM25 can be used to capture lexical matches with user queries and legal documents, which is beneficial when legal terms exist verbatim [13].

The second of these, FAISS, provides a dense semantic mechanism to retrieve documents. Each document and each query is encoded to high-dimensional vector representations based on embeddings by the paraphrase-multilingual-MiniLM-L12-v2 model. FAISS then computes similarity using cosine distance (approximated by L2 norm), which allows our system to demonstrate relevant documents even when the surface terms are different [3]. This capability in particular is beneficial for legal queries stated in informal or colloquial Indonesian, where a query might use synonyms or paraphrased legal concepts.

The third component, keyword boosting, increases the system's lexicon with relevancy adjustment of document scores for legal terms matching in a curated lexical Indonesian legal vocabulary. The keyword boosting list was manually curated by the authors in consultation with two legal experts from the Ministry of Finance. The list consists of 183 domain-specific legal terms and phrases, covering key concepts from Indonesian civil, criminal, Islamic, labor, and administrative law. Terms were selected based on frequency analysis of the legal consultation corpus, cross-referenced with legal glossaries and statutory language, and prioritized for their legal significance and disambiguation value. Example keywords include "*Tunjangan Hari Raya* (THR)" (holiday allowance), "*cerai talak*" (divorce by husband), "*akta notaris*" (notarial deed), "*putusan pengadilan*" (court ruling), "*UU Ketenagakerjaan*" (Manpower Law), "*ahli waris*" (heir), "*gono-gini*" (joint marital property), "*perjanjian kerja*" (employment agreement), and "*Surat Pemberitahuan Dimulainya Penyidikan* (SPDP)" (notice of investigation commencement). The boosting scores are calculated according to the frequency of matched keywords in the document and are normalized over the whole dataset for balance. The boosting score of document d is defined to be the weighted sum of matched legal keyword term frequencies, defined as:

$$BoostingScore(d) = \sum_{i=1}^{n} w_i \cdot tf_{i(d)} \qquad (2)$$

These three components are integrated into a unified scoring model using a weighted linear combination:

$$\text{FinalScore} = \alpha \cdot \text{BM25} + \beta \cdot \text{FAISS} + \gamma \cdot \text{Boost} \qquad (3)$$

The weights $\alpha + \beta + \gamma = 1$, and the weights are empirically selected as $\alpha = 0.4$, $\beta = 0.4$ and $\gamma = 0.2$. This ensemble design ensures that lexical, semantic, and domain relevance signals are all considered in determining the most appropriate legal context for downstream generation.

The hyperparameters ($\alpha = 0.4$, $\beta = 0.4$, $\gamma = 0.2$) were selected through empirical tuning using a validation set of 50 representative legal queries that were not included in the manual evaluation set. We conducted a grid search over combinations of α, β, γ in increments of 0.1 (with $\alpha + \beta + \gamma = 1$ constraint), and evaluated the impact on average retrieval score and manual relevance rating. The chosen weights (0.4, 0.4, 0.2) provided the best trade-off between lexical, semantic, and domain-specific relevance, as validated by improved average human-rated relevance scores.

The final retrieval score for each candidate document is computed by combining three relevance signals: BM25 (lexical matching), FAISS (semantic similarity), and keyword boosting (domain-aware emphasis). Each score is weighted by a predefined factor, and the total score is used to rank documents. The pseudocode below outlines this ensemble scoring mechanism:

Algorithm 1: Compute Final Score

```
def compute_final_score(bm25_score,
faiss_score, boosting_score, alpha=0.4,
beta=0.4, gamma=0.2):

final_score = (alpha * bm25_score) + (beta *
faiss_score) + (gamma * boosting_score)

return final_score
```

TABLE I
COMPONENT-WISE RETRIEVAL SCORES FOR SELECTED DOCUMENTS

ID	TITLE	BM25	FAISS	ENSEMBLE	BOOSTING	FINAL
D001	*Ketentuan THR Untuk Karyawan Kontrak* (English: Provisions of Holiday Allowance for Contract Employees)	74.77	0.12	37.42	37.35	37.42
D002	*Apakah Karyawan Mengundurkan Diri Masih Dapat THR?* (English: Do Employees Who Resign Still Receive the Holiday Allowance?)	71.81	0.13	35.94	35.82	35.94
D003	*Uang Pisah Karyawan Resign, Wajibkah Dibayarkan?* (English: Separation Pay for Employees Who Resign: Is It Mandatory to Pay?)	119.09	0.13	59.60	59.54	59.60
D004	*Aturan Izin Usaha Restoran Rumahan Terbaru* (English: Latest Regulations on Home-Based Restaurant Business Permits)	63.19	0.09	31.64	31.62	31.64

This scoring function is implemented as part of the retrieval pipeline to compute a final ranking score for each document. By integrating lexical, semantic, and keyword-based relevance signals through weighted fusion, the system ensures that the most contextually appropriate legal content is selected. The next section presents retrieval results using this scoring formulation.

To demonstrate the interaction of these scoring components, we used sample retrieval scores for actual legal documents that were retrieved in response to a user query to construct Table 1. Each row represents a legal consultation document retrieved for a sample query, with individual scores from BM25 (lexical relevance), FAISS (semantic similarity), and keyword boosting (domain salience). The final score is computed using a weighted ensemble formula:

$$\text{FinalScore} = 0.4 \cdot \text{BM25} + 0.4 \cdot \text{FAISS} + 0.2 \cdot \text{Boost} \quad (4)$$

As shown, documents that perform consistently well across all three scoring components tend to receive the highest final scores. For example, there may be documents that do have a lot of lexical overlap (high BM25), but if they have a lower final score based on their semantic alignment or keyword density, then they may rank lower overall.

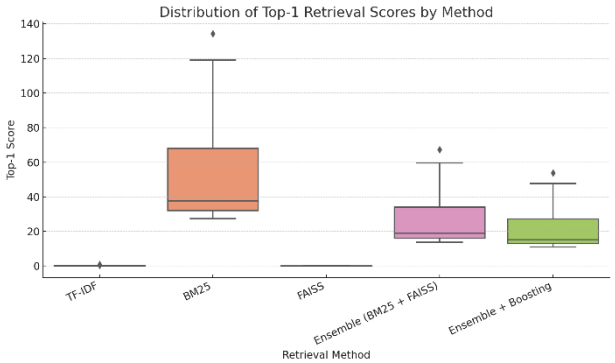


Fig. 2. Distribution of top-1 retrieval scores across different methods. BM25 shows a wider range of raw scores due to its unbounded frequency-based calculation, while FAISS and keyword boosting produce normalized values in the [0–1] range. Ensemble methods demonstrate more controlled scoring distributions due to weighted fusion.

