JURIS.AI: A Legal Assistant For Understanding The Consumer Protection Act, 2019

Adithya Vikas A, Aparajitha P, Aravindhan S S, and R. K. Kapilavani

Abstract

The democratization of legal knowledge is a vital step toward ensuring that individuals can make informed decisions about their rights and responsibilities. This project, JURIS.AI, aims to make legal knowledge—particularly Indian consumer protection rights—accessible to all, regardless of legal expertise. Leveraging Natural Language Processing (NLP) and Machine Learning (ML), the project personalizes responses, provides comprehensive legal information, and breaks down complex legal terminology.
By incorporating Large Language Models (LLMs) for interactive Q&A, and using techniques like Parameter-Efficient Fine-Tuning (PEFT) and Retrieval-Augmented Generation (RAG), the system ensures accurate, context-sensitive answers. Key features include a robust legal knowledge base, text-to-speech and speech-to-text accessibility, and step-by-step legal guidance. Notably, the system's performance is measured using the ROUGE score, improving from 0.69 (QLoRA) to 0.84 (QLoRA + RAG). This tool empowers users with varying levels of legal literacy to navigate the legal landscape confidently, offering practical legal assistance without specialized training.

Index Terms

Quantized Low Rank Adaptation (QLoRA), Retrieval-Augmented Generation (RAG), Parameter Efficient Fine Tuning (PEFT), Large Language Models (LLMs), ROUGE, Legal NLP, Consumer Protection Act.

I. INTRODUCTION

Recent advancements in large language models (LLMs) have created new opportunities for applications in specialized domains such as law and legal services [15], [16]. Traditional legal services often present barriers to accessibility due to complexity, cost, and specialized language [17]. Conversa- tional AI systems offer promising solutions to these challenges by providing user-friendly interfaces for legal guidance [3], [7]. JURIS.AI represents a novel approach to legal assis- tance through the integration of a domain-specific fine-tuned Mistral-7B model and retrieval-augmented generation (RAG). While generalpurpose LLMs have demonstrated impressive capabilities in natural language tasks, they often lack spe- cialized knowledge required for reliable legal applications [24]. Furthermore, these models may generate plausible but incorrect information when addressing domain-specific queries [16]. Our work builds upon key innovations including Song et al.'s [1] unified prompting in legal intelligence applications and the application of chatbots for legal guidance explored by Socatiyanurak et al. [3]. The integration of retrieval-augmented generation, pioneered by Lewis et al. [10] and further developed in recent literature [20], addresses critical limitations of standalone LLMs by grounding responses in verified legal sources. To optimize JURIS.AI's performance, we implement QLoRA fine-tuning on the Mistral-7B model, leveraging techniques from Rahman et al. [4] on fine-tuning with prompt engineering. This approach enables efficient adaptation of the model to legal domain knowledge while maintaining computational feasibility. This paper presents JU- RIS.AI, a specialized legal chatbot that combines domainspecific fine-tuning with retrieval-augmented generation to provide accurate, context-aware legal assistance. We evaluate its performance across various legal domains, demonstrating significant improvements in accuracy, relevance, and user satisfaction, while addressing ethical considerations regarding the responsible deployment of AI systems in legal contexts [15].

A. Literature Review

Zhuoyang Song et al. [1] propose a parallel learning approach for legal intelligence using a HANOI framework, highlighting how unified prompting with machine learning can improve legal decision-making. Their work leverages Transformer models to enhance decision-support systems, illustrating the growing importance of advanced NLP methods in complex legal domains.

Building on the theme of model enhancement, Yuxiang Zhou et al. [2] introduce TopicBERT, a neural language model that integrates topic modeling with fine-tuned BERT for sentiment classification in legal texts. By capturing thematic structures within documents, TopicBERT refines search and retrieval processes in legal case databases, thereby improving the relevance of legal precedents.

In a more application-oriented study, Vorada Socatiyanurak et al. [3] develop LAW-U, an AI-driven legal chatbot for sexual violence victims. Their results underscore the significance of simplifying legal terminology for non-expert users. This chatbot demonstrates how conversational AI can deliver practical legal support, thereby expanding access to justice in specialized contexts.

All authors are with the Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Sriperumbudur, Tamil Nadu, India.

Correspondence: adithya272003@gmail.com

Outside of direct legal applications, S. M. W. Rahman et al. [4] focus on zero-shot crypto sentiment analysis, employing prompt engineering and fine-tuned language models. Although aimed at financial text, their methodology provides insights into how large language models can be adapted for specialized tasks with minimal domain-specific data—a concept relevant for legal systems dealing with niche regulations or complex statutes.

In the public sector, Jiawei Han et al. [5] examine large language models for digital government services, revealing a need for highly optimized, transparent, and user-friendly AI solutions. Their work aligns with the broader push for intelligible and accountable AI in administrative or governmental contexts, which often handle sensitive public data and legal procedures.

From a personalization standpoint, Lei Chen et al. [6] propose user-specific adaptive fine-tuning for cross-domain recommendations. While their focus is on recommendation systems, the concept of tailoring model outputs to individual user profiles can be extended to legal contexts, where user queries may demand context-sensitive or personalized information.

Early strides in AI-driven legal chatbots are exemplified by Shubhashri G et al. [7], who developed LAWBO to simplify complex legal tasks through an interactive system. Their findings stress the importance of designing intuitive user interfaces and demonstrate the feasibility of AI tools in helping non-expert users navigate legal processes.

