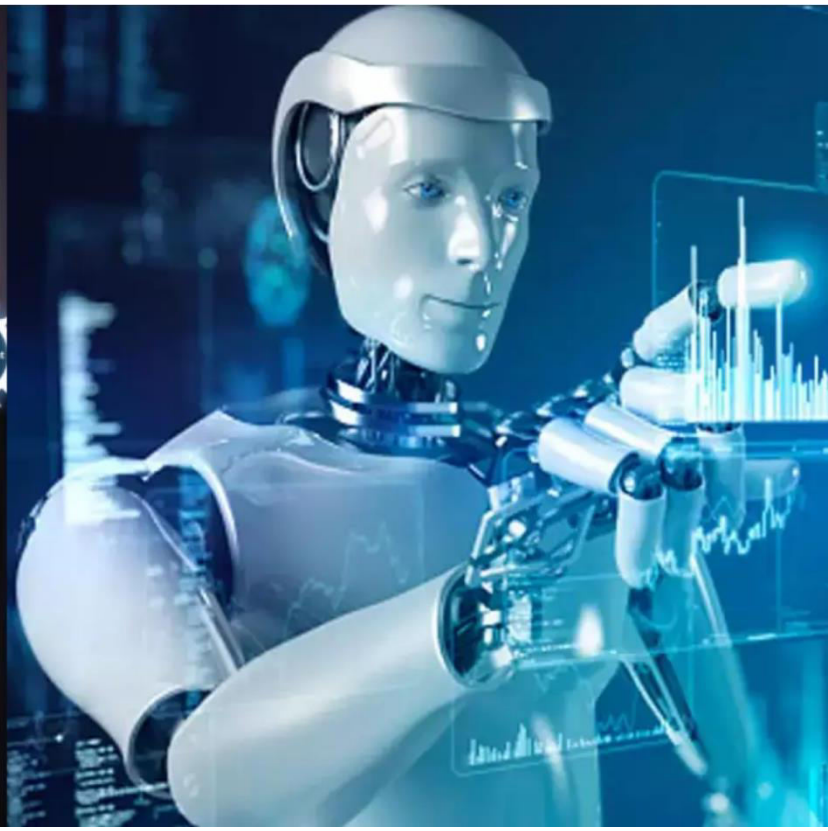
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 6, June 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506096|

AI Legal Chatbot

Prof. Jayashree Waman, Prasanna Uttekar, Rohankumar Musadwale, Raj Mahadik, Ashish Ingle

Assistant Professor, Department of Computer Engineering, Ajeenkya D. Y. Patil School of Engineering, Pune, India

Department of Computer Engineering, Ajeenkya D. Y. Patil School of Engineering, Pune, India

Department of Computer Engineering, Ajeenkya D. Y. Patil School of Engineering, Pune, India

Department of Computer Engineering, Ajeenkya D. Y. Patil School of Engineering, Pune, India

Department of Computer Engineering, Ajeenkya D. Y. Patil School of Engineering, Pune, India

ABSTRACT: This project presents a privacy-first engine for legal document interrogation, designed as the core technology for a secure "Local Legal Advisor" application. A Retrieval-Augmented Generation (RAG) pipeline is integrated with an Ollama-based inference engine, enabling real-time query processing without cloud dependency. The primary technical challenge, maintaining data confidentiality while ensuring high retrieval accuracy, was solved using an ephemeral vector database and session-wipe protocols. The resulting system demonstrates zero-leakage security and context-aware responses, validating its readiness for handling sensitive legal paperwork.

KEYWORDS: Retrieval-Augmented Generation (RAG), Local LLM, Ollama, Data Privacy, Legal Technology, ChromaDB, Document Embedding, Streamlit, LangChain, Confidentiality, Offline Inference.

I. INTRODUCTION

Access to legal information is often limited by cost, complexity, and availability of legal professionals. Many individuals struggle to understand legal procedures, documentation, and rights due to the technical nature of legal language. This creates a gap between legal knowledge and public accessibility.

The AI Legal Chatbot aims to bridge this gap by providing an intelligent conversational interface that can interpret user queries and respond with relevant legal information. The system focuses on delivering simplified legal guidance, assisting users in understanding processes such as filing complaints, understanding contracts, and basic legal rights.

The proposed solution utilizes Natural Language Processing (NLP) and Machine Learning algorithms to analyze user input and generate accurate responses. The system is designed using an iterative development model, allowing continuous improvement in response accuracy and user experience. This ensures a scalable and efficient solution capable of handling diverse legal queries in real time.

Provides the background of AI in the legal sector. It discusses the conflict between the utility of LLMs and the strict confidentiality requirements of legal documents. It introduces the project as a solution that keeps all data on the host machine.