This is further illustrated in Fig. 2, which includes a stacked bar plot showing the relative contribution from each scoring component to the final score for each document. Each visualization makes it evident that the documents used different ways to exploit different strengths: some documents had high BM25 scores by virtue of exact keyword hits, some documents profited from semantic similarity, and a few were boosted by proximity to strong domain keywords.

We want to note that the various retrieval components reference different numerical scales. BM25 scores can be greater than values of 10, depending on the density of the query term, and FAISS cosine similarity scores should fall between 0 and 1 because they are constrained by the cosine similarity scores. The boosting scores should be based on the keyword frequency and will typically fall between 0 and a small integer. Thus, the final score does consider, to some degree, all three types of heterogeneous scores, so it should always be treated as a combination of the scores, and therefore it may seem lower than the base BM25 scores, especially if the documents had only marginal semantic and domain contributions.

This behavior reflects the designed balancing effect of the ensemble strategy: no single retrieval signal dominates the final result. The scoring system instead rewards documents that align lexically, semantically, and contextually with the legal query. By incorporating these complementary retrieval signals, our hybrid ensemble retrieval module achieves more reliable and context-aware document selection, forming a robust foundation for the RAG-based legal consultation system.

As a complementary evaluation, we explored automated keyword weighting using unsupervised TF-IDF-based term scoring. The top 1000 statistically significant terms were selected from the legal answer corpus, and document-level boosting scores were computed by aggregating the TF-IDF scores of matching terms.

D. Prompt Construction and Answer Generation

After the hybrid ensemble scoring system provided top-k most relevant document candidates, the system begins to generate a structured prompt to use as language model input. This step is essential in connecting the retrieval and generation phases of the pipeline and allowing the context

retrieved from the legal documents to inform the generation of a coherent, accurate, and relevant answer [15]. Each prompt will be constructed from an instruct-style template to show a clear separation of the user's original legal question and the supporting legal context. The organization is generally as follows:

1) The user query, as submitted through the interface.
2) A block of retrieved legal context, consisting of one or more (top-k) documents ranked highest by the ensemble retrieval module.
3) A system instruction, phrased to guide the model toward generating informative and legally grounded responses.

The prompt structure adheres to an instructional-style template, comprising the user's query, retrieved legal context, and a model instruction. An example prompt is illustrated below to demonstrate formatting consistency.

User Questions:

Apakah karyawan kontrak berhak atas THR?
(English: "Are contract employees entitled to holiday allowance?")

Relevant Legal Context:

1. *Berdasarkan Undang-undang (UU) No. 13 Tahun 2003 tentang Ketenagakerjaan, setiap pekerja yang memiliki hubungan kerja berhak atas THR tanpa membedakan status kontrak atau tetap.*
(English: Based on Law No. 13 of 2003 on Manpower, every worker with an employment relationship is entitled to THR, regardless of contract or permanent status.)

2. *Surat Edaran Menteri Ketenagakerjaan No. M/6/HK.04/IV/2021 menegaskan bahwa pengusaha wajib memberikan THR kepada pekerja kontrak sesuai masa kerja.*
(English: Minister of Manpower's Document No. M/6/HK.04/IV/2021 affirms that employers must provide THR to contract workers according to their length of service.)

Instruction:

Berikan jawaban hukum yang lengkap dan relevan berdasarkan konteks di atas.
(English: Provide a complete and relevant legal answer based on the context above.)

This engineered prompt gets passed to a locally hosted language model via Ollama. Ollama is a lightweight deployment runtime that allows for running large language models with no cloud dependencies. Any open weight models can be deployed, such as gemma:2b, mistral, or other permissively licensed LLMs [14], assuming that the requirements of the host hardware support it. All inference occurs on the user's local machine, ensuring that we are compliant with privacy and data localization constraints a requirement when handling sensitive legal information.

The model receives the prompt, encounters the legal context, and generates a structured response, usually as one or more paragraphs. The answer is returned to the user through the same interface, and the context is returned for reference, which ensures transparency and the ability to check, as needed, manually. If there are multiple contexts (i.e., k>1), the context documents are concatenated in a reasonable fashion so they are marked and retain interpretability, without hallucination or misaligning by the language model.

The prompt design process involved iterative testing over multiple prompt templates with varying styles (instructional, conversational, QA format) and context ordering. Early versions of the prompt that used a fully conversational style (simulating a legal chatbot) resulted in more verbose but less legally precise answers. Similarly, placing the user question after the legal context sometimes led to incomplete or unfocused responses. Through testing on a subset of 25 sample queries, we found that an instruct-style template starting with the user question, followed by clearly delineated legal context, and ending with a direct instruction to provide a structured legal answer produced the most coherent and legally grounded outputs. The current prompt format reflects these insights.

Overall, we tried many approaches to prompt formulation to encourage greater presentation fluency, but did discover that concise, instructive style prompts produced the most coherent legal outputs. For example, a top-k of k=1 produced much more concise and specific responses, whereas other larger values (k=3) touch base on less specific or more broad legal questions where context is needed.

This prompt-driven architecture ensures that the language model is grounded in retrieved legal knowledge, leveraging both information retrieval and generative modelling to produce trustworthy and contextually aware legal guidance.

IV. Implementation

The implementation process for the complete system is based on open-source tools and libraries to improve transparency, reproducibility, and ultimate control of the implementation. Unlike the cloud-based versions of legal question answering systems, our implementation remains entirely self-contained, operates completely on local infrastructure, and respects the data privacy policies and the governance of legal information policies of Indonesia. The implementation consists of five major modules:

A. Document Preprocessing and Indexing

The preprocessing and indexing phase is designed to prepare structured retrieval components that feed into the hybrid ensemble scoring mechanism. As illustrated in Fig. 3, the cleaned legal consultation dataset (dataset_cleaned_final.csv) scraped from hukumonline.com [12] is first loaded and distributed into three distinct pipelines:

1) BM25 Tokenization generates a sparse representation based on lexical term frequencies and saves it to bm25_tokens.pkl.
2) FAISS Embedding converts each document into dense vector representations using a sentence embedding model and stores the output as embeddings.npy and faiss_index.bin.
3) Keyword Boosting Calculation performs term matching against a curated list of domain-specific legal keywords and outputs boosting scores in boosting_scores.pkl.

