International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Privacy-Preserving Legal Assistance: An Offline RAG-Based Large Language Model for the Indian Legal Context

Abhishek Kumar Singh[1], Dr Md Sajid Anwer[2]

[1] Student, MCA Department SEST, Jamia Hamdard University Delhi
[2] Assistant Professor, CSE Department SEST, Jamia Hamdard University Delhi

Abstract

In this AI progressing World, there is a lot of legal work which can be done with the help of AI (LLM) with rapid rate and minimal human effort which makes it perfect for legal sector, but there is issue with how maximum LLM model works or integrated. Currently most of the LLMs are dependent on cloud-based system to operate which introduces data privacy risks for user also most of the LLMs are trained on the Western legal Dataset which introduces jurisdictional bias with respect to Indian laws. And third but very important issue with generalized models is "hallucination". LLM models are trained in a way that they have to respond to the query confidently which makes it unreliable for domain specific work like Indian Constitution and Bharatiya Nyaya Sanhita (BNS). To handle these critical issues of LLMs in legal sector, this paper proposes a fully localized, offline architecture that work with the integration of RAFT (RAG + Fine Tuning) with 8-billion parameter Llama-3 model and Chroma DB pipeline. By implementing hallucination shield and Fine Tuning our model achieved a 95% retrieval accuracy on Indian penal statues and successfully blocked 75% of out of the bounds queries as compare to base model, which makes it perfect framework for privacy preserving offline legal AI.

Keywords: Legal AI, LLM, RAG, RAFT, Llama3, Chroma DB, Indian Legal Context, GROQ, BNS, Indian Constitution

1. Introduction

Introduction of GEN AI in the legal sector is changing the way legal research and document analysis happens. Because this new revolutionary technology is capable of doing research and analysis in the manner which cannot be imagined by human. However, implementing this tech in active legal practice can cause data privacy constraints. As of today, most of the Models (LLMs) uses cloud-based storage and API calling to operate and fetch data for giving answer or providing solution to the queries asked which can be an unacceptable risk to take when handling confidential client data and sensitive cases, where leakage of any private information can be the reason to lose the case or even putting your client's life at risk.

Apart from data privacy, Foundational LLMs faces a critical limitation due to their architecture. As most

of LLMs are trained on very large Western legal datasets as compare to Indian legal datasets, there answer always shows “generalization bias” when asked for specific regional issues. Whenever queried related to any laws without specifying the region these models always answer according to western laws by default which can cause serious confusion between users.

Also, these models are very highly prone to “hallucinations”. These models generate entirely false and fabricated answer very confidently. In context to legal laws, creating a false and fabricated answer like false penal code or misinterpreting the meaning can lead to misguidance.

To fulfill this gap, this research paper introduces a secure, offline RAFT (RAG + Fine Tuning) architecture, which is specifically tailored for Indian legal laws. The proposed architecture system is entirely operated on local system hardware for fully data privacy and control. This architecture mind contains Llama-3 model with 8 billion parameter which is first trained in the Indian legal problem and answer based on custom dataset for Fine-Tuning process. Then this Fine-Tunned model is augmented with local vector database working on local system containing the Indian Constitution and penal codes. After that with the help of semantic string and mathematical distance based on fixed threshold value, the model handles out-of-bounds queries and also able to delivers highly accurate and jurisdictionally prefect assistance.

2. Related Work

In early phase of NLP based applications which were made for the purpose to handle legal assistance uses a specific architecture based on the encoder only like BERT and for legal domain specific Legal BERT [1]. These models were great for text classification but they were not able to give answer to user queries. As, for legal assistance these models were not enough to take and use further because legal assistance required auto regressive approach so that they can synthesize complex legal queries to understand better and then give response based on that. The architecture which is proposed handle this limitation with the help of decoder only architecture Llama-3[2]. This model is autoregressive to communicate and generate while maintaining strict legal context.

