Department of Electrical and Computer Engineering
North South University

Senior Design Project

A Multi-Agent RAG system for Legal Information Retrieval in Bangladesh

Md. Sifat Haque Zidan **2212768042**

Abdullah Al Raiyan **2212712042**

Rahil Mehnaz **2121843642**

Faculty Advisor

Dr. Mohammad Rashedur Rahman

Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh

Summer, 2025

LETTER OF TRANSMITTAL

August 2025

To

Dr. Mohammad Abdul Matin

Chairman,

Department of Electrical and Computer Engineering

North South University, Dhaka

Subject: Submission of Capstone Project Report on “A Multi-Agent RAG system for Legal Information Retrieval in Bangladesh”

Dear Sir,

Respectfully, as part of our BSc program, we would like to submit our capstone project report on “A Multi-Agent RAG system for Legal Information Retrieval in Bangladesh.” This paper introduces a Multi-Agent Retrieval Augmented Generation (RAG) system specifically designed for Bangladeshi legal information retrieval. The system employs a novel two-agent architecture: a Clarification Agent that refines user queries through Large Language Model (LLM)-driven clarity assessment and interactive questioning, and a primary RAG Agent. This research demonstrates a scalable and potentially cost-effective approach to democratizing legal knowledge, which is particularly vital in low-resource settings such as Bangladesh.
Please review this report and provide your valuable feedback. We hope it provides a clear understanding of the problem and findings, and we trust it will be informative and helpful.

We will be highly obliged if you kindly receive this report and provide your valuable judgment. It would be our immense pleasure if you find this report useful and informative to have an apparent perspective on the issue.

Sincerely Yours,

Md. Sifat Haque Zidan

ECE Department

North South University

Abdullah Al Raiyan
ECE Department
North South University

Rahil Mehnaz
ECE Department
North South University

APPROVAL

Md. Sifat Haque Zidan (ID # 2212768042), Abdullah Al Raiyan (ID # 2212712042), Rahil Mehnaz (ID # 2121843642) from Electrical and Computer Engineering Department of North South University have worked on the Senior Design Project titled "A Multi-Agent RAG system for Legal Information Retrieval in Bangladesh" under the supervision of Dr. Mohammad Rashedur Rahman, in partial fulfillment of the requirement for the degree of Bachelors of Science in Engineering and has been accepted as satisfactory.

Supervisor's Signature

…….

Dr. Mohammad Rashedur Rahman

Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

Chairman's Signature

…….

Dr. Mohammad Abdul Matin

Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

DECLARATION

This is to declare that this project is our original work. No part of this work has been submitted elsewhere partially or fully for the award of any other degree or diploma. All project related information will remain confidential and shall not be disclosed without the formal consent of the project supervisor. Relevant previous works presented in this report have been properly acknowledged and cited. The plagiarism policy, as stated by the supervisor, has been maintained.

ACKNOWLEDGEMENTS

The authors would like to express their heartfelt gratitude towards their project and research supervisor, Dr. Mohammad Rashedur Rahman, Professor, Department of Electrical and Computer Engineering, North South University, Bangladesh, for his invaluable support, precise guidance, and advice pertaining to the experiments, research and theoretical studies carried out during the course of the current project and also in the preparation of the current report.

Furthermore, the authors would like to thank the Department of Electrical and Computer Engineering, North South University, Bangladesh, for facilitating the research. The authors would also like to thank their loved ones for their countless sacrifices and continual support.

ABSTRACT

A Multi-Agent RAG system for Legal Information Retrieval in Bangladesh

Access to legal information presents a significant impediment in developing nations like Bangladesh, where a substantial portion of the populace lacks legal literacy and access to professional counsel. This paper introduces a Multi-Agent Retrieval Augmented Generation (RAG) system specifically designed for Bangladeshi legal information retrieval. The system employs a novel two-agent architecture: a Clarification Agent that refines user queries through Large Language Model (LLM)-driven clarity assessment and interactive questioning, and a primary RAG Agent. The RAG Agent executes a comprehensive process involving document retrieval from a meticulously curated legal corpus, relevance grading, conditional web search, and final answer synthesis. This corpus, compiled from official government websites, legal blogs, and scholarly articles, ensures accurate representation of the Bangladeshi legal landscape. The multi-agent RAG pipeline, through structured collaboration between these autonomous agents, is designed to improve the relevance and contextual accuracy of responses to complex legal queries. A conversational interface built with Streamlit facilitates accessible, realtime user interaction. The Agentic RAG model outperforms baseline and general-purpose legal bots (using GPT-4o-mini and Gemma 3B) on 20 diverse legal queries, showing better contextual relevance, accuracy, and fewer hallucinations. It also supports cost-effective local deployment via the Ollama framework. This research demonstrates a scalable and potentially cost-effective approach to democratizing legal knowledge, which is particularly vital in low-resource settings such as Bangladesh.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Background and Motivation

In many developing nations, access to justice remains a fundamental issue due to limited awareness of legal rights, scarcity of legal professionals, and high legal costs. Bangladesh is no exception. Despite having a structured legal system, most citizens lack basic legal literacy, particularly in rural and underprivileged regions. Legal issues involving land disputes, family matters, and administrative queries frequently go unresolved not due to a lack of legal infrastructure, but rather because people do not know where or how to seek help.

The traditional process of consulting a lawyer or reading legal texts is often inaccessible to the average person, either because of economic constraints or because legal documents are written in technical language. As a result, there exists a considerable information gap between the legal framework of the country and the people it aims to serve. Bridging this gap requires innovative and scalable interventions, especially given the increasing penetration of digital technology across Bangladesh.

Recent advancements in artificial intelligence, especially in the field of Natural Language Processing (NLP), offer promising opportunities to close this gap. Large Language Models (LLMs) such as OpenAI's GPT [20], Meta's Llama, and Google's Gemma have demonstrated impressive performance in various domains, including healthcare, education, and legal advisory. These models can understand, summarize, and generate human-like text, allowing them to act as digital assistants capable of answering domain-specific questions. However, despite their power, LLMs are prone to hallucinations—confidently producing incorrect information—which can be particularly harmful in the legal context. Furthermore, general-purpose LLMs are not always trained on local laws and regulations, making their outputs less relevant or accurate for region-specific queries.

Retrieval-Augmented Generation (RAG) has emerged as an effective technique [11] to address these limitations. RAG combines the language understanding capabilities of LLMs with a retriever module that fetches relevant documents from a trusted corpus. This approach grounds the generated responses in factual information and improves the reliability of outputs. Building upon this concept, our project introduces a multi-agent RAG-based chatbot system customized for the Bangladeshi legal domain [18].

In addition to the aforementioned challenges, the evolving digital ecosystem in Bangladesh also presents an opportunity to empower citizens through technology-driven legal literacy

initiatives. With the proliferation of smartphones and increasing internet penetration, particularly in urban and semi-urban areas, there is an untapped potential to deliver legal support services digitally. However, digitization alone does not solve the problem; information must be accurate, accessible in natural language, and tailored to the Bangladeshi legal framework. This is where intelligent systems like chatbots, integrated with advanced NLP capabilities, come into play.

Moreover, legal systems are inherently complex and vary significantly across jurisdictions. General-purpose models trained on datasets from Western legal systems often fail to account for the nuances of Bangladeshi law, resulting in culturally or legally irrelevant responses. A domain-specific, locally informed model is essential for effective deployment. Studies in legal NLP show that specialized chatbots tailored to national legal systems yield higher user trust and satisfaction [6][17].

Additionally, legal processes are often procedural and context-dependent. A user's question might lack sufficient detail to produce a reliable answer without clarification. By incorporating a Clarification Agent, the system mimics a real legal advisor's behavior—asking clarifying questions to better understand the problem before attempting to answer. This improves the precision and contextual relevance of the final output. Furthermore, the retrieval component in our system draws from a curated knowledge base, which includes authoritative Bangladeshi sources such as the Laws of Bangladesh [16], Supreme Court rulings [26], and legal commentary websites [3][15], ensuring that the answers are both factual and jurisdictionally appropriate.

Our project not only targets individual users but also aims to assist overburdened legal aid centers and NGOs in Bangladesh that struggle to provide one-on-one guidance to the growing volume of clients. By integrating this multi-agent system into their workflow, these organizations could automate initial screening, deliver preliminary legal information, and flag critical cases for human follow-up. Ultimately, this enhances the efficiency and scalability of legal aid in resource-constrained environments.

The intersection of AI and law, while promising, must be approached with a strong emphasis on reliability, explainability, and ethical responsibility. Our system addresses these by grounding responses in traceable sources, supporting human feedback, and providing architecture-level mechanisms to reduce misinformation and hallucination. Through this work,

we aim to show that modern AI techniques, when carefully adapted to local contexts, can make a transformative impact in ensuring justice and legal empowerment.

1.2 Purpose and Goal of the Project

This project aims to design and implement a Multi-Agent Retrieval-Augmented Generation (RAG) chatbot for legal information retrieval in the Bangladeshi context. The system focuses on:

- Improving query clarity using a Clarification Agent
- Ensuring accurate retrieval from a curated legal knowledge base
- Reducing hallucinations by introducing relevance grading and conditional web search
- Providing responses through a conversational user interface (UI)

The goal is to make legal information more accessible, reliable, and cost-effective for the general public.

1.3 Organization of the Report

This report is structured as follows:

- Chapter 2 reviews related work in legal AI systems.
- Chapter 3 describes how legal data was collected, processed, and prepared.
- Chapter 4 explains the methodology, including the multi-agent architecture and tools used.
- Chapter 5 presents results from automated evaluations and human expert assessments.
- Chapter 6 provides conclusions, limitations, and suggestions for future work.

Chapter 2 Literature Review

The integration of artificial intelligence and natural language processing technologies into the legal domain has gained significant momentum in recent years with various approaches such as improving the accessibility, reliability, and efficiency of legal services. The development of legal chatbots and AI-driven legal assistance systems has been a growing research focus in recent years. Various approaches have been proposed to address the challenges of legal knowledge retrieval, accurate response generation, and user trust. In this section, we will review some related and most relevant works in this domain:

2.1 ChatLaw: A Multi-Agent Collaborative Legal Assistant

Jiaxi Cui et al. [8] introduced a Mixture-of-Experts (MoE) architecture combined with a multi-agent collaboration framework to address hallucination and accuracy issues in legal question-answering systems in the Chinese legal domain. By integrating a retrieval-augmented generation (RAG) model and specialized expert networks based on legal sub-domains, ChatLaw outperformed GPT-4 across multiple benchmarks, emphasizing the role of specialization and grounded retrieval in improving legal LLM performance.

2.2 Building a Legal Dialogue System

Mudita Sharma et al. [9] proposed an architecture leveraging AWS Lex and AWS Lambda to build a legal chatbot capable of assisting users in navigating legal services. Due to the scarcity of public legal dialogue datasets, they collected data via crowdsourcing and user interactions. Their hierarchical bot design first gathered essential user information before providing appropriate legal guidance, demonstrating practical strategies for data collection and bot structuring in lowresource domains.

2.3 The Use of Chatbots in Providing Free Legal Guidance

Ogunsan Isaac et al. [10] critically examined the ethical, legal, and regulatory challenges associated with legal chatbots like DoNotPay. They highlighted the opportunities for increasing access to justice but warned about the risks of misinformation, liability issues, and

the potential for unethical practices. The study called for stronger oversight, transparency, and user protection mechanisms when deploying AI in legal advisory contexts.

2.4 Technology Acceptance Model for Lawyer Robots with AI

Ni Xu et al. [11] extended the classic Technology Acceptance Model to understand the factors influencing user adoption of legal chatbots. The study found that perceived ease of use and usefulness were stronger predictors of trust and acceptance than mere regulatory compliance. It underscores the importance of focusing on user experience and credibility when designing legal conversational agents.

2.5 Improving Access to Justice with Legal Chatbots

Marc Queudot et al. [12] developed two retrieval-based chatbots: one aimed at helping immigrants navigate legal procedures, and another for corporate employees seeking legal compliance information. Their work demonstrated that conversational agents could significantly improve accessibility to specialized legal knowledge, particularly for vulnerable or underserved populations.

2.6 A Chatbot for Specialized Domain

Egidia Cirillo et al. [13] tackled the problem of unstructured legal text by proposing a semi-automated pipeline using generative AI. Their approach involved parsing, indexing, and validating large legal corpora to make them usable for retrieval-augmented generation models. This structuring method is critical for improving the accuracy, relevance, and auditability of legal information retrieval systems.

2.7 Transforming legal text interactions

Mohammed Maree et al. [14] presented the development of a legal chatbot trained on a newly created dataset covering Palestinian cooperative law. The system achieved 82% accuracy in answering legal queries and was validated against human expert judgments. This work illustrates how fine-tuning LLMs with domainspecific and localized datasets can yield effective legal assistance tools in niche legal sectors.