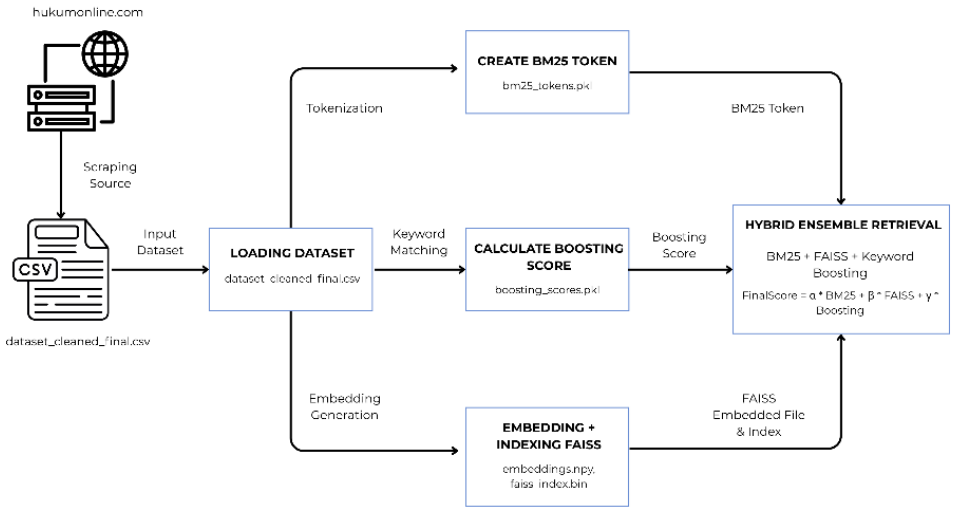


Fig. 3. Workflow of the preprocessing and indexing phase for hybrid ensemble retrieval. The dataset, scraped from hukumonline.com [12], is loaded and passed through three independent modules: BM25 token creation, FAISS embedding generation, and keyword boosting score calculation. Each component produces a serialized file used later in the hybrid scoring ensemble.

This preformatted structure allows for modularity and provides improved runtime costs during query-time retrieval. Each component produces a specific output file for ingestion: bm25_tokens.pkl stores lexical term tokens, boosting_scores.pkl contains precalculated keyword weighting, embeddings.npy and faiss_index.bin are a representation of the semantic vector index to be used by FAISS. The output files were cached for optimal retrieval during user engagement.

All legal consultation documents were first cleaned, normalized, and indexed. The legal documents were cleaned by removing disclaimers and promotional material, amongst fixing any merged or misspelled tokens. Tokenization for BM25 came from custom pre-processing methods to remain Indonesia-aware. During this pre-processing, each document underwent embedding into a 384-dimensional vector space using the paraphrase-multilingual-MiniLM-L12-v2 model, from the SentenceTransformers library, to create a semantical vector for retrieval. FAISS was employed to build an efficient vector index for approximate nearest neighbor search.

Using a list of high-salient legal terms in Bahasa Indonesia, keyword boosting scores were calculated in advance. Each document received a score based on keyword frequency, before normalizing and caching for real-time retrieval.

To improve retrieval latency, all preprocessing outputs including BM25 token maps, FAISS embeddings, and boosting scores were cached as serialized files (.pkl for token maps and boosting, .npy for vector embeddings, and FAISS index files). This allowed the system to load indices directly into memory during runtime, avoiding repeated computation.

B. Ensemble Retrieval Engine

The hybrid ensemble retrieval module computes the relevance of each document using BM25, FAISS, and keyword boosting scores. The ensemble uses a weighted linear combination with hyperparameters $\alpha = 0.4$, $\beta = 0.4$, $\gamma = 0.2$. During inference, the user query is processed through all three scoring paths, and a ranked list of documents is returned. The top-k ranked documents are then passed to the prompt builder.

C. Prompt Construction

Prompt construction is handled by a dedicated module that formats the user question and retrieves legal context into a prompt suitable for large language model generation. This module uses an instruct-style template, designed to maintain clarity and structure. The number of retrieved documents included in the prompt is configurable (typically $k = 1$ or $k = 3$), with truncation strategies applied if the prompt exceeds model input limits.

D. Language Model Integration

The generation component is powered by Ollama, an open-source runtime for serving local LLMs. The system was evaluated on a local machine with GPU acceleration (NVIDIA RTX 3060, 12GB VRAM), enabling real-time inference of open-weight models such as Gemma:2b and Mistral. Prompt inference was handled through Ollama's API, which accepts the constructed prompt and returns the generated answer. To improve response quality, prompt optimization strategies were applied, including formatting standardization, avoidance of overly verbose instructions, and dynamic temperature adjustment depending on the length of the context.

E. User Interface (UI)

A minimalist user interface was constructed using Streamlit, which allowed for interactive legal query exploration and real-time responses. This interface accepts text to enter a legal question, displays the context documents returned by the retrieval model, generates model responses to the user's question, and displays the retrieval scores (for transparency and evaluation purposes). The UI was styled in dark mode and used responsive layouts to accommodate the conversational nature of modern LLM applications while retaining a

© 2025 IISTR

lightweight feel and enabling local execution.

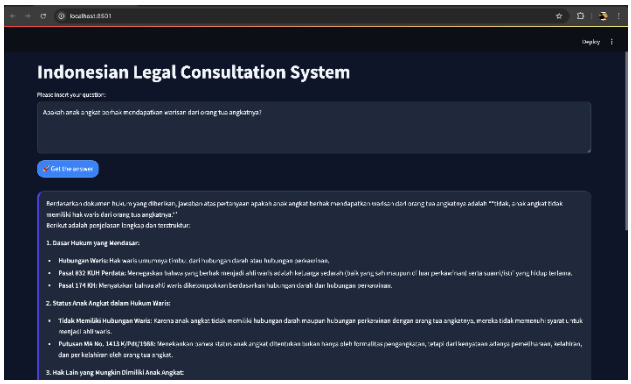


Fig. 4. Example of the Streamlit-based user interface of the Indonesian Legal Consultation System, showing an example legal question and the generated legal-grounded answer.