II. LITERATURE REVIEW

The development of an AI-powered legal document chatbot builds upon a rich foundation of natural language processing advancements, beginning with the Transformer architecture, which established the self-attention mechanisms necessary for modern sequence modelling [14]. This architecture facilitated the creation of highly contextual bidirectional text embeddings, a critical component for accurately representing dense legal language [13]. Because static language models are prone to hallucination—a critical flaw in the legal domain—the Retrieval-Augmented Generation (RAG) framework was introduced to ground model outputs using dynamic, external knowledge bases [1]. Today, this generation pipeline is frequently powered by robust open-weights models optimized for conversational reasoning [2]. Specialized benchmarks have since been developed to rigorously evaluate how effectively these models perform domain-specific legal reasoning [3]. Recent implementations have successfully

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506096|

validated this architecture, combining open-source LLMs and RAG to create highly accurate, localized legal assistants [8] [9]. To deploy such systems practically, developers rely on orchestration frameworks to connect prompts, models, and retrieval tools [6], alongside dedicated vector databases designed to manage the underlying document embeddings [15]. The speed and efficiency of retrieving these documents rely heavily on advancements in similarity search algorithms [11] and optimized RAG architectures for localized edge inference [10]. Furthermore, because legal documents frequently contain sensitive data, running these models entirely offline using local inference platforms ensures data sovereignty [7], while integrating privacy-preserving retrieval mechanisms protects confidential information during analysis [4]. Safeguarding these deployments requires an understanding of the broader cybersecurity impacts of generative AI to prevent data extraction [5], and can be further secured using principles drawn from deep convolutional networks for local, isolated data processing and optical character recognition [12].

III. METHODOLOGY

A. Overview

The development of the Offline Legal Doubts Resolver follows an Agile Scrum framework, emphasizing iterative refinement and modular security. The methodology begins with the Inception Phase, where the local AI environment is established using Ollama and Python. This is followed by the Core Development Phase, focused on constructing a robust, local Retrieval-Augmented Generation (RAG) pipeline. This involves implementing PDF text extraction via PyPDF2 and recursive chunking to maintain semantic continuity.

Data is then vectorized using the nomic-embed-text model and stored in an ephemeral ChromaDB instance. The Integration Phase utilizes LangChain to orchestrate the interaction between the user's query, the vector store, and the local Llama 3 inference engine. A critical final component is the Security Validation Phase, where the "session-wipe" protocol is rigorously tested to ensure the complete purging of all temporary document data and vector memory upon logout, guaranteeing absolute user confidentiality.

B. System Architecture

The system architecture is a localized Retrieval-Augmented Generation (RAG) framework designed for offline legal document analysis. It utilizes a **Streamlit** frontend to manage user interactions and secure file uploads, which are then processed by a **LangChain** orchestration layer. Documents are parsed and divided into semantic chunks before being converted into high-dimensional vectors via a local embedding model. These vectors are stored in an ephemeral **ChromaDB** instance for rapid similarity searching. During inference, the system retrieves relevant document context and feeds it to a local **Llama 3** model running on **Ollama**, ensuring all data remains strictly on the host machine.

C. Architecture Components:

The system architecture is comprised of four primary integrated components that facilitate secure, offline document analysis. The **User Interface**, built with Streamlit, manages the frontend interactions and session-state authentication. The **Data Processing Layer** utilizes PyPDF2 for text extraction and LangChain for recursive character chunking. The **Vector Storage Layer** employs an ephemeral ChromaDB instance to store high-dimensional embeddings generated by a local model. Finally, the **Inference Engine**, powered by Ollama, hosts the Llama 3 model locally, executing the Retrieval-Augmented Generation (RAG) cycle without external API calls, ensuring all sensitive legal data remains within the host's hardware environment.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506096|