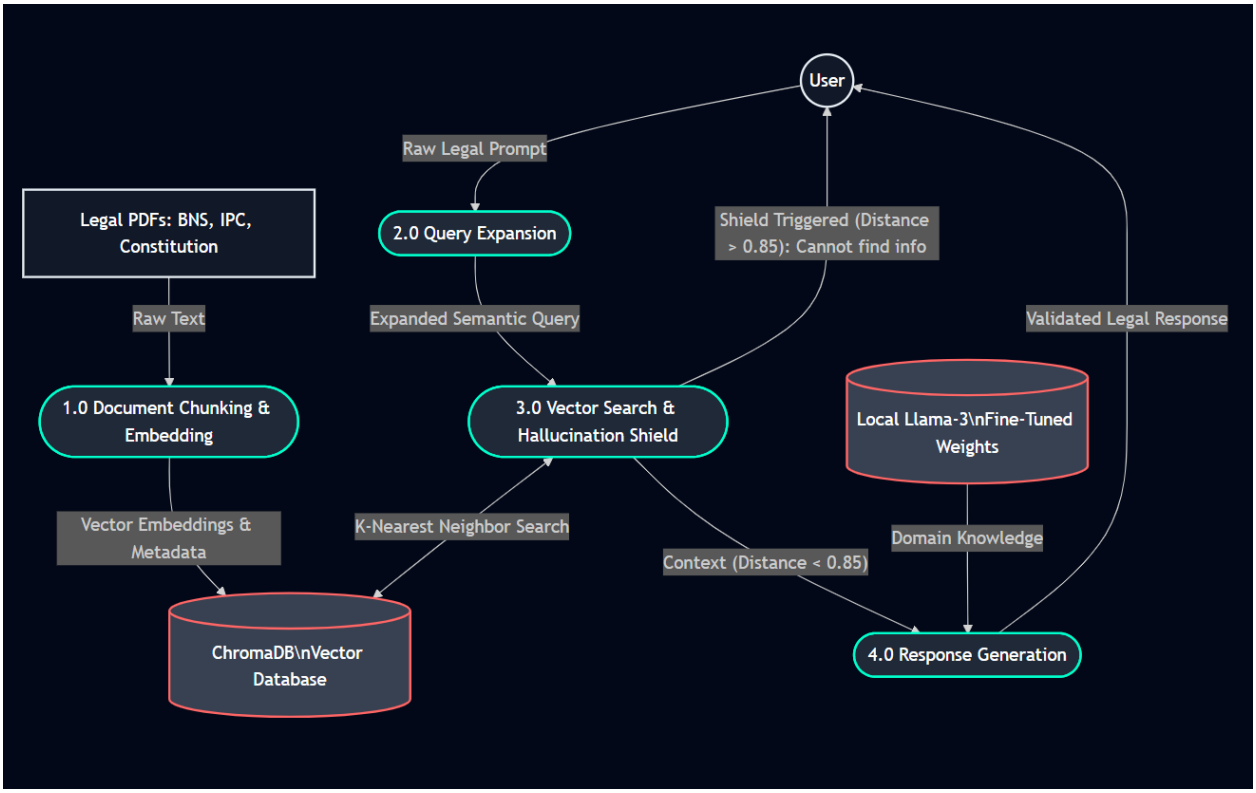
With advanced LLMs such as OpenAI’s various model series which are Pre-trained Transformer model able to handle this issue still, these types of models are totally dependent on cloud, their deployment in legal practice is restricted due to privacy rules and regulations. These models transfer the case sensitive data to third party that strictly breaks the client trust and privacy regulation at the same time. With all this advancement these LLMs still hallucinate if not enough information they get from cloud. Our proposed model specifically resolves these issues as it totally run offline, no need to call any API to any third party to process any data which saves client privacy and most importantly it possesses a mathematical distance based “hallucination shield” in between the Retrieval Augmented Generation (RAG) system [5] to block the hallucination behavior of the model.