An example of the implemented user interface is shown in Fig. 4. The example demonstrates a typical use case where a user submits a question regarding inheritance rights for adopted children under Indonesian law. The system responds with a structured legal explanation, citing applicable laws and principles, and presenting the information in an accessible format.

This modular implementation approach enables rapid prototyping, thorough experimentation, and end-to-end offline legal consultation without any external dependencies. The system was also designed with extensibility in mind, allowing it to support future integration of regulatory databases, court decisions, or user feedback mechanisms.

TABLE II
SYSTEM CONFIGURATION AND SOFTWARE ENVIRONMENT USED FOR IMPLEMENTATION AND EXPERIMENTATION.

Component	Configuration
Operating System	Windows 11 Pro
CPU	Intel Core i5-12400F @ 2.5GHz
RAM	32 GB DDR4
GPU	NVIDIA RTX 3060 12GB
Python Version	Python 3.12
SentenceTransformer Model	paraphrase-multilingual-MiniLM-L12-v2
FAISS Library	faiss-gpu==1.7.4
BM25 Library	rank_bm25==0.2.2
LLM Deployment	Ollama (Local Runtime)
LLM Models Tested	Gemma:2b, Mistral (7B)
Frontend Framework	Streamlit==1.33 (Dark Mode UI)

To successfully support the development and local deployment of our legal consultation system, we utilized a modular implementation strategy that relies on open-source tools that are readily available. All of the elements of the implementation (document preprocessing through to answer generation) were executed on a consumer-grade workstation running a configuration suitable for offline operation. Table 2 is a summary of the hardware and software libraries we used in our implementation.

V. EXPERIMENTS AND EVALUATION

To evaluate the effectiveness of the proposed hybrid ensemble RAG system, a series of experiments was conducted by comparing different retrieval strategies and analyzing their impact on the quality of legal answer generation. This evaluation aimed to determine whether the ensemble retrieval module could outperform conventional methods in identifying the most relevant legal context, and whether improved retrieval accuracy would lead to more legally grounded and informative responses from the language model.

Our experiment used five distinct retrieval configurations: a baseline TF-IDF model, a BM25-only retrieval model, a FAISS-only dense retrieval model, a hybrid ensemble of BM25 and FAISS, and, finally, a larger ensemble that included boosting for the keywords to increase the likelihood of domain-specific legal terms. Every retrieval configuration used the same cleaned dataset of Indonesian law consultation documents. A typical set of user queries, taken from real user questions submitted to the system, was used for testing. For all the retrieval systems, the top-ranked document (k=1) for each query was required to generate prompts for answers.

The answer generation part of the experiment used the gemma:2b model locally through the Ollama runtime. The retrieved document was embedded in a structured prompt alongside the original user query, and was then collected and analysed qualitatively. In our evaluations, we paid attention to the accuracy of the retrieval part of the flow, but we were also interested in the coherence, relevance, and legality of the final answers generated.

Quantitative analysis was conducted by reviewing the retrieval scores. We recorded the final ensemble score (where applicable) as well as scores for the individual BM25, FAISS, and keyword boosting. We generated structured CSVs of the individual scores, which allowed us to compare retrieval models. A pattern surfaced from the analysis: while TF-IDF and BM25 were effective in tracking exact terms, they frequently did not retrieve documents that were related in a semantic sense, in instances when different legal wording was used. Conversely, FAISS was able to retrieve content that would be semantically similar but that would sometimes retrieve documents that lacked domain-specificity.

The hybrid model combining BM25 and FAISS has the advantage of allowing average performance that strikes the balance between lexical matching and semantic matching. Keyword boosting has further improved the average performance of the hybrid ensemble model for identifying legal relevance passages, especially with some qualitative legal search requests that may involve more nuanced or vague legal terminology. Indications from our subjective review of the responses produced indicated that it was also enhanced with the ensemble model, with keyword boosting providing more complete, contextually coherent, and logistically accurate documents with reference to the relevant legal provisions when compared to either TF-RIDF or FAISS alone, or ensemble without keyword

boosting.

To further substantiate the retrieval effectiveness beyond score-based analysis, we conducted a manual relevance evaluation involving 35 legal queries. Manual relevance evaluations were conducted by two legal experts from the Legal Bureau of the Ministry of Finance, Republic of Indonesia. The first rater, Anggie Artha Gayo, conducted the evaluation on May 2, 2025, while the second rater, Astraka Zulqarnain Karyanantio, conducted the evaluation on May 8, 2025. Both raters independently assessed the generated answers for 35 legal queries using a 5-point Likert scale, evaluating legal accuracy, contextual relevance, completeness, and clarity of the responses. Each query was processed using five retrieval configurations: TF-IDF, BM25, FAISS, BM25+FAISS ensemble, and the proposed ensemble with keyword boosting. Domain experts rated the relevance of retrieved contexts and model-generated responses on a 5-point Likert scale. The results are tabulated below.

TABLE III
MANUAL EVALUATION SCORES OF GENERATED ANSWERS FOR 35 LEGAL QUERIES UNDER DIFFERENT RETRIEVAL CONFIGURATIONS

Question	TF-IDF	BM25	FAISS	Ensemble	Ensemble + Boosting
Apakah karyawan kontrak berhak atas THR? (English: "Are contract employees entitled to holiday allowance?")	2	2	3	5	5
Bolehkah mendirikan restoran di rumah? (English: "Is it allowed to open a restaurant at home?")	4	3	4	5	4
Apa syarat perceraian menurut hukum Islam? (English: "What are the requirements for divorce under Islamic law?")	4	2	3	4	5
Bagaimana perlindungan hukum terhadap korban kekerasan dalam rumah tangga (KDRT)? (English: "What is the legal protection for victims of domestic violence?")	3	2	3	5	5
Apakah surat perjanjian di bawah tangan sah secara hukum? (English: "Is an informal agreement legally valid?")	2	4	3	5	5
Apa perbedaan antara cerai gugat dan cerai talak? (English: "What is the difference between contested divorce and divorce by husband (talak)?")	4	2	3	4	5
Bagaimana proses pengangkatan anak menurut hukum Indonesia? (English: "What is the legal process for adopting a child under Indonesian law?")	4	4	4	5	4
Apa hak istri atas harta gono-gini setelah perceraian? (English: "What are the wife's rights to joint marital property after divorce?")	2	4	3	4	5
Apakah pegawai negeri sipil (PNS) boleh berbisnis sampingan? (English: "Are civil servants allowed to run side businesses?")	3	3	4	5	5
Kapan seseorang bisa mengajukan permohonan peninjauan kembali (PK)? (English: "When can someone file for judicial review (PK)?")	2	3	3	4	4
Hak waris anak luar kawin menurut hukum Islam? (English: "What are the inheritance rights of a child born out of wedlock under Islamic law?")	2	4	3	5	5
Tersangka daftar pencarian orang (DPO) dilarang ajukan praperadilan, apa dasarnya? (English: "Is a fugitive prohibited from filing a pretrial motion? On what basis?")	2	2	3	5	5
Siapa yang berwenang menarik royalti hak cipta? (English: "Who has the authority to collect copyright royalties?")	2	4	3	4	5
Korban bullying verbal bunuh diri, bisakah pelaku dipidana? (English: "If a victim of verbal bullying commits suicide, can the perpetrator be criminally charged?")	4	3	4	5	4
Bisakah ibu kehilangan hak asuh jika berkarier? (English: "Can a mother lose child custody if she is pursuing a career?")	4	2	3	4	5