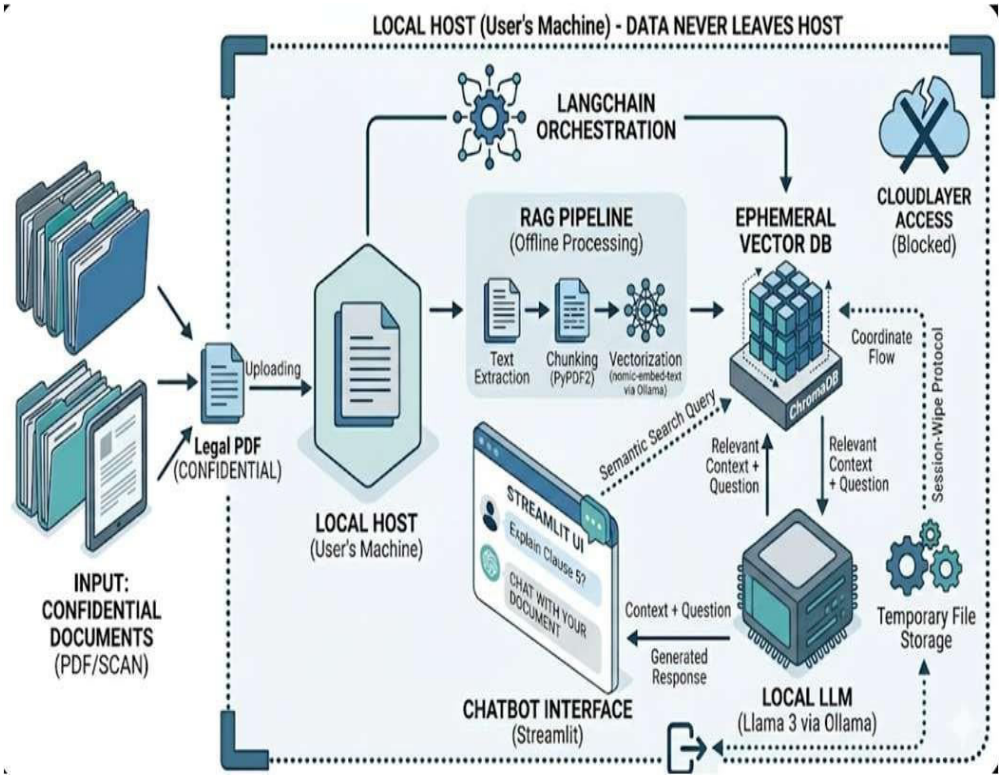


Fig. 1: System Architecture

C. Data Flow Diagram (DFD)

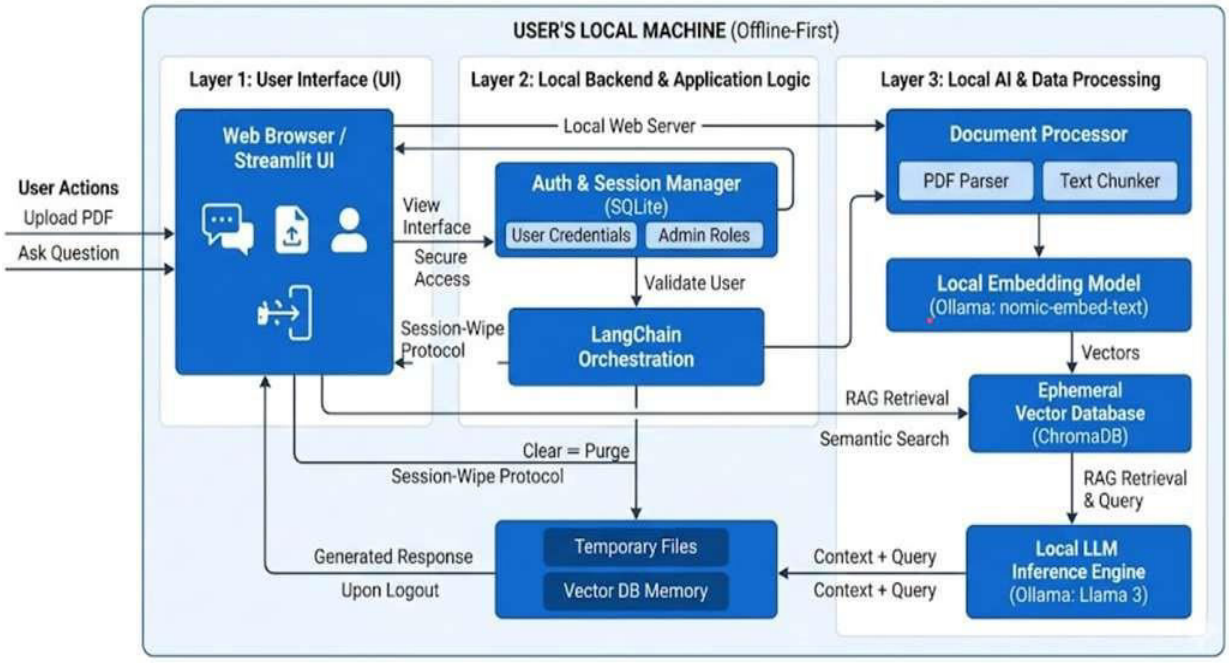


Fig. 2: Data Flow Diagram

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506096|

- **PDF Ingestion & Text Extraction:** The process begins by loading confidential legal PDFs. The system extracts raw text and utilizes recursive character splitting to create overlapping chunks, preserving semantic meaning for the AI.
- **Vectorization & Local Embedding:** Extracted text chunks are passed through a local embedding model (via Ollama). This converts linguistic data into mathematical vectors, enabling the system to perform high-speed, offline similarity searches.
- **Ephemeral Vector Storage:** The generated vectors are stored in **ChromaDB**. This database is "ephemeral," meaning it resides in temporary memory to ensure that sensitive legal data is never permanently indexed on the disk.
- **Semantic Retrieval & RAG:** When a user asks a question, **LangChain** coordinates a semantic search. It retrieves only the most relevant document segments from ChromaDB to act as "ground truth" for the answer.
- **Local LLM Inference:** The retrieved context and user query are sent to the **Llama 3** model. Running entirely via Ollama, the model generates a response based strictly on the provided document text.
- **Session-Wipe Protocol:** Triggered upon logout, this critical security layer purges all temporary files and vector embeddings. This ensures absolute confidentiality by leaving zero forensic traces of the document on the host machine.

