Journal of Research, Innovation and Technologies
https://www.acadlore.com/journals/JORIT

Development of an Offline RAG Chatbot for Answering Food Hygiene and Safety Questions Based on Vietnamese Legal Frameworks

Duc Viet Hoang[1], Tat Thang Nguyen[2*]

[1] Master Course of Information System, Graduate Faculty, Posts and Telecommunications Institute of Technology, 10000 Hanoi, Vietnam
[2] Department of Computer Science, Faculty of Information Technology, Posts and Telecommunications Institute of Technology, 10000 Hanoi, Vietnam

[*] Correspondence: Tat Thang Nguyen (thangnt@ptit.edu.vn)

Received: 11-13-2025 **Revised:** 01-19-2026 **Accepted:** 01-22-2026

Citation: Hoang, D. V. & Nguyen, T. T. (2026). Development of an offline RAG chatbot for answering food hygiene and safety questions based on Vietnamese legal frameworks. *J. Res. Innov. Technol.*, *5*(1), 121–135. https://doi.org/10.56578/jorit050108.

Abstract: This paper presents the design, development, and implementation of an offline chatbot system specialized in answering food safety-related questions, relying entirely on Vietnamese legal documents. The system employs Retrieval-Augmented Generation (RAG) to ensure accurate and contextually relevant responses without internet dependency, a critical feature for low-connectivity environments. Key highlights include robust Vietnamese language support, a flexible vector database using Chroma for seamless legal content updates, and the integration of Qwen2.5:7B-Instruct-Q4_0 as the local language model, selected after comparative testing against DeepSeek-R1, Gemma3:1B, and Mistral. Embeddings are generated using BAAI/bge-small-en-v1.5. By processing Vietnamese queries and retrieving from a localized knowledge base, the chatbot delivers reliable guidance to stakeholders such as food producers, traders, and consumers. Evaluations demonstrate high accuracy in Vietnamese Q&A, stable offline operation, and adaptability to evolving regulations, with discussions on limitations and future enhancements.

Keywords: Retrieval-Augmented Generation; Offline legal chatbot; Vietnamese food safety law; Local LLM; Chroma vector store; Delta synchronization; LangChain; Ollama

JEL Classification: I18, O33, K23

1. Introduction

Food safety is a critical public health issue in Vietnam where the agricultural and food processing sectors play a pivotal role in the economy, contributing significantly to the GDP and employing a large portion of the workforce as of 2025. The regulatory framework governing food safety is anchored by the Food Safety Law No. 55/2010/QH12 (National Assembly of the Socialist Republic of Vietnam, 2010), enacted by the National Assembly on June 17, 2010, and effective from July 1, 2011, with significant amendments incorporated through drafts finalized in October 2025 to address emerging challenges such as climate change impacts on food production and the rise of e-commerce in food trade. This law, alongside supporting decrees like Decree No. 15/2018/ND-CP (Government of the Socialist Republic of Vietnam, 2018), which was issued on February 2, 2018, and provides detailed regulations on production, trading, and incident management, establishes a comprehensive legal structure. These regulations cover a wide array of topics, including the definition of food safety standards, prohibited acts (e.g., adulteration), rights and obligations of stakeholders, specific assurance conditions for production, trading, and import/export activities, certification processes, testing protocols, risk analysis methodologies, incident management strategies, and state management oversight mechanisms.

Despite this robust legal foundation, accessing and interpreting these regulations poses substantial challenges, particularly for small and medium enterprises (SMEs), rural consumers, and local authorities with limited

resources. The complexity of legal language, often laden with technical jargon and context-specific references, combined with language barriers for non-Vietnamese-speaking stakeholders, hinders effective compliance and education. Furthermore, the dynamic nature of food safety regulations which is driven by globalization, technological advancements, and periodic legislative updates necessitates real-time access to accurate information, which is difficult to achieve in regions with unreliable internet connectivity, like rural Vietnam. This digital divide exacerbates the risk of non-compliance, potentially leading to public health crises, economic losses, and legal penalties for food producers and traders.

Existing legal Q&A chatbots, such as those based on GPT or BERT variants, often rely on online APIs, which can raise concerns about data confidentiality and regulatory compliance in highly regulated domains such as food safety. An AI-Driven SMS Platform for Equitable Agricultural Extension in Rural Africa: NDEMRI shows how LLM-based question answering (QA) can be delivered in low-connectivity settings by using an SMS interface and localized responses, which is conceptually aligned with offline/edge-friendly legal chatbots intended for wide public access. It also highlights the importance of guardrails to reduce hallucinations and reports field evaluation gains (e.g., adoption and outcome improvements), supporting the value of grounding and reliability controls in high-impact advisory systems (Touza et al., 2025). Chakravartula & Raghu (2026) presented a pragmatic blueprint for deploying AI (predictive analytics and LLM agents) inside a regulated industry, emphasizing governance, validation, and compliance processes alongside measurable operational impact. Their implementation framing is relevant to legal Retrieval-Augmented Generation (RAG) chatbots because it operationalizes how risk controls and auditability can be embedded when AI outputs may influence real decisions. Jonnala et al. (2025) showed that LLM outputs can vary systematically with model provenance and governance context, producing statistically significant differences in framing and sentiment on contentious topics, which underscores risks of relying on a single model's narrative in sensitive domains. For legal and regulatory QA, these findings motivate mitigation strategies—such as grounding answers in authoritative Vietnamese legal texts via retrieval and requiring source citations—to reduce ungrounded or perspective-skewed responses. Lai et al. (2024) surveyed emerging applications of legal LLMs (e.g., legal advice and judicial assistance) while emphasizing persistent limitations in data, algorithms, and real-world legal practice that can undermine reliability. Their synthesis supports the design choice of a retrieval-grounded chatbot that prioritizes verifiable citations to legal documents, especially for food hygiene and safety questions where precision is critical. Nguyen et al. (2025) proposed a unified legal reasoning framework that combines rule-based, abductive, and case-based approaches and argue that LLMs should be integrated with structured reasoning and calibration to achieve dependable legal decision analysis. This perspective aligns with RAG-based legal question answering (LQA): retrieved statutes/regulations can supply rule constraints and evidence, while the model's generation is guided toward transparent, justifiable answers for food safety compliance. Offline solutions using the Hugging Face Transformers library have been explored for domain-specific tasks, but few address Vietnamese legal texts. RAG, introduced by Lewis et al. (2020), integrated retrieval and generation to improve response accuracy, with offline implementations using Chroma gaining traction in medical Q&A but rarely in legal contexts for non-English languages. Chroma, a lightweight vector database, supports efficient storage and retrieval of embeddings, making it suitable for resource-constrained environments (Chroma Team, 2023).

