PAPER NAME

Final_Report.pdf

WORD COUNT 5628 Words	CHARACTER COUNT 34493 Characters
PAGE COUNT 24 Pages	FILE SIZE 1.2MB
SUBMISSION DATE Mar 26, 2024 1:16 PM GMT+5:30	REPORT DATE Mar 26, 2024 1:17 PM GMT+5:30

● 10% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 9% Internet database
- 6% Publications database
- Crossref database
- Crossref Posted Content database
- 7% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 10 words)

Neural Networks and Deep Learning
Machine Learning Optimization Algorithms

A Project Report on

NyaayAI: RAG-Powered LLM for Legal Assistance

Submitted By

Sahil Shah	Rohit Sonawane	Saket Sultania
70321019091	70321019106	70321019109

Under The Guidance Of

Prof. Archana Nanade Prof. Abhay Kolhe

Department of Computer Engineering

Shri Vile Parle Kelavani Mandal's

Mukesh Patel School of Technology & Management Engineering

March 26, 2024

1. Introduction

India's legal system is grappling with a staggering backlog of cases, with **44,181,460** pending cases per the National Judicial Data Grid (NJDG) [1] and a concerning statistic of **~70%** of these cases being older than one year. This situation underscores the pressing need for accessible and reliable legal information, particularly for the public, to foster a better understanding of the legal system and potentially alleviate the burden on the judiciary.

As highlighted in a lead article published by the Harvard Law School titled "The Implications of AI for Legal Services and Society," [2] one of the most promising applications of artificial intelligence (AI) in the legal domain is providing general legal information to the public. This project aims to leverage the power of large language models (LLMs) and recent advancements in natural language processing (NLP) to develop an intelligent and user-friendly legal aid chatbot, explicitly targeting the provision of general legal information tailored to the Indian context.

However, providing free public online access to "public legal information" of the Indian legal system is a complex endeavor, as highlighted by the National Law University of Delhi's analysis [3]. The complexities arise from the intricate nature of Indian legal information, encompassing legislation, case law, treaties, reports proposing legal reforms, and legal scholarship. Despite considerable progress in providing free access to some types of legal information, the current state needs to catch up to international standards.

A free access legal information system in India must cater to diverse audiences, including legal professionals, government administrators, small to medium enterprises (SMEs), non-governmental organizations (NGOs), students, academics, and the public. While these audiences may have varying levels of legal expertise and access to commercial legal resources, they all require substantial use of legal resources. Due to the high costs, most groups need help accessing commercial online legal services. Existing free resources from the government have significant limitations. While most of these audiences can navigate the English-based legal system and have primary internet access, making such a service genuinely relevant to the public in India remains a more arduous task, given the disparities in internet access and literacy levels, particularly concerning the English-based legal system.

This project aims to democratize access to legal knowledge and empower individuals to understand better and navigate the legal system by developing an open-source, AI-powered legal aid chatbot. The chatbot's ability to provide accurate and tailored responses in natural language can significantly improve legal literacy and access to justice, particularly for underprivileged communities, ultimately contributing to a more efficient and equitable legal system.

2. Problem Statement

The project aims to develop a comprehensive solution to address the lack of accessible and accurate legal information for diverse audiences in India. The project aims to solve the lack of accessible and accurate legal information for diverse audiences in India. The solution is comprised of the following components:

2.1 Open-Source Large Language Model (LLM):

Proprietary LLMs are not easily accessible or affordable for non-commercial use, which limits their availability for developing open-source legal information systems. An open-source LLM is needed to serve as a foundation for building specialized legal AI applications.

2.2 Fine-tuning on Indian Legal Data:

To provide accurate and relevant legal information specific to the Indian context, the LLM must be fine-tuned on a comprehensive corpus of Indian legal documents, such as legislation, case law, and other relevant sources. However, acquiring, curating, and preprocessing high-quality legal data for fine-tuning can be challenging and time-consuming.

2.3 Retrieval Augmented Generation (RAG) and Vector Database:

Fine-tuning legal data can improve the LLM's understanding of the domain. However, access to relevant legal documents and context may only limit its ability to provide accurate and reliable responses. Implementing an RAG system that combines the language model with a vector database of legal documents can enhance the system's ability to retrieve and incorporate relevant information for generating responses. However, developing an effective vector database and retrieval mechanism for legal documents poses data organization, indexing, and efficient retrieval challenges.

By framing the problem statement in terms of these components, the project can address the specific challenges associated with each aspect of the application, ultimately leading to developing a comprehensive, open-source, and sustainable legal aid chatbot tailored to the Indian context.