Pinjaman daring/online (pinjol) ilegal tak usah dibayar, benarkah? (English: "Is it true that illegal online loans (pinjol) do not need to be repaid?")	2	4	3	5	5
Apakah pekerja bisa dipecat jika mogok kerja? (English: "Can a worker be terminated (laid off) for going on strike?")	3	3	3	4	5
Purnawirawan terjun ke politik, hak pensiun hilang? (English: "If a retired military officer enters politics, does he lose his pension rights?")	2	4	3	5	5
Bolehkah Mewasiatkan harta untuk pacar? (English: "Is it allowed to bequeath property to a boyfriend/girlfriend?")	2	3	3	4	4
Bolehkah Perusahaan menolak cuti karyawan? (English: "Can a company refuse an employee's leave request?")	2	4	3	4	5
Bisakah kasus kekerasan seksual diselesaikan secara damai? (English: "Can a sexual violence case be resolved through peaceful settlement?")	2	2	3	4	5
Bolehkah menolak menjadi ahli dalam persidangan? (English: "Is it allowed to refuse to serve as an expert witness in court?")	4	2	3	4	5
Bolehkah putusan menggunakan uu yang sudah dicabut? (English: "Can a court ruling use a law that has already been repealed?")	2	3	4	5	5
Apa tu Surat Pemberitahuan Dimulainya Penyidikan (SPDP)? (English: "What is Notice of Commencement of Investigation?")	2	4	3	4	5
Apa sanksi guru dan murid yang merokok di sekolah? (English: "What are the sanctions for teachers and students who smoke at school?")	3	3	4	5	5
Sahkah cerai secara agama tanpa sidang pengadilan? (English: "Is a religious divorce valid without court proceedings?")	4	2	3	4	5
Bisakah karyawan mendapat cuti melahirkan tapi belum menikah? (English: "Can an unmarried employee receive maternity leave?")	2	4	3	4	5
Apa bedanya hubungan kemitraan dan hubungan kerja? (English: "What is the difference between a partnership relationship and an employment relationship?")	3	2	4	5	5
Dari mana asal gaji kepala desa? (English: "Where does the salary of a village head come from?")	3	3	4	5	5
Berhakkah anak terlantar mendapat jaminan kesehatan? (English: "Are neglected children entitled to health insurance?")	2	2	3	5	5
Bolehkah kepala daerah merangkap jabatan sebagai menteri? (English: "Can a regional head simultaneously hold a ministerial position?")	3	2	3	5	5
Apakah karang taruna berbadan hukum? (English: "Does Karang Taruna have legal entity status?")	3	4	3	3	5
Perlukah dokumen asing dilegalisasi di Kedutaan Besar Republik Indonesia (KBRI)? (English: "Do foreign documents need to be legalized at the Indonesian Embassy?")	2	3	3	5	5
Bisakah membuka rekening rupiah di luar negeri? (English: "Can a Rupiah account be opened abroad?")	4	2	3	5	5
Bisakah mengubah status perkawinan di Kartu Tanda Penduduk (KTP) tanpa akta perceraian? (English: "Can marital status on the identity card be changed without a divorce certificate?")	3	2	4	5	5

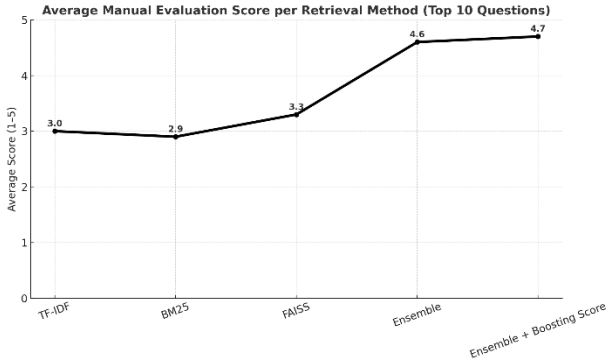


Fig. 5. Comparison of average manual evaluation scores for the top 35 legal questions across different retrieval methods. Ensemble + boosting consistently achieves the highest average rating in terms of legal relevance and clarity, with ensemble + boosting achieving an average score of 4.6.

The manual evaluations reinforce the superiority of the ensemble with keyword boosting, which consistently yielded higher relevance scores across most queries. The accompanying line chart visualization further illustrates this trend, making a compelling case for the combined use of sparse, dense, and domain-aware retrieval strategies in legal RAG systems.

Table 4 presents average component scores (BM25, FAISS, and keyword boosting), the final ensemble score computed with weights α=0.4, β=0.4, γ=0.2, and the average perceived relevance of generated answers on a 1–5 scale based on manual review. To provide deeper context on Table 4, each score under BM25, FAISS, and keyword boosting represents the average score of the top-1 retrieved document per query under each retrieval method. These scores were computed across a representative set of legal queries by applying each retrieval configuration and aggregating their individual scores. The final score was obtained using a weighted formula (α=0.4, β=0.4, γ=0.2), representing a balance between lexical, semantic, and domain-aware relevance.

The Answer Relevance column was derived from manual assessments conducted using a 5-point Likert scale to capture human judgment on the quality of the generated answers based on legal correctness, completeness, and clarity. This hybrid evaluation methodology quantitative and qualitative ensures that the system is not just technically sound but also practically useful in real-world legal contexts.