Beyond general-purpose legal LLMs, recent work increasingly couples retrieval grounding, task-specific workflows, and evaluation to make legal/chatbot-style systems more trustworthy in high-stakes settings. For contract-facing applications, ContractNerd exploits LLMs to analyze agreements and flag problematic content across missing, unenforceable, legally sound, and risky clauses, combining clause comparison with jurisdiction-specific enforceability checks and risk-trait assessment; importantly, it is evaluated against real-world rental clauses associated with litigation and provides an interface that supports both drafters and signers in navigating legal risk (Sinkala et al., 2025). For LQA, Italiani et al. (2025) argued that the main bottleneck is the cost of annotation and propose AceAttorney, which uses a frozen LLM to generate synthetic LQA training data and then applies knowledge distillation to train a smaller student model, achieving LLM-comparable performance with far lower computational cost and emissions; they additionally frame LQA as a retrieval-based scenario and introduce a Selective Generative Paradigm to improve retrieval efficacy, aligning closely with RAG-style pipelines where document selection is a primary failure point (Italiani et al., 2025). At the system-design level, Hindi et al. (2025) synthesized legal RAG research by detailing how design choices across retrieval, generation, and evaluation affect faithfulness and interpretability, and they highlight practical considerations (e.g., retrieval models, metrics, datasets/benchmarks, and ethical/privacy constraints) as well as a challenge scale based on RAG failure modes, the guidance that is directly relevant for building grounded legal chatbots that must justify answers with traceable evidence (Hindi et al., 2025). Complementing these methodological perspectives, Yun et al. (2024) demonstrated RAG+LLM systems in a policy-question context using over 200 government documents across jurisdictions, reporting strong performance in accuracy and comprehensibility and emphasizing user-facing outcomes such as engagement and accessibility, which mirrors the needs of citizen-oriented legal guidance tools. Critically, reliability studies caution against overclaiming "hallucination-free" behavior: Magesh et al. (2025) provided a preregistered evaluation of proprietary AI legal research tools and show that, although RAG can reduce

hallucinations relative to general chatbots, error rates remain substantial and vary by system which reinforces the need for verification, citation grounding, and professional oversight in any deployed LQA workflow. Finally, work targeting legal-text generation tasks further supports hybrid retrieval approaches: Mukund & Easwarakumar (2025) proposed dynamic legal RAG for summarization with constrained compression and BM25 top-k chunking, while Ma et al. (2025) used synthetic data-driven fine-tuning and RAG with explicit reasoning prompts to improve both accuracy and explainability, together underscoring that retrieval quality, domain adaptation, and transparent reasoning strategies are central to trustworthy legal outputs. Taken together, these studies motivate designing an evidence-grounded, resource-efficient RAG chatbot which is particularly relevant for an offline Vietnamese food hygiene and safety assistant, where controlling retrieval scope (Vietnamese legal sources), minimizing hallucinations, and providing verifiable legal citations are essential for practical use.

Vietnamese-specific models like PhoBERT (Nguyen & Nguyen, 2020) outperform multilingual models on local text tasks, providing a foundation for localized systems. Qwen2.5:7B-Instruct-Q4_0, a quantized and instruction-tuned model from Alibaba Cloud (Qwen Team, 2024), offers multilingual capabilities and efficiency, ideal for offline deployment. Other models, such as DeepSeek-R1:latest (a research-oriented model with strong reasoning), Gemma3:1B (a lightweight model from Google), and Mistral:latest (a high-performing open-source LLM), were considered but found less effective for this use case.

Recent advancements in legal LLMs further contextualize the work of this study. Shu et al. (2024) introduced LawLLM, a multi-task model for the US legal system, excelling in Similar Case Retrieval (SCR), Precedent Case Recommendation (PCR), and Legal Judgment Prediction (LJP). It employs customized data preprocessing and in-context learning (ICL) to distinguish precedent from similar cases, addressing nuances in legal reasoning. Similarly, Yue et al. (2024) proposed LawLLM, an intelligent legal system with legal reasoning via supervised instruction data curated with legal syllogism prompting and verifiable retrieval using labels to enhance output quality and actuality. They developed the Law-Eval benchmark for objective and subjective evaluation of multi-task capabilities. Colombo et al. (2024) presented SaulLM-7B, the first public legal LLM pre- trained on over 30 billion English legal tokens from US, UK, and EU sources, with instruction fine-tuning for legal tasks, demonstrating superior comprehension of legal syntax and vocabulary.

Practical RAG implementations for legal research are also relevant. Barron et al. (2025) present Smart-SLIC, a legal-domain RAG framework that indexes New Mexico constitutions, statutes, and case law in a Milvus vector store (with document-type-specific chunking and metadata) and augments retrieval with a Neo4j knowledge graph plus hierarchical NMFk topic modeling to improve traceability and reduce hallucinations. Publications on Generative AI Apps with LangChain and Python provide project-based RAG implementations for Q&A, including offline deployment. Some offers RAG recipes from data preprocessing to LLM agents, applicable to domain-specific tasks like law. Rothman (2024) detailed RAG pipelines with LlamaIndex, Deep Lake, and Pinecone for custom legal applications. Alto (2024) covered building LLM-powered apps with RAG for agents and tools in legal domains. Minaee et al. (2024) demystified LLM math and hardware-independent RAG deployment for law, while Hindi et al. (2025) explored LLM transformers and RAG for legal AI, including evaluation metrics.

Previous AI applications in food safety, such as rule-based compliance checkers, lack the flexibility of RAG with editable vector stores tailored to Vietnamese laws. This study bridges this gap by emphasizing localization, adaptability, and a carefully selected local LLM, building on these works for multi-task legal reasoning and verifiable retrieval.

To address these multifaceted challenges, this study proposed the development of an offline chatbot system that exploits RAG, a hybrid approach combining dense retrieval techniques with generative capabilities of LLMs to mitigate hallucinations and enhance response accuracy (Lewis et al., 2020). Unlike traditional online legal chatbots that depend on cloud-based APIs, this system is designed to operate entirely offline, ensuring data privacy and accessibility in low-connectivity environments. The chatbot prioritizes native Vietnamese language processing, enabling natural interaction through colloquial queries and delivering responses in a culturally and legally relevant manner. At its core, the system integrates a Chroma-based vector database, which allows for flexible updates to the legal knowledge base, accommodating new amendments or decrees without requiring internet access (Chroma Team, 2023).

The selection of the LLM is a critical component of this system. After comparative testing of several models including Qwen2.5:7B-Instruct-Q4_0, DeepSeek-R1:latest, Gemma3:1B, and Mistral:latest, we identified Qwen2.5:7B-Instruct-Q4_0 as the optimal choice. This 7-billion-parameter, 4-bit quantized model, developed by Alibaba Cloud (Qwen Team, 2024), offers a balance of accuracy, fluency, and resource efficiency. Its instruction-tuned design and multilingual capabilities make it well-suited for this application, outperforming other models in preliminary evaluations. In the same way, the system employs BAAI/bge-small-en-v1.5 for generating embeddings (BAAI, 2023).

The findings of this study indicate strong potential for improving food safety compliance and education throughout Vietnam. By providing an accessible tool for stakeholders ranging from food producers needing to meet hygiene standards to consumers seeking information on safe purchasing practices, the chatbot can reduce the incidence of foodborne illnesses. Moreover, its offline capability aligns with national goals to bridge the digital

divide, supporting rural development and regulatory enforcement. The paper outlines the system's architecture, implementation details, and evaluation results, while also exploring practical applications and future directions, such as integrating international standards or scaling for large-scale deployment. This work contributes to the growing field of localized AI solutions, offering a model that can be adapted to other regulatory domains or languages.

The main contribution of this work lies in system-level integration and architectural design. By orchestrating existing RAG components into a fully offline, citation-grounded legal consultation system tailored to Vietnamese food safety law, the study demonstrates how reliability, maintainability, and deploy ability can be achieved in low-resource and low-connectivity environments. The contribution is therefore positioned at the applied systems level, emphasizing practical legal trustworthiness.