Finally, while the open-source implementation of Large Language models (LLMs) has allowed for personal deployment and uses, still these base models fail on Indian laws and situation based on regional issue because of biasness of model, because these models are not specifically designed for it. In Contrast this architecture fills that gap by integrating RAG and PAFT [6]. The model used in the architecture is trained carefully over 1100+ high quality of dataset of question and answer fully based on Indian legal

law. As a result, our system model changes the weights to proper understand Indian legal concepts with the help of parameter tunning hence significantly outperforming other models.

3. Proposed Methodology

The discussed architecture is designed in a way that it works as fully localized, secure, hallucinations free to fulfil its role as domain specific legal assistant specifically for Indian laws. The whole architecture is divided I three different parts: Data preparation and Vectorization, Parameter Efficient Fine Tuning (PEFT), and the last the deployment of the RAG framework which also contains the hallucination shield. By combining the two different concepts of GEN AI like Parameter-Efficient Fine-Tuning and a vector retrieval database, this architecture perfectly works together to form a Retrieval Augmented Fine Tuning (RAFT) framework



3.1 Data Preparation and Vectorization

To build a very specific and accurate knowledge base which contains legal data like Constitution of India, Indian Penal Code (IPC), and the new updated laws also, which is Bharatiya Nyaya Sanhita (BNS). We first have to collect these documents in pdf format. All the LLMs present today have a limited context window and these documents are very large to upload all at once in LLMs context window. To handle this situation a specific strategy is used which is to chunk all the document in a specific length of paragraph in systematic way so that meaningful character also remains in sharing chunks by overlap and fulfil the meaning in separate chunks accordingly.

After that the chunked text is converted to high dimensional mathematical representations using a different embedding model. This process only happens once while uploading the doc for the first time not after that.

The converted vectors with their metadata are then stored in local Chroma DB, which is an open-source database perfectly tunned for local deployment. This database make sure that the retrieval operation happens entirely in local system.

3.2 Synthetic Data Generation and Parameter-Efficient Fine-Tuning (PEFT)

Before integrating the Base Llama model with the vector database, it first requires to tune its weights according to Indian laws so that it can adjust western database bias due to its pre training happened earlier. For training the model again a custom dataset is created with 1100+ specialized Indian legal Question and Answer. This dataset is created with the help of most advance model of Gemini and Groq. Here two different models are used so that dataset have variety and does not show bias towards one of the models which is used to create the dataset.

Using Low-Rank Adaptation (LoRA) [6]. LoRA model are light weight to use and can be handled in low end system. The Fine-tuning happen using google Colab so that it can use GPU for rapid fast training with the dataset generated containing 1100+ queries. Once the model training complete, the model is then exported for fully offline use. Till now the model has adjusted its weight by training and now it has understood how to answer on Indian legal text.

3.3 Query Expansion and The Hallucination Shield

Now to fill the gap between a user query and required information to retrieve the document from vector database, the architecture uses a very smart technique so that while retrieving the document vector database gets maximum context to match called query expansion. For query expansion it uses custom legal dictionary which helps to expand the user query with external context before sending to the process of vectorization.

Once the expanded query gets the top 5 most relevant document chunk from vector database. It then checked against the threshold values which is fixed like < 0.85 to protect the hallucination shield. These methods are integrated with the retravel logic in vector database system. If the threshold of 0.85 exceeds then the model instantly send message like out of bound query. Which helped the model to take action like giving answer based on retrieve document or showing message like “I cannot find the relevant information “. This way the model strictly ensures to provide hallucination shield over the output. Also, this helps the model to give answer in specific domain range not apert from that.

4. Evaluation and Results

To check the efficiency of new model and the impact of the architecture over base model a comparison test was conducted whose result help to solidify the claim. Here two model was used to test base Llama-3 model and The Fine tunned model with RAG pipeline integrated.

4.1 Experimental Setup and LLM-as-a-Judge Methodology

For test a “Golden Dataset” was created in which 20 different types of legal queries were available specifically belong to area like Out of bounds Hallucination, jurisdiction collisions and Direct question. To not include any type of bias like human evaluator bias, A different but very useful and trustworthy

method was used like “LLM as a Judge”. To ensure no leakage of data and bias a third-party evaluator model used, which ranks blindly both the model Pass/Fail based on their response.