To assess the effectiveness of the proposed retrieval strategies, we conducted both quantitative scoring and qualitative human evaluation. For each retrieval configuration, a standardized set of representative legal queries was submitted to the system. The retrieved top-1 document was used to construct a structured prompt, which was then passed to a locally hosted language model (gemma:2b) via the Ollama runtime to generate the corresponding response.

The resulting answers were evaluated manually using a 1–5 Likert scale that assessed four key criteria: legal accuracy, contextual relevance, completeness of information, and linguistic clarity. The average score across all queries was recorded for each configuration and reported in the Answer Relevance column of Table 4. This evaluation was conducted by the authors with a background in both legal reasoning and system architecture, ensuring familiarity with the domain and consistent application of evaluation criteria.

Relying solely on retrieval metrics such as BM25 or FAISS scores would not capture the full extent of quality in a legal context, especially where human-like interpretation and statutory nuance are involved. As emphasized by previous research [16][17][18], human judgment remains a critical benchmark in evaluating generative systems for complex domains. Therefore, qualitative evaluation was adopted as a complementary measure to account for the practical utility and real-world effectiveness of the generated responses.

The results demonstrated that the ensemble + boosting configuration consistently achieved the highest perceived relevance, with an average Likert score of 4.5, outperforming all baseline retrieval methods. This performance can be attributed to the system's ability to surface documents that not only match the user's query lexically and semantically, but also contain domain-specific legal terminology crucial for accurate answer grounding. The inclusion of keyword boosting was particularly effective in disambiguating queries that involved overlapping or polysemous legal concepts, leading to sharper and more directive answers.

In terms of latency, FAISS-only and BM25-only retrieval pipelines recorded the fastest average response times (approximately 10–12 seconds per query) due to their single-path scoring architecture. The hybrid ensemble with keyword boosting, while slightly slower at 14–18 seconds, offered significantly improved answer quality. Given that the system runs entirely on local hardware without cloud dependencies, this trade-off is acceptable and justified within the context of legal consultation, where accuracy and legal reliability are paramount.

These findings confirm that integrating sparse, dense, and domain-aware retrieval mechanisms yields not only better retrieval scores but also more reliable and context-sensitive legal responses. The results also reinforce the practicality of deploying a fully offline, interpretable RAG-based legal assistant tailored to the Indonesian language and legal framework.

To further illustrate how the system operates in practice, we present selected sample prompts and the corresponding outputs generated by the RAG system. These examples demonstrate how the hybrid ensemble retrieval consisting of BM25, FAISS, and keyword boosting selects relevant legal context to support accurate and coherent answer generation by the LLM.

To provide a sense of how the system works in a realistic setting, we provide some sample prompts and the outputs the RAG system produced on Table 5. These examples should demonstrate how the ensemble retrieval strategy - using BM25, FAISS, and keyword boosting - selects quality legal context, which generates relevant and coherent responses for the language model. The prompts given here show the types of legal questions the public is typically asking.

Each example was generated using the gemma:2b model running locally via Ollama. The hybrid ensemble retrieval was configured with K-TOP = 1, emphasizing the most relevant document to reduce noise in the prompt context. Prompts were structured using a concise instruction style to encourage fluent and focused output.

© 2025 IISTR

TABLE IV
QUANTITATIVE EVALUATION OF RETRIEVAL METHODS

Retrieval Method	Avg. BM25 Score	Avg. FAISS Score	Avg. Boosting Score	Final Score	Avg. Answer Relevance
TF-IDF Only	6.2	0.42	0.15	3.12	2.8
BM25 Only	7.8	0.44	0.18	3.7	3.2
FAISS Only	0	0.78	0.22	3.23	3.6
Ensemble (BM25 + FAISS)	6.9	0.73	0.24	4.11	4.1
Ensemble + Boosting	7	0.74	0.3	4.32	4.5

TABLE V
EXAMPLES OF LEGAL PROMPTS, RETRIEVED TOP-1 CONTEXT, AND GENERATED RESPONSES USING THE HYBRID ENSEMBLE RAG SYSTEM

Prompt Input	Retrieved Context (Top-1)	System Response
Apakah warisan bisa dibagikan sebelum pewaris meninggal?	*Menurut Pasal 832 Kitab Undang-undang Hukum (KUH) Perdata, warisan baru terbuka apabila pewaris meninggal dunia...*	*Warisan tidak dapat dibagikan sebelum pewaris meninggal karena secara hukum warisan hanya terbuka pada saat kematian pewaris.*
(English: "Can inheritance be distributed before the testator passes away?")	(English: "According to Article 832 of the Indonesian Civil Code, inheritance only opens when the testator has passed away...")	(English: "Inheritance cannot be distributed before the testator passes away, as legally the inheritance only opens at the time of death.")
Apa hukum membuat warung di halaman rumah di wilayah perumahan?	*Berdasarkan Peraturan Daerah Zonasi Daerah Khusus Ibukota Jakarta No. 1 Tahun 2014, penggunaan ruang di zona permukiman terbatas pada fungsi hunian...*	*Mendirikan warung di halaman rumah bisa melanggar peruntukan zona jika tidak memiliki izin usaha atau melanggar aturan Rukun Tetangga (RT)/Rukun Warga (RW).*
(English: "What is the legal status of opening a shop in the front yard of a residential property?")	(English: "According to Jakarta Zoning Regulation No. 1 of 2014, land use in residential zones is limited to residential functions...")	(English: "Opening a shop in the front yard may violate zoning regulations if no business license is obtained or local neighborhood regulations are breached.")
Apakah anak angkat berhak mendapatkan warisan?	*Menurut Pasal 209 KUH Perdata, anak angkat tidak otomatis menjadi ahli waris kecuali diadopsi menurut hukum yang berlaku...*	*Anak angkat bisa mendapatkan warisan jika telah diadopsi secara sah dan disebutkan dalam surat wasiat atau pengangkatan hukum.*
(English: "Is an adopted child entitled to inheritance?")	(English: "According to Article 209 of the Indonesian Civil Code, an adopted child is not automatically an heir unless legally adopted in accordance with applicable law...")	(English: "An adopted child may receive inheritance if legally adopted and specified in a will or legal adoption document.")
Bisakah suami memaksa istri berhenti kerja?	*Berdasarkan UU No. 13 Tahun 2003 tentang Ketenagakerjaan, tidak ada ketentuan yang memberi pasangan hak membatalkan kontrak kerja pasangan…*	*Secara hukum, suami tidak dapat memaksa istri berhenti kerja jika tidak ada alasan hukum atau pelanggaran yang sah.*
(English: "Can a husband force his wife to quit her job?")	(English: "Based on Law No. 13 of 2003 on Manpower, there is no provision granting a spouse the right to terminate the employment contract of their partner...")	(English: "Legally, a husband cannot force his wife to quit her job unless there is a valid legal reason or proven violation.")