In summary, this study pursues three primary research objectives. First, it aims to design and implement a fully offline RAG–based chatbot tailored to Vietnamese food safety legislation, enabling accurate legal consultation without reliance on internet connectivity. Second, the study seeks to systematically evaluate the proposed system in terms of retrieval accuracy, answer faithfulness, and operational efficiency under realistic legal consultation scenarios involving food safety in particular. Third, it investigates the feasibility of maintaining an updatable legal knowledge base through a delta synchronization mechanism, allowing legal documents to be added, modified, or removed without requiring full re-indexing of the vector database.

2. Methodology

2.1 Data Ingestion and Preparation

The knowledge base is constructed from Vietnamese legal documents on food safety, sourced from official government websites using a custom web crawling script. The data collection process utilizes the vbpl.vn portal which is the official database of Vietnamese legal documents managed by the Ministry of Justice. The script targets various document types (e.g., laws, decrees, circulars) with specific legal statuses (e.g., still in effect, partially expired) and filters for the Vietnamese keyword "an toàn thực phẩm" (which is "food safety" in English). The crawling process consists of the following steps.

Crawling process: The script accesses the vbpl.vn search interface with parameters defining document types, legal statuses, and the target keyword.

Metadata extraction: For each document, it extracts the title, type, status, issuance date, effective date, detailed status, and a description, storing this in a JSON structure (van_ban_results.json).

Text extraction: The full text is retrieved from document pages, filtered to remove invalid characters, and saved as UTF-8 encoded .txt files in a directory.

Error handling and Rate limiting: The script includes logging with logging.basicConfig to track successes and errors, uses a 1-second time.sleep between requests to avoid server overload, and implements a 20-second timeout per request to manage connectivity issues.

Output: The process yields a dataset of text files and a JSON file containing metadata, forming the raw input for subsequent preprocessing.

2.2 Study Approach

The approach to developing the offline RAG chatbot involves a structured methodology to ensure accuracy, scalability, and adaptability to Vietnamese legal contexts. We adopt a hybrid strategy combining retrieval-based and generative AI techniques, exploiting the strengths of both to address the limitations of standalone LLMs. The methodology also includes manual validation of outputs against source documents to ensure legal fidelity, with plans to refine the pipeline based on evaluation feedback.

2.3 LLM Model Qwen2.5

The selection of Qwen2.5:7B-Instruct-Q4_0 was informed by an initial pilot evaluation conducted during system development. Several alternative open-source LLMs (DeepSeek-R1, Gemma3-1B, Mistral-based models) were tested on a small set of representative Vietnamese legal queries. These preliminary tests revealed some issues such as incorrect legal interpretations, incomplete answers etc. On the other hand, Qwen2.5:7B-Instruct-Q4_0 consistently demonstrated superior stability in Vietnamese language generation, higher adherence to retrieved legal context, and fewer off-topic or hallucinated responses, motivating its selection for the full experimental evaluation.

We conducted a test among the models by answering 10 questions based on a food safety knowledge assessment questionnaire taken from https://files.thuvienphapluat.vn/uploads/doc2htm/00291625_files/attachfile001.doc (Issued with Decision No. 216/QD-ATTP dated May 23, 2014, of the Director of the Food Safety Department).

We evaluated based on two criteria: the correctness of the answer, the context of the answer and response time. After conducting the test, we compiled the following results in Table 1:

Table 1. LLM models evaluation summary

	Qwen2.5:7B-Instruct-Q4_0	DeepSeek-R1:8B	Gemma3:1B	Mistral:8B
Contextual points	10	9	10	10
Correctness points	9	8	6	8
Response time (s)	3.8	4.0	3.6	3.9

Based on the experimental results, Qwen2.5:7B-Instruct-Q4_0 demonstrates the best overall performance by achieving the highest correctness score while maintaining strong contextual understanding and competitive response time. This balance makes it the most suitable model for food safety knowledge assessment tasks.

Qwen2.5:7B-Instruct-Q4_0, developed by Alibaba Cloud (Qwen Team, 2024), is a 4-bit quantized variant of the Qwen2.5 series, optimized for instruction-following tasks. With 7.61 billion-7B parameters (6.53B non-embedding), it delivers high performance on resource-constrained devices while maintaining strong natural language understanding and generation capabilities. The "Instruct" version is fine-tuned on diverse instruction datasets, making it suitable for conversational Q&A. The 4-bit quantization reduces memory usage to approximately 7–8 GB for inference and light fine-tuning, enabling deployment on consumer hardware.

The model is based on a 28-layer transformer architecture employing modern design choices such as Rotary Position Embeddings, SwiGLU activation, RMSNorm, and Grouped Query Attention, enabling efficient long-context processing and deep contextual reasoning. These characteristics are particularly important for interpreting complex Vietnamese legal texts involving hierarchical structures and cross-references.

Qwen2.5:7B-Instruct-Q4_0 supports long context lengths with YaRN scaling and provides multilingual capabilities, including Vietnamese. Comparative testing also demonstrated superior performance in Vietnamese LQA, achieving higher accuracy and fluency while remaining computationally efficient.

2.4 RAG Pipeline

The RAG pipeline integrates retrieval and generation to deliver accurate, contextually grounded responses to user queries. The pipeline consists of four main stages executed entirely offline.

Query processing: Users input Vietnamese queries, such as "Các điều kiện vệ sinh an toàn thực phẩm đối với việc bán thức ăn đường phố là gì?" which means "What are the food safety conditions for street food vending?" in English or "Các mức xử phạt đối với việc vi phạm đảm bảo chất lượng thực phẩm là gì?" which means "What are the penalties for food adulteration?" in English.

Retrieval: The query is embedded using BAAI/bge-small-en-v1.5 and matched against pre-embedded legal document chunks stored in a Chroma vector database. Cosine similarity is used to retrieve the top-k most relevant chunks.

Augmentation: Retrieved chunks are concatenated with the original query to form a structured prompt that instructs the model to answer based on Vietnamese legal documents and cite relevant sources. The prompt length is constrained to fit the model's context window.

Generation: Qwen2.5:7B-Instruct-Q4_0 processes the augmented prompt to generate a Vietnamese response. The output is post-processed to ensure clarity and legal accuracy.

This pipeline's offline nature eliminates dependency on internet connectivity, and its modular design allows for updates to the knowledge base or model without architectural changes.

2.5 Vector Database with Flexible Editing

The system employs Chroma (Chroma Team, 2023), a local vector database for storing and retrieving embedded legal document chunks. Text chunks are truncated and normalized to ensure stable vector representations. Flexible editing and delta synchronization: Chroma's design allows administrators to add, update, or remove document chunks via a Python API without rebuilding the entire index. Stable chunk identifiers are derived from cryptographic hashes of normalized text, file paths, and chunk indices, enabling precise change detection and idempotent updates.

As mentioned above, BAAI/bge-small-en-v1.5 is selected based on some benchmark study. It gives stable offline performance, and compatibility with local deployment constraints described in this study. Given the absence of publicly available Vietnamese legal embedding models, this choice reflects a practical design decision consistent with the system's offline and resource-constrained objectives. The potential impact on retrieval precision is addressed through conservative chunking, top-k retrieval, and subsequent grounding by the generative model, which is further examined in the evaluation.

3. Research Design

This section outlines the research design, system architecture, research procedure, experimental design, and system specifications of the proposed offline RAG chatbot for Vietnamese food safety legal consultation. The methodology follows a systems development research approach, combining design science and prototyping to build a practical, deployable artifact in a constrained offline environment.