On the efficiency front, Mauricio Fadel Argerich et al. [8] address the computational challenges of running large language models at scale. They propose strategies to optimize inference—essential for real-time or high-volume legal applications where quick turnaround and resource conservation are priorities.

In a related effort to streamline legal processes, Amato et al. [9] present an intelligent conversational agent for the legal domain. Their agent leverages domain-specific knowledge and advanced natural language understanding to assist in case analysis and legal research, indicating the growing potential for AI to facilitate legal workflows.

Patrick Lewis et al. [10] introduce Retrieval-Augmented Generation (RAG) for knowledge-intensive NLP tasks. By combining a retrieval mechanism with generative models, RAG ensures that AI-driven outputs remain grounded in reliable source material, thereby reducing factual errors. This approach is especially pertinent to legal AI, where precision and correctness are paramount.

Building on foundational language model architecture, T. Le Scao et al. [11] introduce BLOOM, a 176B-parameter multilingual language model designed for open access. While not specifically targeting legal applications, BLOOM's extensive multilingual capabilities offer potential for cross-jurisdictional legal systems where multiple languages must be processed and understood simultaneously.

Expanding the transfer learning paradigm, C. Raffel et al. [12] explore the text-to-text transformer (T5) framework, which unifies various NLP tasks through a consistent format. Their work demonstrates how pre-trained models can be effectively adapted to specialized domains through transfer learning—a principle

applicable to legal text processing where domain adaptation is crucial for accurate interpretation of legal language.

In the open-source domain, S. Zhang et al. [13] present OPT, a series of open pre-trained transformer language models. Their research emphasizes responsible AI development through transparency and accessibility, paralleling growing concerns about explainability and accountability in legal AI systems where decisions may have significant real-world implications.

Focusing on conversational capabilities, R. Thoppilan et al. [14] introduce LaMDA, a language model specifically designed for dialogue applications. Their approach to maintaining context across extended conversations is particularly relevant for legal chatbots that must track complex case details and user circumstances throughout interaction sessions.

Addressing regulatory perspectives, D. Necz [15] examines the legal and ethical frameworks governing chatbots in legal markets. This research highlights the tension between technological innovation and professional regulation, outlining considerations that must inform the development of systems like JURIS.AI to ensure compliance with ethical standards and legal practice rules.

A. Perlman [16] explores broader societal implications of generative AI in legal services, identifying both opportunities for improved access to justice and potential risks to professional standards. This work underscores the importance of balancing innovation with responsibility when deploying AI in legal contexts.

Focusing specifically on access to justice, M. Queudot et al. [17] evaluate how legal chatbots can bridge gaps in legal service provision. Their research demonstrates measurable improvements in legal resource accessibility while acknowledging limitations of current technologies—insights directly applicable to the positioning and scope definition of JURIS.AI.

Z. Misquitta et al. [18] provide a practical implementation study of a law chatbot, documenting technical architecture and user interaction patterns. Their findings on interface design and query processing offer valuable lessons for optimizing user experience in legal AI applications.

Taking a risk-management approach, S. Migliorini [19] analyzes potential legal liabilities associated with commercial AI chatbots. This framework for risk assessment is essential for JURIS.AI's development, informing both technical safeguards and user disclosure protocols.

Recent advances in retrieval-augmented generation are comprehensively reviewed by Y. Gao et al. [20], who catalog methods for integrating external knowledge with generative capabilities. Their taxonomy of RAG implementations provides crucial guidance for selecting appropriate architectures when developing knowledge-intensive applications like legal assistance systems.

S. Gupta et al. [21] further extend RAG research by mapping its evolution and identifying emerging trends. Their forwardlooking analysis suggests potential directions for optimizing retrieval mechanisms in specialized domains such as law, where the volume and complexity of source materials present unique challenges.

J. Chen et al. [22] contribute empirical benchmarking of LLMs in retrieval-augmented generation tasks, providing per-formance metrics across different models and retrieval strategies. Their comparative approach informs model selection and parameter optimization for JURIS.AI's implementation.

Exploring knowledge enhancement, S. Xu et al. [23] propose integrating knowledge graphs with RAG systems through dual-pathway approaches. This structured representation of legal relationships could significantly improve JURIS.AI's ability to navigate complex legal hierarchies and precedent relationships.

J.G. Meyer et al. [24] examine broader implications of LLMs in academic contexts, including challenges of accuracy and ethical use. Their findings on verification strategies and educational applications have parallels in legal training and research applications.

Finally, Z. Wan et al. [25] survey efficiency techniques for large language models, addressing computational constraints that affect real-world deployment. These optimization strategies are crucial for ensuring JURIS.AI's scalability and responsiveness in practical legal assistance scenarios.

II. RESEARCH GAP AND MOTIVATION

A. Research Gap

The Consumer Protection Act 2019 is a comprehensive legislation that addresses a wide range of consumer issues, including unfair trade practices, product liability, and con- sumer grievances [15], [19]. However, a significant gap exists between the availability of this legal information and the ability of the general public to access and understand it. Existing legal resources, such as government websites and legal handbooks, are often written in complex legal jargon, making it difficult for laypeople to comprehend the informa- tion [16], [17]. Moreover, while there are some platforms that provide legal advice, they often lack the ability to personalize responses based on individual queries [3], [9]. These platforms typically offer generalized information, which may not be applicable to a user's specific situation. As a result, many individuals are left without the necessary guidance to address their consumer grievances effectively [17].