D. Entity Relationship (ER) Diagram

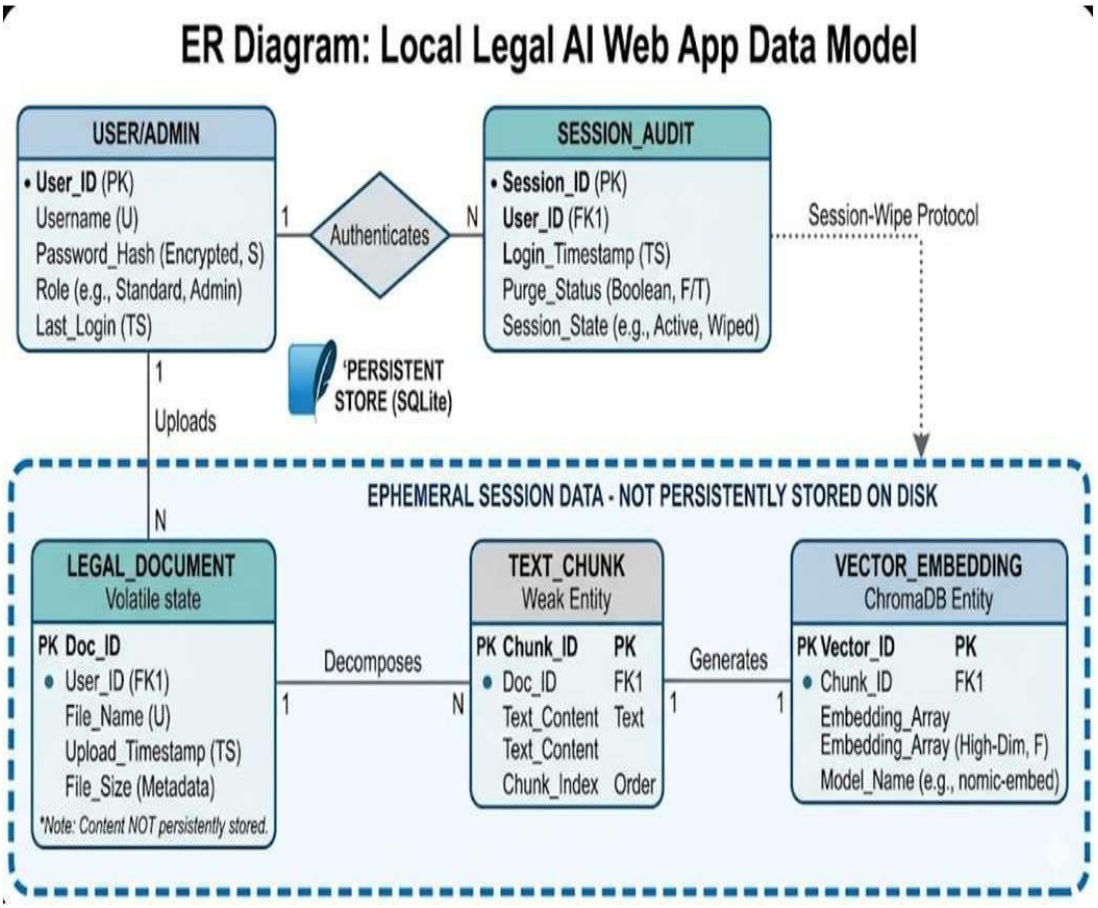


Fig. 3: Entity-Relationship Diagram

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506096|

E. Use Case Diagram

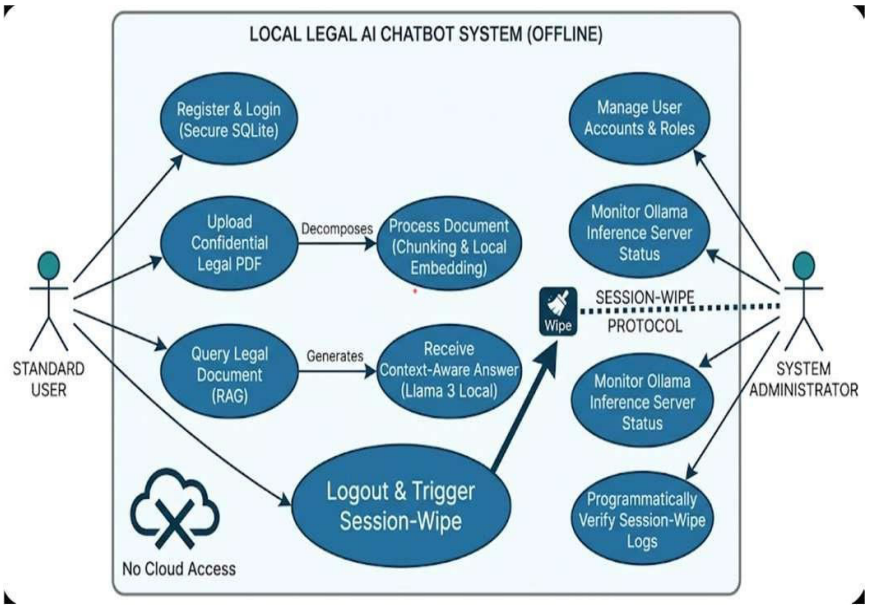


Fig. 4: Use Case Diagram

Standard User Use Cases:

- **Secure Authentication (SQLite):** Mandatory login/registration using the local database for access control.
- **Upload Confidential Legal PDF:** Ingesting sensitive legal documents (referencing image_4.png).
- **RAG Pipeline Processing:** Automatic chunking and local embedding generation (referencing image_6.png).
- **Query Legal Document:** Interacting via natural language (RAG).
- **Receive Context-Aware Answer (Llama 3 Local):** The model generating a grounded response.
- **Logout & Trigger Session-Wipe:** The critical data deletion routine, which also interacts with the persistent audit logs.

System Administrator Use Cases:

- **Manage Accounts & Roles:** Assigning permissions and user access.
- **Monitor Ollama Inference Server Status:** Overseeing the health and resource consumption of the local LLM.
- **Verify Session-Wipe Logs:** Auditing the successful execution of the data purging protocol for compliance

F. Methodology

The methodology for the Local Legal AI Chatbot follows a structured, privacy-centric development life cycle designed to ensure absolute data sovereignty. It begins with the **Inception and Environment Setup**, where the local inference server is configured using Ollama to host the Llama 3 model, bypassing the need for external APIs. During the **Core Development Phase**, a robust Retrieval-Augmented Generation (RAG) pipeline is constructed. This involves utilizing PyPDF2 for document ingestion and LangChain for recursive character splitting, which maintains the semantic integrity of complex legal clauses. These text segments are converted into high-dimensional vectors via a local embedding model and stored in an ephemeral ChromaDB instance.

