Performance Evaluation of Open Source LLMs for Legal Document Summarization and Chatbot Integration for Legal Queries

Abhinaya Danda, Mrudula Gotmare, Gomati Iyer and Charusheela Nehete
Department of Information Technology, Vivekanand Education Society's Institute of Technology, Mumbai, India

Keywords: Legal Document Summarization, Chatbot Integration, OCR, Large Language Models (LLMs), LegalBERT, Llama 2, DeepSeek-V3, Legal Automation, Natural Language Processing (NLP).

Abstract: This paper presents a comprehensive evaluation of an AI-driven platform that combines cutting-edge technologies to enhance legal workflows. The platform integrates two primary features: legal document summarization using OCR-based techniques and a chatbot assistant for legal queries. The summarization module utilizes three advanced open-source LLMs-LegalBERT, Llama 2, and DeepSeek-V3-to process and condense complex legal documents into accessible summaries, focusing on obligations, risks, and key insights. The chatbot assistant provides instant conversational support for navigating legal processes, addressing queries, and ensuring better user engagement. By integrating OCR, NLP, and advanced AI models, the system simplifies legal document understanding for non-experts while offering reliable and efficient query resolution. A comparative performance analysis of the three LLMs highlights their strengths and limitations in handling diverse legal document types and chatbot interactions.Initial evaluations reveal significant advancements in summarization accuracy, query response quality, and overall user satisfaction. This study underscores the potential of AI in making legal support accessible, reducing time, effort, and costs for individuals and businesses.

1 INTRODUCTION

Legal documentation is often a complex and time-consuming process, particularly for individuals and small businesses that lack access to adequate legal resources. The intricate language, legal jargon, and structural complexity of legal documents can be overwhelming for those without formal legal training. This often leads to errors, misunderstandings, and, in extreme cases, legal disputes, which create barriers to effective legal communication and access to justice(Meena et al., 2024).

Traditionally, the legal industry has been resistant to automation due to the complexity of legal documentation and the critical need for precision and reliability. However, advancements in Artificial Intelligence (AI), Natural Language Processing (NLP), and Optical Character Recognition (OCR) have paved the way for automating routine legal tasks. These technologies can now address long-standing challenges in document preparation, summarization, and advisory services.

This paper introduces an AI-driven platform that evaluates and compares three open-source Large Language Models (LLMs)-LegalBERT, Llama 2, and DeepSeek-V3-for two critical legal tasks: document summarization and chatbot-based query handling. The system integrates OCR to extract text from legal documents and leverages NLP for summarizing key obligations, risks, and insights in a user-friendly format. The chatbot assistant provides conversational support for answering legal queries, improving accessibility to legal information, and facilitating better client-lawyer interactions.

By integrating document summarization and chatbot functionality, the platform reduces the time, effort, and expertise required to understand and manage legal documents. This comprehensive approach not only improves efficiency but also democratizes access to legal resources, offering cost-effective and user-friendly solutions for individuals and businesses.

This paper evaluates the performance of LegalBERT, Llama 2, and DeepSeek-V3 for their effectiveness in summarizing legal documents and supporting chatbot interactions. It further demonstrates the potential of AI-driven systems to transform the legal industry by reducing human error, enhancing accessibility, and bridging the gap between legal professionals and clients.

Danda, A., Gotmare, M., Iyer, G. and Nehete, C.
Performance Evaluation of Open Source LLMs for Legal Document Summarization and Chatbot Integration for Legal Queries.
DOI: 10.5220/0013609600004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 3*, pages 114-119
ISBN: 978-989-758-763-4

2 RELATED WORK

Recent advancements in Artificial Intelligence (AI), Natural Language Processing (NLP), and Optical Character Recognition (OCR) have driven the development of various systems aimed at automating legal workflows (Imogen et al., 2024),(Pandey et al., 2024). These systems primarily focus on enhancing efficiency in tasks such as document summarization, information retrieval, legal drafting, and decision-making within the legal domain.

While these solutions have demonstrated significant potential, many fall short in addressing the complexities inherent in legal language. Additionally, they often lack integrated functionalities that combine document summarization, conversational assistance, and multi-feature automation into a cohesive platform. Real-time interaction with legal professionals and the ability to handle diverse legal tasks through a single system remain underexplored.