2.8 LAWBOT: A Smart User Indian Legal Chatbot

Nikita et al. [15] introduced LawBot, a legal chatbot designed to simplify access to Indian legal resources. Combining machine learning and IR techniques, LawBot assists users by providing relevant legal advice and references. The project highlighted how technology can help bridge the gap between legal services and common citizens in complex legal environments.

2.9 An Intelligent Conversational Agent for the Legal Domain

Flora Amato et al. [16] introduced CREA2, an intelligent agent designed to help users navigate legal concepts, draft legal documents, and propose solutions in disputes such as divorces, inheritances, and corporate divisions. The agent leverages natural language processing (NLP) and semantic search, utilizing SBERT for evaluating query relevance in an unsupervised manner. CREA2 demonstrates how conversational agents can reduce legal professionals' workloads and improve user access to legal support across digital platforms.

2.10 Legal Solutions - Intelligent Chatbot Using Machine Learning K. R. A et al. [17]

proposed an AI-powered legal chatbot that empowers users with fundamental legal knowledge, real-time attorney consultations, and personalized legal instructions based on location and financial status. By integrating advanced machine learning and NLP techniques, the chatbot autonomously retrieves and processes legal information, aiming to democratize legal resource access and provide rapid, context-specific support to a wide range of users.

2.11 Development of a legal Document AI-Chatbot

Devraj et al. [21] guides beginners in basic legal Chabot development by explaining the processes from the start, gives an idea about LangChain, introduces the technique Cosine Similarity, the design of their bot, frontend backend giving us the idea of what our bot might look like and how LangChain is connected, and how the texts are broken into small chunks are explained easily.

2.12 RAGged Edges: The Double-Edged Sword of Retrieval-Augmented Chatbots

Feldman et al. [23] explores how Retrieval-Augmented Generation (RAG) can counter hallucinations by integrating external knowledge with prompts. Their results show that RAG increases accuracy in some cases but can still be misled when prompts directly contradict the model's pre-trained understanding. In this paper, they tested the effectiveness of context prompting as used in RAG to determine its effectiveness when compared to the same prompt without context, which we applied in our agentic RAG. Their research has shown that in the presence of complex or misleading search results, a RAG system may often get things wrong. A missing section may lead to hallucinations; an unconventional placement of dates may result in the time from one event being attributed to another.

2.13 RAGAS: Automated Evaluation of Retrieval Augmented Generation

For evaluation, es et al. [24] has shown that the predictions from RAGAs are closely aligned with human predictions, especially for faithfulness and answer relevance.

While the studies mentioned above have made significant contributions in the field of legal chatbots, our work introduces a few unique aspects that set it apart. First and foremost, this work focuses on Bangladeshi legal information, which is often overlooked in many global chatbot systems. Unlike the systems discussed earlier, which generally provide legal guidance in a broad sense, this chatbot is specifically designed for Bangladesh's legal context. Another important difference is the use of a multi-agent system, which allows the chatbot to handle more complex, multi-turn conversations with users. It also goes beyond basic legal information by incorporating features like conditional web searches and relevance checking, which ensure that users get responses based on the most up-to-date legal data. Additionally, this chatbot includes clarification prompts, a feature that helps users refine their questions, improving the quality of the legal advice given. These tailored features make this work stand out as a more specialized and region-specific solution, particularly suited to the needs of Bangladeshi users, which differentiates it from the more generalized systems found in previous research.

Table 1 Literature review comparison

Work (Year)	Field of Work	Methodology	Limitations	Uses RAG	Multi-Agent	Clarification Prompts	Relevance Filtering
ChatLaw (2024)	Legal QA (Chinese Law)	Multi-Agent RAG, Knowledge Graphs	Domain-specific	✓	✓	✗	✓
Legal Dialogue (2021)	Legal Chatbot Dev	AWS Lex, Lambda, Hierarchical Bot	No RAG or evaluation	✗	✗	✓	✗
DoNotPay Ethics (2025)	Legal Ethics & Regulation	Critical Review	No system implemented	✗	✗	✗	✗
TAM for Lawyer Bots (2022)	AI Legal Trust Studies	TAM Survey, Statistical Modelling	Theoretical only	✗	✗	✗	✗
Justice Chatbots (2020)	Immigration & Compliance	Retrieval-based, compliance chatbots	Narrow scope	✗	✗	✗	✗
Legal Structuring (2024)	Legal Corpus Engineering	Generative Parsing + Indexing	Requires structured corpus	✗	✗	✗	✓
Palestinian Bot (2023)	Cooperative Law (Palestine)	Fine-tuned LLM + Expert Validation	Narrow law scope	✗	✗	✗	✗
LAWBOT (2024)	Indian Legal Chatbot	IR + ML	Lacks multi-turn, no RAG	✗	✗	✗	✗
CREA2 (2023)	Legal NLP	SBERT + Semantic Search	Not real-time	✗	✗	✗	✓
Legal ML Bot (2023)	General Legal NLP	ML + NLP	No grounding; basic responses	✗	✗	✗	✗
LangChain Bot (2023)	LangChain Legal Example	LangChain, Cosine Similarity	Basic setup, no evaluation	✓	✗	✗	✗

RAGged Edges (2024)	RAG Evaluation Research	Prompt Testing, RAG Diagnosis	Analytical only	✓	✗	✗	✗
RAGAS (2023)	LLM Evaluation Metrics	Metric Framework	Evaluation-only framework	✗	✗	✗	✗
Proposed Work (2025)	Bangladeshi Legal IR Chatbot	Multi-Agent RAG using LangChain, LangGraph, Ollama	Domain-Specific Bias	✓	✓	✓	✓

Chapter 3 Data Collection and Preparation

In the development of a legal information retrieval system tailored to the Bangladeshi legal context, data is a foundational pillar. Unlike general-purpose chatbots, legal domain models require high-quality, jurisdiction-specific, and well-structured legal data. This chapter elaborates on the end-to-end process of data acquisition, preprocessing, embedding generation, and storage strategies used in the proposed multi-agent RAG system, with techniques grounded in both domain knowledge and modern NLP best practices.

3.1 Data Source

The dataset for this project was assembled from diverse legal resources that are publicly accessible and representative of Bangladeshi law. Specifically, the sources include:

- Bangladesh Supreme Court Judgments [26]
- Official Government Legislative Portal - Laws of Bangladesh [16]
- Legal analysis articles from legal blogs such as BD Law Post [3] and BD Law Help [15]

These sources collectively cover statutory law, constitutional judgments, and interpretive articles that are essential for a legal question-answering system. The documents were primarily in PDF format, featuring legal jargon, multi-column layouts, and occasional scanned elements.

To extract meaningful and accurate content from these PDFs, we used a hybrid extraction pipeline:

- PyPDFDirectoryLoader from langchain.document_loaders [13] to load documents and prepare them for processing.
- fitz (from PyMuPDF) [21] to handle complex or poorly structured PDFs for better text fidelity.

This hybrid strategy ensured reliable text extraction even in cases of formatting inconsistencies or embedded scanned images.

3.2 Data Preparation

Given the unstructured nature of legal PDFs, extensive preprocessing and structural reorganization were necessary to make the data usable for semantic retrieval. The following steps were applied.

3.2.1 Text Chunking and Normalization

Legal texts are typically verbose and context-rich, making them unsuitable for direct input into language models. Therefore, long documents were split into semantically coherent segments using:

- RecursiveCharacterTextSplitter from LangChain [13], which breaks down documents based on newline (\n\n) and sentence boundaries (.) while preserving contextual coherence.

Chunk size was controlled (e.g., 512–1024 tokens) to balance semantic integrity with efficient computation. Following chunking, normalization steps were applied:

- Stripping whitespace and special characters
- Removing headers/footers and irrelevant page elements
- Standardizing section headings, case citations, and clause references

These steps ensured the processed text retained legal meaning while being machine-readable.

3.2.2 Embedding Generation

For semantic retrieval and similarity search, the preprocessed text chunks were transformed into vector representations using OpenAI's Embedding API (text-embedding-ada-002) [20]. Each chunk was passed through this embedding model, converting it into a high-dimensional vector.

These embeddings were crucial for enabling accurate legal information retrieval by capturing the underlying semantic meaning of text rather than relying solely on keyword overlap.

3.2.3 Metadata Tagging and Indexing

To improve retrieval accuracy and support context-aware filtering, each document chunk was enriched with metadata:

- Source Type (e.g., “Statute,” “Judgment,” “Article”)
- Legal Domain (e.g., “Criminal Law,” “Family Law,” “Property Law”)
- Date of Publication
- Document ID and Section References

This metadata was indexed alongside the embedding vectors, supporting features such as scoped retrieval and filtered search based on jurisdiction or legal context.

3.2.4 Vector Store Construction

After embedding and tagging, the data was stored in a vector database:

- ChromaDB [4] was selected for its lightweight design and seamless integration with LangChain.
- Each record in Chroma contained: the original chunk text, its vector embedding, and associated metadata.

The vector store acts as the knowledge base for the RAG Agent, enabling fast and accurate dense retrieval when answering user queries.

3.2.5 Clarification-Responsive Structuring

In anticipation of multi-agent collaboration (particularly involving the Clarification Agent), additional structuring was applied:

- Retrievability flags were added to chunks likely to contain definitions, procedural norms, or statutory references.
- Query expansion keywords were tagged to assist the Clarification Agent in follow-up questioning.

These design elements were essential to support clarification and disambiguation tasks in the system [7].

3.2.6 Web Search Integration Logic

While the system's primary responses are grounded in the local vector database, a conditional fallback to web search was implemented using the TavilySearchResults Tool. If no sufficiently relevant chunks are retrieved, the system:

- Reformulates the query using the Rewriter LLM
- Performs a live web search for supplemental context
- Integrates and grades the results for factual consistency using a structured LLM grader

This mechanism ensures robustness and up-to-date legal information, especially for edge cases not covered in the core dataset.

3.2.7 Summary of the Workflow

The entire data pipeline is illustrated in Figure 1, detailing how documents flow through the loader, chunker, embedder, vector store, and ultimately power the Clarification and RAG agents.

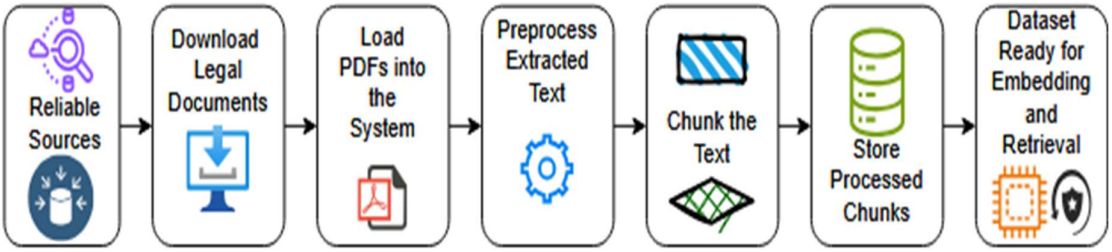

Figure 1 Data Preparation Workflow.

Chapter 4 Methodology

In this work, a Multi-Agent Retrieval-Augmented Generation (RAG) chatbot system specialized for Bangladeshi legal information retrieval. The system integrates modular agent-based reasoning via LangChain and LangGraph frameworks, which provide the foundation for controlled clarification flows, high-precision retrieval, and context-grounded answer generation. Furthermore, the methodology ensures compatibility with both cloud and local environments by incorporating Ollama for cost-effective execution. This hybrid design aims to maintain scalability while making the technology practical for use in low-resource regions such as rural Bangladesh. The overall methodology can be divided into several key stages:

4.1 Multi-Agent Architecture Using LangGraph

The innovative backbone of this system is the LangGraph-powered StateGraph, which defines how agent nodes interact through state transitions and conditional logic. Each node within the graph corresponds to an LLM-driven operation such as query assessment, rewriting, retrieval, or answer synthesis. LangGraph's flexibility allows for branching logic based on agent output, enabling dynamic flow control in real-time interactions [14].

4.1.1 Clarification Agent

Serving as the system's entry point, the Clarification Agent is responsible for improving input quality and legal specificity. Its importance lies in reducing the risk of irrelevant or inaccurate answers caused by poorly formed queries.

The agent performs two primary tasks:

- Query Clarity Assessment: A lightweight LLM (referred to as assessment_llm) analyzes the user's query to determine whether it is complete and specific enough to proceed directly to retrieval.
- Clarifying Follow-Up Generation: If the query is vague, the agent invokes a second model (question_gen_llm) to formulate clarifying questions aimed at eliciting essential information (e.g., case type, relevant law, specific jurisdiction).

The workflow proceeds as follows:

1. Receive user's initial input.
2. Apply structured LLM prompt to assess clarity.
3. If determined unclear, trigger clarification question.
4. Await user response and re-assess.
5. Upon receiving a clear query, forward it to the RAG Agent.

This process ensures that only queries with sufficient legal grounding proceed through the pipeline, thereby improving retrieval efficiency and answer quality.