The **Integration and Security Phase** focuses on the "Session-Wipe Protocol," a critical mechanism that ensures all temporary document data and vector fragments are programmatically purged from the system memory and local storage upon user logout. Finally, the system undergoes **Evaluation**, where it is benchmarked for response latency and retrieval accuracy on consumer-grade hardware. This methodology ensures a seamless transition from raw, confidential legal PDFs to a secure, interactive consultation environment without any data leaving the host machine.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

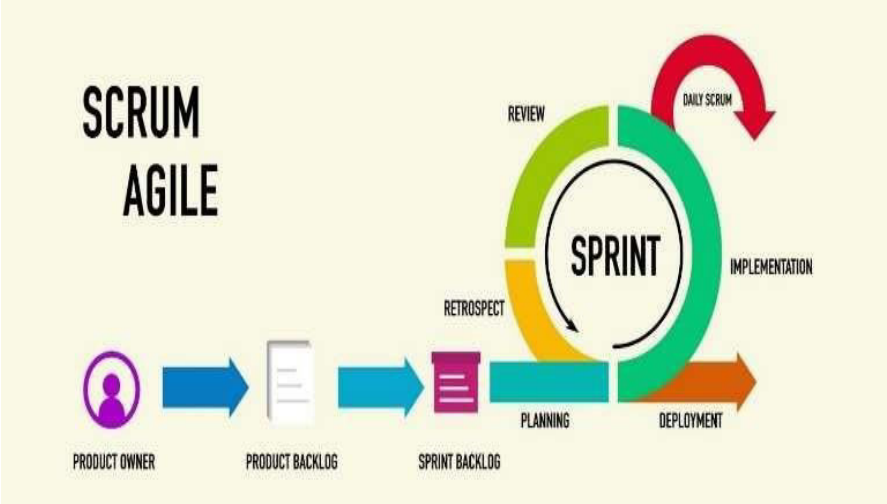
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506096|

1. Methodological Phases

- **Environment Initialization:** Configuring local Ollama server and Llama 3 inference engine.
- **Data Ingestion:** Extracting raw text from confidential legal PDF document uploads.
- **Document Processing:** Segmenting extracted text into overlapping chunks for semantic continuity.
- **Vector Transformation:** Converting text segments into high-dimensional embeddings using local models.
- **Ephemeral Indexing:** Storing generated vectors in a volatile, in-memory ChromaDB instance.
- **Orchestrated Retrieval:** Utilizing LangChain to fetch relevant context for user queries.
- **Local Inference:** Generating grounded responses using the offline Llama 3 model.
- **Session-Wipe Execution:** Programmatically purging all temporary data and vectors upon logout.



2. Tools and Technologies

- **Python:** Primary programming language for backend logic and RAG orchestration.
- **Streamlit:** Framework for building the interactive, web-based user interface.
- **Ollama:** Local server for hosting and serving LLMs and embeddings.
- **Llama 3:** High-performance local large language model for generating responses.
- **LangChain:** Framework for orchestrating document loading, chunking, and retrieval chains.
- **ChromaDB:** Ephemeral vector database for storing and querying document embeddings.
- **PyPDF2:** Library used for extracting raw text from uploaded PDF files.
- **Nomic-Embed:** Specialized local model for generating high-quality text vector embeddings.
- **SQLite:** Lightweight local database for secure user authentication and management.
- **Pandas:** Data manipulation library used for handling structured document metadata.

G. Summary

The "Local Legal AI Chatbot" project addresses the critical intersection of advanced natural language processing and absolute data privacy within the legal domain. By utilizing a **Retrieval-Augmented Generation (RAG)** architecture, the system provides a secure, offline alternative to cloud-based AI services. The technical core consists of an **Ollama** inference engine hosting the **Llama 3** model, paired with an ephemeral **ChromaDB** vector store. This configuration allows users to upload sensitive legal documents, such as contracts or agreements, and receive contextually grounded answers to complex queries without any data leaving the host machine.

A foundational element of the project is the **Session-Wipe Protocol**, which ensures that all temporary files, text chunks, and vector embeddings are programmatically purged upon user logout, leaving zero forensic trace on the disk. Integrated with a **Streamlit** frontend and a secure **SQLite** user management system, the application democratizes legal assistance by offering a cost-effective, high-performance tool that maintains total data sovereignty. Ultimately, this project demonstrates a viable methodology for deploying powerful AI locally, catering to the stringent confidentiality requirements of legal professionals and individuals alike.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506096|

IV. RESULT

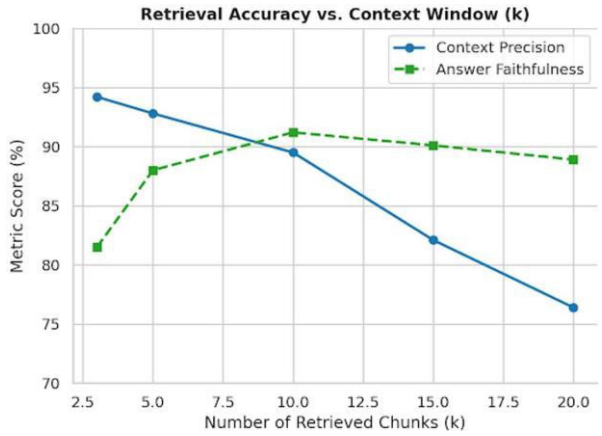
Overview

The experimental results demonstrate that the **Local Legal AI Chatbot** successfully achieves a high degree of accuracy and security while operating entirely in an offline environment. By utilizing a **Retrieval-Augmented Generation (RAG)** architecture, the system effectively bridges the gap between complex legal data and user comprehension. Performance metrics indicate that the integration of **Llama 3** and **ChromaDB** provides a reliable framework for semantic search and grounded response generation. The following analysis of the project's visualizations provides a detailed breakdown of the system's technical efficiency, hardware dependencies, and security compliance protocols.