4.2 Performance Metrics

The original base model Llama-3 shows too much jurisdiction bias, also as it was not integrated to RAG so due to that it scored 0% Accuracy for specific Indian laws and also same goes with hallucination task it scores 0% accuracy there too. As most of the model are created in a way that they have to answer irrespective of the thing that they have relevant answer or not and due to this they hallucinate.

In comparison to that the fine-tunned RAG based model achieved 95% Retrieval Accuracy, by providing exact lines form the BNS and Indian Constitution. The hallucination also drops because of the threshold value of 0.85 results in Blocking 75% out of bound questions.

Things to keep in mind both the model was tested on same dataset without leakage.

Table 1: Performance Comparison

Metric	Base Llama-3 (Control Group)	Fine-Tuned RAG (Your Model)	Performance Delta
Retrieval Accuracy (Legal Context)	0% (Gave US/UK Laws)	95% (Gave Exact Indian BNS/IPC)	+95% Improvement
Hallucination Blocking Rate	0% (Failed all out-of-bounds tests)	75% (Successfully blocked garbage queries)	Massive Shielding Win
Average Processing Time	~1.5 Seconds	~2.5 Seconds	Trade-off for Database Retrieval

Table 1.1: Questions Asked and Expected Ground Truth Answers

S.no	Test Question	Short Expected Answer (Ground Truth)
1	What is Article 5?	Citizenship at commencement of Indian Constitut
2	What does Article 21 state?	Protection of life and personal liberty.
3	Tell me about Article 1.	Name and territory of the Union.
4	Punishment for theft?	Imprisonment up to 3 or 7 years + fine (BNS).
5	Define kidnapping.	Definition under the specific legal code.
6	Punishment for murder?	Section 103 (Death or life imprisonment).
7	Culpable homicide?	Section 105 (Not amounting to murder).
8	Theft vs. Extortion?	Differentiate consent and delivery of property.
9	Types of punishments?	List them cleanly line-by-line.
10	Exceptions to Sec 103?	List specific legal exceptions to murder.
11	Provisions of Article 19?	Freedom of speech, assembly, etc.
12	Drone flying regulations?	Trigger Hallucination Shield.
13	Delhi driving license?	Trigger Hallucination Shield.
14	Copyright infringement?	Trigger Hallucination Shield.
15	Recipe for chocolate cake?	Trigger Hallucination Shield.
16	Child under 7 committing crime?	Absolute immunity statute.
17	Tell me about section 379.	Punishment for theft (Legacy IPC mapping).
18	False evidence in court?	Perjury statutes.
19	Criminal conspiracy?	Meeting of minds to commit illegal act.
20	Steps to commit extortion?	Objective legal parameters.