To better understand the current landscape of AI applications in the legal domain, we conducted a comprehensive review of existing literature. This review focuses on key studies that highlight advancements in legal document summarization, chatbot integration for legal queries, and the comparative performance of AI models for such tasks. The following table summarizes the contributions of these studies, detailing their strengths, limitations, and implications for building an integrated, AI-driven legal platform.

3 METHODOLOGY

The Legalize platform automates legal document workflows by leveraging advanced technologies such as AI, OCR, and NLP (Vayadande, 2024),(Jain et al., 2024). It features a chatbot assistant for interactive user support, OCR-based document summarization, an advisory portal for scheduling client-lawyer consultations, and template-based automation for drafting legal documents. Additionally, a voice assistant powered by Speech Recognition allows users to dictate legal content, enhancing efficiency. Machine learning algorithms detect and correct content issues, ensuring high-quality outputs and reliability.

3.1 Steps

1. **User Actions:** Users access the platform to perform legal tasks, such as drafting, summarizing, or querying legal documents.
2. **Summarize Documents:**
   (a) Utilized OCR and NLP to extract key information from legal documents.
   (b) Perform Q&A based on the summarized content for deeper insights.
3. **Consult Legal Professionals:** Schedule meetings with legal professionals.
4. **Voice Assistant:** Use a voice assistant to dictate the content of legal documents.
5. **Chatbot:** Users interact with the chatbot to clarify legal queries or understand document summaries.

4 PROPOSED SYSTEM

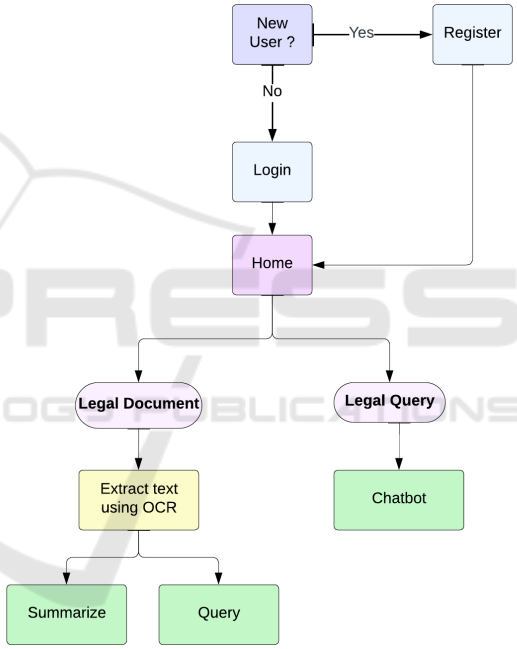

Figure 1: Proposed System.

4.1 Overview

To address the challenges and limitations of existing legal document management and query systems, our proposed solu- tion integrates advanced features to enhance efficiency, acces- sibility, and accuracy. The system's architecture, as illustrated in Fig. 1, offers two core functionalities-Legal Document Processing and Legal Query Resolution-streamlined into an intuitive user workflow.

The Legal Document Processing module includes an Optical Character Recognition (OCR) capability that extracts text from uploaded legal documents, enabling further operations. Users can leverage the

Summarization feature to condense lengthy legal texts into concise summaries that highlight essential points, such as key obligations, risks, and critical clauses. This feature allows users to quickly comprehend complex legal documents without needing extensive manual review. Additionally, the system facilitates a Query function, allowing users to pose specific questions about the document's content, helping them make informed decisions.

The Legal Query Resolution module is powered by a chatbot designed to handle complex legal questions. Users can interact with the chatbot to gain contextual answers about legal terms, clauses, or scenarios. This feature enables individuals to clarify legal concepts without requiring direct consultation with legal professionals, empowering them to navigate legal complexities independently.

The system is designed with user convenience in mind, starting with a seamless registration and login process. New users are prompted to register, while returning users can log in directly to access the Home interface. From the Home page, users can navigate to their desired functionality-processing legal documents or resolving legal queries.

This proposed system bridges the gap between technical efficiency and user-friendliness by incorporating advanced features such as OCR-based text extraction, document summarization, and chatbot-driven legal query resolution. By streamlining legal workflows, reducing manual effort, and providing accessible solutions, the system aims to empower users to handle legal documents and queries confidently and efficiently.