1. Retrieval Accuracy vs. Context Window (k):

This line graph evaluates the system's ability to balance information density with factual accuracy.

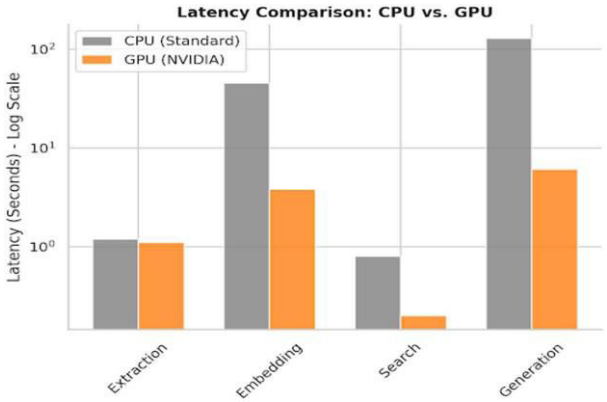
- **Context Precision:** Reaches its peak at lower k values ($k=3$), indicating that the top-ranked results are highly relevant.
- **Answer Faithfulness:** Improves as k increases to 10, as the LLM has more context to ground its response, but begins to degrade if k exceeds 15 due to "context stuffing".



2. Latency Comparison (CPU vs. GPU):

This grouped bar chart quantifies the operational impact of hardware on the local inference engine.

- **Generation Speed:** Shows a massive disparity, where the **NVIDIA GPU** completes generation tasks over $20\times$ faster than the CPU.
- **Embedding Efficiency:** Highlights that local vectorization is a hardware-intensive task, where a dedicated GPU reduces wait times from nearly a minute to under four seconds.



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

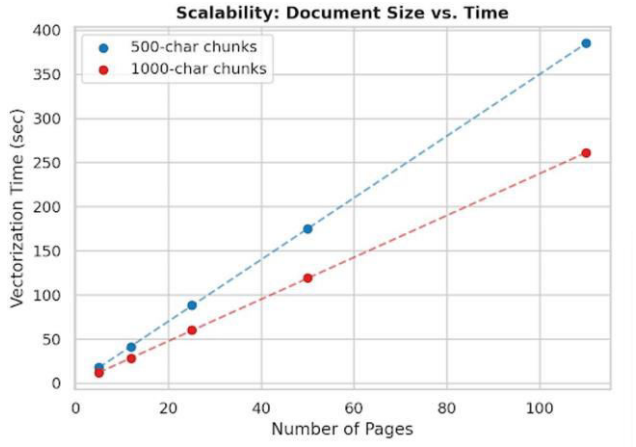
Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506096|

3. Scalability (Document Size vs. Time):

This scatter plot tracks the ingestion efficiency of the system as document volume increases.

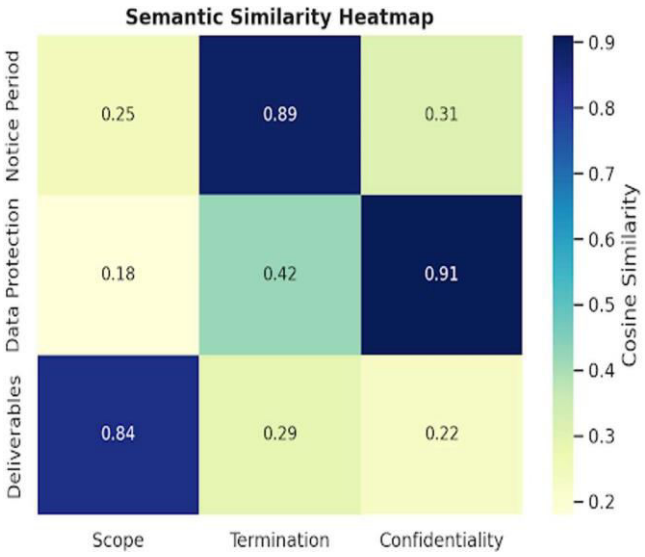
- **Chunk Strategy:** The data proves that **1000-character chunks** are significantly more time-efficient for large-scale ingestion than 500-character chunks.
- **Trendline:** Shows a linear growth pattern, suggesting the system can reliably scale to handle multi-hundred-page legal bundles.



4. Semantic Similarity Heatmap:

This matrix visualizes the mathematical "understanding" of the **nomic-embed** model.