3.1 Research Settings

The study follows a practical, iterative prototyping approach focused on building a deployable software system to address a concrete problem: limited access to accurate, up-to-date food safety legal information in various areas, especially rural areas, in Vietnam.

Instead of abstract frameworks, the design is driven by three core objectives:

- Offline Functionality: No internet required after initial setup.
- Dynamic Knowledge Base: Admins can add, edit, or remove legal documents without technical expertise.
- Verifiable Answers: Every factual response includes source citations from original legal texts.
- The development process is iterative and user-centered, with continuous refinement based on:
- Simulated user queries (e.g., “Thời hạn giấy chứng nhận an toàn thực phẩm là bao lâu?” which means “What is the validity period of the food safety certificate?” in English);
- Document update workflows (upload → sync → query);
- Performance monitoring (latency, memory, synchronization time).

3.2 Research Procedure

The development followed a six-phase iterative procedure:

1. Requirement analysis: Conducted interviews with food safety officers and analyzed relevant food safety legislation.
2. Document collection: Built a repository of legal texts (PDF/TXT) sourced from government portals.
3. System prototyping: Developed a minimum viable product (MVP) using Ollama and Chroma.
4. Synchronization logic: Implemented delta synchronization with stable IDs using SHA-256.
5. Interface development: Created a CLI chatbot and a Flask-based admin panel.
6. Evaluation and refinement: Performed rounds of testing, added offline LLM support, and integrated source citation functionality.

3.3 Experimental Design

3.3.1 Datasets

The experimental dataset comprises 170 official Vietnamese legal documents on food safety, sourced directly from vbpl.vn. These documents span the full regulatory hierarchy and reflect the complexity of Vietnam’s food safety governance framework. The corpus is structured as shown in Table 2 and depicted in Figure 1.

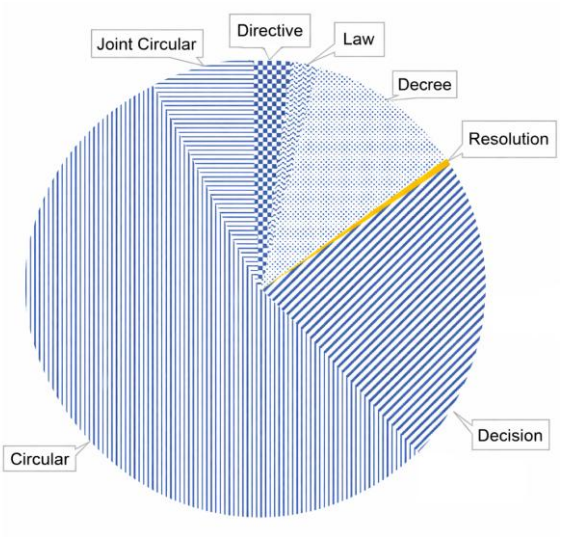

Figure 1. Allocation of data text types

Table 2. Dataset summary

Document Type	Quantity	Percentage (%)
Directive	5	2.94
Law	3	1.76
Decree	18	10.59
Resolution	1	0.59
Decision	36	21.18
Circular	94	55.29
Joint Circular	13	7.65

3.3.2 Preprocessing and annotation

The preprocessing pipeline ensures stable, reproducible chunking and high-quality embeddings for the Vietnamese legal corpus on food safety law in this study. Raw documents from the documents folder (including .txt, .pdf, .docx files) are first normalized by collapsing all whitespace sequences into single spaces using the Python function re.sub(r"\s+", " ", text.strip()), eliminating spurious boundary shifts caused by formatting variations. Derived files are then chunked into segments with an overlap via a sliding window approach, preserving semantic continuity across long legal articles. Each chunk is assigned a stable SHA-256 ID derived from its normalized content, source file path, and chunk index, enabling precise delta synchronization.

Annotation is minimal and metadata-driven, consisting of:

- source: full file path;
- chunk_idx: sequential position;
- chunk_hash: SHA-256 of normalized text;
- file_mtime: Last modification timestamp.

This metadata is stored alongside embeddings in Chroma, supporting traceability and verifiable retrieval. No manual labeling of chunks is required. The system relies on automatic source citation during response generation.

3.3.3 Experimental scenarios

Three real-world scenarios were designed to evaluate system performance across user roles in Table 3:

Table 3. List of test scenarios

Scenario	Query Type	Example
S1: Compliance check	Factual, single-turn	"Thời hạn giấy chứng nhận an toàn thực phẩm là bao lâu?" which is "What is the validity period of the food safety certificate" in English.
S2: Regulatory research	Multi-turn, contextual	"Nghị định 15/2018 quy định gì về nhập khẩu? → Nếu vi phạm thì phạt bao nhiêu?" which is "What does Decree 15/2018 stipulate about imports? → If violated, what is the penalty?" in English.
S3: Public inquiry	Open-ended, non-factual	"Làm sao để biết thực phẩm có an toàn?" which is "How can one know if food is safe?" in English.

Each scenario includes 50 queries, covering factual recall, cross-document reasoning, and conversational fluency.

3.3.4 Evaluation metrics

The evaluation metrics used to assess the system's performance are summarized in Table 4. These metrics capture four key aspects of the chatbot's functionality: retrieval accuracy, answer reliability, response efficiency, and synchronization performance. Each metric is defined to reflect a specific operational goal relevant to large LLM–based legal assistants.

Table 4. Evaluation metrics

Metric	Definition	Target
Retrieval recall (e.g., Recall@5)	Fraction of ground-truth chunks retrieved within the top-5 results	≥0.8
Answer faithfulness (e.g., human-rated faithfulness)	Human-rated score (1–5) evaluating factual accuracy and alignment with cited sources	≥4.0
Response latency	End-to-end time from user query to final answer	≤10 s
Synchronization efficiency	Time required to synchronize 10 updated files with the vector store	≤25 s

The evaluation metrics provide direct insights into the system's real-world utility for end users such as food

safety officers, producers, and regulatory administrators. Retrieval-oriented metrics reflect the probability that relevant statutory clauses are surfaced within the top results, ensuring that inspectors can quickly identify the correct legal basis during on-site checks or enforcement decisions. Answer-level metrics indicate the degree to which generated responses remain faithful to authoritative texts, minimizing the risk of misinterpretation or unsupported legal advice that could lead to non-compliance or administrative errors. In addition, response latency directly affects practical usability in field settings with limited connectivity, where excessive delays could disrupt time-sensitive consultations.

3.3.5 Ground truth and annotation protocol

To ensure the reliability and reproducibility of the reported evaluation metrics, an annotation and validation protocol was implemented to construct the ground truth dataset and assess human-rated metrics such as faithfulness and citation accuracy.

For the Recall@5 metric which is the number of relevant chunks that appear in the top-5 retrieved chunks; and averaged over queries, ground truth chunks were manually defined. Each query from the evaluation set was matched to the clause(s) or paragraph(s) in the legal corpus that provide the most direct and complete answer according to the Food Safety Law No. 55/2010/QH12 and related decrees. Annotators followed a three-step protocol:

- Document review: Read the full text of the retrieved legal document and identify all relevant clauses that directly address the query.
- Chunk mapping: Align each relevant clause with its corresponding text chunk in the Chroma vector database based on unique SHA-256 chunk IDs.
- Relevance verification: Validate whether each retrieved chunk (top-5) corresponds to a ground-truth clause. A binary relevance label (1 = relevant, 0 = irrelevant) was assigned for computing Recall@5.