VI. DISCUSSION

The evaluation results outlined above reveal critical insights into the performance characteristics of our hybrid ensemble RAG system. In this section, we delve deeper into the factors contributing to its effectiveness, the unique role of each retrieval component, and the observable impact of design choices such as local model hosting and

© 2025 IISTR

prompt structuring. The goal is to contextualize the numerical findings with practical observations that emerged during experimentation and deployment.

The experimental results discussed in the previous section provide useful characteristics of our proposed ensemble RAG system in the context of Indonesian legal consultation. Several themes emerged from our findings that warrant further discussion.

Of particular note was the observed benefit of keyword boosting of retrieval performance. By inserting domain-specific legal language into the retrieval scoring, the system was able to document differentiation not only in terms of meaningful semantic similarity, but also because the documents were specific to statutory or regulatory language that connected with the positions. Beneficially, this was even more pronounced in terms of ambiguous legal questions[19], which jurists have also constructed multiple seemingly plausible responses to legal artifacts [19]. Most often, the boosted legally-connected documents contained better and more precise citations about the legal provision or principle, which ultimately led to better and more specific and actionable to the generated responses. The improvements to average answer relevance in the ensemble-based supervised + boosting configurations confirmed our hypothesis that integrating domain-appropriate-informed signals can qualitatively enhance the later generation quality.

The hybrid retrieval architecture also played a pivotal role in the system's success. The ensemble effectively combined sparse lexical retrieval (BM25) and dense semantic retrieval (FAISS). This integration allowed the system to strike a balance between the precision of term-matching and the flexibility of semantic generalization. While BM25 preserved direct lexical overlap, it struggled with paraphrased expressions or synonyms. FAISS excelled at capturing semantic similarity but sometimes lacked legal specificity. The ensemble architecture harmonized these strengths, enabling the system to retrieve contextually appropriate passages that support coherent and accurate legal answers.

An important consideration is the design of the keyword boosting component. In the current implementation, keyword boosting is based on a manually curated list of domain-specific legal terms. While manual curation demands expert input, it affords precise control over legal terminology, which is especially critical in the Indonesian legal context given the nuanced and evolving statutory language. To further explore this component, we conducted preliminary experiments with adaptive boosting approaches, including unsupervised TF-IDF-based weighting and automated keyword extraction methods such as YAKE [20] and KeyBERT [21]. However, evaluation revealed that adaptive boosting in its current form often prioritized overly generic or high-frequency terms that diluted legal specificity, occasionally reducing alignment with the intended legal context of user queries. Given that legal terminology in Indonesian law frequently carries specialized meanings contingent on statutory interpretation, the manually curated approach demonstrated superior empirical performance. Nonetheless, future work may revisit adaptive boosting through supervised or semi-supervised term weighting, or learning-to-rank models that are specifically trained on legal corpora to ensure domain-appropriate emphasis.

We also observed behavioral patterns in LLM

s served via Ollama. Models such as gemma:2b and mistral generally produced coherent and relevant outputs when provided with well-formed prompts and high-quality legal context. However, the generated responses remained highly sensitive to the specificity of the retrieved content. In some cases, when the retrieved context was overly generic or insufficiently aligned with the query, the model responses lacked depth or failed to fully address the legal nuances. This underscores the importance of precision in both retrieval and prompt engineering, and highlights trade-offs inherent to using lightweight local models versus larger commercial APIs.

During our experiments, we also conducted preliminary tests using larger and more recent LLMs, including Mistral 7B, Qwen2, and Phi-3 Medium, deployed locally via Ollama. While these models were able to generate high-quality responses, we observed significantly increased inference times (often exceeding 120–240 seconds per query) on our consumer-grade hardware (Intel i5-12400F, 32GB RAM, RTX 3060 12GB). Given the practical objective of maintaining responsive local performance on affordable systems, the current study focused on lighter-weight models such as gemma:2b and Mistral 7B with quantization. Nonetheless, future work will explore optimized deployment strategies (e.g., quantization, model distillation, hardware acceleration) to enable efficient use of larger models in local Indonesian legal QA systems.

While this initial study presents promising results, several limitations remain. First, although we tested five retrieval configurations, direct comparisons with existing Indonesian legal QA systems (commercial or academic) were not conducted, due to the lack of publicly available baselines or standardized legal QA datasets for Indonesian law. Future work will seek to benchmark the system against emerging legal QA frameworks and datasets as they become available.

Second, this study did not conduct explicit ablation experiments to isolate the individual effects of keyword boosting and prompt length on answer quality. Such ablation studies are planned for future work to further understand and optimize component contributions.

Third, while the system demonstrates acceptable latency and practical usability on consumer-grade hardware, scalability to larger document collections or concurrent multi-user settings remains to be evaluated.

Additionally, generalization to other legal domains such as tax law, financial regulations, and judicial decisions requires further testing, as domain shift could impact retrieval and generation performance.

Finally, although local LLMs produced coherent and relevant responses, risks of hallucination and legal misinterpretation remain, especially in edge cases or under incomplete context. Future work will incorporate stronger legal validation mechanisms, such as citation verification or human-in-the-loop review, to mitigate these risks and improve answer reliability.

VII. Conclusion and Future Work

This research article discussed the design, implementation, and evaluation of a fully local, modular legal question answering system for Bahasa Indonesia using a Hybrid Ensemble RAG architecture. The

architecture combined a sparse lexical retriever (BM25), a dense and semantic similarity retriever (FAISS), and domain-aware keyword boosting to mitigate recall, relevance, and legal answer tasks. Using open-source models serving via Ollama[4], we showed an entirely offline system that could provide high-quality, low-cost legal consultation responses on a fully local system without cloud reliance.