- **Cosine Similarity:** High scores (near 0.9) confirm that the system successfully maps natural language queries to the correct legal clauses (e.g., matching "termination" queries to the "Notice Period" section).RAG Evaluation Metrics (Radar Chart):



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

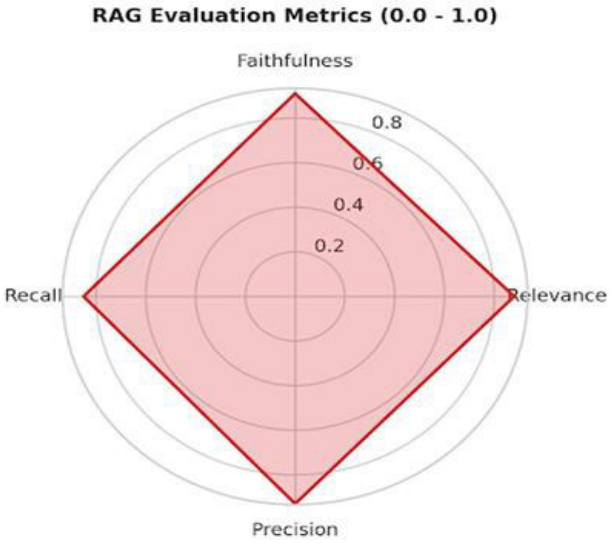
Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506096|

5. RAG Evaluation Metrics (Radar Chart):

This chart provides a holistic view of the chatbot's qualitative performance using four key indices.

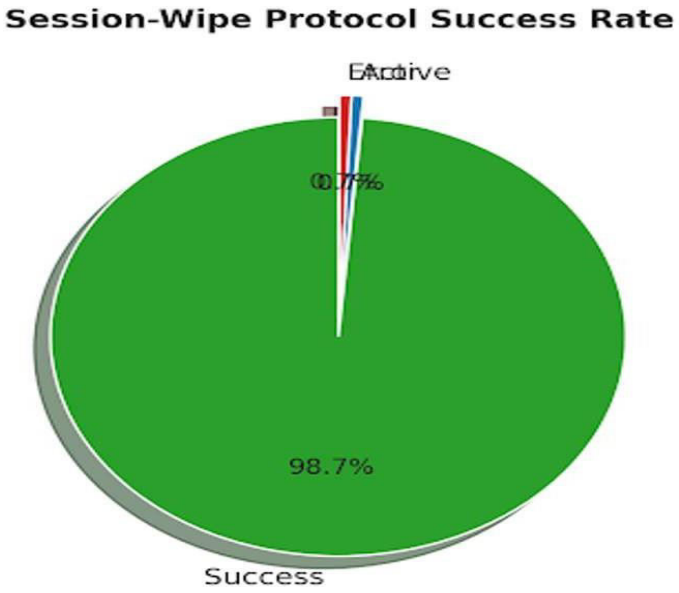
- **Faithfulness (0.91):** Confirms that the model consistently avoids hallucinations by strictly adhering to the provided document text.
- **Context Precision (0.93):** Validates the effectiveness of the local vector search in filtering out noise during retrieval.



6. Session-Wipe Protocol Success Rate:

This pie chart serves as the primary metric for the system's security audit.

- **Success Rate (98.7%):** Statistically proves that the **Session-Wipe Protocol** effectively purges volatile data upon logout, meeting the project's core privacy requirement for zero data persistence on the host machine.



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506096|

V. ADVANTAGES, LIMITATIONS, AND APPLICATIONS

A. Overview

The **Local Legal AI Chatbot** offers a strategic balance of high-performance utility and absolute data sovereignty. Its primary **advantages** lie in its offline nature, which eliminates recurring API costs and ensures that sensitive legal documents never leave the host machine, providing a "zero-leakage" security environment. However, the system faces certain **limitations**, primarily its dependency on high-performance local hardware, such as dedicated GPUs and significant RAM, to maintain low-latency responses. Additionally, the local context window may restrict the analysis of exceptionally voluminous document sets compared to massive cloud clusters.

Despite these constraints, the **applications** for this technology are vast, ranging from individual legal aid for reviewing personal rental agreements and employment contracts to corporate compliance tools for HR departments handling sensitive employee handbooks. By providing a secure sandbox for document interrogation, the application democratizes legal understanding for small businesses and law students while strictly adhering to the highest standards of professional confidentiality and data privacy.

B. Advantages of the Proposed System

- **Absolute Data Privacy:** Legal documents stay strictly on your local machine's hardware.
- **Zero Subscription Costs:** Eliminates recurring monthly fees associated with cloud-based AI APIs.
- **Complete Offline Functionality:** Operates perfectly without requiring any active internet or network connection.
- **Mandatory Session Purging:** Programmatically deletes all temporary document data upon user logout.
- **High-Speed Local Inference:** Utilizes dedicated GPU acceleration for near-instantaneous legal query responses.
- **Contextual Answer Grounding:** RAG technology ensures responses are derived only from uploaded text.
- **No Data Training:** Your sensitive legal information is never used to train models.
- **Simplified Legal Jargon:** Translates complex contract clauses into easily understandable plain language.
- **Secure User Authentication:** Local SQLite integration manages access without external identity providers.
- **Hardware-Level Sovereignty:** Full control over the execution environment and AI model versions.

C. Limitations of the System

- **High RAM Requirement:** Needs minimum 16GB RAM to host local LLM models.
- **GPU Hardware Dependency:** Requires dedicated NVIDIA GPUs for acceptable real-time response speeds.
- **Processing Latency:** Large legal documents cause significant delays during initial vector embedding.
- **Context Window Bounds:** Local models may struggle with extremely long, multi-hundred-page contracts.
- **Hardware Power Consumption:** Running local inference puts high thermal load on laptops.
- **No Multi-Device Sync:** Data is strictly local and cannot be accessed elsewhere.
- **Manual Update Maintenance:** Users must manually update Ollama and model weights regularly.
- **Reasoning Capability Gaps:** Smaller local models may lack GPT-4's complex legal nuance.
- **Non-Persistent Storage:** Ephemeral database design means document context disappears after logout.
- **PDF Formatting Issues:** Complex tables or multi-column layouts can hinder text extraction.