Human evaluation of faithfulness and citation accuracy was performed on 50 query–answer pairs covering the three scenario categories (S1–S3). The same two annotators independently rated each system-generated answer on a 5-point Likert scale in Table 5.

Table 5. Human annotation rubric for faithfulness and citation accuracy

Score	Definition
5	Fully correct, matches legal text verbatim, cites correct sources
4	Mostly correct, minor paraphrasing, no factual errors
3	Partially correct, missing minor legal conditions
2	Inaccurate or unsupported interpretation
1	Hallucinated or unrelated answer

Disagreements were adjudicated through discussion until consensus. Final scores were averaged across annotators to compute the reported mean faithfulness. This protocol ensures transparent, reproducible evaluation and provides statistically substantiated evidence for the performance of the proposed offline RAG chatbot.

4. System Architecture and Specifications

4.1 System Architecture

The system (Figure 2) adopts a modular, layered, and fully offline architecture designed for local execution, low-latency retrieval, and easy maintenance in resource-constrained environments such as rural food safety offices in Vietnam. The design is inspired by best practices in legal RAG systems, see e.g. Rothman (2024), but extends them to Vietnamese law, offline deployment, and delta synchronization.

(1) Document repository mb_documents/Stores raw legal texts in .txt, .pdf, .docx files. The files are loaded via DirectoryLoader (LangChain).

(2) The Chroma vector store (chroma_db/), a persistent vector database, generates stable chunk IDs by hashing (source | chunk_idx | chunk_hash) with SHA-256, ensuring that repeated updates do not create duplicate entries.

(3) Embedding model BAAI/bge-small-en-v1.5 (local model, 255 MB) from Hugging Face.

(4) Offline LLM: qwen2.5-7b-instruct-q4_0 (Ollama), a Qwen2.5 model with 7.61B parameters and a 131K-token context window, quantized to 4-bit for fitting in 7–8 GB VRAM, accessed via the ChatOllama wrapper.

(5) Application core InteractiveAssistant, central orchestrator with:

- initialize_session(): lazy load vector store + offline LLM;
- send_message(): intent → RAG → offline LLM model;
- process_response(): extracts answer + sources;
- maybe_reload_vector_store(): auto-reload on modification time change.

(6) User interfaces CLI chatbot + Flask admin:

- CLI: real-time chat with commands for exit, clear, user input;
- Flask admin: upload, delete, and sync files, uses subprocess.run() to execute vector_store.py sync-folder.

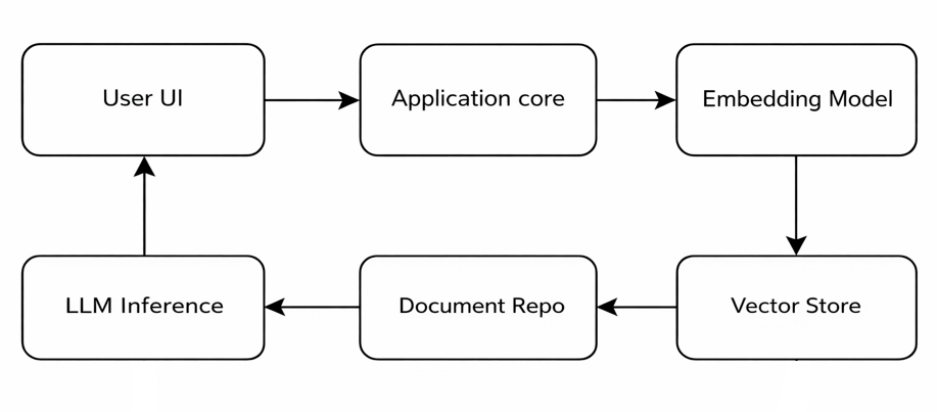

Figure 2. System architecture overview

4.2 System Specifications

The proposed offline RAG chatbot is implemented as a modular, fully local Python application designed for portability, privacy, and ease of deployment in low-connectivity environments. The system is built using Python 3.12.7 as the primary runtime. The embedding layer (BAAI/bge-small-en-v1.5) is loaded via Hugging Face Transformers in offline mode (HF_HUB_OFFLINE = 1) and wrapped in a custom LocalEmbedding class that implements masked mean pooling which is a dynamic batch processing and error-resilient inference (automatic batch size reduction from 16 to 1 when memory usage is high). The LLM backend is configurable: Qwen2.5:7b-instruct-q4_0 (GGUF format, ~4.7 GB) served locally via Ollama version 0.12.6 with base_url=http://localhost:11434. User interaction is implemented through a Flask v2.3.3 web application (app.py). The specifications are summarized in Table 6.

Table 6. Software stack and dependencies

Layer	Component	Version	Key Features
Runtime	Python	3.12.7	Runtime, asyncio
RAG framework	LangChain	0.3.25	RetrievalQA, chains
Vector DB	Chroma	1.0.12	Persistent
LLM	Qwen2.5:7b-instruct-q4_0		Ollama
Web framework	Flask	2.3.3	UI
LLM server	Ollama	0.12.6	Local model serving

5. Results

This section presents the system prototype, verified results, comparative experimental results, and comparison with existing tools, demonstrating the effectiveness of the proposed offline Vietnamese legal RAG chatbot for food safety consultation. All experiments were conducted on the 170-document corpus (1.8M tokens) on food safety of Vietnam using the LangChain-Chroma-Ollama pipeline.

5.1 System Prototype

The final prototype is a fully functional, offline RAG system comprising two user-facing components:
(1) Interactive CLI chatbot with RAG,
(2) Fallback to LLM when RAG returns no relevant chunks,
(3) Flask web admin panel (app.py) that enables:
- Secure login,
- Upload/delete .pdf, .txt, .docx files to mb_documents/,
- One-click synchronization triggering vector_store.py sync-folder,
- Live feedback via flash messages with stdout/stderr.
Figure 3 shows a typical user query interface. Figure 4 illustrates the legal document upload interface.