B. Motivation

The motivation behind JURIS.AI stems from the need to create a tool that can provide personalized, easy-to-understand legal information to consumers [7], [18]. By focusing on the Consumer Protection Act 2019, JURIS.AI seeks to empower consumers by making legal knowledge more accessible and actionable. This project aims to address the following key challenges:

1) Complexity of Legal Terminology: Legal language is often difficult for laypeople to understand, which can prevent them from accessing their rights.
2) Limited Access to Legal Resources: Many individuals do not have access to affordable legal advice or resources, leaving them uninformed about their rights and responsibilities.
3) Lack of Personalization in Legal Platforms: Existing legal platforms often provide generic information, which may not be relevant to the user's specific query.

By addressing these challenges, JURIS.AI aims to transform the way consumers interact with the law, making it easier for them to understand their rights and take appropriate action when those rights are violated.

C. Related Work

In recent years, there has been an increasing interest in the use of artificial intelligence (AI) and machine learning (ML) to simplify and automate legal processes. Several AI-based platforms have been developed to assist with legal research, contract analysis, and case prediction. However, most of these platforms are designed for legal professionals and do not cater to the general public.

D. AI in Legal Research

Platforms like ROSS Intelligence and CaseText have rev- olutionized legal research by using AI to streamline the process of finding relevant case law and legal statutes. These platforms use NLP to analyze legal documents and provide users with precise answers to their queries. However, they are primarily targeted at lawyers and legal researchers, and their user interfaces are not designed for individuals without a legal background.

E. Consumer Rights Assistance Platforms

In the realm of consumer rights, there are a few plat- forms that offer legal advice and information. Websites like Consumer Complaints India provide users with the ability to file complaints and seek redressal for consumer grievances. However, these platforms often lack personalization and are limited to providing general information. They do not offer detailed, section-specific guidance based on the Consumer Protection Act 2019.

F. JURIS.AI: Filling the Gap

JURIS.AI builds upon the advancements made in AI-based legal research and consumer rights platforms, but it differentiates itself by offering a personalized, interactive legal assistant specifically designed for consumers. By utilizing state-of-theart models like QLoRA and RAG, JURIS.AI is able to provide precise,

contextually relevant answers to user queries, making legal information more accessible to the general public.

III. PROPOSED METHODOLOGY

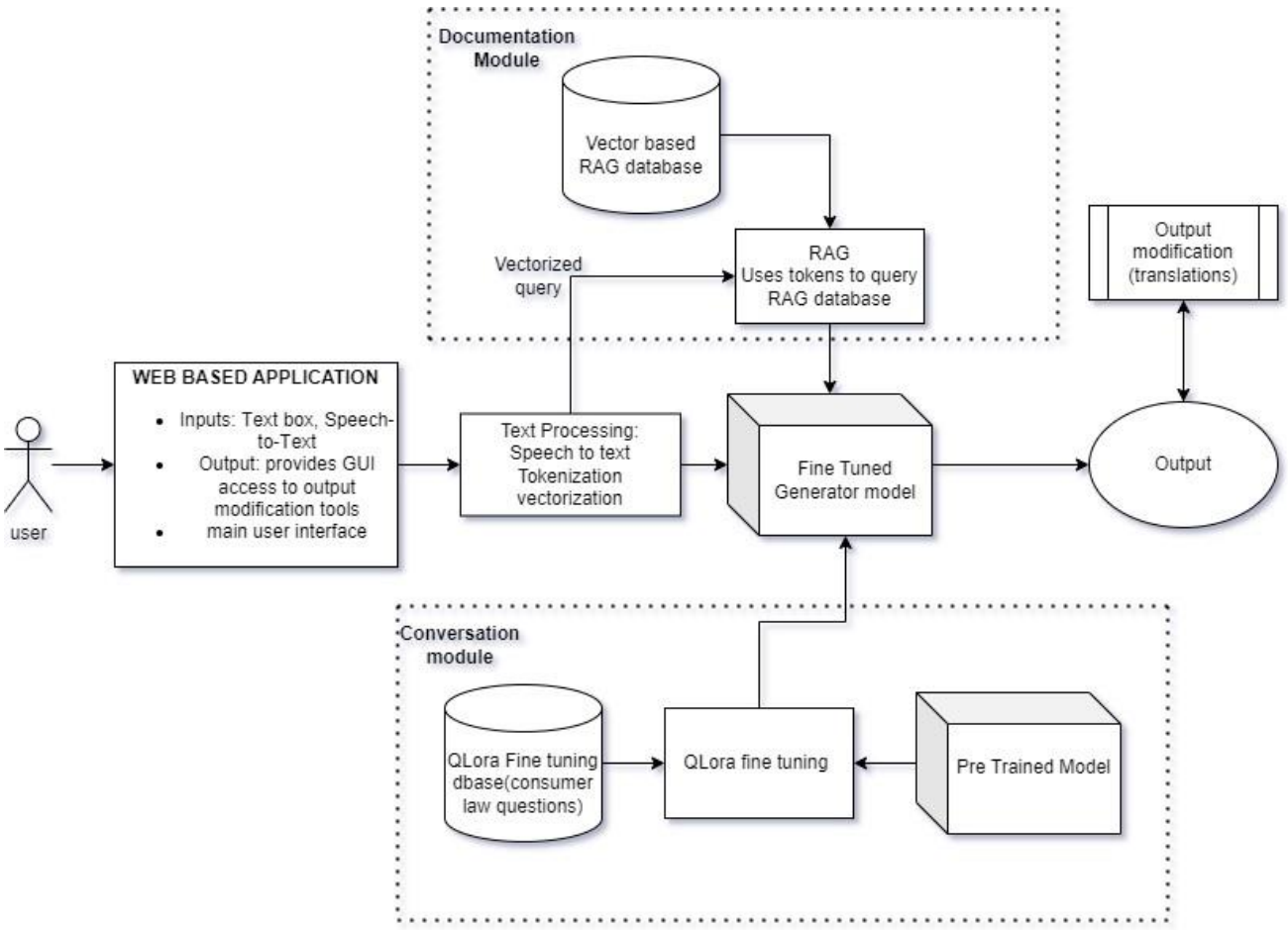

Fig. 1: System architecture of JURIS.AI showing major modules such as speech-to-text, legal reasoning via RAGLoRA, and response generation.