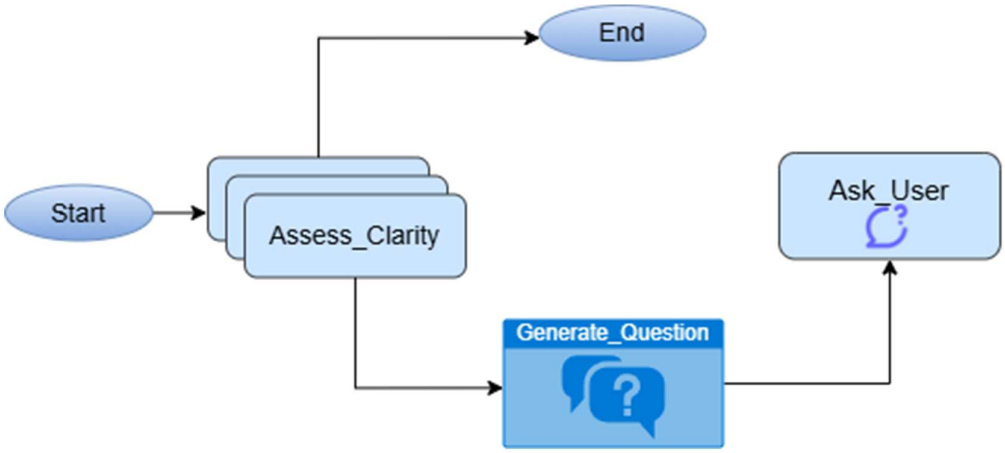

Figure 2 Clarification Agent Workflow.

4.1.2 Retrieval-Augmented Generation Agent (RAG Agent)

Once the query is clarified and rewritten, the RAG Agent takes over. It is responsible for the main tasks of document retrieval and grounded answer synthesis.

Core functionalities include:

- Vector-Based Retrieval: The rewritten query is used to perform a dense similarity search in ChromaDB, which houses the embedded legal dataset [4]. The vector store was prepared using OpenAI's text-embedding-ada-002 model [20], and documents were chunked using LangChain's text splitting utility.

- Relevance Grading: Retrieved chunks are passed to a structured_llm_grader, which scores them based on their alignment with the query's intent. Only top-ranked chunks are selected for the generation phase [7].
- Conditional Web Search: If the internal corpus lacks relevant content, the RAG Agent conditionally invokes the TavilySearchResults tool to fetch real-time web data. The use of LangGraph's add_conditional_edges() ensures this step is only triggered when needed [14].
- Answer Synthesis: A final generative model (either GPT-4o-mini or Gemma-3B, depending on deployment) generates a complete response using the most relevant content chunks. Prompts are crafted to enforce citation, legal tone, and factual consistency.

This structured pipeline reduces the likelihood of hallucination and improves legal reliability, as demonstrated in system evaluations [7].

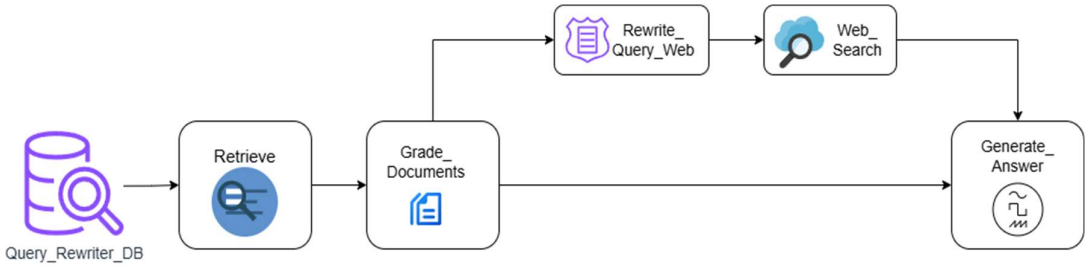

Figure 3 RAG Agent Workflow.

4.2 Query Rewriting

Before initiating the retrieval phase, the system employs a Query Rewriter Node, powered by a dedicated rewriter_llm. This component is crucial for aligning the user's natural language input with the formal, legal language of the corpus [7].

Often, legal documents contain highly structured terminology that differs from everyday speech. For example, a user query like "How can I get a divorce?" may be rewritten as "What are the legal provisions for divorce under Bangladeshi family law?" This transformation helps maximize semantic similarity between the query and the embedded legal text.

The rewriter operates based on predefined prompt templates that instruct the LLM to retain the user's intent while formalizing phrasing, incorporating legal references, and eliminating ambiguity. This step significantly boosts the accuracy of the retrieval process by improving vector similarity scores.

4.3 Conditional Web Search

Another important part of the system is the structure for conditional web search. If the local database provides enough information, the system answers directly. And if not, it can be extended to perform a real-time web search for additional information. This decision-making logic was handled using a should_search function and conditional routing inside the LangGraph [14].

4.4 Language Model and Tool Handling

The core intelligence of the proposed legal chatbot system is built on a carefully orchestrated combination of large language models (LLMs) and external modular tools. These components operate collaboratively and dynamically throughout the multi-agent architecture, enabling the system to perform complex legal reasoning, retrieval, and generation tasks with high efficiency and domain accuracy.

For language understanding and text generation, OpenAI's GPT 4o mini and Gemma3:4b via Ollama was used through the LangChain's ChatOpenAI wrapper [20].

- **GPT-4o-mini**: This is a high-performance proprietary model by OpenAI, used in cloud deployments. It is known for its reasoning capabilities and response fluency. The model is accessed through LangChain's ChatOpenAI wrapper [20].
- **Gemma-3B via Ollama**: This is an open-source language model optimized for local execution, particularly suited for low-resource settings where API access is not feasible. Ollama allows containerized deployment of Gemma, supporting real-time inference on standard hardware [7].

These models serve different purposes depending on the context—GPT-4o-mini is often used when cloud connectivity is available and high-end reasoning is desired, while Gemma-3B enables privacy-preserving and cost-effective local use.

The system also incorporates several LangChain-defined tools, each designed to execute a specific task within the agent workflow:

- **retrieval_tool**: This component uses vector similarity search within ChromaDB [4] to extract top-ranked legal text chunks related to the input query. It ensures that the generated response is grounded in relevant source material.
- **search_tool**: Powered by the TavilySearchResults API [7], this tool enables real-time web search when the vector database cannot fully satisfy a query. It supports dynamic knowledge augmentation.
- **answer_tool**: Once the relevant context is collected, this tool facilitates natural language generation, formatting the answer in a legally coherent, user-friendly manner. It leverages the system's main LLM.

These tools are not statically bound to any single agent but are instead dynamically invoked based on reasoning logic derived from the user's query state. LangGraph manages the control flow, ensuring that only the necessary tools are triggered per interaction path. This allows the system to operate efficiently, minimizing latency and optimizing resource usage.

In essence, the combination of multiple models and specialized tools gives the system flexibility, resilience, and depth—empowering it to understand, clarify, retrieve, and respond to legal queries across a wide range of complexity and ambiguity.

4.5 User Interface

To facilitate real-time human interaction, a Streamlit-based user interface was developed [25].

Key features include:

- Query Input: A text input box for user legal questions.
- Clarification Interaction: Dialog interface that captures and integrates follow-up clarification.
- Answer Output: Displays the system's response along with references to source documents.

The UI supports both web deployment and local execution. It is designed for ease of use in both urban and rural environments, with potential for kiosk-based deployment in legal aid centers.

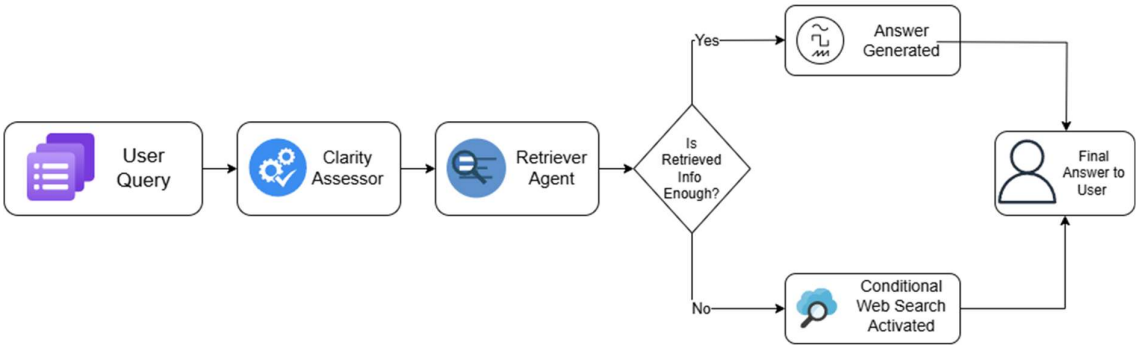

Figure 4 Detailed Agent Workflow.

4.6 Local Deployment

To address scalability and cost concerns, the system was extended for local deployment using the Ollama framework. This enabled testing with open-source LLMs like Gemma3-4B without dependency on commercial APIs. The local setup runs on a mid to high-end GPU or CPU environment and mirrors the cloud-based agent pipeline.

- Model: Gemma-3B was used as the local alternative to OpenAI's models.
- Hardware Requirements: Deployment was tested on machines with 8–16GB RAM, confirming feasibility for consumer-grade systems.
- Benefits: Enables cost savings, ensures data privacy, and removes reliance on external APIs.

This setup is particularly valuable for use in remote regions or institutions with limited internet access or budget constraints.

4.7 Evaluation

The methodology includes a dual-mode evaluation pipeline consisting of automated benchmarking and human expert review.

4.7.1 Automated Evaluation – LLM As A Judge

LLM As A Judge To evaluate performance, a set of 20 diverse legal queries representative of the most common user concerns in the Bangladeshi legal do main were used. These queries include procedural legal actions, statutory interpretations, policy clarifications, and case law queries. The same query set was used across all models to ensure consistency. A comparative testing against two models are included:

– GPT-4o Mini

– Gemma3-4B

20 legal queries were used to assess:

- Answer Relevancy
- Faithfulness to Retrieved Context
- Contextual Precision and Recall
- G-Eval Correctness Score [7]
- Hallucination Rate

Results showed substantial improvements over standard RAG architectures in all categories.

4.7.2 Human Evaluation – Legal Expert Review

Two independent expert evaluators assessed each generated response using a diverse set of 15 legal question-answer pairs, all designed by the practicing legal expert. Both evaluators are native Bangladeshi speakers with high proficiency in English and possess strong knowledge of Bangladeshi law. The evaluation team—one male and one female, with an average age of 30, consists of a practicing lawyer at the Supreme Court of Bangladesh and an undergraduate law student. Each response was independently rated by both evaluators. The evaluation was

conducted voluntarily as part of an academic exercise. The test involved a comparative assessment between the publicly available ChatGPT-4o and the proposed RAG-based system, with ChatGPT-4o selected due to its widespread use and popularity as a source for legal queries. Evaluation criteria included:

- Legal Correctness
- Linguistic Quality
- Source Traceability

The RAG system outperformed GPT-4o baselines in most metrics, affirming its suitability for domain-specific legal assistance [7].

4.8 Methodological Advantages

The proposed methodology offers several distinct advantages over traditional single-agent or end-to-end LLM systems:

- Agentic Modularity: Each agent handles a well-defined role, making the system interpretable and easier to debug or improve.
- Transparent Execution: All state transitions and decisions are traceable within LangGraph logs.
- Flexible Deployment: Operates effectively in both online (OpenAI) and offline (Ollama) environments.
- Legal Domain Alignment: The system is trained and evaluated specifically on Bangladeshi law content, ensuring contextual appropriateness.

This architecture forms a strong foundation for future expansions, including the addition of multilingual support, voice interfaces, or broader legal corpus coverage.