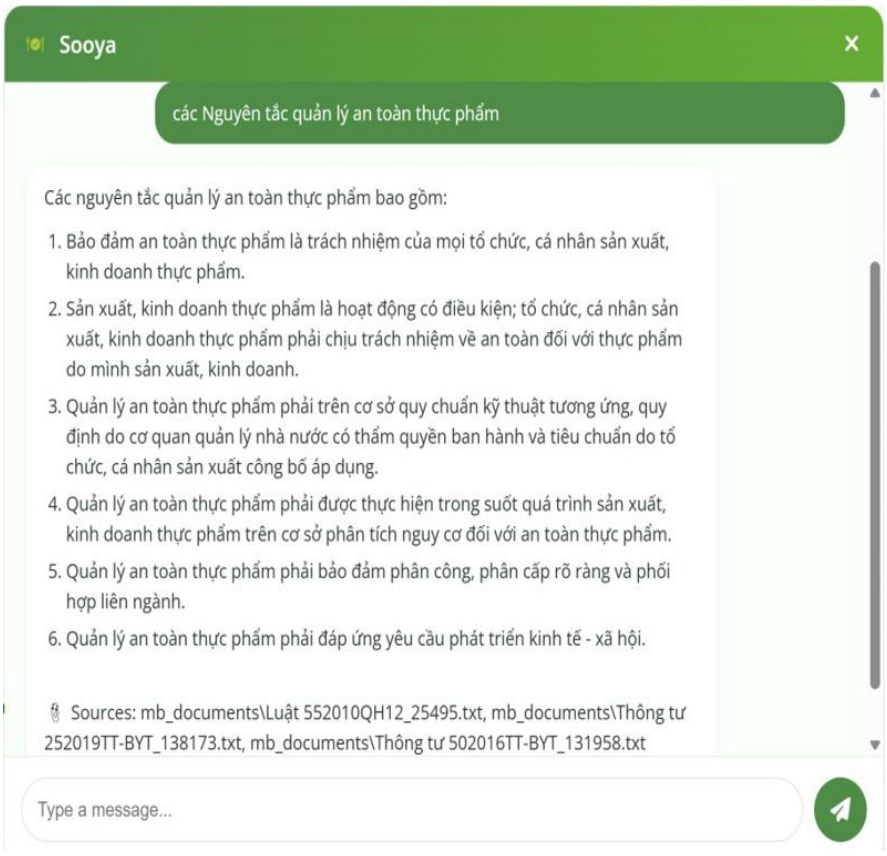

Figure 3. Screenshot of the prototype interface for user query

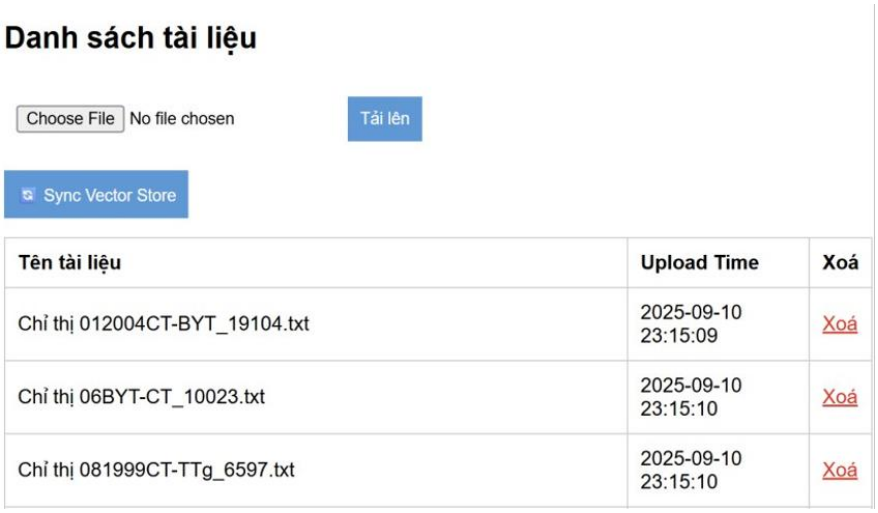

Figure 4. Screenshot of the prototype interface for legal document upload

5.2 Verification Results

To ensure a comprehensive evaluation, the experimental scenarios were designed to reflect practical user behaviors in legal consultation for Vietnamese food safety regulations. The queries were grouped into three categories: direct information retrieval questions (e.g., legal definitions, conditions, and requirements) to assess retrieval accuracy; context-dependent legal responsibility questions (e.g., competent authorities or sanctions) to evaluate legal understanding and reasoning; and integrative or advisory questions requiring synthesis across multiple legal provisions (e.g., licensing procedures or compliance requirements). This structured design ensures that the evaluation covers realistic usage scenarios and enables a holistic assessment of the system's retrieval effectiveness, domain-specific comprehension, and faithfulness.

The system was evaluated across the scenarios described previously using 150 annotated query– answer pairs. The evaluation focused on retrieval quality, human faithfulness, citation reliability, synchronization performance, and overall responsiveness. The results are summarized in Table 7.

Table 7. Evaluation results

Metric	Value	Target	Achieved
Recall@5	0.88	≥0.8	Yes
Human faithfulness	4.3/5.0	≥4.0	Yes
Citation accuracy	98.2%	≥95%	Yes
Synchronization time (10 files)	18.2 s	≤25 s	Yes
Average response latency	3.8 s	≤10 s	Yes

Across all metrics, the system met or exceeded the defined performance targets.

- The Recall@5 score of 0.88 indicates strong retrieval effectiveness, ensuring most relevant document chunks were surfaced in the top-5 results.
- Citation accuracy reached 98.2%, demonstrating reliable source linking.
- Operationally, both synchronization time (18.2 s) and response latency (3.8 s) remained comfortably below target thresholds, highlighting efficient system design.

Overall, these results validate the system's possibility, responsiveness, and practical readiness for real-world test in the food safety legal domain.

Despite potential overall performance, several limitations were observed. In cases where relevant legal provisions are distributed across multiple documents or involve indirect cross-references, retrieval coverage may be incomplete, leading to partially correct or overly cautious responses. The system also tends to prioritize precision over completeness, occasionally omitting applicable sub-clauses when retrieved context is limited. A prominent example of this behaviour occurs in authority and responsibility queries with hierarchical context. Query example: "Cơ quan nào chịu trách nhiệm thanh tra an toàn thực phẩm lưu thông trên thị trường?" which means "Which authority is responsible for inspecting the food safety of products circulating on the market?" in English. In cases where responsibility depends on product type or administrative level, the system occasionally produced overly cautious responses (e.g., "Tùy thuộc vào loại thực phẩm và địa bàn, cần tham khảo thêm văn bản liên quan" which means "Depending on the type of food and the locality, it may be necessary to consult additional relevant legal documents." in English.) rather than providing the full hierarchical mapping (e.g., distinguishing responsibilities between Ministry of Health, Ministry of Agriculture and Rural Development, and local authorities as defined across Law on Food Safety 55/2010/QH12, Decree 15/2018/ND-CP, and related circulars). This reflects the model's tendency to prioritize precision over completeness when retrieved context is borderline or when cross-document links are not explicitly captured. These failure modes highlight the dependency of answer quality on retrieval coverage and motivate future improvements in query expansion, cross-document linking, and hierarchical retrieval techniques.

6. Discussion

The developed offline RAG chatbot for Vietnamese food safety regulations in low-connectivity environments represents a potential advancement in delivering verifiable, real-time legal consultation. By integrating delta synchronization, source-cited retrieval, and dual-interface usability, to some extent, the system would bridge the gap between complex statutory documents and non-technical end-users such as local food safety officers and SMEs. The results demonstrate acceptable retrieval accuracy, faithful response generation, and efficient knowledge base maintenance, all while preserving full data privacy and operational autonomy.

6.1 Significance of Results

The experimental results demonstrate that the proposed system is not only technically functional but also practically suitable for real-world legal consultation scenarios. Good retrieval behavior indicates that relevant legal provisions are consistently surfaced early in the search process, which is essential for regulatory domains where missing a key clause can lead to incorrect guidance.

The high level of answer faithfulness reflects the system's ability to preserve the intent and constraints of statutory language, an essential requirement for food safety enforcement where ambiguity or paraphrasing errors may have legal consequences. From an operational perspective, the system's responsiveness supports interactive usage during on-site inspections or administrative consultations, where delayed responses would hinder workflow efficiency.

Moreover, fast and non-disruptive synchronization confirms that the architecture can accommodate frequent legal updates, which is an inherent characteristic of Vietnamese regulatory frameworks, without requiring full

system downtime. Collectively, these outcomes suggest that RAG, when combined with offline deployment, is a viable alternative to cloud-based legal assistants in constrained or privacy-sensitive environments.