The architecture we propose as a model can be seen in Fig. 1, and the specifics of it will be addressed in the subsequent paragraphs.

A. Data Collection and Preprocessing

The first step in building JURIS.AI involves the collection of a comprehensive dataset that includes legal documents, case studies, and relevant sections of the Consumer Protection Act, 2019. This data is then preprocessed to ensure it is in a format suitable for training the machine learning models.

The data is categorized and processed as follows:

1) Legal Documents: The primary source of data for JURIS.AI consists of legal documents related to the Consumer Protection Act, 2019. These documents include the text of the Act itself, as well as related case law and legal interpretations.
2) Case Studies: In addition to legal documents, the dataset includes case studies that illustrate how specific sections of the Consumer Protection Act have been applied in real-world scenarios. These case studies provide valuable context and help the model understand the practical implications of the law.
3) Preprocessing: The collected data is preprocessed using standard Natural Language Processing (NLP) techniques, such as tokenization, stopword removal, and lemmatization. This ensures that the data is clean and structured for effective model training.

B. Model Training and Fine-Tuning Process

The core architecture of JURIS.AI relies on a multi-model framework combining QLoRA (Quantized Low-Rank Adaptation) and RAG (Retrieval-Augmented Generation). This allows the legal assistant to understand user queries better, retrieve relevant legal data, and provide precise answers based on the Consumer Protection Act, 2019.

C. Quantized Low-Rank Adaptation (QLoRA)

QLoRA is a model adaptation technique designed to make large models more efficient for fine-tuning by reducing their memory footprint without sacrificing accuracy. JURIS.AI uses QLoRA to adapt pre-trained legal language models to the specific context of Indian consumer law, reducing the training complexity while maintaining high performance. The QLoRA model operates as follows:

1) Quantization: Large pre-trained models often have bil- lions of parameters, making them memory-intensive. QLoRA reduces the number of bits used for floating-point precision (from 32 to 8), significantly decreasing memory usage. This is achieved by mapping each parameter ω to a scaling factor s using Equation 1:

$$\omega\tilde{} = \text{round}\left(\frac{\omega}{s}\right) \tag{1}$$

2) Low-Rank Approximation: By applying a low-rank decomposition to the weight matrices in the neural network Equation 2, QLoRA approximates the original model, focusing only on the most significant components. This allows the model to adapt to the target domain (consumer law) without the need to retrain the entire model from scratch. Let W represent the weight matrix in a pre-trained model. During QLoRA fine-tuning, it is decomposed as:

$$W = UV^T + \epsilon \tag{2}$$

where U and V are low-rank matrices, and ϵ represents the residual error. The low-rank approximation enables faster and more efficient learning, especially when combined with quantization.

3) Gradient Update Equation:During backpropagation Equation 3, the gradients for U and V are computed as:

$$\Delta U = \eta\left(\frac{\delta L}{\delta U}\right), \Delta V = \eta\left(\frac{\delta L}{\delta V}\right) \tag{3}$$

where η is the learning rate and L is the loss function (e.g., cross-entropy loss).

D. Retrieval-Augmented Generation (RAG)

JURIS.AI incorporates the RAG model to enable precise and context-aware legal responses. It significantly enhances the system's ability to generate exact and context-aware legal responses by incorporating an external retrieval mechanism. RAG combines a retrieval mechanism with a generative model (such as GPT-based models). In this process, the model retrieves relevant legal documents or sections of the Consumer Protection Act, 2019, from a pre-indexed knowledge base and then generates responses based on both the retrieved text and the query. RAG operates in two stages:

1) 1. Retrieval: Given a query qqq, the retriever searches through a knowledge base DDD and identifies the most relevant documents $d_1,d_2,...,d_n$. The probability of each document being relevant is computed as Equation 4:

$$P(d_i|q) = \frac{exp(score(q,d_i))}{\sum_j exp(score(q,d_j))} \tag{4}$$

where score(q,d_i) measures the similarity between the query and document d_i.

2) Generation: After retrieval, the generative model uses the query qqq and the retrieved documents $d_1,d_2,...,d_n$ in Equation 5 to generate the final response rrr. The probability of generating a token r_t is:

$$P(r_t|r_{<t},q,d_1,d2,....,d_n) \tag{5}$$

The complete response is then obtained by Equation 6:

$$p(r|q,d_1,d_2,....,d_n) = {}^{Y}P(r_t|r_{<t},q,d_1,d2,....,d_n) \tag{6}$$

where r_t is the token at time step t in the generated response. This dual-stage process allows JURIS.AI to generate responses that are both legally accurate and tailored to the user's specific needs, enhancing the overall quality of the assistant.