4.9 Work Timeline (Gantt Chart)

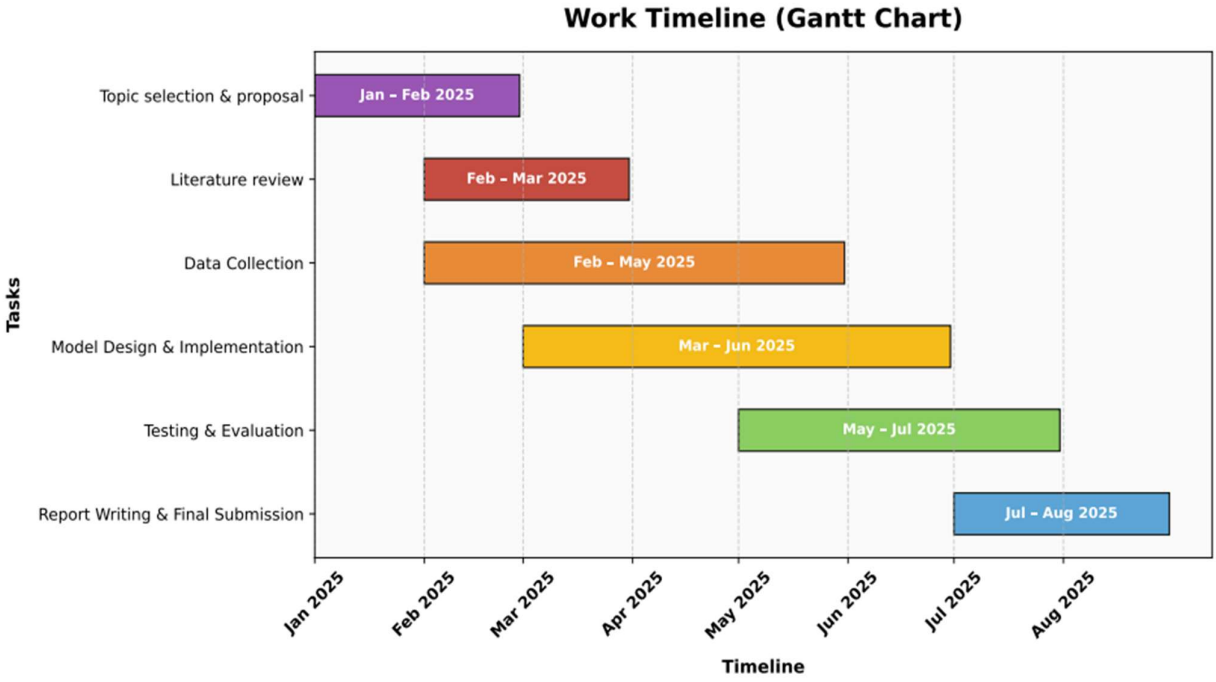

Figure 5 Work Timeline (Gantt Chart)

Chapter 5 Results and Discussion

The developed multi-agent RAG system resulted in a functional chatbot capable of addressing legal queries specific to the Bangladeshi context. An example interaction showcasing the system's ability to handle a user query regarding the Right to Information (RTI) Act is presented in Figure 5.

To assess the performance of the proposed work, the evaluation is conducted using two distinct criteria: (1) Deepeval framework for automated evaluation and (2) Evaluation by legal experts based on 15 questions formulated by the evaluators themselves.

5.1 LLM As a Judge

To quantitatively assess the performance and reliability of the system, the Deepeval framework was employed [8] for evaluation. A curated test suite comprising 20 distinct legal questions, along with expected answers relevant to the laws and legal landscape of Bangladesh, was generated via Deepeval's 'Generate from document' feature. A separate script was used to create the test cases. The evaluation focused on a set of standard RAG metrics designed to measure different facets of the system's retrieval and generation quality. The specific metrics evaluated include:

- **Answer Relevancy:** Measures how pertinent the generated answer is to the user's query. It is calculated as the ratio of relevant statements to the total statements in the answer:

$$Answer\ Relevancy = \frac{\text{Number of Relevant Statements}}{\text{Total Number of Statements}}$$

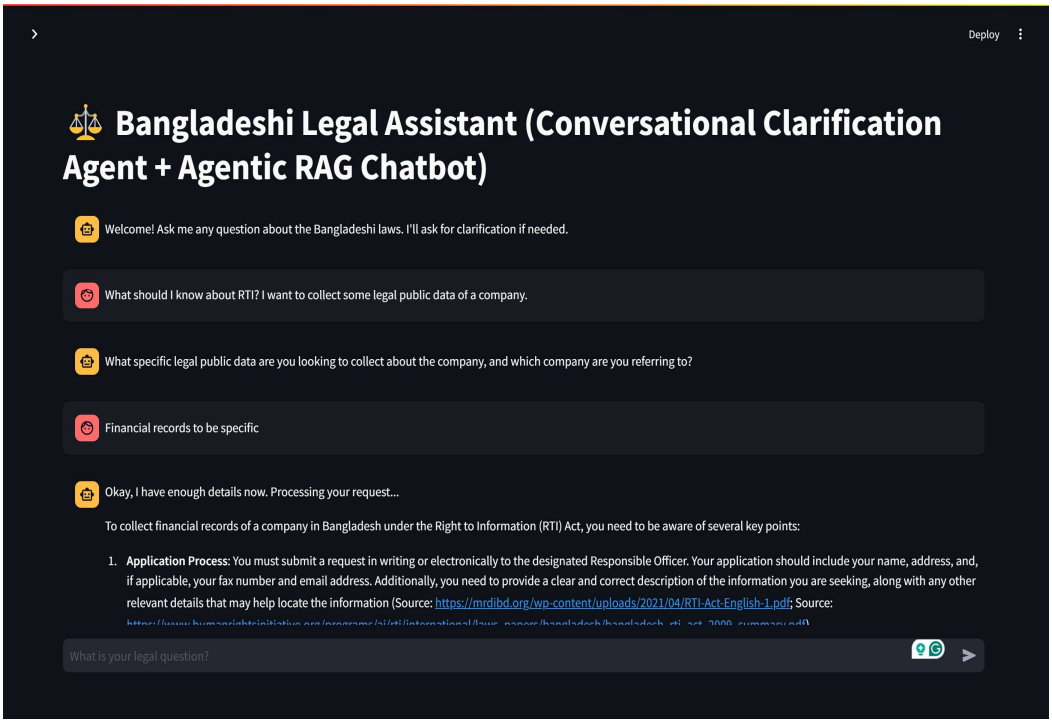

Figure 6 Example interaction demonstrating a query ('What should I know about RTI? I want to collect some legal public data of a company.') and the corresponding detailed response generated by the multi-agent system.

- **Faithfulness:** Assesses whether the generated answer is factually consistent with the retrieved context documents. It is calculated based on the claims made in the answer versus the provided context:

$$Faithfulness = \frac{\text{Number of Truthful Claims}}{\text{Total Number of Claims}}$$

- **Contextual Precision:** Evaluates the quality of the retrieval component by measuring the proportion of relevant documents among the retrieved set, weighted by their rank:

$$Contextual\ Precision = \frac{1}{Number\ of\ Relevant\ Nodes} \sum_{k=1}^{\infty} \frac{Number\ of\ Relevant\ Nodes\ Up\ to\ Position\ k}{k} \times rk$$

where r_k indicates if the node at rank k is relevant.

- **Contextual Recall:** Measures the extent to which the retrieved context contains all the necessary information required to formulate the ideal answer (often compared against a ground truth or expected output):

$$Contextual\ Recall = \frac{\text{Number of Attributable Statements}}{\text{Total Number of Statements}}$$

(Note: Deepeval calculates this based on the expected output).

Contextual Relevancy: Assesses the overall relevance of the retrieved context passages to the input query, calculated as:

$$Contextual\ Relevancy = \frac{\text{Number of Relevant Statements}}{\text{Total Number of Statements}}$$

(Note: Deepeval calculates this based on the input query and the retrieved context).

– Correctness (GEval): G-Eval is a framework that uses LLM-as-a-judge with chain-of-thoughts (CoT) to evaluate LLM outputs based on ANY custom criteria. The G-Eval metric is the most versatile type of metric deepeval has to offer, and is capable of evaluating almost any use case with human-like accuracy.

Usually, a GEval metric will be used alongside one of the other metrics that are more system specific (such as ContextualRelevancyMetric for RAG, and TaskCompletionMetric for agents). This is because G-Eval is a custom metric best for subjective, use case specific evaluation.

Since G-Eval is a two-step algorithm that generates chain of thoughts (CoTs) for better evaluation, in deepeval this means first generating a series of evaluation-steps using CoT based on the given criteria, before using the generated steps to determine the final score using the parameters presented in an LLMTestCase.

(Note: Although GEval is great in many ways as a custom, task-specific metric, it is NOT deterministic.)

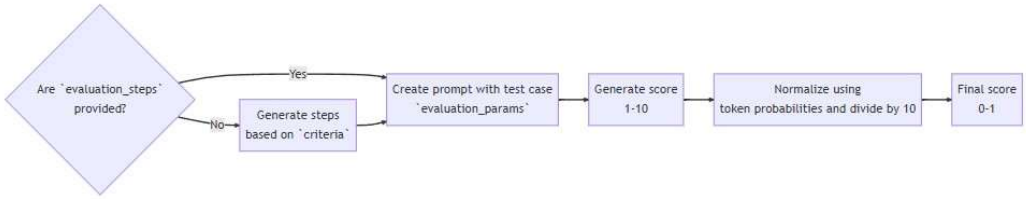

Figure 7 A flow-chart showcasing the steps to process GEval score. [7]

- **Hallucination:** The hallucination metric uses LLM-as-a-judge to determine whether your LLM generates factually correct information by comparing the actual output to the provided context:

$$Hallucination = \frac{\text{Number of Contradicted Contexts}}{\text{Total Number of Contexts}}$$

Deepeval utilizes an LLM-as-a-judge approach for these metrics, often providing reasoning alongside the numerical scores.

Using Deepeval, the system was conducted across 20 test case queries. A comparative testing against two models were also included:

- GPT-4o Mini
- Gemma3-4B

The comparisons between The Proposed RAG using GPT-4o-mini and Vanilla RAG using GPT-4o-mini are:

Table 2 Performance comparison of the proposed RAG system using GPT-4omini against the Vanilla RAG using GPT-4o-mini. Higher scores indicate better performance for all metrics except for the Hallucination Score, where a lower score is better.

Metric	Proposed RAG performance (AVG)	Vanilla RAG (AVG)	Performance difference	% improvement
Answer Relevancy	0.96	0.94	0.01	0.57
Faithfulness	0.90	0.80	0.10	10.1
Contextual Precision	0.96	0.88	0.08	8.45
Contextual Recall	0.93	0.88	0.05	4.66

Contextual Relevancy	0.61	0.54	0.06	6.59
Correctness (Geval)	0.75	0.66	0.09	8.73
Hallucination Score	0.14	0.28	-0.13	13.5

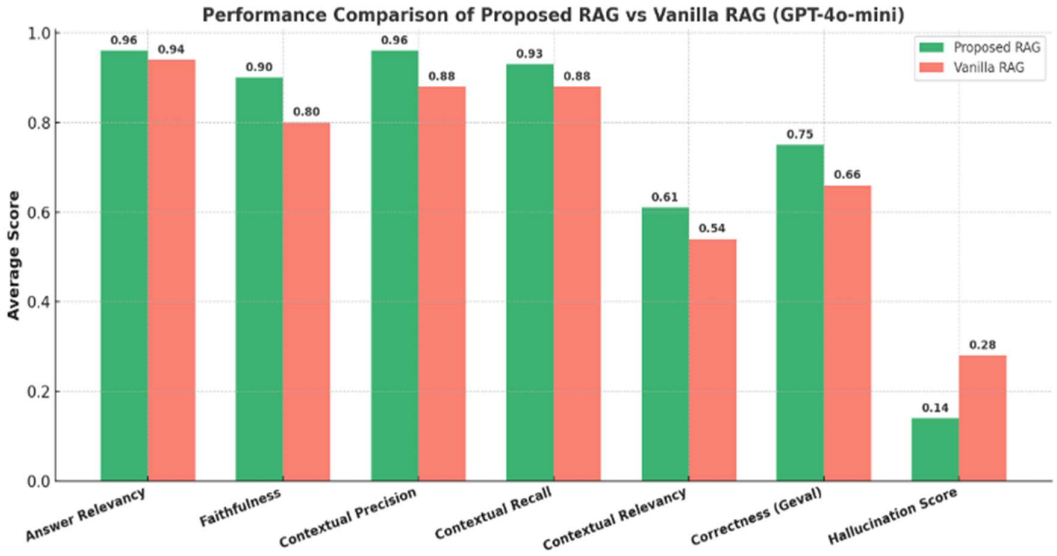

Figure 8 Performance Comparison of Proposed RAG vs Vanilla RAG(GPT-4o-mini)

Also the comparisons between The Proposed RAG using GPT-4o-mini and Vanilla RAG using GPT-4o-mini are shown in the given bar chart:

Now, the comparisons between The Proposed RAG using locally run Gemma34b via Ollama and Vanilla RAG using Gemma3-4b are:

Table 3 Performance comparison of the proposed RAG system using Gemma34b against the Vanilla RAG using Gemma3-4b.