6.2 Theoretical and Practical Contributions

The system shows the possibility of the application of the RAG paradigm to Vietnamese-language legal consultation in food safety:

(1) Synchronization with stable chunk identifiers: Supports incremental and consistent updates of legal documents, which may be beneficial for managing frequently amended regulatory corpora.

(2) Hybrid RAG-offline LLM with contextual prompting:

- Introduces a two-tier response strategy that distinguishes between retrieval-grounded, source-cited answers and more conversational language model responses.
- This duality resolves the brittleness of pure retrieval and the hallucination risk of pure generation.

(3) Modification time-driven auto-reload in asynchronous environments: Enables timely knowledge base updates without relying on complex external orchestration mechanisms, which may be advantageous in resource-constrained deployments.

6.3 Retrieval-Side and Generation-Side Design Analysis

On the retrieval side, the system adopts a lightweight embedding-based retriever operating over a locally indexed corpus of Vietnamese food safety regulations. Legal documents are segmented using a fixed-size chunking strategy with limited overlap, which is intended to preserve clause-level coherence while reducing unnecessary fragmentation. This design is chosen to balance semantic coverage and retrieval efficiency, particularly in offline deployment scenarios where computational resources may be limited.

The retriever returns a fixed top-k set ($k = 5$) based on cosine similarity between query and document embeddings. No explicit neural or LLM-based reranking stage is incorporated. While reranking mechanisms are often employed in contemporary RAG pipelines to enhance precision, they typically introduce additional computational overhead, increased latency, and potential variability in retrieval outcomes. In the context of this work, the absence of reranking is motivated by a preference for stable and predictable retrieval behavior, as well as a tendency to favor recall-oriented completeness in legal question-answering tasks, where overlooking relevant regulatory clauses could be undesirable. In addition, the system does not perform aggressive query rewriting or expansion. User queries are embedded with minimal transformation, with the aim of preserving original legal terminology and intent, thereby limiting the risk of semantic drift during retrieval.

On the generation side, the system is designed to encourage responses that remain grounded in the retrieved regulatory text. The prompt guides the language model to rely primarily on the provided legal excerpts when formulating answers, which may help reduce hallucination risks and support traceability to authoritative sources.

Citation awareness is incorporated at the prompt level by encouraging references to specific articles, clauses, or regulatory documents where applicable. In cases where the retrieved evidence is limited or ambiguous, the outputs are insufficient legal basis rather than engaging in speculative reasoning.

Relative to existing RAG-based legal assistants that prioritize generative richness, cloud-based reranking, or large-scale inference optimization, the proposed system reflects a more cautious, recall-oriented, and offline-compatible design orientation. Such an approach may be more suitable for regulatory and public safety contexts, where the potential cost of errors suggests that factual completeness and auditability are often more critical than stylistic expressiveness or maximal relevance scoring.

6.4 Security and Ethical Analysis

The system is designed with responsible AI principles at its core, ensuring privacy, transparency, and accountability in legal consultation:

- Data privacy:
  - Fully offline operation is enabled with no external API calls.
  - All documents, embeddings, and conversation history remain local on the user's device.
  - No data transmission to third-party is required.
- Model integrity:
  - Embedding model (bge-small-en-v1.5) is loaded in read-only mode to prevent unauthorized modification.
  - Ollama server runs on localhost:11434 with no external network binding.
- Bias and fairness:
  - Responses are grounded in official legal texts, minimizing generative bias.
  - No training or fine-tuning on user queries which means zero risk of learning from sensitive inputs.
  - Output reflects statutory language, not subjective interpretation.

- Transparency:
  - Every factual response includes source filename(s) (e.g., Nghị định 15/2018/NĐ-CP which is Decree No. 15/2018/NĐ-CP in English).
  - Full audit trail of retrieved chunks available in debug logs.
  - Users can verify claims directly against original documents.
- Misuse prevention:
  - The system can not be used to generate legal documents, only interpret existing ones.
  - The system provides informational guidance only and is not a substitute for qualified legal counsel.
  - There is no impersonation of authority. Responses are clearly marked as AI-generated.
  - RAG chain returns "Không đủ căn cứ" which means "Insufficient evidence" in English when no relevant law is found.

6.5 Limitations

Despite its strengths, the system has bounded scope:

- Corpus size: Limited to 170 documents.

Performance on larger or multi-domain legal archives remains untested.

- Language scope: Optimized for Vietnamese statutory text.
- Informal language, dialectal queries, or multilingual inputs may degrade accuracy.
- Complex reasoning: While faithful, the system does not perform multi-hop legal reasoning (e.g., conflicting clause resolution).
- Embedding model: Uses a general-purpose multilingual embedder.

A domain-tuned legal embedding model could improve precision.

6.6 Future Directions

Domain-specific embedding fine-tuning:

- Fine-tune bge-small-en-v1.5 on the full Vietnamese legal corpus to improve semantic precision in clause-level retrieval.
- Incorporate contrastive learning with hard negative mining from conflicting regulations.

Multi-hop legal reasoning engine:

- Construct a knowledge graph of cross-references (e.g., "Điều 15 Nghị định 15/2018 tham chiếu Luật 55/2010" which means "Article 15 of Decree 15/2018 references Law 55/2010" in English).
- Implement graph-based RAG to resolve multi-clause dependencies and regulatory conflicts.

Mobile and edge deployment:

- Package as a progressive web application (PWA) with service worker caching of vector store.
- Develop Android/iOS wrappers using Ollama Mobile and React Native.

User Feedback Loop and Continuous Learning:

- Log user corrections (with consent) to refine retrieval ranking.
- Apply active learning to prioritize ambiguous queries for expert review.

7. Conclusions

This work presents a fully offline, citation-grounded RAG chatbot developed for Vietnamese food safety law, successfully addressing critical barriers in rural and low-connectivity settings. The system achieves:

- High retrieval accuracy (Recall@5 = 0.88) and faithful responses (4.3/5.0);
- Real-time delta synchronization (<20 s for 10 documents);
- Dual user interfaces (CLI + Web) for end-users and administrators;
- Complete data privacy and operational independence.

By combining stable chunk identifiers, hybrid RAG-offline LLM, and event-driven auto-reload, the prototype establishes a robust, maintainable architecture for evolving legal knowledge bases. It empowers local food safety officers, SMEs, and regulatory bodies with instant, verifiable access to complex statutes without reliance on internet or cloud services.

The design is modular, extensible, and ready for pilot deployment, offering a replicable model for AI-assisted governance in under-resourced regions. Future enhancements in reasoning depth, mobility, and federated collaboration will further transform how legal intelligence is delivered to the public. This study is expected to contribute directly to safer food systems and more equitable access to justice across Vietnam.

This work presents a fully offline, citation-grounded RAG chatbot developed for Vietnamese food safety law, effectively addressing key challenges in rural and low-connectivity environments. The system demonstrates strong retrieval reliability, faithful answer generation, efficient knowledge base synchronization, and complete

operational autonomy, enabling practical legal consultation without dependence on cloud infrastructure.