E. Instruction Fine-Tuning

Fine-tuning a model to understand legal jargon and complex consumer rights is a delicate process. JURIS.AI employs instruction fine-tuning to improve the model's ability to follow user instructions and respond accurately based on legal standards. This involves training the model on a dataset of legal questions and answers, such as previous case studies, legal precedents, and annotated sections of the Consumer Protection Act, 2019.

F. The JURIS Framework

The JURIS Framework encapsulates the core methodolo- gies employed in JURIS.AI, integrating quantized fine-tuning (QLoRA) and retrieval-augmented generation (RAG) into a unified legal AI system. The framework is designed to optimize model efficiency, enhance retrieval precision, and generate legally coherent responses. At its core, the system operates in two tightly connected stages:

1) Model Adaptation (QLoRA Fine-Tuning): he legal language model is adapted to the Consumer Protection Act, 2019, using low-rank adaptation and quantization (8-bit precision), enabling efficient domain-specific learning while reducing memory overhead.
2) Contextual Response Generation – RAG Retrieval- Augmented Reasoning: A retrieval mechanism is integrated with the generative model to ensure responses are not only fluent but also grounded in verifiable legal documents, mini- mizing hallucinations in AI-generated legal advice. The JURIS Framework effectively bridges data preprocessing, efficient model adaptation, and retrieval-augmented response genera- tion to create an AI-powered legal assistant tailored for con- sumer law queries. By leveraging QLoRA's low-memory fine- tuning and RAG's contextual retrieval, the system achieves high precision, computational efficiency, and verifiable legal guidance. This approach ensures that JURIS.AI remains scalable, interpretable, and practically deployable for real-world legal applications.

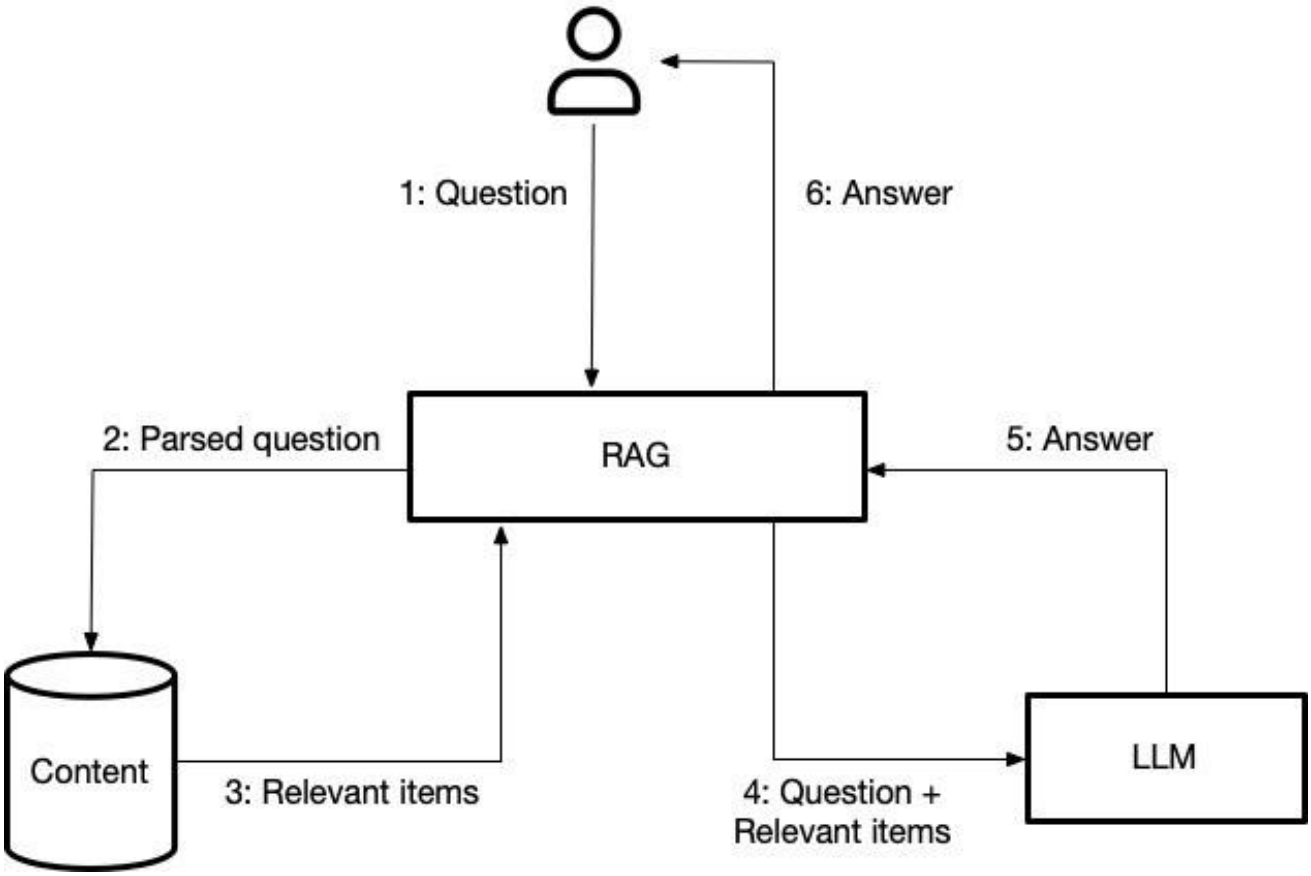

Fig. 2: The JURIS Framework

Fig. 2 expresses the functionality of RAG where a user poses a question to the model which is parsed and processed by the vector store which then retrieves relevant documents, this is then piped into the LLM which combines factual data and generates a conversational style of content after understanding the user requirements.