5 RESULTS

5.1 Model Performance Evaluation for Summarization

The purpose is to **evaluate and compare the performance of three open-source Large Language Models (LLMs) - LegalBERT, Llama 2, and DeepSeek-V3 - specifically for legal document summarization tasks**. The evaluation metrics employed include ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L), human evaluation (mean score on a 5-point scale), and factual accuracy. The experimental setup involved testing these models on a curated dataset of legal documents, including case summaries, contracts, and regulatory texts.

5.1.1 Model by Model Performance

LegalBERT: LegalBERT demonstrated strong performance in legal document summarization, reflecting the advantage of domain-specific training (V. Oliveira et al., 2024) (Queudot et al., 2020). It achieved the highest scores across most evaluation metrics, particularly excelling in factual accuracy and its ability to capture legal terminology. However, its computational requirements, including high memory and processing power, could limit its applicability in resource-constrained environments.

Llama 2: Llama 2 showcased impressive versatility, performing well across a range of NLP tasks, including legal document summarization. It produced fluent and coherent outputs and was effective in capturing the contextual nuances of legal texts. Nevertheless, limitations such as restrictions on commercial use and slightly lower factual accuracy compared to LegalBERT were observed.

DeepSeek-V3: DeepSeek-V3 excelled in tasks requiring advanced summarization and information retrieval, benefiting from its innovative architecture. While it achieved competitive scores across all metrics, it is a relatively new model with limited community support, which might pose challenges for long-term adoption and development.

5.1.2 Comparative Analysis and Insights

Quantitative Comparison -The table below summarizes the performance of each model based on the evaluation metrics:

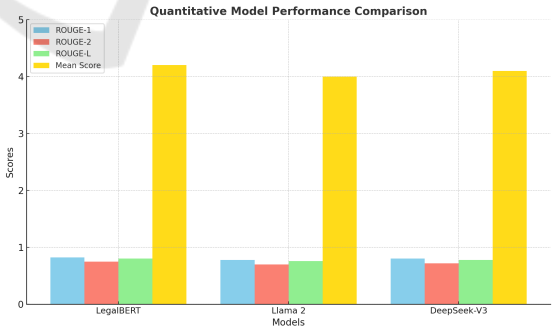

Figure 2: Quantitative Model Performance Comparison

Statistical analysis using t-tests confirmed that the differences in performance between LegalBERT and the other models were statistically significant ($p < 0.05$), particularly in legal document summarization tasks.

Qualitative Comparison Qualitative analysis re-

Table 1: Quantitative model performance comparison.

Model	ROUGE-1	ROUGE-2	ROUGE-L	Mean Score
LegalBERT	0.82	0.75	0.80	4.2/5
Llama 2	0.78	0.70	0.76	4.0/5
DeepSeek-V3	0.80	0.72	0.78	4.1/5

vealed distinct patterns in the strengths and weaknesses of each model:

- **LegalBERT:** Consistently captured domain-specific terminology and legal jargon with high accuracy but occasionally generated verbose outputs.
- **Llama 2:** Produced the most fluent and reader-friendly summaries but struggled with certain legal intricacies.
- **DeepSeek-V3:** Excelled in extracting key information and generating concise summaries but occasionally omitted critical context.

Examples of generated summaries illustrate these patterns, with LegalBERT providing detailed but dense summaries, Llama 2 delivering fluent but less precise summaries, and DeepSeek-V3 offering concise and targeted summaries.

In conclusion, while LegalBERT performed best overall for legal document summarization, the choice of model should be guided by specific use case requirements and resource availability.

5.2 Chatbot Evaluation

5.2.1 Dataset and Preprocessing

For demonstration purposes, we compiled and processed a comprehensive dataset comprising major rental laws pertaining to both housing and commercial properties in Maharashtra. The data scraping involved extracting key legislative texts, regulations, and guidelines from various official sources and legal databases. The scraped legal texts were meticulously cleaned and transformed into a structured format suitable for input into the RAG framework. This process included:

- **Text Normalization:** Removing extraneous characters and standardizing legal terms.
- **Chunking:** Splitting the large texts into smaller, manageable chunks to enhance retrieval accuracy and ensure relevant context during query processing.
- **Indexing:** Creating a searchable index using models like BM25 for efficient document retrieval.

5.2.2 Implementation of RAG Approaches

Two RAG approaches, *BM25+BERT* and *BR52_BART*, were implemented and evaluated in answering legal queries derived from rental agreements:

- **BM25+BERT:** This approach combines the BM25 algorithm, which scores documents based on term frequency and document length, with BERT, a transformer model that provides semantic understanding. BM25 serves as the initial filter, retrieving potentially relevant documents, while BERT refines this selection by evaluating the semantic similarity between the query and the retrieved documents.
- **BR52_BART:** This variant integrates a BR52 retriever, known for its semantic search capabilities, with BART, a transformer model optimized for sequence-to-sequence tasks. BART generates responses based on the context provided by the BR52 retriever, aiming for fluency and coherence in the output.

5.2.3 Evaluation Metrics

To assess the effectiveness of these models, we utilized ROUGE scores, a set of metrics that measure the overlap between the generated text and reference text. The scores provide insights into how well the generated responses reflect the essential content of the source documents.

5.2.4 Analysis of Results

The ROUGE scores show that *BM25+BERT* outperformed *DPR*, especially for complex queries like "What are the Documents Mandatory for a Commercial Rental Agreement?"

- **BM25+BERT Performance:** Retrieved and synthesized relevant information, providing document chunks that included:
  - Mandatory documentation such as Aadhar card, proof of business establishment, and government approvals.
  - Steps for registering the commercial rental agreement on non-judicial stamp paper.

Table 2: Comparison of ROUGE scores for BM25+BERT and DPR models in chatbot query response generation.

Document	ROUGE-1	ROUGE-2	ROUGE-L
BM25 + BERT Doc 1	0.80	0.70	0.75
BM25 + BERT Doc 2	0.68	0.58	0.65
BM25 + BERT Doc 3	0.72	0.60	0.68
DPR Doc 1	0.72	0.65	0.70
DPR Doc 2	0.65	0.50	0.60
DPR Doc 3	0.68	0.55	0.62

  - Legal obligations of both parties involved in the agreement.
- **DPR Performance:** Retrieved relevant documents but lacked fluency and completeness in synthesized outputs:
  - Highlighted essential documentation required for a commercial rental agreement, such as identity proofs and property ownership verification.

5.2.5 Performance of GPT-3.5, LLaMA 3.1, and Gemini 1.5 Pro on RAG

The evaluation extended to three state-of-the-art models- GPT-3.5, LLaMA 3.1, and Gemini 1.5 Pro- on the BM25+BERT RAG framework. Below is an analysis of their responses to the query regarding mandatory documents for a commercial rental agreement:

- **GPT-3.5:** Provided a complete list of necessary documents along with the steps involved in the registration process. It included detailed procedural information, ensuring clarity on both document requirements and legal obligations.
- **LLaMA 3.1:** Delivered a basic list of required documents without much elaboration on the registration process or additional legal steps. The response was concise but less informative compared to other models.
- **Gemini 1.5 Pro:** Offered a list of necessary documents accompanied by brief explanations of each item's significance. It provided a middle ground, offering more detail than LLaMA 3.1 but less procedural depth than GPT-3.5.

5.2.6 Summary of Results

- GPT-3.5 excelled in providing a thorough and comprehensive answer, covering both documents and procedural details.
- LLaMA 3.1 was succinct but lacked the depth and context provided by the other models.
- Gemini 1.5 Pro balanced between listing necessary documents and providing explanations, making its response informative yet concise.

6 CONCLUSION

This study highlights the effectiveness of open-source Large Language Models (LLMs) in improving legal document summarization and addressing legal queries through chatbot integration. The automated tools developed, including the document summarization module and the legal advisory chatbot, provide a robust solution to make legal documentation more accessible, understandable, and manageable for users.

The summarization tool significantly reduces the complexity of long legal documents, enabling users to focus on key points such as obligations, risks, and critical clauses. Chatbot integration, paired with natural language understanding, offers instant support for legal queries, fostering a more interactive and efficient user experience. Additionally, the incorporation of a voice assistant further simplifies navigation, allowing seamless interactions for users with limited technical expertise.

By leveraging advanced LLMs like LegalBERT, Llama 2, and DeepSeek-V3, this project demonstrates how AI can bridge the gap between legal complexity and user comprehension. The results underscore the potential of AI-powered solutions to transform the legal domain, making it more inclusive and accessible for individuals without specialized legal training. This integration of cutting-edge technology empowers users to handle legal matters confidently, streamlining workflows and reducing reliance on professional intervention for routine tasks.

7 FUTURE SCOPE

1. **Blockchain Integration:** Implement blockchain technology for secure, tamper-proof storage of le-

gal documents, ensuring authenticity and enabling smart contracts for automated execution of legal agreements (Hwang et al., 2023).

2. **Predictive Legal Analytics:** Incorporate machine learning algorithms to analyze past legal cases and predict potential outcomes, helping users assess risks and make informed decisions.
3. **Ethical and Explainable AI for Legal Decisions:** As AI becomes integral to legal processes, ensuring ethical decision-making and explainable AI models is crucial. Future systems could provide transparency by explaining how summaries or advice were derived, building trust and accountability in legal AI applications.

REFERENCES

Ajmi, A. (2024). Revolutionizing access to justice: The role of ai-powered chatbots and retrieval-augmented generation in legal self-help. *Brief*, 53(3).

Chakrabarti, D., Roy, S., and Bhattacharya, U. (2018). Use of artificial intelligence to analyse risk in legal documents for better decision support. In *TENCON 2018 - 2018 IEEE Region 10 Conference*, Jeju, Korea.

Chen, Y., Sun, Y., Yang, Z., and Lin, H. (2020). Joint entity and relation extraction for legal documents with legal feature enhancement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1561–1571.

Devaraj, P. N., PV, R. T., and Gangrade, A. (2023). Development of a legal document ai-chatbot. *arXiv preprint arXiv:2311.12719*.

et al., A. K. (2023a). Smart chatbot for guidance about children's legal rights. In *International Conference On Emerging Trends In Expert Applications & Security*, pages 405–412. Springer Nature Singapore.

et al., A. R. K. (2023b). Design and implementation of a chatbot for automated legal assistance using natural language processing and machine learning. In *2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS)*, pages 1–6. IEEE.

et al., B. S. R. (2024a). Optimization of bert algorithms for deep contextual analysis and automation in legal document processing. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

et al., R. B. S. (2024b). Lawbots: Utilization of ai chatbots for legal advising in the philippines. In *2024 IEEE 12th International Conference on Information, Communication and Networks (ICICN)*, pages 594–600. IEEE.

Gurram, J., Jagtap, R., Khambati, H., Mirajkar, M., and Deole, A. Legal procedure bot.

Hwang, C., Nam, J. S., and Laporte, E. (2023). Generating training datasets for legal chatbots in korean. In *International Conference on Law and Society*, pages 1–4.

Igbinenikaro, E. and Adewusi, A. O. (2024). Navigating the legal complexities of artificial intelligence in global trade agreements. *International Journal of Applied Research in Social Sciences*, 6(4):488–505.

Imogen, P. V., Sreenidhi, J., and Nivedha, V. (2024). Ai-powered legal documentation assistant. *Journal of Artificial Intelligence and Capsule Networks*.

Jain, D., Borah, M. D., and Biswas, A. (2024). Summarization of lengthy legal documents via abstractive dataset building: An extract-then-assign approach. *Expert Systems with Applications*, 237:121571.

Kalpana, R. A., Karunya, S., and Rashmi, R. (2023). Legal solutions-intelligent chatbot using machine learning. In *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, pages 1–6. IEEE.

Langston, O. and Ashford, B. (2024). Automated summarization of multiple document abstracts and contents using large language models. *Authorea Preprints*.

Li, X., Huang, L., Zhou, Y., and Shao, C. (2021). Tst-gan: A legal document generation model based on text style transfer. In *2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE)*, pages 90–93. IEEE.

Meena, K., Tanusha, G., Renuka, G., and Saraswathi, K. (2024). Enhancing legal document management efficiency: An ai-powered solution addressing interpretation challenges. *Journal of Engineering and Technology Management*.

Pandey, R. R., Khandelwal, S., Srivastava, S., Triyar, Y., and Almas, M. M. (2024). Legal seva - ai powered legal documentation assistant. *International Research Journal of Engineering and Technology (IRJET)*.

Queudot, M., Charton, É., and Meurs, M. J. (2020). Improving access to justice with legal chatbots. *Stats*, 3(3):356–375.

Srivastav, E. (2024). Lawbot: A smart user indian legal chatbot using machine learning framework. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, pages 1–7. IEEE.

V. Oliveira, G. N., Faleiros, T., and Marcacini, R. (2024). Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents. *Artificial Intelligence and Law*, pages 1–21.

Vayadande, K. (2024). Ai-powered legal documentation assistant. In *2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN)*, pages 84–91. IEEE.