Metric	**Proposed RAG performance (AVG)**	**Vanilla RAG (AVG)**	**Performance difference**	**% improvement**
Answer Relevancy	0.92	0.77	0.14	14.4
Faithfulness	0.70	0.67	0.02	2.53
Contextual Precision	0.89	0.85	0.03	3.44

Contextual Recall	0.96	0.91	0.05	4.69
Contextual Relevancy	0.56	0.51	0.04	4.61
Correctness (Geval)	0.62	0.61	0.01	0.57
Hallucination Score	0.22	0.27	-0.05	5.00

Also the Performance comparison of the proposed RAG system using Gemma34b against the Vanilla RAG using Gemma3-4b are shown in the given bar chart:

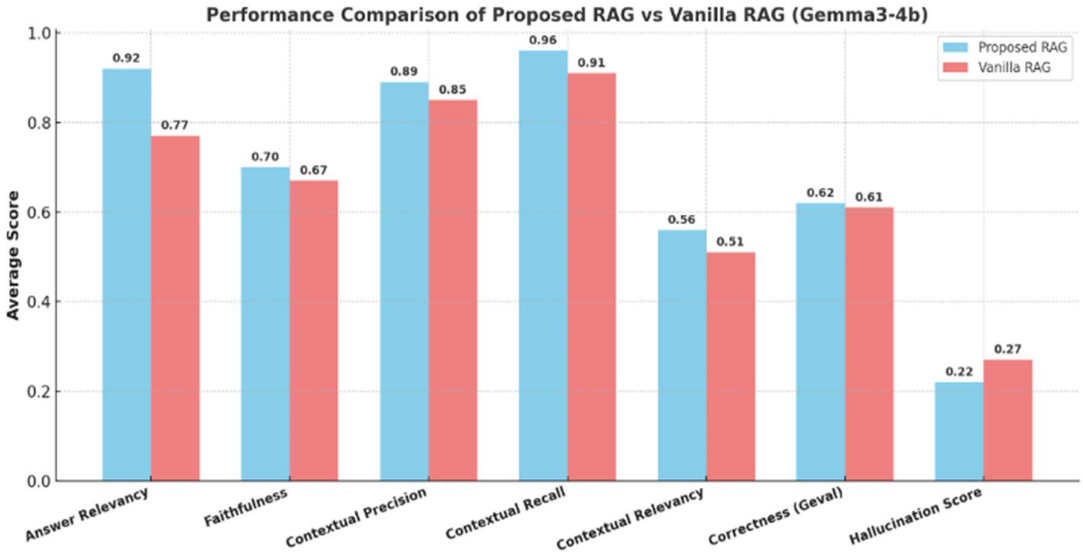

Figure 9 Performance Comparison of Proposed RAG vs Vanilla RAG(Gemma3-4b)

The proposed RAG system using GPT-4o-mini outperforms Vanilla RAG using GPT-4o-mini in 6 out of 7 metrics, with notable gains in correctness, faithfulness, and factual grounding, making it a more trustworthy and accurate solution.

It has proved that the proposed model has:

- High Faithfulness (+10.1%): This shows the proposed model's responses are more grounded in the source material, a critical feature in factual QA systems.
- Lower Hallucination Rate: A significantly lower hallucination score (0.14 vs. 0.28) indicates the proposed system is more reliable and factual, with a 13.5% improvement.
- Better Contextual Understanding: Metrics like Contextual Precision and Recall (8.45% and 4.66% gains respectively) suggest better grasp and representation of contextually relevant information.

While the performance gap with Gemma3-4B is narrower in some metrics (like Faithfulness), the proposed RAG model using Gemma3-4B still consistently leads in most areas, particularly in generating more relevant and accurate responses.

It has also proved that the proposed model has:

- **Answer Relevancy (+14.4%):** The proposed model provides much more relevant answers, which is key in user-facing applications.
- **Higher Contextual Recall and Relevancy:** It captures more meaningful context from source documents.
- **Reduced Hallucination:** A lower hallucination score again indicates greater factual reliability than Gemma3-4B.

Across both comparisons:

- The proposed RAG model consistently outperforms other Vanilla RAG models on critical NLP benchmarks.
- It minimizes hallucinations, enhances answer relevancy, and maintains strong contextual fidelity.
- The improvements in faithfulness, precision, and correctness demonstrate it is well-suited for real-world applications requiring accuracy, trustworthiness, and contextual understanding.

5.2 Legal Expert Evaluation

For human evaluation, each evaluator rates responses on a 1–5 scale based on two criteria: correctness, assessed by comparing responses to ground-truth answers, and quality, evaluated based on grammar and writing style. The meaning of the scores is illustrated in a corresponding Figure 7 [18].

Human Evaluation Rating Criteria
1 – Incorrect and very poor quality
2 – Mostly incorrect and poor quality
3 – Partially correct and average quality
4 – Mostly correct and good quality
5 – Fully correct and excellent quality

Figure 10 Interpretation of human evaluation scores (1–5) [18]

The final score for each response is calculated by averaging the individual scores assigned by the evaluators across all 15 questions.

Table 4 Performance comparison of the proposed RAG system against chatGPT-

Evaluator	Proposed RAG score (AVG) out of 5	chatGPT-4o score (AVG) out of 5	Performance difference	% improvement
01	3.67	3.52	0.15	15
02	3.73	3.60	0.13	13

And also the Performance comparison of the proposed RAG system against chatGPT-4o are shown in the given bar chart:

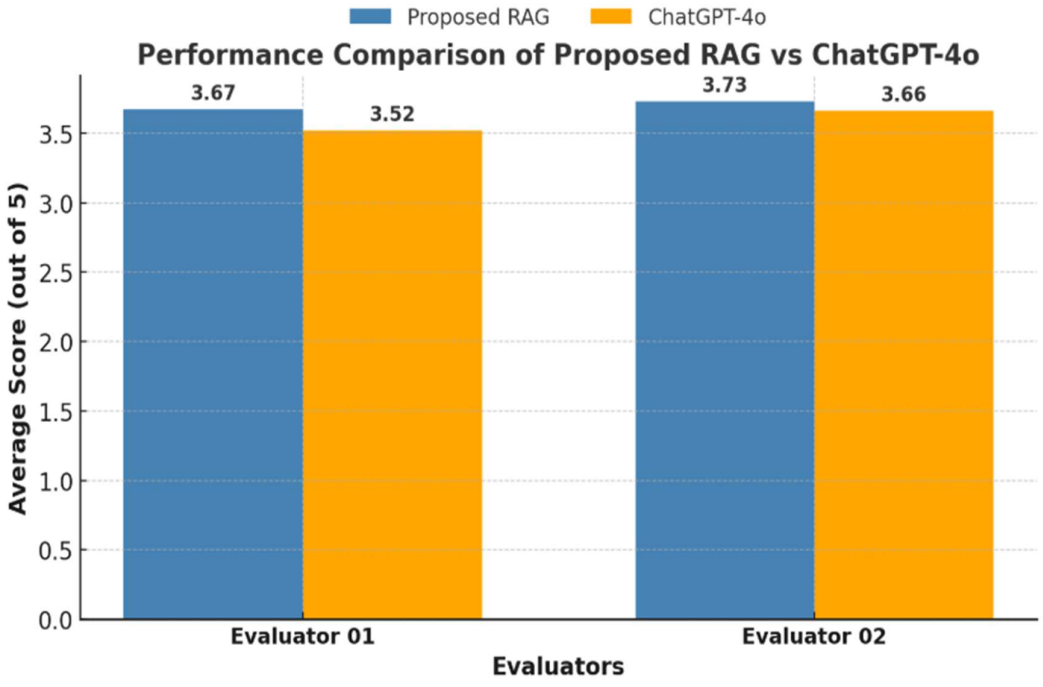

Figure 11 Performance Comparison of Proposed RAG vs ChatGPT-4o

This consistency across multiple human evaluators demonstrates that the proposed system is not just marginally better but reliably outperforms chatGPT4o in practical settings. The evaluators also provided feedback based on the tests:

Evaluator 1 (Practicing Supreme Court Lawyer) For ChatGPT-4o, criminal law responses leaned heavily on U.S Law, whereas Proposed RAG answers were relevant to Bangladeshi Law. context. The custom chatbot scored slightly higher at 3.67, though answers lacked depth. More localized data and analytical detail were recommended for improvement.

Evaluator 2 (Law Student) The chatbot effectively answers and asks clarifying questions, which improves precision. However, some technical followup questions may be too complex for general users lacking legal knowledge. The language can also be vague or difficult to understand. Simplifying both language and legal complexity is advised to improve accessibility.

5.3 Discussion

The development of this multi-agent RAG system represents a significant step towards enhancing legal information accessibility in Bangladesh. Our findings, based on a comparative evaluation against baseline models, demonstrate the tangible benefits of employing a sophisticated multi-agent architecture with LangGraph to address the complexities of legal query processing.

The core strength of our approach is the division of labor between a Clarification Agent and a RAG Agent is validated by the quantitative results. Unlike monolithic RAG systems, our agentic pipeline consistently outperforms a vanilla RAG setup. The evaluation showed significant improvements with the GPT-4omini model, including a 10.1% increase in Faithfulness, an 8.7% increase in Correctness (GEval), and a 13.5% reduction in Hallucination. These gains underscore the effectiveness of the architecture's components: the Clarification Agent improves input query precision, the rewriter node optimizes retrieval, and the relevance grader filters noise from the context. This multi-step process directly mitigates the risks of generating irrelevant or factually incorrect information, a critical requirement in the legal domain and a known challenge for RAG systems as noted by Feldman et al. [11].

And the human evaluation results clearly demonstrate the superiority of the proposed RAG system over the ChatGPT-4o model that is available for public use online, in terms of overall performance. Both human evaluators consistently rated the RAG system higher, with average scores of 3.73 and 3.67 compared to 3.60 and 3.52 for ChatGPT-4o. This reflects an absolute performance gain of up to 0.7 and a relative improvement ranging from 13% to 15%. Such gains indicate that the RAG model provides more accurate and reliable responses, particularly in domain-specific tasks like Bangladeshi legal information retrieval.

A key contribution of this work is the successful implementation and validation of the system using a locally deployed open-source model, Gemma3-4b, via the Ollama framework. While the absolute performance was lower than with GPT-4o-mini, the agentic structure still provided a clear advantage over a vanilla RAG using the same model, most notably a 14.4% improvement in Answer Relevancy. This demonstrates that the architectural benefits are modelagnostic and provides a viable path for creating a cost-effective, scalable, and privacy-preserving solution suitable for low-resource settings like Bangladesh, addressing major limitations of relying solely on proprietary, API-based models.

Despite the promising results, several limitations remain. The system's effectiveness is fundamentally tied to the quality and currency of its curated dataset [3,16,26]. Finally, the trade-off between the high performance of proprietary models and the accessibility of open-source models highlights the need for careful consideration during large-scale deployment.

Nonetheless, the multi-agent framework presented here offers a flexible and powerful paradigm. By breaking down the complex task into discrete, managed steps, the system achieves a demonstrable improvement in reliability and adaptability, providing a validated blueprint for leveraging advanced AI to address critical information access gaps in specialized domains like Bangladeshi law.

Chapter 6 Conclusions

6.1 Summary

Access to clear and accurate legal information remains a significant barrier in Bangladesh. This paper addressed this challenge by designing, implementing, and evaluating a novel Multi-Agent Retrieval-Augmented Generation (RAG) chatbot tailored for the Bangladeshi legal domain. By leveraging the LangGraph framework, we developed a sophisticated two-agent architecture that separates query clarification from the core retrieval and generation process.

Our quantitative evaluation demonstrated the system's superiority over standard RAG baselines. The agentic architecture delivered significant improvements in critical metrics, most notably enhancing Faithfulness, increasing overall Correctness, and reducing Hallucinations. Furthermore, we validated this approach on both a leading proprietary model (GPT-4o-mini) and a locally deployed open-source model (Gemma3-4b via Ollama), proving the architecture's versatility and establishing a clear path toward a cost-effective, scalable solution.

The key contribution of this work is the empirical evidence that a structured, multi-agent RAG system provides quantifiable gains in reliability and factual accuracy within a low-resource, specialized domain. This study moves beyond a theoretical proposal to offer a validated blueprint for democratizing legal knowledge in Bangladesh. It showcases an adaptable and robust approach that bridges a critical information gap, offering a promising and practical direction for future legal technology development in similar contexts worldwide.

6.2 Limitations & Future Work

Despite its strong performance, the system has a few limitations:

- Dependence on the quality and coverage of the vector store
- Occasional clarification failure when queries are too vague
- No Bengali language support in current LLMs

Building upon the promising comparative results, several avenues for future work can further enhance the capabilities and reliability of this multi-agent RAG system for Bangladeshi legal information.

- Enhanced Evaluation Framework: While 'Deepeval' [8] provided strong comparative metrics, a more comprehensive evaluation is crucial.
  - Multi-faceted RAG Evaluation: Incorporating frameworks like RAGAS [10] can provide more granular diagnostics of the retrieval and generation components.
- Knowledge Base Expansion and Maintenance: The system's accuracy is intrinsically tied to its knowledge base.
  - Automated Data Ingestion Pipeline: Developing a semi-automated pipeline to continuously identify, ingest, and index new laws, amendments, and court judgments [16,26] is essential for maintaining currency.
  - Diversification of Sources: Expanding the corpus to include more case law summaries, legal FAQs, and authoritative commentaries [3] will broaden the system's contextual understanding.
- Advanced Agent Architectures and Capabilities: The 'LangGraph' framework [14] allows for further sophistication.
  - Specialized Legal Analysis Agent: An agent could be developed to perform more complex legal reasoning, such as identifying analogous case law [24] or interpreting statutes within a specific factual context.
  - Persistent Memory for Contextual Conversations: Integrating persistent memory [27] would enable more coherent, multi-turn dialogues, allowing the system to retain user context for more personalized assistance.
- Model Optimization and Fine-tuning:
  - Optimizing Open-Source Models: Further experimentation with efficient open-source LLMs using Ollama, including quantization and optimization techniques, is needed to narrow the performance gap with proprietary models.
  - Domain-Specific Fine-tuning: Fine-tuning an open-source model like Gemma on a curated dataset of Bangladeshi legal text and question-answer pairs could significantly boost its understanding of local legal terminology and reasoning patterns, offering a powerful combination of performance and cost-effectiveness.
- User Interface and Accessibility:

- Enhanced UI/UX: Improving the ‘Streamlit’ interface [25] with features like chat history, source document linking, and user feedback mechanisms will enhance trust and usability.
- Multilingual Support: Developing support for Bengali would dramatically increase the system’s accessibility and impact across Bangladesh.