G. RAGLoRA Methodology

1) Dataset Preparation: Converted legal documents related to the Consumer Protection Act 2019 into standardized format using PDF extraction tools.
2) Training Data Generation: Utilized ChatGPT to create question-answer pairs from the standardized legal texts for fine-tuning purposes.
3) Document Restructuring: Engineered documents with strategically overlapping content and associated questions to optimize RAG retrieval performance.
4) Query Pipeline Development: Implemented a multi-stage query processing system that: - Captures initial user input Reformulates input into an optimized retrieval query - Retrieves relevant legal information from the document corpus Combines retrieved information with the original query.
5) Model Fine-tuning: Applied QLoRA (Quantized Low- Rank Adaptation) techniques to fine-tune the Mistral-7B model on the generated question-answer pairs.
6) Response Generation System: Integrated the retrieval results with the fine-tuned language model to produce factually accurate, conversational responses in accessible language.

H. User Interaction Design

The user interface of JURIS.AI is designed to be simple and intuitive, allowing individuals with varying levels of legal literacy to easily interact with the system. Users can input their queries in natural language, and JURIS.AI will respond with relevant legal information, including the specific sections of the Consumer Protection Act that apply to their situation.

1) Natural Language Input: Users can ask questions in plain English (or other supported languages), without needing to know any legal terminology.
2) Section-Specific Responses: Based on the user's query, JURIS.AI will retrieve and present the relevant sections of the Consumer Protection Act, along with a simplified explanation of how those sections apply to the user's case.
3) Actionable Recommendations: In addition to providing legal information, JURIS.AI will also offer recommendations on what steps the user can take next, such as filing a complaint with the appropriate consumer forum.

I. Personalization Features

JURIS.AI uses machine learning algorithms to personalize its responses based on the user's input. This ensures that the information provided is not only accurate but also tailored to the specific needs of the user. 1. User Context: JURIS.AI takes into account the user's context, such as the nature of the consumer grievance, and provides information that is relevant to their specific situation. 2. Adaptive Learning: Over time, JURIS.AI will learn from user interactions and refine its responses to provide even more accurate and personalized information.

J. Accessibility Features

JURIS.AI is designed to be accessible to all users, regardless of their level of legal literacy or physical ability. To achieve this, the platform includes several accessibility features:

1) Text-to-Speech: Users who prefer to listen to the information can use the text-to-speech feature, which reads out the legal information provided by JURIS.AI.
2) Speech-to-Text: For users who may have difficulty typing, the speech-to-text feature allows them to input their queries using voice commands.

K. Evaluation and Feedback

The effectiveness of JURIS.AI in processing consumer law queries was evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric. Two distinct phases of experimentation were conducted:

1) Baseline Phase (QLoRA Only): The first phase assessed system performance with a Quantized Low-Rank Adaptation (QLoRA) model in isolation.
2) Hybrid Phase (QLoRA + RAG): The second phase integrated Retrieval-Augmented Generation (RAG) into the QLoRAbased model, aiming to improve factual grounding and context specificity. ROUGE measures the degree of overlap between gener- ated text and reference text. We report ROUGE-1 Overlap of unigrams (single words), ROUGE-2 Overlap of bigrams (pairs of consecutive words), and ROUGE-L Longest Common Subsequence (LCS) overlap.
3) ROUGE-N: ROUGE-N (where N = 1,2,...) captures the n-gram overlap in terms of recall. For a single reference summary S and generated text G, the recall-based ROUGE-N is typically defined as:

$$ROUGE - N = \frac{\sum_{n \in S} \min(Cnt_G(n), Cnt_S(n))}{\sum_{n \in S} Cnt_S(n)} \tag{7}$$

where $Cnt_G(n)$ in Equation 7 is the number of times an n-gram n appears in the generated text G, $Cnt_S(n)$ is the number of times an n-gram n appears in the reference summary S, and the *min* function ensures only overlapping n-grams are counted.

4) ROUGE-L: ROUGE-L leverages the Longest Common Subsequence (LCS) concept using Equations 8,9,10. It measures how well the generated text's token sequence aligns with that of the reference. Let LCS(G,S) be the length of the LCS between the generated text G and reference S. Then:
Precision (P_L):

$$P_L = \frac{\text{LCS}(G,S)}{\text{Length}(G)} \tag{8}$$

Recall (R_L):

$$R_L = \frac{\text{LCS}(G,S)}{\text{Length}(S)} \tag{9}$$

F-measure (F_L):

$$F_L = \frac{(1+\beta^2) \cdot P_L \cdot R_L}{R_L + \beta^2 \cdot P_L} \tag{10}$$

When $\beta = 1$, precision and recall are weighted equally.

L. Quantitative Analysis

The baseline implementation utilizing only the QLoRA model demonstrated strong initial performance, achieving a ROUGE1 score of 0.69. This indicates that while QLoRA enabled efficient fine-tuning and adaptation to legal texts, it had certain limitations in retrieving highly relevant legal references.

With the integration of RAG for retrieval-augmented generation, the system exhibited a notable improvement, achieving a final ROUGE-1 score of 0.84. This represents an approximate 9.6% increase in accuracy, demonstrating that the hybrid QLoRA+RAG approach significantly enhances both the contextual relevance and precision of generated legal responses. The RAG module effectively mitigated hallucinations by grounding responses in retrieved legal documents, thus aligning the system's answers with established statutes and case law.