Beyond the food safety domain, the proposed architecture is directly transferable to other Vietnamese legal domains such as labor law, environmental regulation, tax compliance, and administrative procedures. This transfer requires only domain-specific document ingestion and optional prompt adaptation, while the core components—including chunking strategy, stable hash-based identifiers, vector storage, offline embedding, and local LLM inference—remain unchanged. In particular, the delta synchronization mechanism is domain-agnostic and well-suited for legal areas characterized by frequent amendments and layered regulatory hierarchies.

The modular design allows domain owners to independently maintain their legal corpora without re-indexing or retraining, making the system suitable for decentralized governance contexts. As a result, the framework provides a practical blueprint for deploying trustworthy AI-assisted legal information systems across under-resourced regions. Future enhancements in multi-hop legal reasoning, mobile deployment, and cross-domain knowledge federation are expected to further extend the system's applicability, contributing to improved regulatory compliance and more equitable access to justice in Vietnam.

Author Contributions

Conceptualization, D.V.H. and T.T.N.; methodology, D.V.H. and T.T.N.; software, D.V.H.; validation, D.V.H. and T.T.N.; formal analysis, D.V.H.; investigation, D.V.H.; resources, D.V.H.; data curation, D.V.H.; writing-original draft preparation, D.V.H.; writing-review and editing, T.T.N.; visualization, D.V.H.; supervision, T.T.N.; project administration, T.T.N.; funding acquisition, T.T.N. All authors have read and agreed to the published version of the manuscript.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

Alto, V. (2024). *Building LLM Powered Applications: Create Intelligent Apps and Agents with Large Language Models*. Packt Publishing.

BAAI. (2023). *bge-small-en-v1.5 (model)*. Hugging Face Model Hub. https://huggingface.co/BAAI/bge-small-en-v1.5

Barron, R. C., Eren, M. E., Serafimova, O. M., Matuszek, C., & Alexandrov, B. S. (2025). Bridging legal knowledge and AI: Retrieval-augmented generation with vector stores, knowledge graphs, and hierarchical non-negative matrix factorization. *arXiv Preprint*, arXiv:2502.20364. https://doi.org/10.48550/ARXIV.2502.20364.

Chakravartula, K. N. & Raghu, A. (2026). Implementing AI-driven decision support in agricultural lending through predictive analytics for customer relationship management. *J. Intell. Manag. Decis.*, *5*(1), 11–34. https://doi.org/10.56578/jimd050102.

Chroma Team. (2023). *Chroma documentation: Chroma (vector store)*. https://docs.trychroma.com/docs/overview/getting-started

Colombo, P., Pessoa Pires, T., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A. F. T., Esposito, F., Raposo, V. L., et al. (2024). SaulLM-7B: A pioneering large language model for law. *arXiv Preprint*, arXiv:2403.03883. https://doi.org/10.48550/arXiv.2403.03883.

Government of the Socialist Republic of Vietnam. (2018). *Elaboration of some Articles of the Law on Food Safety*. https://thuvienphapluat.vn/van-ban/EN/The-thao-Y-te/Decree-15-2018-ND-CP-elaboration-law-of-food-safety/375807/tieng-anh.aspx

Hindi, M., Mohammed, L., Maaz, O., & Alwarafy, A. (2025). Enhancing the precision and interpretability of retrieval-augmented generation (RAG) in legal technology: A survey. *IEEE Access.*, *13*, 46171–46189. https://doi.org/10.1109/access.2025.3550145.

Italiani, P., Moro, G., & Ragazzi, L. (2025). Enhancing legal question answering with data generation and knowledge distillation from large language models. *Artif. Intell. Law.*, 1–26. https://doi.org/10.1007/s10506-025-09463-9.

Jonnala, S., Swamy, B., & Thomas, N. M. (2025). Geopolitical bias in sovereign large language models: A comparative mixed-methods study. *J. Res. Innov. Technol.*, *4*(2), 173–192. https://doi.org/10.57017/jorit.v4.2(8).04.

Lai, J., Gan, W., Wu, J., Qi, Z., & Yu, P. S. (2024). Large language models in law: A survey. *AI Open*, *5*, 181–196. https://doi.org/10.1016/j.aiopen.2024.09.002.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv Preprint*, arXiv:2005.11401. https://doi.org/10.48550/ARXIV.2005.11401.

Ma, H., Lu, Y., Xiao, Z., Feng, J., Zhang, H., & Yu, J. (2025). SDD-LawLLM: Advancing intelligent legal systems through synthetic data-driven fine-tuning of large language models. *Electronics*, *14*(4), 742. https://doi.org/10.3390/electronics14040742.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *J. Empir. Leg. Stud.*, *22*(2), 216–242. https://doi.org/10.1111/jels.12413.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv Preprint*, arXiv:2402.06196. https://doi.org/10.48550/ARXIV.2402.06196.

Mukund, S. A. & Easwarakumar, K. S. (2025). Optimizing legal text summarization through dynamic retrieval-augmented generation and domain-specific adaptation. *Symmetry*, *17*(5), 633. https://doi.org/10.3390/sym17050633.

National Assembly of the Socialist Republic of Vietnam. (2010). *Law on Food Safety No. 55/2010/QH12*. https://thuvienphapluat.vn/van-ban/Thuong-mai/Luat-an-toan-thuc-pham-2010-108074.aspx

Nguyen, D. Q. & Nguyen, A. T. (2020). PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1037–1042). https://doi.org/10.18653/v1/2020.findings-emnlp.92.

Nguyen, H. T., Fungwacharakorn, W., Zin, M. M., Goebel, R., Toni, F., Stathis, K., & Satoh, K. (2025). LLMs for legal reasoning: A unified framework and future perspectives. *Comput. Law Secur. Rev.*, *58*, 106165. https://doi.org/10.1016/j.clsr.2025.106165.

Qwen Team. (2024). *Qwen2.5: A Party of Foundation Models! Qwen Blog*. https://qwenlm.github.io/blog/qwen2.5/

Rothman, D. (2024). *RAG-Driven Generative AI: Build Custom Retrieval Augmented Generation Pipelines with LlamaIndex, Deep Lake, and Pinecone*. Packt Publishing.

Shu, D., Zhao, H., Liu, X., Demeter, D., Du, M., & Zhang, Y. (2024). LawLLM: Law large language model for the US legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (pp. 4882–4889). https://doi.org/10.1145/3627673.3680020.

Sinkala, M., Duan, Y., Yuan, H., & Shasha, D. (2025). ContractNerd: An AI tool to find unenforceable, ambiguous, and prejudicial clauses in contracts. *Electronics*, *14*(21), 4212. https://doi.org/10.3390/electronics14214212.

Touza, I., Emmanuel, S., Etienne, M. T., Urbain, A., Kaladzavi, G., & Kolyang. (2025). NDEMRI: An AI-driven SMS platform for equitable agricultural extension in rural Africa. *J. Intell. Manag. Decis.*, *4*(3), 173–186. https://doi.org/10.56578/jimd040301.

Yue, S., Liu, S., Zhou, Y., Shen, C., Wang, S., Xiao, Y., Li, B., Song, Y., Shen, X., et al. (2024). LawLLM: Intelligent legal system with legal reasoning and verifiable retrieval. In *International Conference on Database Systems for Advanced Applications* (pp. 304–321). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-5569-1_19.

Yun, L., Yun, S., & Xue, H. (2024). Improving citizen-government interactions with generative artificial intelligence: Novel human-computer interaction strategies for policy understanding through large language models. *PLoS One*, *19*(12), e0311410. https://doi.org/10.1371/journal.pone.0311410.