– Cost-Performance Analysis for Deployment: A thorough analysis of the cost-performance trade-offs between different models (e.g., GPT-4o-mini vs. a fine-tuned Gemma) is required to develop a sustainable deployment strategy for a public-facing service.

Chapter 7 Budget

Not Applicable (N/A).

Chapter 8 Complex Engineering Problems and Activities

8.1 Complex Engineering Problems

Table 5 Complex Engineering Problems

Range of Complex Engineering Problem Solving		
Complex Engineering Problems have characteristic P1 and some or all of P2 to P7		
Attributes		A Multi-Agent RAG system for Legal Information Retrieval in Bangladesh (Python, LangChain, ChromaDB, Ollama, Streamlit, LLMs)
P1	Depth of knowledge required (K3-K8)	The project required knowledge of NLP, information retrieval, multi-agent system design & local LLM deployment, understanding RAG architecture, embeddings & vector database(Chroma) along with Python based frameworks like LangChain & Ollama.
P2	Range of conflicting requirements	none
P3	Depth of analysis required	Analysis was needed to choose retrieval thresholds & evaluation metrics. Various LLMs were compared to minimize hallucination while preserving factual correctness.
P4	Familiarity of issues	The project involved new domains such as multi-agent coordination, local LLM inference and DeepEval-based evaluation which extended beyond standard coursework topics.
P5	Extent of applicable codes	It followed AI ethics guidelines avoiding misinformation and bias.

P6	Extent of stakeholder involvement	Direct involvement from potential users (law students & lawyers) informed the system design to ensure social usability and accessibility.
P7	Interdependence between subsystems	The system integrates Clarification Agent, Retriever Agent and Streamlit UI, each interacting dynamically for query refinement and response synthesis.

8.2 Complex Engineering Activities

Table 6 Complex Engineering Activities

Range of Complex Engineering Activities		
Complex activities means (engineering) activities that have some or all of the following characteristics		
Attributes		A Multi-Agent RAG System for Legal Information Retrieval in Bangladesh (Python, LangChain, ChromaDB, Ollama, Streamlit, LLMs)
A1	Range of resources	The project involves human resources, computational resources (local GPU/CPU systems for model inference), software tools (LangChain, ChromaDB, Ollama, Streamlit) and research datasets (legal corpora, PDF sources).
A2	Level of interactions	As this is a group project, all the team members work together by resolving their internal conflict towards solving an issue.
A3	Innovation	While choosing a topic, we did extensive background review in order to add new contributions to our project. Also, while implementing the project we learned a few new and creative ways to solve different technical problems. The project adds innovation by solving a specific problem of legal information

		retrieval in an educational & useful setting. The project is innovative in applying a multi-agent RAG pipeline for legal information in a low-resource setting like Bangladesh. It introduced Clarification and RAG Agents and domain-specific corpus retrieval offering a new approach to legal information accessibility.
A4	Consequences to society/ environment	Our system positively impacts society by democratizing access to legal knowledge, promoting awareness and reducing dependency on costly legal consultation.
A5	Familiarity	Involved tackling unfamiliar challenges such as model hallucination, multi-agent synchronization & context-based retrieval. Required experimentation and literature review to identify effective solutions.

References

1. A, K.R., S, K., & R, R. (2023). Legal solutions - intelligent chatbot using machine learning. In 2023 International Conference on Disruptive Technologies (ICDT), pp. 374–378. IEEE. https://doi.org/10.1109/ICDT57929.2023.10150955

2. Amato, F., Fersini, E., Gatta, M., & Sciarrone, C. (2023). An intelligent conversational agent for the legal domain. In Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management KDIR, pp. 236–243. SCITEPRESS – Science and Technology Publications. https://doi.org/10.5220/0012194000003598

3. BD Law Post. (n.d.). BD law post article. Available at: https://www.bdlawpost.com/search/label/Articles?max-results=10

4. Chroma. (2024). Chroma documentation. Available at: https://docs.trychroma.com/docs/overview/introduction

5. Cirillo, E., Fersini, E., Gatta, M., & Muscio, A. (2024). A chatbot for specialized domain. In: Intelligent Systems and Applications. IntelliSys 2023, Lecture Notes in Networks and Systems, vol. 851. Springer, Cham. https://doi.org/10.1007/978-3-031-47721-7_12

6. Cui, J. et al. (2024). ChatLaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. arXiv preprint arXiv:2401.02914. Available at: https://arxiv.org/abs/2401.02914

7. DeepEval. (n.d.). Deepeval documentation. Available at: https://deepeval.com/docs/metrics-llm-evals

8. DeepEval Documentation. (n.d.). Evaluation introduction. Available at: https://www.deepeval.com/docs/evaluation-introduction

9. Devaraj, P.N. et al. (2023). Legal chatbot using LangChain. arXiv preprint arXiv:2311.12719. Available at: https://arxiv.org/abs/2311.12719

10. Es, S. et al. (2023). RAGAS: Automated evaluation of retrieval augmented generation. arXiv preprint arXiv:2309.15217. Available at: https://arxiv.org/abs/2309.15217

11. Feldman, P., Foulds, J.R. & Pan, S. (2024). RAGged edges: The double-edged sword of retrieval-augmented chatbots. arXiv preprint arXiv:2403.01193. Available at: https://arxiv.org/abs/2403.01193

12. Isaac, O.J., Nwabueze, J.: The use of chatbots in providing free legal guidance: Benefits and limitations. ResearchGate (2025), further publication details needed for precise citation type. Original note: [Provide more specific publication details if available, e.g., journal, conference]

13. LangChain. (n.d.). LangChain documentation. Available at: https://python.langchain.com/docs/how_to/#document-loaders

14. LangGraph. (n.d.). LangGraph documentation. Available at: https://langchain-ai.github.io/langgraph/

15. LAW HELP BD. (2024). Penal codes section. Available at: https://lawhelpbd.com/category/cpc/

16. Laws of Bangladesh. (n.d.). Laws of Bangladesh. Available at: http://bdlaws.minlaw.gov.bd/

17. Maree, M., Abu-Qauod, R. & Tuffaha, B. (2023). Transforming legal text interactions: Leveraging NLP and LLMs for legal support in Palestinian cooperatives. Artificial Intelligence and Law. https://doi.org/10.1007/s10506-023-09358-0

18. Muhammad Rafsan Kabir, Rafeed Mohammad Sultan & F.R.M.R.A.S.M.N.M.S.R. (2025). LegalRAG: A hybrid RAG system for multilingual legal information retrieval. arXiv preprint arXiv:2504.16121. Available at: https://arxiv.org/pdf/2504.16121

19. Nikita et al. (2024). LAWBOT: A smart user Indian legal chatbot using machine learning framework. In 2024 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT), pp. 1–6. IEEE. https://doi.org/10.1109/CISCT61344.2024.10554004

20. OpenAI. (2024). OpenAI platform: Embeddings guide. Available at: https://platform.openai.com/docs/guides/embeddings

21. PyMuPDF. (n.d.). PyMuPDF documentation. Available at: https://pymupdf.readthedocs.io/en/latest/

22. Queudot, M., Canton, J. & Masse, J.J. (2020). Improving access to justice with legal chatbots. Information, 11(5), p.276. https://doi.org/10.3390/info11050276

23. Sharma, M. et al. (2021). Building a legal dialogue system: Development process, challenges and opportunities. arXiv preprint arXiv:2109.12654. Available at: https://arxiv.org/abs/2109.12654

24. Siino, M., Falco, M., Croce, D. & Rosso, P. (2025). Large language models for the legal domain: A survey. IEEE Access, 13, pp.19533–19554. https://doi.org/10.1109/ACCESS.2025.3533217

25. Streamlit. (n.d.). Streamlit documentation. Available at: https://docs.streamlit.io/

26. Supreme Court of Bangladesh. (n.d.). Supreme Court of Bangladesh. Available at: https://www.supremecourt.gov.bd/web/indexn.php

27. Weng, L. (2023). LLM-powered autonomous agents. Lil'Log Blog. Available at: https://lilianweng.github.io/posts/2023-06-23-agent/

28. Xu, N., Wang, K.J. & Li, C.Y. (2022). Technology acceptance model for lawyer robots with AI: A quantitative survey. In: HCI International 2022 – Late Breaking Work. Lecture Notes in Computer Science, vol. 13526. Springer, Cham. https://doi.org/10.1007/978-3-031-22107-7_27

Appendix

A1: Test cases: Validation results-Agent RAG:

Test-case 1: RTI Law Application - 1

==========

Metrics Summary

- Correctness (GEval) (score: 0.3362479981138105, threshold: 0.5, strict: False, evaluation model:

gpt-4o-mini, reason: The actual output identifies several tenets of the RTI law, but it does not address

the specific application strategies outlined in the expected output, such as understanding the law,

following up on requests, and advocating for wider usage. While it provides relevant information about

the law's principles, it lacks the practical guidance necessary for proficient application and successful

information retrieval as requested in the input., error: None)

- Answer Relevancy (score: 1.0, threshold: 0.5, strict: False, evaluation model: gpt-4o-mini, reason:

The score is 1.00 because the response directly addressed the essential tenets of the RTI law without

including any irrelevant statements., error: None)

- Faithfulness (score: 0.7777777777777778, threshold: 0.5, strict: False, evaluation model: gpt-4o-

mini, reason: The score is 0.78 because the actual output inaccurately suggests that the RTI law has

broader restrictions on withholding information, while it actually limits exemptions. Additionally, it fails

to clarify that authorities are only required to provide publishable parts of requested information, not

the entirety if some parts are exempt., error: None)

- Contextual Precision (score: 1.0, threshold: 0.5, strict: False, evaluation model: gpt-4o-mini,

reason: The score is 1.00 because all relevant nodes are ranked higher than the irrelevant node. The first

six nodes provide essential insights into the tenets of the RTI law, such as 'Maximum information

disclosure' and the importance of training, while the seventh node is ranked last and states that it does

not address the promotion of RTI among journalists, which is a narrower focus. This clear distinction in

relevance supports the perfect score., error: None)

- Contextual Recall (score: 0.75, threshold: 0.5, strict: False, evaluation model: gpt-4o-mini, reason:

The score is 0.75 because while several sentences in the expected output, such as those about preparing

precise requests and following up on applications, are well -supported by node(s) in retrieval context,

some sentences lack direct references to specific parts of the context, which slightly lowers the overall

alignment., error: None)

- Contextual Relevancy (score: 0.6222222222222222, threshold: 0.5, strict: False, evaluation model:

gpt-4o-mini, reason: The score is 0.62 because while some relevant statements outline essential tenets

of the RTI law, such as 'maximum information disclosure' and 'ensuring transparency and

accountability', many statements in the retrieval context focus on limitations and applications of the law

rather than its core principles, making the overall relevance moderate., error: None)

A2: GEMMA3 4B – Agent RAG:

Test case 1: RTI Law Application – 1 (3 docs)

- Correctness [GEval] (score: 0.5318400999946388, threshold: 0.5, strict: False, evaluation model:

gpt-4o-mini, reason: The actual output addresses the main question by outlining essential tenets of the

RTI law, but it diverges significantly from the expected output. While it provides relevant details about

the law, it lacks the emphasis on practical application steps, follow-up actions, and advocacy for wider

usage that are present in the expected output. Additionally, the actual output does not improve upon

the expected output by offering new insights or perspectives., error: None)

- Answer Relevancy (score: 1.0, threshold: 0.5, strict: False, evaluation model: gpt-4o-mini, reason:

The score is 1.00 because the response directly addressed the essential tenets of the RTI law without

including any irrelevant statements., error: None)

- Faithfulness (score: 0.8571428571428571, threshold: 0.5, strict: False, evaluation model: gpt-4o-

mini, reason: The score is 0.86 because the actual output incorrectly suggests that the RTI law can be

used to address utility connection issues, while the retrieval context clarifies that it primarily pertains to

transparency and accountability, such as investigating teacher attendance., error: None)

- Contextual Precision (score: 1.0, threshold: 0.5, strict: False, evaluation model: gpt-4o-mini,

reason: The score is 1.00 because all relevant nodes are ranked higher than the irrelevant node. The first

node discusses the 'right to information' and emphasizes the importance of understanding the

provisions of the RTI Act, which aligns with the need for 'Understanding the Law'. The second node

outlines principles such as 'Maximum information disclosure' and 'Easing access to information', directly

relating to 'Clear Application'. The irrelevant node, ranked third, does not provide relevant information

regarding the tenets of RTI law for proficient application, focusing instead on specific examples and

processes., error: None)