This result underscores the importance of integrating retrieval-based methods with fine-tuned generative models when dealing with legally sensitive AI applications. The combination of QLoRA's efficiency in domain adaptation and RAG's knowledgegrounded retrieval mechanism is pivotal in ensuring both performance and accuracy in delivering personalized legal assistance.

M. Qualitative Assessment

Beyond quantitative metrics, qualitative evaluation revealed notable improvements in response quality following RAG integration. The system demonstrated enhanced capabilities in:

1) Processing queries containing complex legal terminology,
2) Addressing ambiguous scenarios with greater precision, and
3) Providing contextually relevant responses aligned with legal frameworks

The RAG component proved particularly effective in mapping user queries to appropriate legal precedents and relevant sections of the Consumer Protection Act, 2019. This capability significantly enhanced the system's ability to deliver comprehensive and accessible legal assistance, even when confronted with nuanced inquiries. These findings underscore the value of hybrid architectures in legal AI systems, where the combination of foundation models and retrieval mechanisms can yield superior results compared to single-model approaches.

N. Comparative Analysis: QLoRA vs. QLoRA+RAG

A systematic comparison between the standalone QLoRA model and the integrated QLoRA+RAG architecture revealed distinct performance characteristics. While the base QLoRA model demonstrated proficiency in generating contextually relevant responses, it occasionally struggled with precise legal citations and complex dispute resolution.

The integration of RAG significantly enhanced these capabilities by implementing an efficient retrieval system that bridges the gap between generative language models and structured legal data.

The enhanced architecture showed marked improvements in the following areas:

- Citation accuracy and specificity
- Complex legal query resolution
- Access to relevant statutory references

This analysis confirms that the integration of RAG effectively addresses the limitations of standalone language models in legal applications while preserving their inherent strengths in natural language understanding.

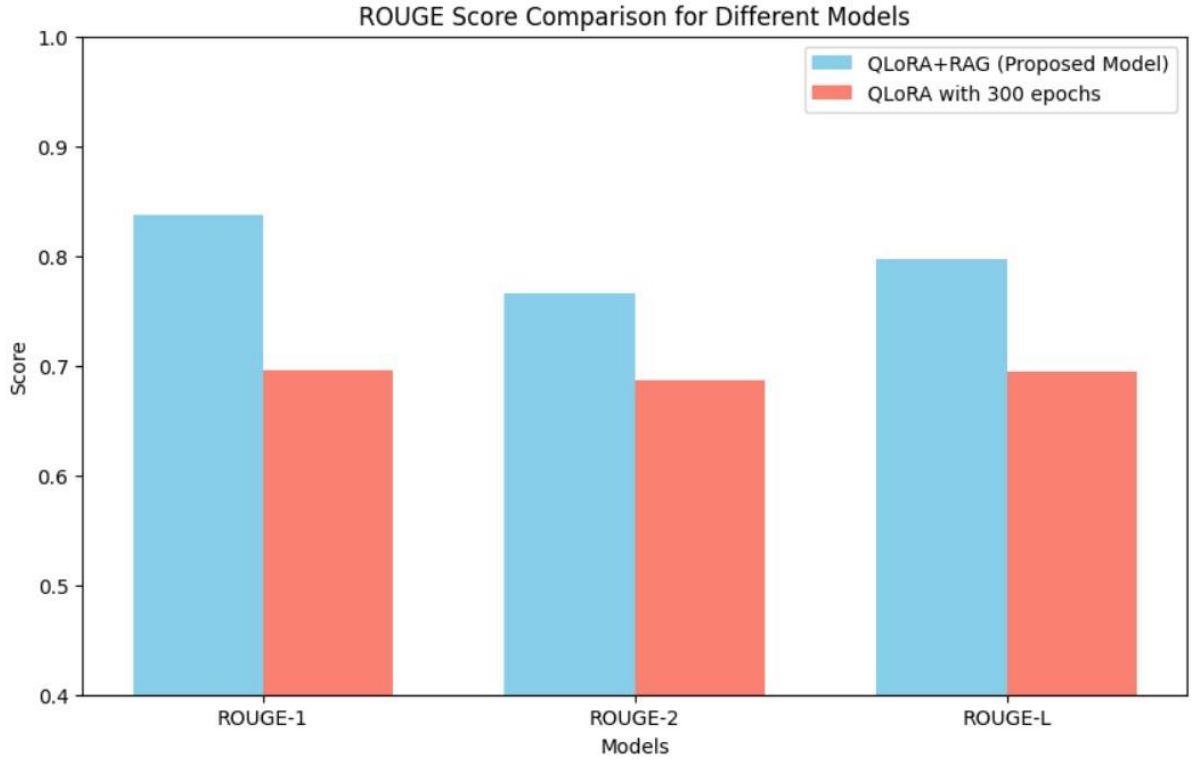

Fig. 3: ROUGE Comparison: Proposed Model vs. QLoRA Model

IV. CONCLUSION

The JURIS.AI project demonstrates that advanced Natural Language Processing (NLP) techniques, particularly Quantized Low-Rank Adaptation (QLoRA) and Retrieval-Augmented Generation (RAG), can effectively bridge the

MODEL	ROUGE-L SCORE (%)
Mistral 7B with QLoRA and RAG (Proposed Model)	79.77
HANOI [1]	78.38
GCALLM [5]	41.57