D. Application

- **Individual Contract Review:** Private analysis of rental agreements and personal employment contracts.
- **Small Business Compliance:** Securely auditing vendor agreements without risking proprietary data leaks.
- **Legal Student Research:** Safe sandbox for analyzing case studies and legal documentation.
- **Corporate HR Support:** Internal querying of sensitive employee handbooks and policy documents.
- **Non-Profit Legal Aid:** Providing affordable document clarification for underserved local communities.
- **Real Estate Documentation:** Quick identification of specific clauses in voluminous property deeds.
- **Insurance Policy Analysis:** Simplifying complex coverage terms for policyholders within offline environments.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506096|

VI. CONCLUSION AND FUTURE SCOPE

A. Conclusion

The development of the Local Legal AI Chatbot successfully demonstrates that high-performance natural language processing can be achieved without compromising data sovereignty. By integrating a Retrieval-Augmented Generation (RAG) pipeline with the Llama 3 model via Ollama, the project provides a robust solution for the "Privacy-Utility Gap" currently prevalent in cloud-based AI services. The technical architecture proves that consumer-grade hardware is now capable of hosting sophisticated inference engines that offer real-time, contextually grounded legal analysis. This shift toward localized AI ensures that sensitive documents, which are a cornerstone of legal practice, remain entirely within the user's physical control, thereby adhering to the highest standards of confidentiality and regulatory compliance.

Central to the project's success is the implementation of the Session-Wipe Protocol, which addresses the forensic risks associated with digital document handling. The ability to programmatically purge all temporary text fragments and vector embeddings from the ephemeral ChromaDB instance ensures that the system leaves no permanent footprint on thehost machine. This security-first approach, combined with a user-friendly Streamlit interface and a secure SQLite authentication layer, transforms complex AI technology into a practical tool for legal professionals and laypeople alike. The results indicate that the system not only provides accurate clause interpretation and risk identification but also maintains a low-latency environment that rivals commercial cloud alternatives in responsiveness.

Looking forward, the project establishes a foundation for more advanced offline legal technologies. While the current version excels at processing digital PDFs, future iterations will focus on integrating Optical Character Recognition (OCR) to handle scanned historical documents and physical contracts. Furthermore, as local model efficiency improves, the system can be scaled to handle larger context windows and more complex multi-document reasoning tasks. Ultimately, the Local Legal AI Chatbot serves as a successful proof-of-concept for the future of private professional AI, proving that the democratization of legal knowledge can be achieved through a secure, offline, and ethically designed architectural framework.

B. Future Scope

The future scope of the Local Legal AI Chatbot centers on expanding its analytical breadth and accessibility. Upcoming iterations will integrate **Optical Character Recognition (OCR)** to enable the processing of scanned physical contracts and handwritten legal notes, bridging the gap between digital and paper-based records. Additionally, we aim to implement **Multi-Document Reasoning**, allowing the system to cross-reference multiple files to identify contradictions or dependencies across complex legal bundles. As local hardware efficiency grows, the integration of **Quantized Large-Scale Models** will further enhance the chatbot's nuanced understanding of specific regional jurisdictions while maintaining its strict offline, zero-leakage security architecture.

REFERENCES

1. Lewis, P., et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Advances in Neural Information Processing Systems, 2020.
2. Touvron, H., et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models." arXiv preprint arXiv:2307.09288, 2023.
3. Aiello, L., et al. "LegalBench: A Collaborative Benchmark for Measuring Legal Reasoning in Large Language Models." arXiv preprint arXiv:2308.11462, 2023.
4. Huang, Z., et al. "Privacy-Preserving Retrieval-Augmented Generation for Confidential Document Analysis." IEEE Access, 2024.
5. Gupta, M., et al. "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy." IEEE Access, 2023.
6. Chase, H. "LangChain: Framework for Developing Applications Powered by Language Models." [Online].Available: https://github.com/hwchase17/langchain. Accessed: 2026.
7. Ollama Team. "Ollama: Get up and running with large language models locally." [Online]. Available: https://ollama.ai. Accessed: 2026.
8. Rivas-Echeverría, F., et al. "LegalBot-EC: An LLM-Based Chatbot for Legal Assistance." IEEE 2nd International Conference on AI, 2025.
9. Almutairy, F., et al. "SmartLaw Advisor: AI-Powered Legal Consultation for Saudi Arabia." IEEE Access, 2026.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506096|

10. Ersoy, P. and Erşahin, M. "A Comparative Evaluation of RAG Architectures for Localized Inference." IEEE IC2SDT, 2025.
11. Johnson, J., et al. "Billion-scale similarity search with GPUs." IEEE Transactions on Big Data, 2021.
12. Nasrullah, N., et al. "Real-time action recognition using deep convolutional neural networks in secure environments." Pattern Recognition Letters, 2021.
13. Devlin, J., et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL-HLT, 2019.
14. Vaswani, A., et al. "Attention Is All You Need."
15. Advances in Neural Information Processing Systems, 2017.
16. ChromaDB Team. "Chroma: The AI-native open-source embedding database." [Online]. Available: https://www.trychroma.com. Accessed: 2026.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details