-Contextual Recall (score: 0.8888888888888888, threshold: 0.5, strict: False, evaluation model:

gpt-4o-mini, reason: The score is 0.89 because most sentences in the expected output are well-

supported by corresponding nodes in the retrieval context, such as the emphasis on understanding the

RTI Act and the need for proactive follow-up, which align closely with the context provided. However,

there is a slight gap as one sentence discusses the tenets without a direct reference, preventing a

perfect score., error: None)

- Contextual Relevancy (score: 0.5882352941176471, threshold: 0.5, strict: False, evaluation model:

gpt-4o-mini, reason: The score is 0.59 because while there are relevant statements about the essential

tenets of the RTI law, such as 'maximum information disclosure' and 'ensuring transparency and

accountability', many statements in the retrieval context focus on unrelated aspects, like the limitations

of the RTI law and assistance to NGOs, which detracts from the overall relevance., error: None)

- Hallucination (score: 0.0, threshold: 0.5, strict: False, evaluation model: gpt-4o-mini, reason: The

score is 0.00 because there are no contradictions between the actual output and the provided context,

indicating complete alignment and accuracy., error: None)

A3: GEMMA3 4B – Normal RAG:

Test case 1: RTI Law Application - 1

==========

Metrics Summary

- Correctness [GEval] (score: 0.46800396769536456, threshold: 0.5, strict: False, e valuation model:

gpt-4o-mini, reason: The actual output identifies several relevant tenets of the RTI law, such as the

importance of clearly defining information needs and understanding the purpose of the law. However, it

lacks key components from the expected output, including the need for clear application, follow-up,

persistence, and advocacy for wider usage. While it provides some useful insights, it does not fully align

with the expected output's comprehensive list of essential tenets., error: None)

- Answer Relevancy (score: 0.9230769230769231, threshold: 0.5, strict: False, evaluation model:

gpt-4o-mini, reason: The score is 0.92 because while the output effectively identifies essential tenets of

the RTI law, it includes an irrelevant statement about deadlines that detracts slightly from its focus on

application and retrieval., error: None)

- Faithfulness (score: 0.8571428571428571, threshold: 0.5, strict: False, evaluation model: gpt-4o-

mini, reason: The score is 0.86 because the actual output incorrectly states that the RTI law is primarily

useful for tracking progress on requests, while the retrieval context clarifies that it is not a

comprehensive solution for obtaining utility connections., error: None)

- Contextual Precision (score: 0.4777777777777777, threshold: 0.5, strict: False, evaluation model:

gpt-4o-mini, reason: The score is 0.48 because while there are relevant nodes that provide essential

tenets of the RTI law, they are not ranked higher than several irrelevant nodes. Specifi cally, the first

node ranks highest but states that 'Chapter Two The applied aspects of the RTI law' does not provide

specific tenets, which detracts from the overall relevance. The second node also ranks high and

discusses RTI in relation to school management without outlining essential tenets. In contrast, the

relevant nodes, which include important tenets like 'Clear Application' and 'Education and Training,' are

ranked lower, leading to a lower contextual precision score., error: None)

- Contextual Recall (score: 0.75, threshold: 0.5, strict: False, evaluation model: gpt-4o-mini, reason:

The score is 0.75 because while several sentences in the expected output effectively connect to specific

nodes in the retrieval context, such as the importance of cl ear applications and proactive follow-ups,

some sentences lack direct references to the nodes, which slightly diminishes the overall contextual

recall., error: None)

- Contextual Relevancy (score: 0.5, threshold: 0.5, strict: False, evaluation model: gpt-4o-mini,

reason: The score is 0.50 because while some relevant statements highlight the RTI law's purpose, such

as 'The RTI law ensures the Bangladeshi citizen his or her right to get the information on all activities...'

and 'Maximum information disclosure', the majority of the retrieval context focuses on unrelated

specifics and practical applications that do not address the essential tenets of the RTI law., error: None)

- Hallucination (score: 0.0, threshold: 0.5, strict: False, evaluation model: gpt-4o-mini, reason: The

score is 0.00 because there are no contradictions between the actual output and the provided context,

indicating complete alignment and accuracy., error: None)

A4: Normal RAG Test:

Test case 1: RTI Law Application – 1

==========

Metrics Summary

- Correctness [GEval] (score: 0.49694192985047236, threshold: 0.5, strict: False, evaluation model:

gpt-4o-mini, reason: The actual output identifies several key tenets of the RTI law, such as clarity of

information requests and the lack of requirement to justify requests, which align with the input's focus

on proficient application. However, it lacks several components present in the expected output, such as

the importance of following up on applications and the need for education and training, leading to a

partial but incomplete response., error: None)

- Answer Relevancy (score: 1.0, threshold: 0.5, strict: False, evaluation model: gpt-4o-mini, reason:

The score is 1.00 because the response directly addressed the essential tenets of the RTI law without

including any irrelevant statements., error: None)

- Faithfulness (score: 0.8571428571428571, threshold: 0.5, strict: False, evaluation model: gpt-4o-

mini, reason: The score is 0.86 because the actual output incorrectly states that the RTI law permits

access to attendance registers and educational statistics, which are not referenced in the retrieval

context., error: None)

- Contextual Precision (score: 0.3333333333333333, threshold: 0.5, strict: False, evaluation model:

gpt-4o-mini, reason: The score is 0.33 because only one relevant node is ranked higher than two

irrelevant nodes. The first node ranks first but states that 'Chapter Two The applied aspects of the RTI

law does not provide specific tenets or guidance for proficient application and successful information

retrieval,' making it irrelevant. The second node, ranking second, mentions that 'The second document

discusses the RTI law's utility in various contexts but lacks clear tenets or principles for its application,'

further contributing to the lower score. In contrast, the third node, which ranks third, effectively

outlines essential tenets by stating that 'The third document outlines the importance of knowing what

information to request and provides examples of information requests,' justifying its relevance. Thus,

the score reflects that while there is some relevant information, it is overshadowed by the presence of

irrelevant nodes ranked higher., error: None)

- Contextual Recall (score: 0.5, threshold: 0.5, strict: False, evaluation model: gpt-4o-mini, reason:

The score is 0.50 because while some sentences in the expected output, such as those about preparing

precise requests and following up on applications, are well-supported by node(s) in retrieval context,

others like education and training, networking, and the concluding remarks lack direct references,

leading to a mixed alignment., error: None)

- Contextual Relevancy (score: 0.3157894736842105, threshold: 0.5, strict: False, evaluation model:

gpt-4o-mini, reason: The score is 0.32 because the majority of the retrieval context is irrelevant,

discussing topics like 'teachers attendance register' and 'drop out rates' which do not relate to the RTI

law. However, relevant statements such as 'The RTI law can be a useful tool in this regard' and 'Using the

RTI, the following information may be collected: project proposal, work order, tender process' provide

some connection to the essential tenets of the RTI law., error: None)

- Hallucination (score: 1.0, threshold: 0.5, strict: False, evaluation model: gpt-4o-mini, reason: The

score is 1.00 because the actual output completely fails to align with the context, neglecting personal

experiences and collective action related to the RTI law, leading to significant contradictions., error:

None)

B1: Test results (GPT-4o-mini):

	A	B	C	D	E	F	G	H	I
1		AGENTIC RAG							
2	TEST NI	Test Case Name	GEval Score	AnswerRelevancy Score	Faithfulness Score	ContextualPrecision Sco	ContextualReca	ContextualRelevancy Score	Hallucination Score
3	1	Test-case 1: RTI Law Application - 1	0.336247998	1	0.777777778	1	0.75	0.622222222	0.666666667
4	2	Test case 2: Cheque Dishonour - 1	0.807762112	1	1	1	1	0.833333333	0
5	3	Test case 3: Cheque Dishonour - 3	0.862245933	1	0.923076923	0.8875	1	0.846153846	0
6	4	Test case 4: Cheque Dishonour 5 (web -searched)	0.808608119	0.666666667	1	1	1	0.5	0.666666667
7	5	Test case 5: Cheque Dishonour 6 (5 docs)	0.85621765	0.9	0.777777778	1	1	0.846153846	0
8	6	Test case 6: Domestic Violence test-case 1 (5 docs)	0.788967608	1	0.95	1	1	0.214285714	0
9	7	Test case 7: Domestic Violence test-case 2 (2 docs)	0.894043388	1	0.727272727	1	1	0.625	0
10	8	Test case 8: Domestic Violence test-case 3 (3 docs)	0.911209799	1	1	1	1	0.4	0
11	9	Test case 9: Domestic Violence test-case 4 (5 docs)	0.693537202	1	1	1	1	0.576923077	0
12	10	Test case 10: Domestic Violence test-case 5 (5 docs)	0.75621765	1	1	1	1	0.846153846	0
13	11	Test case 11: Domestic Violence test-case 6 (web search)	0.138794259	0.666666667	1	0.5	0	0.1	1
14	12	Test case 12: Domestic Violence test-case 7 (3 docs)	0.812791328	1	1	1	1	0.555555556	0
15	13	Test case 13: RTI - 1 (5 docs)	0.634438459	1	1	1	0.7	0.813559322	0
16	14	Test case 14: RTI - 2 (5 docs)	0.707723822	1	0.714285714	0.8875	1	0.236842105	0
17	15	Test case 15: Citizenship Act 1951 - 1 (5 docs)	0.805207553	1	0.888888889	1	1	0.80952381	0
18	16	Test case 16: Citizenship Act 1951 - 2 (5 docs)	0.890088923	1	0.8	1	1	0.5	0
19	17	Test case 17: Civil code - 1 (web search)	0.800670484	1	1	1	1	0.571428571	0.666666667
20	18	Test case 18: Criminal code -1 (2 docs)	0.881307287	1	0.75	1	1	0.818181818	0
21	19	Test case 19: Criminal code -2 (5 docs)	0.881307287	0.833333333	0.625	1	1	0.75	0
22	20	Test case 20: RTI - 3 (5 docs)	0.73473586	1	1	1	1	0.657142857	0
23	21	Test case 21: Citizenship act 1951 - 3 (5 docs)	0.694654772	1	1	1	1	0.724137931	0
24		AVERAGE SCORE	0.747465595	0.955555556	0.901622848	0.96547619	0.92619048	0.611742755	0.142857143
25									
26		NORMAL RAG							
27	TEST NI	Test Case	Correctness [GE	Answer Relevancy	Faithfulness	Contextual Precision	Contextual Reca	Contextual Relevancy	Hallucination
28	1	Test case 1: RTI Law Application-1	0.49694193	1	0.857142857	0.333333333	0.5	0.315789474	1
29	2	Test case 2: Cheque Dishonour-5	0.789968041	1	0.75	1	0.75	0.333333333	0
30	3	Test case 3: Domestic Violence test-case-1	0.726585964	0.833333333	1	1	1	0.666666667	0
31	4	Test case 4: Cheque Dishonour - 3	0.722437743	0.916666667	1	1	1	0.75	0
32	5	Test case - 5: Cheque Dishonour-6	0.678771475	0.916666667	0.833333333	1	1	0.692307692	0
33	6	Test case - 6: Domestic Violence test-case - 2	0.75621765	1	0.888888889	1	1	0.6875	0.5
34	7	Test case - 7: Domestic Violence test-case - 3	0.906489427	1	0.692307692	0.833333333	1	0.6	0.333333333
35	8	Test case - 8: Domestic Violence test-case-4	0.633613333	1	1	1	1	0.3	0
36	9	Test case - 9: Domestic Violence test-case-5	0.740316489	1	1	1	0.857142857	0.666666667	0
37	10	Test case 10: Domestic Violence test-case-6	0.729272895	1	0.75	1	0.8	0.4	0
38	11	Test case 11: Domestic Violence test-case-7	0.757926801	1	0.777777778	1	1	0.764705882	0
39	12	Test case 12: RTI - 1	0.709886124	1	1	0.5	1	0.333333333	1
40	13	Test case 13: RTI - 2	0.563798728	1	0.75	1	1	0.269230769	0
41	14	Test case 14: RTI-3	0.732370362	1	0.777777778	1	0.857142857	0.7	0
42	15	Test case - 15: Citizenship act 1951 - 1	0.798966573	1	0.857142857	1	1	0.333333333	0.333333333
43	16	Test case - 16: Citizenship act 1951 - 2	0.686573688	1	0.666666667	1	0.75	0.375	0.666666667
44	17	Test case - 17: Citizenship act 1951 - 3	0.680289554	1	1	1	1	0.636363636	0
45	18	Test case 18: Civil-code-1	0.146409254	0.714285714	0.25	0.5	1	0.5	1
46	19	Test case - 19: Criminal code -1	0.460441513	0.9	0.666666667	1	0.714285714	0.888888889	1
47	20	Test case - 20: Criminal code-2	0.495081273	0.666666667	0.4	0.333333333	0.25	0.5	0
48	21	Test case - 21: Cheque Dishonour - 1	0.650229105	1	0.9	1	1	0.75	0
49		AVERAGE SCORE	0.660123234	0.949886621	0.800843072	0.880952381	0.87993197	0.545862842	0.277777778
50									
51									
52									
53									
54			GEval Score	AnswerRelevancy Score	Faithfulness Score	ContextualPrecision Sco	ContextualReca	ContextualRelevancy Score	Hallucination Score
55		Proposed RAG performance (AVG)	0.747465595	0.955555556	0.901622848	0.96547619	0.926190476	0.611742755	0.142857143
56		Normal RAG performance (AVG)	0.660123234	0.949886621	0.800843072	0.880952381	0.879931973	0.545862842	0.277777778
57		Performance difference (+ value indicates better proposed Rag performa	0.087342361	0.005668934	0.100779776	0.08452381	0.046258503	0.065879913	-0.134920635
58		% improvement	8.734236054	0.566893424	10.07797758	8.452380957	4.625850343	6.587991329	13.49206349
59									
60									