Fig. 4: ROUGE-L Score (in %) Comparison of Our Model vs. Others

PROPOSED MODEL	BATCH SIZE	NUMBER OF EPOCHS	ROUGE-1 SCORE	ROUGE-2 SCORE	ROUGE-L SCORE
Mistral 7B (QLoRA + RAG)	8	300	0.837083	0.766693	0.797692

Fig. 5: ROUGE Scores of Proposed Model

gap between complex legal resources and non-expert users. By focusing on Indian consumer protection law, the system addresses a pressing need for accessible, personalized legal guidance. Empirical results evidenced by the notable rise in ROUGE scores upon integrating RAG with QLoRA in Figure 3, 4, 5, confirm that retrieval-driven generation significantly boosts both accuracy and context relevance. Moreover, the incorporation of speech-based accessibility features and ac- tionable recommendations extends JURIS.AI's utility to a broad user base. Future work will concentrate on broadening domain coverage, enhancing multilingual capabilities, and refining model interpretability to maintain transparency in high-stakes legal contexts. Ultimately, JURIS.AI underscores the potential of AI-powered solutions to democratize legal knowledge, fostering greater awareness and empowerment among consumers.

REFERENCES

[1] Zhuoyang Song, Min Huang, Qinghai Miao, and Fei-Yue Wang, "Parallel Learning for Legal Intelligence: A HANOI Approach Based on Unified Prompting," IEEE Transactions on Computational Social Systems, vol. 11, no. 2, pp. 2765–2775, Apr. 2024.

[2] Yuxiang Zhou, Lejian Liao, Yang Gao, Rui Wang, and Heyan Huang, "TopicBERT: A Topic-Enhanced Neural Language Model Fine-Tuned for Sentiment Classification," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 1, pp. 380–393, Jan. 2023.

[3] Vorada Socatiyanurak et al., "LAW-U: Legal Guidance Through Artificial Intelligence Chatbot for Sexual Violence Victims and Survivors," IEEE Access, vol. 9, pp. 142834–142844, Sep. 2021.

[4] S. M. W. Rahman, I. Tashdeed, M. Kaur, and H.-N. Lee, "Enhancing Zero-Shot Crypto Sentiment with Fine-Tuned Language Model and Prompt Engineering," IEEE Access, vol. 12, pp. 10146–10159, Jan. 2024.

[5] Jiawei Han et al., "Intelligent Practices of Large Language Models in Digital Government Services," IEEE Access, vol. 12, pp. 8633–8640, Jan. 2024.

[6] Lei Chen et al., "User Specific Adaptive Fine Tuning for Cross Domain Recommendations," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 3, pp. 3239–3252, Mar. 2023.

[7] Shubhashri G, Unnamalai N, Kamalika G, "LAWBO: A Smart Lawyer Chatbot," Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, pp. 348–351, Jan. 2018.

[8] Mauricio Fadel Argerich and Marta Patino-Martˇ´ınez, "Measuring and Improving the Energy Efficiency of Large Language Models Inference," IEEE Access, vol. 12, pp. 80194–80207, Jun. 2024.
[9] Amato, F., Fonisto, M., Giacalone, M., and Sansone, C., "An Intelligent Conversational Agent for the Legal Domain," Information, vol. 14, no. 6, pp. 307–320, May 2023.
[10] Patrick Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, pp. 9459–9474, Dec. 2020.
[11] T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," arXiv preprint arXiv:2211.05100, Nov. 2022.
[12] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, Jan. 2020.
[13] S. Zhang et al., "OPT: Open Pre-trained Transformer Language Models," arXiv preprint arXiv:2205.01068, May 2022.
[14] R. Thoppilan et al., "LaMDA: Language Models for Dialog Applications," arXiv preprint arXiv:2201.08239, Jan. 2022.
[15] D. Necz, "Rules over Words: Regulation of Chatbots in the Legal Market and Ethical Considerations," Hungarian Journal of Legal Studies, vol. 64, no. 3, pp. 472–485, Jun. 2024.
[16] A. Perlman, "The Implications of ChatGPT for Legal Services and Society," Michigan Technology Law Review, vol. 30, no. 1, 2024.
[17] M. Queudot, E. Charton, and M.-J. Meurs, "Improving Access to Justice with Legal Chatbots," Statistics, vol. 3, no. 3, pp. 356–375, Sep. 2020.
[18] Z. Misquitta, A. A. Sawant, A. U. Shaikh, A. S. Patil, and N. Narkar, "Law Chatbot," International Journal For Science Technology And Engineering, vol. 12, no. 5, pp. 164–170, May 2024.
[19] S. Migliorini, "'More than Words': A Legal Approach to the Risks of Commercial Chatbots Powered by Generative Artificial Intelligence," European Journal of Risk Regulation, pp. 1–18, Feb. 2024.
[20] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997, Dec. 2023.
[21] S. Gupta, R. Ranjan, and S. Singh, "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions," arXiv preprint arXiv:2410.12837, Oct. 2024.
[22] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," arXiv preprint arXiv:2309.01431, Sep. 2023.
[23] S. Xu, M. Y. Chen, and S. Chen, "Enhancing Retrieval-Augmented Generation Models with Knowledge Graphs: Innovative Practices Through a DualPathway Approach," Lecture Notes in Computer Science, pp. 398–409, Jan. 2024.
[24] J. G. Meyer et al., "ChatGPT and Large Language Models in Academia: Opportunities and Challenges," Biodata Mining, vol. 16, Jul. 2023.
[25] Z. Wan et al., "Efficient Large Language Models: A Survey," arXiv preprint arXiv:2312.03863, Dec. 2023.