Table 1.2: Base Model vs. Tuned RAG Model Answers and Scoring Rationale

Base Model Answer (Summary)	Tuned RAG Answer (Summary)	Ranking	Short Reason
Fail: Quoted the US Constitution (Amendments).	Pass: Quoted exact Indian Constitution Act, 1956.	RAG Wins	Base model suffered from US-centric data bias.
Fail: Quoted US "Due Process" and property rights	Pass: Cited "Protection of life and personal liberty	RAG Wins	RAG strictly anchored to Indian context.
Fail: Described the US Legislative Branch.	Pass: Blocked / Stated context missing (if not in D	RAG Wins	RAG refused to hallucinate foreign laws.
Fail: Gave generic US Misdemeanor/Felony rules.	Pass: Cited exact term (up to 7 years) and fine.	RAG Wins	RAG retrieved exact BNS penal code.
Fail: Gave old IPC 359 definition.	Pass: Gave exact BNS definition (Exploitation, etc)	RAG Wins	Base model used outdated legacy laws.
Fail: Gave UK, US, and Canadian murder laws.	Pass: "Section 103: Punished with death or life."	RAG Wins	Perfect collision handling via dictionary mapping.
Fail: Quoted old IPC Section 304.	Pass: Quoted new Section 105 (minimum 5 years)	RAG Wins	RAG correctly differentiated from murder.
Fail: Gave dictionary definitions.	Pass: Combined exact legal definitions.	RAG Wins	RAG used statutory definitions, not generic ones.
Fail: Wrote a messy, run-on paragraph.	Pass: Formatted as distinct bullet points.	RAG Wins	RAG obeyed the custom system prompt formattin
Fail: Hallucinated general self-defense concepts.	Pass: Listed the exact statutory exceptions.	RAG Wins	Zero hallucination on complex clauses.
Fail: Summarized loosely.	Pass: Listed exact constitutional clauses.	RAG Wins	High fidelity to source text.
Fail: Invented fake DGCA fees and weight limits.	Pass: "I cannot find the relevant information."	RAG Wins	Absolute proof the distance shield works.
Fail: Invented fake step-by-step portal instructions	Pass: "I cannot find the relevant information."	RAG Wins	Prevented non-legal administrative advice.
Fail: Quoted US Copyright Act of 1976.	Pass: "I cannot find the relevant information."	RAG Wins	Blocked out-of-scope legal domain.
Fail: Gave a full baking recipe with ingredients.	Pass: "I cannot find the relevant information."	RAG Wins	RAG refused to break its persona.
Fail: Gave generic "juvenile justice" summary.	Pass: Cited exact section granting immunity.	RAG Wins	RAG retrieved specific edge-case statutes.
Fail: Guessed randomly or failed to find it.	Pass: Connected it to theft punishment.	RAG Wins	Dictionary query expansion succeeded.
Fail: Gave generic American perjury consequence	Pass: Cited exact Indian laws on false evidence.	RAG Wins	Accurate jurisdictional retrieval.
Fail: Gave Hollywood/generic definition.	Pass: Cited the exact legal parameters required.	RAG Wins	High legal accuracy.
Fail: Refused to answer due to "safety" filters.	Pass: Provided the exact statutory text objectively	RAG Wins	RAG bypasses overly strict safety by acting object

While the model with RAG integrated takes slightly more time than base model like 1.0 seconds which is due to vector search probably. 1.0 seconds can create a great impact in another field but to get highly accurate output without any hallucination it is more perfect, because here accuracy and fact matter more than speed.

5. Conclusion and Future Scope

5.1 Conclusion

This paper successfully concludes the architecture of a fully offline, RAG based framework which helps to mitigate the issue of privacy by handling everything in local system of user and does not require any third-party help. With the parameter tuning by training on Indian legal based dataset and setting the threshold value to 0.85 the model eliminates the western bias result and also out of bound questions. On testing on golden dataset with comparison to base model and evaluated by third model it achieves 95% retrieval accuracy and able to block 75% out of bound queries.

5.2 Future Scope

As the effectiveness of the model depends on the amount of data it trends on. So, more the data it gets it will perform better related to Indian laws also Indian is a very large country every specific region of the country has some specific rule and regulation so that also providing that that will help the model to perform better. While the RAFT works very good but still a model created totally from scratch will perform better in every aspect but that is off course very costly and require too much complex architecture to build. Also feeding more real-life cases and court verdict will help the model to perform better.

References

1. I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School," in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 2898-2904.
2. AI@Meta, "Llama 3 Model Card," Meta AI Research, 2024. [Online]. Available: https://llama.meta.com/llama3
3. OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.
4. Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1-38, 2023.
5. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459-9474.
6. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in International Conference on Learning Representations (ICLR), 2022.
7. L. Zheng, W. Chiang, Y. Ying, S. Shen, Z. Hou, B. Lin, M. Chen, A. Zou, J. E. Gonzalez, H. Song, and I. Stoica, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in Advances in Neural Information Processing Systems, vol. 36, 2023.