B2: Test results (Gemma3-4b):

		AGENTIC RAG							
		Test Case	Correctness [GEval]	Answer Relevancy	Faithfulness	Contextual Precision	Contextual Recall	Contextual Relevancy	Hallucination
3	1	Test case 1: RTI Law Application-1 (3 docs)	0.5318401	1	0.857142857	1	0.888888889	0.588235294	0
4	2	Test case 2: Checque Dishonour-5 (web search)	0.567964775	1	0.8	1	1	0.888888889	0
5	3	Test case 3: Checque Dishonour - 3 (web search)	0.577054913	0.666666667	0.5	1	1	0.692307692	0.333333333
6	4	Test case 4: Checque Dishonour - 6 (web search)	0.583518103	0.666666667	0.75	1	1	0.692307692	0.333333333
7	5	Test case 5: Domestic Violence test-case-1 (3 docs)	0.594654772	1	0.666666667	1	1	0.333333333	0
8	6	Test case 6: Domestic Violence test-case-2 (3 docs)	0.837936007	1	0.25	1	1	0.411764706	0
9	7	Test case 7: Domestic Violence test-case-3 (2 docs)	0.812797545	1	0.75	1	1	0.545454545	0
10	8	Test case 8: Domestic Violence test-case-4 (2 docs)	0.503533257	1	0.5	1	1	0.764705882	0
11	9	Test case 9: Domestic Violence test-case-5 (2 docs)	0.54166007	1	0.666666667	1	1	0.210526316	0
12	10	Test case 10: Domestic Violence test-case-6 (web search)	0.434782172	0.588235294	0.933333333	0.5	0.833333333	0.533333333	0.666666667
13	11	Test case 11: Domestic Violence test-case-7 (3 docs)	0.794792447	0.785714286	0.666666667	1	1	0.652173913	0
14	12	Test case 12: RTI - 1 (3 docs)	0.627411669	1	0.818181818	1	1	0.633333333	1
15	13	Test case 13: RTI - 2 (5 docs)	0.664493708	0.833333333	0.6	0.804166667	1	0.333333333	0
16	14	Test case 14: RTI - 3 (4 docs)	0.569445075	1	0.8	1	1	0.606060606	0
17	15	Test case 15: Citizenship act 1951 â€“ 1 (4 docs)	0.641068665	1	0.714285714	1	1	0.347826087	0.333333333
18	16	Test case 16: Citizenship act 1951 â€“ 3 (3 docs)	0.598153682	1	0.7	0	0.666666667	0.882352941	0.333333333
19	17	Test case 17: Citizenship act 1951 â€“ 2 (4 docs)	0.830459269	0.8	0.4	1	1	0.363636364	0
20	18	Test case 18: Civil-code-1(web search)	0.671951877	1	1	1	1	0.666666667	0.333333333
21	19	Test case 19: Criminal code - 1 (web search)	0.604695529	1	0.666666667	1	1	0.5	0
22	20	Test case 20: Criminal code - 2 (web search)	0.371306951	1	1	0.5	0.8	0.5	1
23		AVERAGE	0.617976029	0.917030812	0.701980519	0.890208333	0.959444444	0.557312046	0.216666667
24									
25									
26		Normal RAG							
27		Test Case	Correctness [GEval]	Answer Relevancy	Faithfulness	Contextual Precision	Contextual Recall	Contextual Relevancy	Hallucination
28	1	Test case 1: RTI Law Application - 1	0.468003968	0.923076923	0.857142857	0.477777778	0.75	0.5	0
29	2	Test case 2: Checque Dishonour - 5	0.637023865	0.666666667	0	1	1	0.428571429	0
30	3	Test case 3: Checque Dishonour-3	0.791391882	0.666666667	0.727272727	1	1	0.772727273	0
31	4	Test case 4: Checque Dishonour-6	0.791391882	0.666666667	0.727272727	1	1	0.772727273	0
32	5	Test case 5: Domestic Violence test-case-1	0.579487901	1	0.933333333	1	1	0.791666667	0
33	6	Test case 6: Domestic Violence test-case-2	0.892237889	1	0.25	1	1	0.666666667	0
34	7	Test case 7: Domestic Violence test-case-3	0.875491498	0.75	0.666666667	0.833333333	1	0.461538462	0
35	8	Test case 8: Domestic Violence test-case-4	0.607300895	0.625	0.928571429	1	1	0.681818182	0.333333333
36	9	Test case 9: Domestic Violence test-case-5	0.607218514	1	1	1	1	0.421052632	0
37	10	Test case 10: Domestic Violence test-case-6	0.376603678	0.666666667	0.75	1	1	0.4	0
38	11	Test case 11: Domestic Violence test-case-7	0.571306952	1	0.75	1	1	0.703703704	0.333333333
39	12	Test case 12: RTI - 1	0.642167763	0.9	0.857142857	0.5	1	0.310344828	1
40	13	Test case 13: RTI - 2	0.576202702	1	0.75	0.916666667	1	0.333333333	0
41	14	Test case 14: RTI - 3	0.614025924	0.857142857	0.714285714	1	1	0.6	0
42	15	Test case 15: Citizenship act 1951 - 1	0.85	0.625	0.571428571	1	1	0.222222222	0.333333333
43	16	Test case 16: Citizenship act 1951-3	0.601963329	1	0.8	0.8875	1	0.611111111	0.333333333
44	17	Test case 17: Citizenship act 1951 - 2	0.623258669	0.8	0.25	1	0.75	0.384615385	0.666666667
45	18	Test case 18: Civil-code-1	0.3	0.6	0.5	0.5	1	0.4375	1
46	19	Test case 19: Criminal code - 1	0.525750294	0.714285714	0.666666667	1	0.75	0.571428571	0.333333333
47	20	Test case 20: Criminal code - 2	0.314025924	0	0.833333333	0	0	0.153846154	1
48		AVERAGE	0.612242676	0.773058608	0.676655844	0.855763889	0.9125	0.511243695	0.266666667
49									
50									
51			GEval Score	AnswerRelevancy Sc	Faithfulness Sc	ContextualPrecision	ContextualRecall Sc	ContextualRelevancy	Hallucination Score
52		Proposed RAG performance (AVG)	0.617976029	0.917030812	0.701980519	0.890208333	0.959444444	0.557312046	0.216666667
53		Normal RAG performance (AVG)	0.612242676	0.773058608	0.676655844	0.855763889	0.9125	0.511243695	0.266666667
54		Performance difference (+ value indicates better proposed Ra	0.005733353	0.143972204	0.025324675	0.034444444	0.046944444	0.046068352	-0.05
55		% improvement	0.573335292	14.39722043	2.532467532	3.444444444	4.694444444	4.606835189	5

C1: Sample User-Bot Interaction Log (Evaluated Session):

#	User Query	Bot Response Summary	Notes/Evaluation	Ranking
1	What is Homicide? Discuss it with classification.	Explained culpable homicide with classification under Bangladeshi Penal Code.	Ranked 3–4; provided accurate, structured response.	4
2	Does a person have right to kill someone?	Asked clarification; later responded on legal justifications such as self-defense.	Clarification loop worked well.	4
3	What is Hurt and Grievous Hurt?	Defined as per Sections 319 and 320 with examples.	Satisfactory legal grounding.	4
4	What is theft? Can a person steal own property?	Defined statutory theft; clarified own-property scenario.	Prompt-based flow effective.	4
5	When does theft become robbery?	Explained based on use of force under Section 390.	Accurate legal distinction.	4
6	What is Dacoity?	Defined with penalty and distinction from robbery.	Good grounding.	4
7	Distinction between misappropriation and breach of trust?	Defined Sections 403 & 405 with legal examples.	Conceptual clarity noted.	4
8	What is House Trespass?	Defined with punishment and categories.	Clear response.	4
9	Define House Breaking. How it differs from Criminal Trespass?	Provided structured comparison.	Well-structured comparison.	4

10	Which act is considered House Breaking?	Listed acts like entering through window, breaking latch, etc.	Practical examples appreciated.	4
11	Land dispute: when can magistrate attach property?	Asked clarifications repeatedly.	Needs grounding improvement.	2
12	Grounds for attachment under section 146?	Unable to retrieve sufficient legal reference.	Gap in retrieval.	1
13	Acts considered as public nuisance?	Gave several contextual examples like keeping dangerous animals, public indecency, etc.	Detailed and contextual.	4
14	Rights of wrongfully confined person?	Explained Section 340, procedures under CrPC.	Sound legal process overview.	4
15	Conducting a search for a missing person?	Gave detailed online GD and follow-up process.	Strong practical value.	4

Evaluator's Observation: The chatbot is capable of answering properly and it asks questions to clarify the situation which is necessary for correct and precise answers but sometimes when it asks technical questions regarding the situation, it may be difficult for the common people to provide with proper information as they lack proper legal knowledge. The language is sometimes vague and difficult to understand. The bot needs to simplify the language and technicality to take it to the level of mass people.

C2: ChatGPT Legal Queries Summary Table:

#	User Query	Bot Response Summary	Notes/Evaluation	Ranking
1	What is Homicide? Discuss it with classification.	Detailed classification of homicide under law.	Accurate, structured response	3
2	Does a person have right to kill someone?	Explained legal justifications such as self-defense.	Confusing wording; could mislead readers.	1
3	What is Hurt and Grievous Hurt?	Defined as per Sections 319–320 of Penal Code.	Well-explained with comparison table.	4
4	What is theft? Can a person steal own property?	Explained theft vs. possession and cited examples.	Good explanation; mentioned case law.	4
5	When does theft become robbery?	Explained escalation from theft to robbery with examples.	Clearly presented legal distinction.	4
6	What is Dacoity?	Defined under Sec. 391 with comparison and examples.	Well-structured and complete.	4
7	Dishonest misappropriation vs. Criminal breach of trust?	Detailed comparison with examples and a table.	Can be improved.	3
8	What is House Trespass?	Explained Sec. 442 and related laws.	Can be improved; a bit lengthy.	3
9	Which act we can consider as House Breaking?	Explained six legal methods under Sec. 445.	Thorough and clearly categorized.	4
10	Procedure to stop breach of peace in land dispute?	Step-by-step guide including Sec. 144, 145 CrPC.	Detailed and procedural.	3
11	When can a magistrate attach the disputed land?	Described conditions and process.	Comprehensive explanation.	4

12	Grounds of attachment under 146; powers of receiver?	Outlined CrPC provisions and receiver powers.	Legally clear and structured.	4
13	Which acts are public nuisance?	Listed common nuisances under Sec. 268.	Very detailed and practical.	4
14	When a wrongfully confined person can be searched?	Conditions and legal safeguards explained.	Clear and sensitive to rights.	4
15	Step-by-step search procedure?	Detailed legal procedure outlined.	Well-organized and accurate.	4

Evaluator's Observation: The best thing about ChatGPT is that it makes things much more lucid and easy to understand. Even people with minimal to no legal knowledge can follow along. Clear explanations in simple language are among its best traits. However, it may sometimes provide researchers with incorrect or partially accurate answers, which can mislead the concerned person.

D: Legal Expert Evaluation Rubric:

Human Evaluation Rating Criteria
1 – Incorrect and very poor quality
2 – Mostly incorrect and poor quality
3 – Partially correct and average quality
4 – Mostly correct and good quality
5 – Fully correct and excellent quality

E: Deployment Requirements:

Component	**Specification**
RAM	8–16 GB
Storage	10–20 GB (for vector DB and models)
CPU	Quad-core (minimum)
GPU (optional)	4 GB VRAM (for open-source LLMs like Gemma3-4B)
Operating System	Windows/Linux/macOS
Python Version	3.9 or higher
Frameworks	LangChain, LangGraph, Streamlit
Vector Store	ChromaDB
Model Serving	Ollama (for local deployment of Gemma3-4B)
PDF Parsing	PyMuPDF (fitz)
Web Search Integration	Tavily API Key

F: Screenshot of UI Interface: