Legal Ai For All: Reducing Perplexity And Boosting Accuracy In Normative Texts With Fine-Tuned Llms And Rag

#1. CH.VENKATA GOPI, #2. M. VENKATA NARENDRA BABU, #3. A. DEVENDRA, #4. BANDLA.KARTHIK #5.Mrs. N. RADHA.M.E.

#1234 UG Students, Department of Artificial Intelligence and Data Science, School of Engineering & Technology, Dhanalakshmi Srinivasan University, Trichy-621112, Tamil Nadu
Email: cherukurigopi120@gmail.com, narendrababumindala@gmail.com, annangidevendra2004@gmail.com, bandlakarthik19@gmail.com

#5 Assistant Professor, Department of Artificial Intelligence and Data Science, School of Engineering & Technology, Dhanalakshmi Srinivasan University, Trichy-621112, Tamil Nadu
Email:radhan.set@dsuniversity.ac.in

----------**********----------

Abstract:

Legal texts are often written in highly technical and formal language, making them difficult for non-experts to understand and interpret correctly. With the growing adoption of artificial intelligence in the legal domain, large language models (LLMs) have shown promise in supporting natural language interaction with normative and regulatory documents. However, general-purpose LLMs frequently suffer from high perplexity, hallucinations, and limited domain reliability when applied to legal texts, which restricts their practical usability in real-world legal information systems.

To address these challenges, this paper proposes a reliable and enhanced legal text interpretation framework that combines fine-grained large language models with Retrieval-Augmented Generation (RAG). The approach leverages supervised fine-tuning on a synthetic yet human-validated legal question–answer dataset derived from official data protection laws and regulations, enabling clause-level and article-level understanding of legal content. In addition, a RAG architecture is integrated to ground model responses in authoritative legal sources, thereby reducing hallucinations and improving factual consistency while maintaining adaptability to evolving legal documents.

Experimental results demonstrate that the proposed framework significantly improves legal question-answering performance compared to baseline models. Fine-tuned models achieve substantial reductions in perplexity, while the RAG-enhanced system attains accuracy levels exceeding 90% across multiple difficulty categories in benchmark evaluations. These findings highlight that combining fine-grained model adaptation with retrievalgrounded generation provides a robust and scalable solution for reliable legal text interpretation, supporting broader accessibility and trustworthy use of AI-driven legal information systems.
.

Keywords —**Large language models (LLM), generative AI, fine tuning, RAG, Ecuadorian law, legal.**

I. INTRODUCTION

The rapid growth of digital legal information has created new opportunities for improving access to legal knowledge through intelligent systems. Traditionally, legal information systems rely on keyword-based search engines and static document retrieval methods. While these approaches enable users to locate legal documents, they often fail to capture the semantic meaning behind user queries. Legal texts typically contain complex sentence structures, domain-specific terminology, and hierarchical legal provisions, which make them difficult for non-experts to interpret. As a result, individuals seeking legal information frequently encounter barriers in understanding laws, regulations, and procedural guidelines.

Recent advancements in **large language models (LLMs)** have significantly improved the capabilities of natural language processing systems. These models demonstrate strong performance in tasks such as text generation, summarization, and question answering. In the legal domain, LLMs enable the development of conversational systems capable of interpreting legal texts and providing natural language explanations to user queries. Legal chatbots powered by LLMs have the potential to assist users in legal research, regulatory understanding, and access to public legal information services.

Despite their advantages, large language models face several challenges when applied to legal contexts. Models trained on general-purpose corpora may lack sufficient exposure to domain-specific terminology and legal reasoning structures. Furthermore, generative models may produce **hallucinated or unsupported statements**, which is particularly problematic in legal applications where factual accuracy and traceability are essential. These limitations highlight the need for mechanisms that ensure responses are grounded in authoritative legal sources.

To address these issues, **Retrieval-Augmented Generation (RAG)** has emerged as an effective approach for integrating external knowledge sources into language model generation. In a RAG framework, relevant documents are retrieved from a legal corpus using semantic search techniques and then incorporated into the generation process. This architecture allows the system to produce responses that are grounded in actual legal texts, improving factual reliability and interpretability. By combining information retrieval with neural text generation, RAG systems provide a scalable solution for knowledge-intensive tasks in legal domains.

Legal chatbot systems designed using retrieval-augmented approaches can support a wide range of applications, including legal question answering, regulatory compliance assistance, and public legal information services. These systems often rely on vector databases and embedding models to store and retrieve legal documents based on semantic similarity. Hybrid retrieval techniques combining dense and sparse representations further improve retrieval accuracy and coverage across diverse queries. By grounding generated responses in retrieved legal documents, such systems reduce hallucinations and increase user trust.

Another important factor in legal chatbot development is **domain adaptation**. Fine-tuning language models using legal corpora, statutes, and regulatory documents enables models to better understand legal terminology and context. In situations where annotated legal datasets are limited, synthetic question–answer pairs derived from authoritative legal documents can be used to expand training data. Human validation of such synthetic datasets ensures accuracy and reliability while enabling scalable dataset construction.

The integration of retrieval mechanisms, domain-adapted language models, and structured legal corpora has enabled the development of advanced legal conversational systems. These systems aim to simplify access to legal information while preserving the accuracy and transparency required in legal contexts. By supporting natural language interaction with complex legal documents, legal chatbots can reduce barriers associated with traditional legal information systems.

As legal frameworks continue to evolve and expand, scalable solutions are required to maintain up-to-date legal knowledge systems. Retrieval-augmented

architectures address this challenge by allowing new legal documents to be indexed and accessed without retraining the entire model. This design improves system maintainability and ensures that responses remain aligned with current legal regulations.

Overall, the integration of large language models with retrieval-based knowledge grounding represents a promising direction for legal artificial intelligence. Such systems have the potential to enhance accessibility, transparency, and efficiency in legal information services. Continued research in this area focuses on improving model reliability, reducing hallucinations, and ensuring responsible deployment in sensitive legal environments.

II. PROBLEM STATEMENT

The rapid growth of large-scale textual knowledge bases has increased the demand for intelligent systems capable of generating accurate, context-aware, and factually grounded responses to natural language queries. Pure large language models rely on parametric knowledge learned during training, which is static and constrained by the training data cutoff. As a result, such models may produce responses that are outdated, incomplete, or factually incorrect, commonly referred to as hallucinations. This limitation becomes critical in knowledge-intensive domains where correctness, traceability, and reliability are essential. The core problem is to design a system that can generate coherent textual responses while remaining aligned with an external and dynamically evolving knowledge source.

In a conventional language model, a response y is generated from an input query q by maximizing the conditional probability of the output given the input. This objective ignores external knowledge sources and can be expressed as follows.

$$y = \text{argmax}_P(y|q)$$

Retrieval-Augmented Generation extends this formulation by incorporating a retrieved document d obtained from a retrieval function applied to the query. The revised objective maximizes the conditional probability of the response given both the query and the retrieved context, enabling factual grounding and improved reliability.

$$y = \text{argmax}_P(y|q,d)$$

III. PROPOSED SYSYTEM

The proposed system is designed to improve the interpretation of complex legal texts by combining domainadapted language modeling with external knowledge grounding. It focuses on enhancing factual reliability and contextual understanding when processing regulatory and statutory documents. The system aims to address limitations of standalone language models in knowledge-intensive environments.

At the core of the system, an open-source large language model is adapted using supervised fine-tuning on structured legal question–answer data. This adaptation enables the model to capture legal terminology, clauselevel semantics, and domain-specific language patterns that are not sufficiently represented in general-purpose training corpora.

To mitigate hallucination and outdated knowledge issues, the proposed system integrates a retrieval-augmented generation mechanism. This mechanism retrieves relevant legal passages from an external corpus and injects them into the generation process, ensuring that responses remain grounded in authoritative and up-to-date legal sources.

The retrieval component employs dense and sparse text representations to support semantic matching between user queries and legal documents. Retrieved content is ranked based on relevance and provided as contextual input to the language model, enabling evidence-aware response generation and improved factual consistency.

The system architecture is modular, separating retrieval, generation, and evaluation components. This modularity allows individual components to be updated or replaced without retraining the entire model, improving scalability and maintainability in environments where legal texts evolve frequently.

Synthetic legal question–answer pairs are used to support scalable dataset construction when annotated legal data is limited. Human validation is applied to ensure correctness and relevance, enabling reliable supervision during fine-tuning while maintaining alignment with official legal documents.

During inference, the system processes natural language queries by first identifying relevant legal context and then generating responses conditioned on both the query and retrieved evidence. This

approach supports coherent, context-aware explanations rather than isolated text generation.

The proposed system is evaluated using benchmark datasets consisting of multiple-choice and open-ended legal questions. Performance metrics such as accuracy and consistency are used to assess improvements over baseline language models without retrieval support.

By decoupling knowledge storage from model parameters, the system enables rapid updates through corpus modification rather than repeated fine-tuning. This design significantly reduces computational cost while preserving the ability to reflect legal amendments and regulatory changes.

Overall, the proposed system provides a reliable and scalable framework for legal text interpretation by combining fine-grained language model adaptation with retrieval-based grounding. It is well suited for deployment in legal information systems requiring accuracy, transparency, and adaptability.

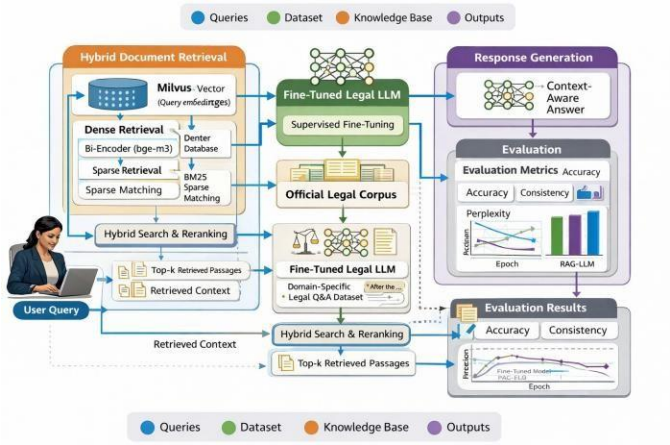
The main contributions of this study can be summarized as follows:

- We propose a retrieval-augmented generation framework that integrates supervised fine-tuning with external knowledge retrieval to improve legal text interpretation.
- We construct a domain-specific legal question–answer dataset derived from authoritative regulatory documents to support fine-grained model adaptation.
- We demonstrate that grounding language model outputs in retrieved legal context significantly reduces hallucinations and improves response reliability.
- We evaluate the proposed system using benchmark legal question-answering tasks and show measurable improvements in accuracy over baseline models.

Advantages of the Proposed System

- Reliability:
  The proposed system grounds generated responses in retrieved authoritative legal documents, significantly reducing hallucinations and improving factual consistency in legal text interpretation.
- DomainAdaptability:
  Supervised fine-tuning on legal question–answer datasets enables the system to better understand domain-specific terminology and clause-level semantics compared to general-purpose language models.
- Accuracy:
  The integration of retrieval mechanisms with fine-grained model adaptation improves accuracy across benchmark legal question-answering tasks when compared to standalone language models.
- MaintainableArchitecture:
  The modular architecture of the system allows individual components to be easily replaced or improved, ensuring long-term maintainability and efficient deployment in evolving legal environments.

IV. SYSTEM ARCHITECTURE



Existing Algorithm

SFT:

Supervised Fine-Tuning is commonly used to adapt large language models for domain-specific tasks. In this approach, a pretrained language model is trained on labeled question–answer datasets to improve its understanding of domain-specific language. In legal applications, SFT enables the model to learn legal terminology, document structure, and contextual patterns present in statutes and regulations. However, SFT-based systems rely primarily on knowledge stored within model parameters, which may lead to hallucinations or

outdated information when legal frameworks change.

Proposed Algorithm

Hybrid Dense–Sparse Retrieval Learning Model

The proposed algorithm integrates retrieval-based learning with supervised fine-tuning to improve accuracy and reliability in legal question-answering systems. The model combines **dense semantic embeddings** with **sparse keyword-based retrieval** to identify relevant legal documents from an external knowledge base. These retrieved documents are then provided as contextual input to the language model during response generation. This hybrid retrieval mechanism improves both recall and precision while ensuring that generated answers remain grounded in authoritative legal sources.

V. PROJECT MODULES

Module 1 : Legal Data Acquisition and Preprocessing

The Legal Data Acquisition and Preprocessing section focuses on collecting authoritative legal documents and transforming raw text into structured representations. This includes cleaning, normalization, segmentation, and preparation of question–answer pairs required for effective downstream learning.

The Domain-Specific LLM Fine-Tuning section describes adapting an open-source large language model using supervised learning on curated legal datasets. This process enables the model to capture legal terminology, clause-level semantics, and contextual dependencies present in normative texts.

The Hybrid Retrieval and Response Generation section integrates dense and sparse retrieval techniques with the fine-tuned language model. Retrieved legal context is injected into the generation process to produce grounded, accurate, and context-aware responses suitable for knowledge-intensive applications.

Module 2 : Domain-Specific LLM Fine-Tuning

The Hybrid Dense–Sparse Retrieval section explains the use of semantic vector embeddings combined with keyword-based matching to retrieve relevant legal passages. This dual strategy improves recall and precision when handling diverse legal queries with varying lexical and semantic structures.

The Context-Aware Response Generation section details how retrieved legal evidence is combined with the input query during inference. Conditioning generation on external context ensures that responses remain factually grounded, coherent, and aligned with authoritative legal sources.

The System Evaluation and Validation section outlines the experimental setup used to assess model performance. Metrics such as accuracy, consistency, and perplexity are employed to compare the proposed system against baseline models and validate its effectiveness.

Module 3 : Hybrid DenseSparse Retrieval

The Knowledge Base Construction section focuses on organizing official legal documents into a structured corpus suitable for retrieval. Legal texts are indexed using dense embeddings and sparse representations to support efficient semantic search and accurate context extraction during inference.

The Query Understanding and Reformulation section handles the analysis of user input to improve retrieval effectiveness. Queries are processed to capture intent and key legal concepts, enabling better alignment between the query and relevant passages in the legal corpus.

The System Integration and Deployment section describes how individual components are combined into a unified pipeline. Emphasis is placed on modular design, scalability, and ease of maintenance to support real-world deployment in evolving legal environments.

Module 4 : Retrieval-Augmented Response Generation

The Data Validation and Quality Control section emphasizes ensuring the correctness and relevance of legal data used in the system. Human verification and consistency checks are applied to minimize noise, preserve legal meaning, and improve the reliability of supervised learning outcomes.

The Model Optimization and Efficiency section discusses techniques used to reduce computational cost while maintaining performance. Parameter-efficient fine-tuning and optimized inference

strategies enable deployment on limited hardware without compromising response quality.
The Error Analysis and Performance Monitoring section examines system outputs to identify common failure patterns. Continuous monitoring supports iterative refinement of retrieval strategies and model behavior, leading to sustained accuracy and robustness over time.

Module 5 : System Evaluation and Validation
The Security and Data Governance section addresses the handling of sensitive legal information within the system. Access control mechanisms and data management policies are considered to ensure that stored legal documents and generated responses comply with ethical and governance requirements.
The Explainability and Transparency section focuses on enabling traceable outputs by linking generated responses to retrieved legal sources. This supports interpretability and helps users understand how conclusions are derived from authoritative legal texts.
The System Scalability and Maintenance section discusses long-term operability of the proposed framework. Design choices emphasize modular upgrades, efficient indexing, and seamless integration of newly added legal documents as regulations evolve.

VI. CONCLUSION

This study has underscored the significant potential of open-source large language models, when combined with fine-tuning techniques and Retrieval-Augmented Generation, to enhance access to and comprehension of legal information within the Ecuadorian context. The research focused particularly on the Ley Orgánica de Protección de Datos Personales and its regulation, shedding light on how these technologies can address real-world legal challenges.
The findings revealed that fine-tuned models achieved substantial reductions in perplexity while markedly improving accuracy on specific legal queries, outperforming their baseline counterparts by a considerable margin. Despite these advancements, Retrieval-Augmented Generation systems demonstrated superior overall performance, particularly due to their adaptability to frequent legislative updates.
Among the key contributions of this work are the development of the first open-source models and specialized legal benchmarks tailored to the Ley Orgánica de Protección de Datos Personales in Ecuador. These resources provide a robust foundation for evaluating large language models in this specific domain and contribute to the democratization of legal knowledge.

FUTUREWORK

Looking ahead, future investigations could explore the adaptation of these models to additional legal fields within Ecuador, incorporate innovative techniques to optimize computational resources, and foster greater interdisciplinary collaboration among experts in artificial intelligence, law, and related disciplines to enhance the versatility and impact of these tools across diverse legal contexts.

REFERENCES

[1] P. S. García-Montero, P. Vizcaíno, I. G. Reyes-Chacón, and M. E. Morocho-Cayamcela, “Legal AI for all: Reducing perplexity and boosting accuracy in normative texts with fine-tuned LLMs and RAG,” *IEEE Access*, 2025.

[2] V. Raghupathi, Y. Zhou, and W. Raghupathi, “Legal decision support: Exploring big data analytics approach to modeling pharma patent validity cases,” *IEEE Access*, 2018.

[3] J. Cui, X. Shen, and S. Wen, “A survey on legal judgment prediction: Datasets, metrics, models and challenges,” *IEEE Access*, 2023.

[4] C. He, T. P. Tan, X. Zhang, and S. Xue, “Knowledge-enriched multi-cross attention network for legal judgment prediction,” *IEEE Access*, 2023.

[5] A. S. Imran, H. Hodnefjeld, Z. Kastrati, N. Fatima, S. M. Daudpota, and M. A. Wani, “Classifying European Court of Human Rights

cases using transformer-based techniques," *IEEE Access*, 2023.

[6] M. Kutbi, "Named entity recognition utilized to enhance text classification while preserving privacy," *IEEE Access*, 2023.

[7] O. A. Cejas, M. I. Azeem, S. Abualhaija, and L. C. Briand, "NLP-based automated compliance checking of data processing agreements against GDPR," *IEEE Transactions on Software Engineering*, 2023.

[8] A. Iftikhar, S. W. U. Q. Jaffry, and M. K. Malik, "Information mining from criminal judgments of Lahore High Court," *IEEE Access*, 2019.

[9] G. Li, Z. Wang, and Y. Ma, "Combining domain knowledge extraction with graph long short-term memory for learning classification of Chinese legal documents," *IEEE Access*, 2019.

[10] A. Munthuli et al., "Transformers for multi-intent classification and slot filling of Supreme Court decisions related to sexual violence law," *IEEE Access*, 2023.

[11] V. Javidroozi, H. Shah, and G. Feldman, "FABS: A framework for addressing the business process change challenges for smart city development," *IEEE Access*, 2023.

[12] L. Yan et al., "Practical and ethical challenges of large language models in education: A systematic scoping review," *British Journal of Educational Technology*, vol. 55, no. 1, pp. 90–112, 2024.

[13] M. U. Hadi et al., "A survey on large language models: Applications, challenges, limitations, and practical usage," *Authorea Preprints*, 2023.

[14] S. Tan and Y. Guo, "A study of the impact of scientific collaboration on the application of large language models," *AIMS Mathematics*, vol. 9, no. 7, pp. 19737–19755, 2024.

[15] M. A. K. Raiaan et al., "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," *IEEE Access*, vol. 12, pp. 26839–26874, 2024.

[16] Anthropic, "The Claude 3 model family: Opus, Sonnet, Haiku," Mar. 2024.

[17] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural legal judgment prediction in English," arXiv:1906.02059, 2019.

[18] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017.

[19] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media, 2009.

[20] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Englewood Cliffs, NJ: Prentice Hall, 2009.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[22] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, pp. 3–26, 2007.