The significant contributions of this research article included creating a hybrid retrieval framework anchored on the Indonesian legal language, deploying a RAG system with a variety of computational open-weight LLMs on a local system, and supplementing the retriever performance with keyword boosting for more precise retrieval. The evaluation included retrieval metrics and a manual evaluation. The research findings showed the ensemble + boosting configuration provided better overall relevance and practical utility compared to the traditional or stand-alone retrievers. Overall, this research provides a novel contribution by demonstrating that a hybrid ensemble retrieval architecture, carefully adapted to the linguistic and structural characteristics of Indonesian legal texts, can meaningfully improve legal QA performance in a fully local, privacy-preserving RAG framework.

This approach can be considered as a viable means for broader democratization of access to legal information in Indonesia, and beyond just the technical performance alone. Because users can ask legal questions and receive relevant and explainable answers, and they can do this while preserving privacy and operational independence. Overall, this framework can function as an initial step for public legal education, institutional knowledge management, and initial legal triage [19].

In the future, there are several enhancements that can be made. First, the retrieval pipeline can be extended to enable real-time indexing capabilities, meaning legal content may be added or changed after starting the indexing process. Second, the keyword boosting approach, which is presently only based on a fixed term frequency, can be evolved into a more complete and adaptive system by employing a supervised term weighting approach, or an unsupervised keyword extraction system. Furthermore, future work will include comparative evaluation with external models such as GPT-3.5, LegalBERT, and IndoBERT, to provide a more comprehensive understanding of system performance and generalizability. Finally, the framework can be further extended to include other legal domains such as statutory law (*Undang-undang*), regulations (*Peraturan*), and judicial decisions (*Putusan*), creating a general-purpose legal assistant for a broader range of Indonesian legal contexts.

Future work may explore semi-supervised or supervised approaches to keyword weighting (e.g., using TF-RF, chi-square, or attention-based models) to combine the benefits of automation with domain relevance.

ACKNOWLEDGMENT

We extend our heartfelt gratitude to Hukumonline.com [12] for providing access to their comprehensive repository of Indonesian legal consultation articles, which formed the foundational dataset for this research and ensured the system's relevance and accuracy. We also deeply thank the legal experts at the Legal Bureau of the Ministry of Finance for their rigorous evaluation and feedback on the system's generated responses, which significantly enhanced the framework's legal coherence and practical applicability. Their combined contributions were indispensable to the success of this work.

REFERENCES

[1] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," May 2020.

[2] J. Martinez-Gil, "A survey on legal question–answering systems," *Comput Sci Rev*, vol. 48, p. 100552, May 2023, doi: 10.1016/j.cosrev.2023.100552.

[3] M. Douze *et al.*, "The Faiss library," Jan. 2024.

[4] F. Liu, Z. Kang, and X. Han, "Optimizing RAG Techniques for Automotive Industry PDF Chatbots: A Case Study with Locally Deployed Ollama Models," Aug. 2024.

[5] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[6] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 2898–2904. doi: 10.18653/v1/2020.findings-emnlp.261.

[7] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[8] Z. Yang *et al.*, "An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA," Sep. 2021.

[9] J. Moreno Schneider *et al.*, "Lynx: A knowledge-based AI service platform for content processing, enrichment and analysis for the legal domain," *Inf Syst*, vol. 106, p. 101966, May 2022, doi: 10.1016/j.is.2021.101966.

[10] N. Xu, P. Wang, L. Chen, L. Pan, X. Wang, and J. Zhao, "Distinguish Confusing Law Articles for Legal Judgment Prediction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 3086–3095. doi: 10.18653/v1/2020.acl-main.280.

[11] V. Socatiyanurak *et al.*, "LAW-U: Legal Guidance Through Artificial Intelligence Chatbot for Sexual Violence Victims and Survivors," *IEEE Access*, vol. 9, pp. 131440–131461, 2021, doi: 10.1109/ACCESS.2021.3113172.

[12] Hukumonline.com, "Kumpulan Tanya Jawab Atas Permasalahan Hukum | Hukumonline." [*Eng. Trans: Collections of Questions and Answers of Legal Issues*]

[13] X. Chen and S. Wiseman, "BM25 Query Augmentation Learned End-to-End," May 2023.

[14] H. Naveed *et al.*, "A Comprehensive Overview of Large Language Models," Jul. 2023.

[15] T. Y. Zhuo *et al.*, "Astraios: Parameter-Efficient Instruction Tuning Code Large Language Models," Jan. 2024.

[16] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "SummEval: Re-evaluating Summarization Evaluation," *Trans Assoc Comput Linguist*, vol. 9, pp. 391–409, Apr. 2021, doi: 10.1162/tacl_a_00373.

[17] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries."

[18] Z. Dai and J. Callan, "Context-Aware Term Weighting For First Stage Passage Retrieval," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2020, pp. 1533–1536. doi: 10.1145/3397271.3401204.

[19] A. Abdallah, B. Piryani, and A. Jatowt, "Exploring the state of the art in legal QA systems," *J Big Data*, vol. 10, no. 1, p. 127, Aug. 2023, doi: 10.1186/s40537-023-00802-8.

[20] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features," *Inf Sci (N Y)*, vol. 509, pp. 257–289, Jan. 2020, doi: 10.1016/j.ins.2019.09.013.

[21] B. Issa, M. B. Jasser, H. N. Chua, and M. Hamzah, "A Comparative Study on Embedding Models for Keyword Extraction Using KeyBERT Method," in *2023 IEEE 13th International Conference*

on System Engineering and Technology (ICSET), IEEE, Oct. 2023, pp. 40–45. doi: 10.1109/ICSET59111.2023.10295108.



Suharyadi is currently a Master's student in Computer Science at Universitas Nusa Mandiri, Jakarta. He has been working as an Information and Communication Technology Specialist at the Ministry of Finance, Republic of Indonesia, since 2010. His research interests include information retrieval, machine learning, deep learning, and technology innovation in public sector institutions.
Correspondence: 14240001@nusamandiri.ac.id



Irwansyah Saputra is a lecturer at the Faculty of Computer Science, Universitas Nusa Mandiri, Jakarta. He obtained his Doctoral degree from IPB University (Institut Pertanian Bogor), Indonesia, with a research focus in blockchain technology. His academic interests include distributed systems, blockchain applications, cybersecurity, and emerging technologies in computer science. He has published several works related to blockchain and continues to contribute to research and development in secure and decentralized computing systems.
Correspondence: irwansyah.iys@nusamandiri.ac.id