3. Literature Survey

The utilization of artificial intelligence (AI) and natural language processing (NLP) to enhance legal services has long been a topic of interest. [4] Recently, it has gained significant attention, with various approaches using traditional NLP methods and the latest transformer-based paradigms. These approaches fall into two categories: Logic-based or Data-centric. While AI applied to the legal domain has a promising future, it also presents some challenges. A possible solution is a Proactive Legal Information Retrieval and filtering system, which could simplify accessing legal texts and provide more relevant and timely data.

However, using conventional deep learning methodologies in law has always left room for improvement, from insufficient data to a need for more structured relationships between sentences and references between articles. A paper has reviewed various AI applications in the legal domain, including tools for legal data search, legal data analytics, and legal intelligence systems. [5] The paper also discusses the need for a generative model and creating a "legalbot" for the legal domain. The results suggest room for improvement in developing these systems, and the paper attempts to determine which deep learning technique works best for each problem application.

To understand the application of Large Language Models (LLM), we delve into the potential impact of LLMs on the legal domain. [6] Introduces the critical features of LLMs and their applicability to legal contexts while raising pertinent questions and issues surrounding their increasing presence in this field. The authors posit that LLMs will necessitate re-evaluating traditional legal approaches, roles, and education methodologies. Although generative models like LLMs have been studied for decades with practical applications in areas like speech recognition and translation, the abundance of textual data generated by LLMs is poised to transform the nature of legal work, requiring a shift in skill and tasks for legal professionals.

There is a growing interest in leveraging artificial intelligence (AI) and natural language processing (NLP) to enhance legal services, focusing on traditional NLP methods and advanced transformer-based paradigms. [7] It categorizes approaches into Logic-based and Data-centric methodologies, discussing their strengths and challenges in interpreting legal texts and queries. By exploring AI-based legal assistance systems and proactive legal information retrieval, the review emphasizes the potential to streamline legal research processes, improve the accuracy of legal advice, and revolutionize the legal industry. Thus, it underscores the significance of incorporating AI and NLP technologies to offer fast, efficient, and accurate legal assistance, aligning closely with the objectives of your project.

Discussing the problem of pending legal cases in India, the paper proposes a Virtual Legal Assistant (VLA) solution, leveraging AI and interactive technologies to provide legal consultation. [8] It parallels existing AI technologies like virtual assistants and discusses relevant legal prediction and case retrieval research. The proposed VLA framework comprises four key components, including Text Analytics, Semantic Networks, Question Generation, and User Interface, emphasizing seamless integration and information exchange. This review is invaluable as it provides insights into the challenges faced in legal proceedings and offers a solution approach through AI-driven assistance. It also outlines the technical components necessary for implementing a Virtual Legal Assistant.

LoRA freezes pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, significantly reducing the number of trainable parameters while maintaining model quality. [9] The review introduces Low-Rank Adaptation (LoRA) as an efficient strategy for fine-tuning large-scale language models like GPT-3, aiming to reduce the computational and memory costs associated with full fine-tuning. This approach offers practical benefits such as reduced GPU memory requirements, faster training time, and seamless task-switching during deployment without introducing additional inference latency. Implementing LoRA in an LLM chatbot could optimize performance and resource utilization, allowing efficient language model adaptation to specific legal tasks while minimizing computational overhead. Additionally, LoRA principles may offer insights into designing and optimizing neural networks with dense layers, enhancing the chatbot's effectiveness.

The review comprehensively examines Retrieval-Augmented Generation (RAG), highlighting its evolution and impact on enhancing Large Language Models (LLMs) by integrating external knowledge bases. [10] It categorizes RAG into Naive, Advanced, and Modular paradigms, elucidating their architectural advancements and contributions to model interpretability. The paper underscores RAG's integration with other AI methodologies like fine-tuning and reinforcement learning, expanding its capabilities. Additionally, it discusses the emergence of hybrid content retrieval methods, offering enriched data sourcing. This review is valuable for understanding RAG's progression, offering insights into its potential applications in various domains. For our project, which involves developing a chatbot with RAG optimization, this review can inform our understanding of RAG's components and advancements, aiding in designing and implementing our chatbot system.

The paper discusses the evolution of natural language processing, focusing on the Retrieval-Augmented Generation (RAG) framework and its modular design's impact on performance. Through experimentation with a unique dataset from the Amazon Rainforest, it highlights the importance of preserving indigenous cultures and biodiversity. [11] Findings reveal that models like GPT and Palm excel under different contexts, emphasizing the significance of tailored datasets and embedding mechanisms in optimizing Large Language Models (LLMs). For an LLM chatbot, the research offers insights into advanced techniques in NLP, including modular framework design and context-aware optimization, which can enhance the chatbot's ability to provide culturally sensitive legal assistance tailored to diverse user needs and contexts.

Showcases implementation of a Retrieval-Augmented Generation (RAG)-based chatbot for credit card-related queries using FAQ data, emphasizing the superiority of an in-house retrieval embedding model over general-purpose public models in accuracy and out-of-domain query detection. [12] The paper introduces a token optimization strategy using Reinforcement Learning (RL), controlled by a policy-based model to decide on context retrieval, resulting in significant cost savings and slight accuracy improvements. This approach is adaptable to any existing RAG pipeline, offering a generic solution for enhancing chatbot performance. This survey provides valuable insights into specialized chatbot development, retrieval model training, and RL-based optimization techniques, which can aid in crafting a more efficient and accurate legal assistance system tailored to user queries.

4. Proposed Work

This project proposes a system combining state-of-the-art Natural Language Processing (NLP) technologies integrated with various algorithms, neural network approaches, and optimization processes to develop a chatbot specifically tuned for legal assistance and information retrieval.

The workflow of the Large Language Model chatbot starts with initial user interaction, where the user submits a query through Chainlit, a web-based interface optimized explicitly for hosting large language models. Upon retrieval of the query, the chatbot initiates the word embedding and tokenization process with the help of the nomic embedding model, where the chatbot retrieves relevant information stored in the vector database using chroma, and this process is done using the Retrieval-Augmented Generation (RAG) mechanism which consists of a corpus of additional relevant data. Simultaneously, the chatbot performs semantic analysis using the open-source Large Language Model. Combining user query analysis and retrieved information, the chatbot generates a response tailored explicitly to the user's initial query.

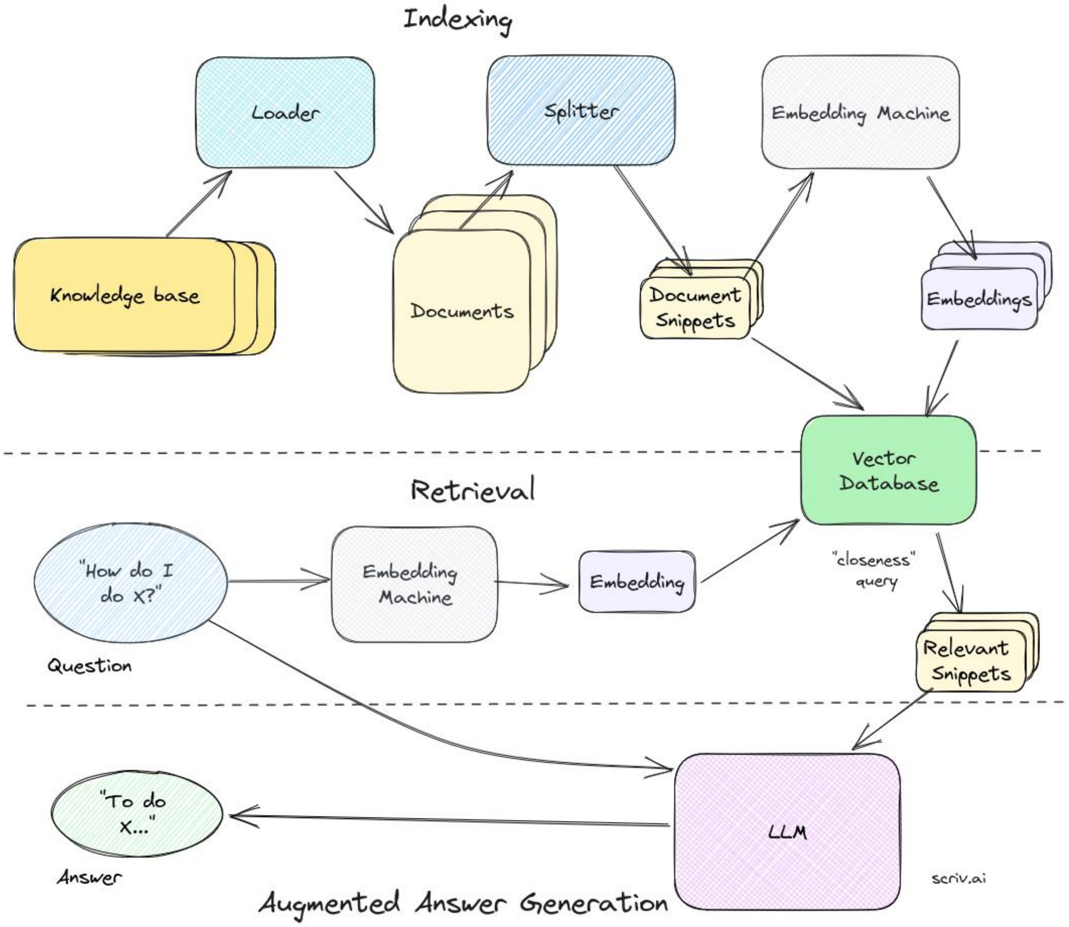


Figure 1 General RAG Architecture

4.1 Dataset Creation and Acquisition:

Dataset creation and acquisition had two primary purposes, which were to create a corpus of data for the Retrieval-Augmented Generation (RAG) mechanism and to optimize the base model of the Large Language Model (LLM) to be able to finetune the model for scalability and lightweight purposes efficiently.

The dataset used for RAG encompassed various sources, such as the Indian Penal Code (IPC), the Code of Criminal Procedure (CrPC), and the Constitution of India. These documents served as the foundational information crucial to the Indian legal system. These datasets were obtained through official government websites in “.pdf” format, and thereby, after the acquisition, the documents were put through a multitude of steps involving pre-processing where tasks like tokenization, removing stop words, and converting the text to lowercase were implemented. The pre-processed documents were indexed for efficient retrieval during the RAG process.

In addition to the curated corpus of documents for RAG, the dataset used to optimize the base LLM model was created by skimming through the web to find open-source datasets containing question-answer query pairs involving legal topics. These datasets were retrieved through sources like Kaggle and Hugging Face, and this process also involved custom data, which tried to cover simulated legal queries, variations of existing questions, or specialized legal scenarios designed to fine-tune the chatbot's performance and address specific use cases.

Using these datasets, the system gained expertise and exposure to a broader range of legal scenarios, enhancing overall effectiveness and efficiency in understanding and responding to users. By integrating curated legal documents and diverse datasets, including open-source and custom-generated data, the training ensured that the LLM chatbot was well-equipped to handle various legal inquiries accurately and relevantly.

Table 1 Overview of Datasets Used for NyaayAI

Dataset Name	Source	Number of Entries	Use Case
Indian Penal Code (IPC)	Legal Documents	Entire	RAG
Constitution of India	Legal Documents	Entire	RAG
Code of Criminal Procedure (CrPC)	Legal Documents	Entire	RAG
Legal Q&A Dataset	Hugging Face	13,324	Optimization
Legal Assistance Kaggle Dataset	Kaggle	5,545	Optimization
Open Legal Knowledge Base	Open Source	5,000	Optimization
Custom-Generated Legal Scenarios	Project-specific	1500	Optimization

4.2 Understanding the LLM model

At the core of the system lies an open-source Large Language Model that is renowned for its processing capability in the domain of natural language understanding and generation, which helps to address an array of tasks ranging from answering questions based on contextually based information to generating coherent and relevant text which is in response to user-generated queries. Among the various Large Language Models considered for integration in our system, Google’s Gemma performed and fit the domain required ideally compared to counterparts like Meta’s Llama2 and various other open-source models.

Despite architectural differences, training data, and implementation, these LLMs share several commonalities.

- Firstly, they are built upon transformer architectures, allowing for efficient sequential data processing through attention mechanisms.
- Secondly, they utilize pre-training followed by fine-tuning on specific tasks, enabling them to learn diverse linguistic patterns and contextual understandings from vast text corpora.
- Additionally, they can generate coherent and contextually relevant text across various topics, showcasing their proficiency in natural language understanding and generation tasks.

Moreover, LLMs are adaptable to various downstream applications, such as language translation, text summarization, and sentiment analysis, owing to their versatility and generalization abilities. Furthermore, they often face ethical and societal concerns regarding biases, misinformation propagation, and data privacy, necessitating ongoing research and development to address these challenges and ensure responsible deployment. Overall, LLMs represent a significant advancement in natural language processing, offering powerful tools for understanding, generating, and analyzing human language at scale.

4.3 Understanding the Retrieval-Augmented Generation (RAG) Mechanism and ChromaDB

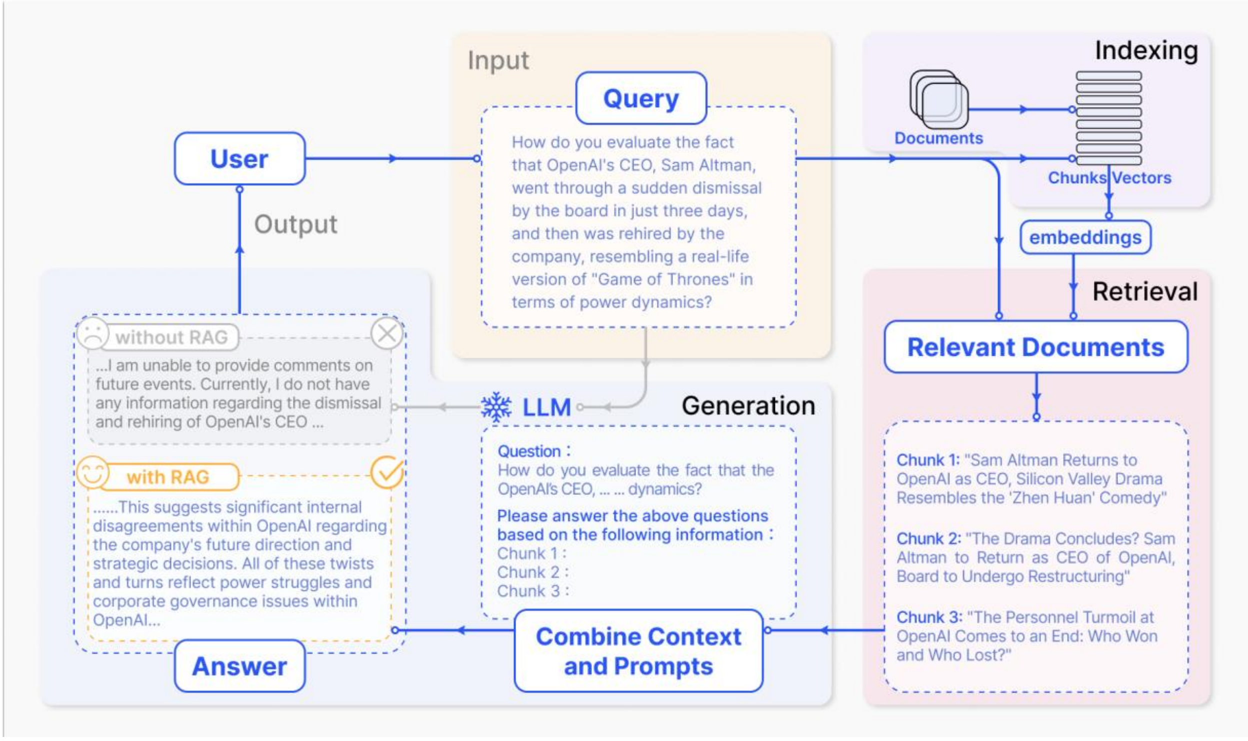
The core functionality of the chatbot relies on the seamless integration of the Retrieval-Augmented Generation (RAG) method with proficient management of document embeddings through Chroma, a high-performance vector database.

4.3.1 Retrieval-augmented generation (RAG):

Retrieval-augmented generation (RAG) is a framework that helps combine the elements of retrieval-based and generative-based models. It extends previous models like T5 (text–to–text transfer transformer) and BART (Bidirectional and Auto-Regressive Transformers), incorporating retrieval mechanisms to enhance performance.

In traditional generative models, the model generates text based entirely on the input received and its learned parameters. Still, on the other hand, retrieval-based models rely solely on retrieving relevant information from a database provided by a given query and then using the retrieved documents to generate appropriate responses.

The RAG methodology combines Large Language Models (LLM) with information extracted from a thorough selection of a corpus of legal documents. The process followed by the RAG mechanism begins by processing user input, which is put through tokenization to split the text into several discrete tokens like words and punctuation; this step puts forth the foundation that enables the system to unravel the semantic nuances embedded with the initial input text.



How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was rehired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?

...I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and rehiring of OpenAI's CEO ...

......This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI...

Question：
How do you evaluate the fact that the OpenAI's CEO, ... dynamics?
Please answer the above questions based on the following information：
Chunk 1 :
Chunk 2 :
Chunk 3 :

Chunk 1: "Sam Altman Returns to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"

Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"

Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"

Figure 2 A representative instance of the RAG process applied to question-answering

Once the tokenization process is completed, the tokens are transformed into a dense vector representation known as Word embedding. These embeddings are the backbone of the system's understanding, capturing intricate semantic relationships between words and enhancing the chatbot's comprehension of the input text. The chatbot gains a deeper understanding of the user's query by extracting and utilizing the true power of word embeddings, facilitating more precise and contextually relevant response generation.

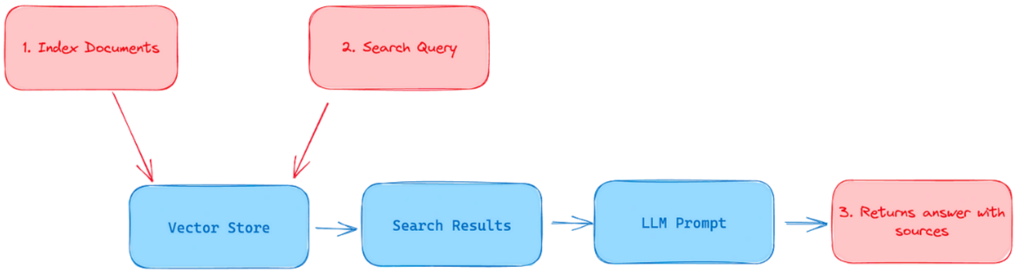
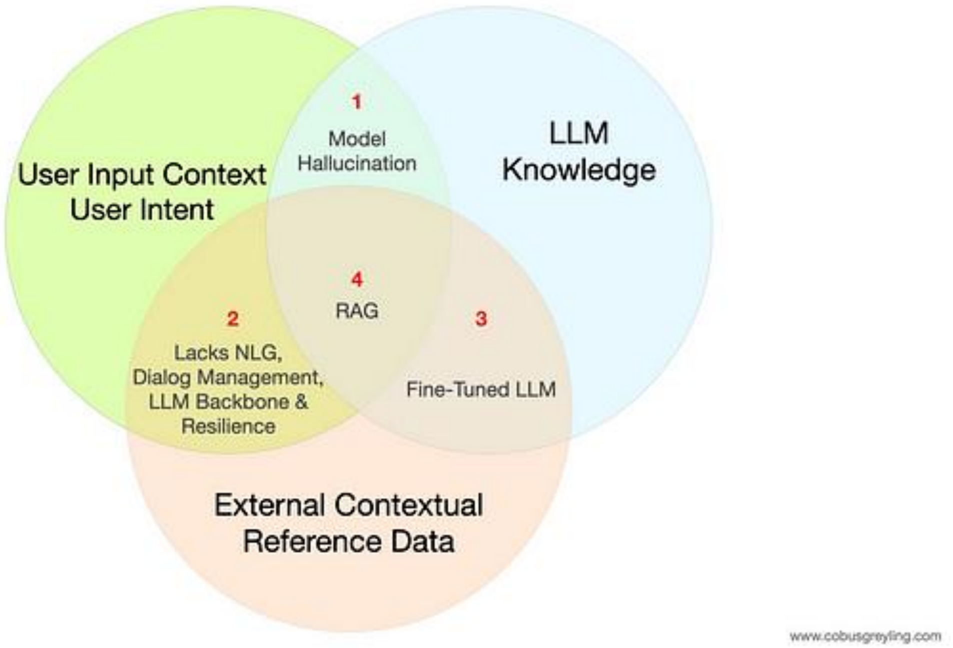


Figure 3 Flowchart of RAG

The true potential of the RAG mechanism shines in its ability to retrieve pertinent information from a designated corpus of legal documents, which includes authoritative sources such as the Indian Penal Code (IPC), the Indian Constitution, and the Code of Criminal Procedure (CrPC). The chatbot navigates the vast expanse of legal documents through advanced retrieval algorithms, retrieving information that aligns seamlessly with the user's query. This integration of external knowledge elevates the chatbot's capabilities, ensuring the accuracy and reliability of the information provided to the user.

Additionally, the system incorporates question-and-answer pairs from open-source datasets to enhance responsiveness and speed, fostering continuous learning and improvement.



www.cobusgreyling.com

Figure 4 Significance of RAG

4.3.2 Chroma Vector Database:

Chroma DB is an open-source vector storage system designed to store and retrieve document embeddings efficiently. The base function of this vector database is to store embeddings associated with metadata for subsequent use by extensive language models.

Chroma can also lay the groundwork for semantic search engines that operate on textual data; it offers an ideal solution for managing large volumes of unstructured and semi-structured data. Through seamless integration, chroma maximizes retrieval efficiency, enabling quick access to stored data from the collection of legal documents.

Chroma’s architecture favors speed and scalability, so Chroma can easily store and manage an extensive collection of document embeddings. By complex indexing algorithms and data structures, chroma can maximize the retrieval process, enabling the chatbot to swiftly seep through the corpus of data and retrieve important information. Chroma also supports a distributed architecture, which helps in the fault tolerance capabilities of the database and provides ready-to-go access to critical data even when system or network failures occur.

Integrating the RAG mechanism with Chroma's efficient document embedding management empowers the chatbot to provide unparalleled legal assistance and information retrieval. The chatbot surpasses traditional constraints by seamlessly combining external knowledge with natural language comprehension, delivering precise and contextually rich responses to user queries.

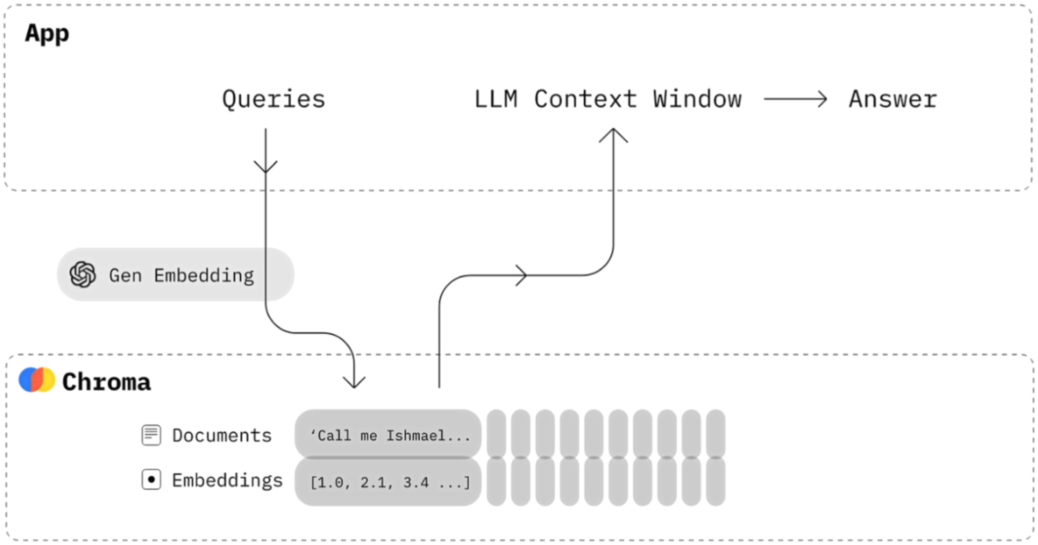


Figure 5 Workflow of Using ChromaDB as Vector Store

4.4 Understanding the Embeddings Model:

The nomic embedding model is used for training advanced methods specifically designed and tuned to handle the unpredictability and quirks of complex language. By utilizing and taking advantage of the advancements of cutting-edge technologies in the domain of neural networks and natural language processing techniques, the model is engrained with the ability to identify and extract minute semantic details hidden in complex documents.

The embedding models are iteratively trained on diverse datasets. So, their ability to retrieve and represent complex knowledge is stored in a vector form, enhancing its effectiveness in the chatbot retrieval tasks. The model also integrates well with other technologies used in the system, such as chroma, a high-performance vector database, enhancing the efficiency of precise information retrieval.

Due to the combined use of the nomic embedding model and chroma, the chatbot can perform exceptionally well and gain unmatched access to accurate complex data extracted from the large corpus of various legal documents. The nomic embedding model is a critical component of the system architecture as it provides the precise expertise to negotiate the complex world of legal jargon. By harnessing the power of advanced natural language processing techniques, the Nomic Embeddings Model elevates the chatbot's capabilities, enabling it to provide unparalleled assistance to users seeking legal information and guidance.

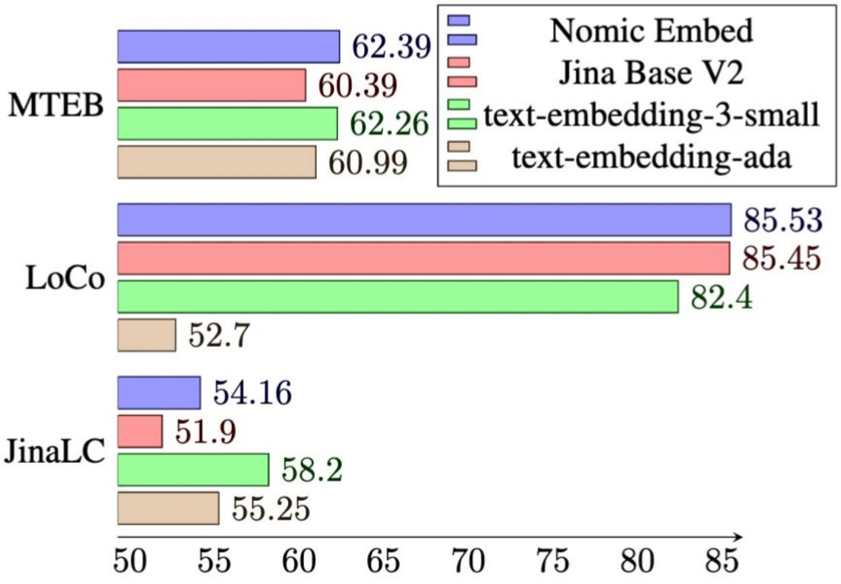


Figure 6 Comparative Analysis of Nomic Embedding

4.5 Understanding the Optimization Process:

In the proposed system, we analyzed many optimization techniques and algorithms needed to improve the performance aspects of the model, like speed, accuracy, and cost-saving, by reducing resource consumption and minimizing downtime. Various optimization algorithms like Parameter Efficient Fine Tuning (PEFT), Low-Rank Adaptation (LORA), and Quantized Low-Rank Adaptation (QLORA) were utilized in the project to enhance the performance and efficiency of the chatbot. These algorithms play a pivotal role in improving the accuracy and responsiveness of the chatbot as they help generate contextually relevant responses to user queries.

4.5.1 Parameter-Efficient Fine-Tuning (PEFT):

PEFT is an optimization algorithm that aims to transform the input data's partially exchangeable features and improve model performance. In the context of the chatbot project, PEFT is used to pre-process and convert input features, such as user queries and legal documents, to capture hidden and underlying semantic relationships better. By optimizing the feature transformation process, PEFT enhances the ability of the chatbot to understand and generate accurate responses to user-specific queries.

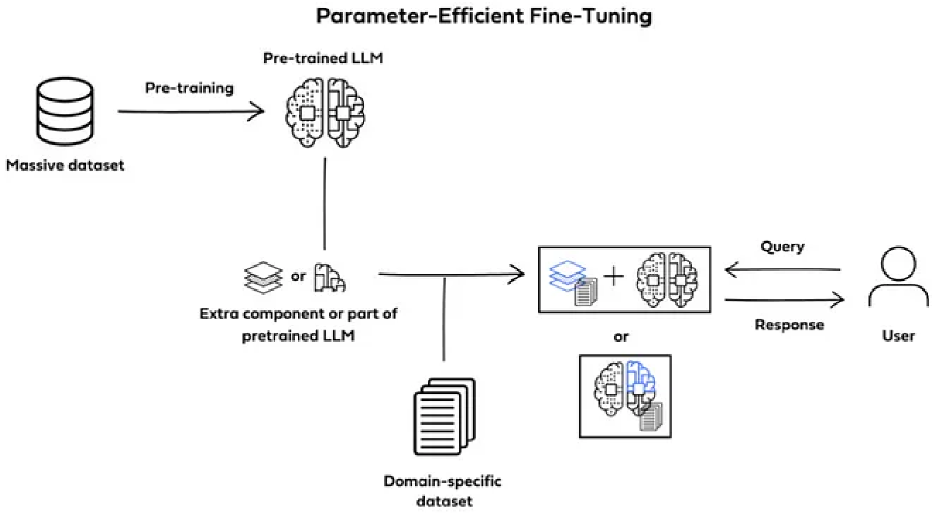


Figure 7 Workflow for PEFT

4.5.2 Low-Rank Adaptation (LORA):

Low-rank adaptation (LORA) is a concept commonly used in the context of machine learning, particularly in the field of domain adaptation. Domain adaptation refers to the scenario where the training data (source domain) and the testing data (target domain) have different distributions. LORA focuses explicitly on adapting a model from a high-dimensional source domain to a low-dimensional target domain.

The basic idea behind LORA is to exploit the low-rank structure inherent in the feature space to perform adaptation effectively. In many real-world scenarios, the target domain may have fewer relevant features or exhibit specific correlations among features that a low-rank representation can capture.

LORA methods typically involve subspace alignment, where the source and target domains are mapped into a common subspace, preserving the low-rank structure. By aligning the subspaces of the source and target domains, LORA aims to mitigate the distribution shift between them and improve the performance of the model on the target domain.

Overall, LORA techniques aim to leverage the inherent structure of data to perform domain adaptation efficiently, especially when dealing with high-dimensional source and low-dimensional target domains.

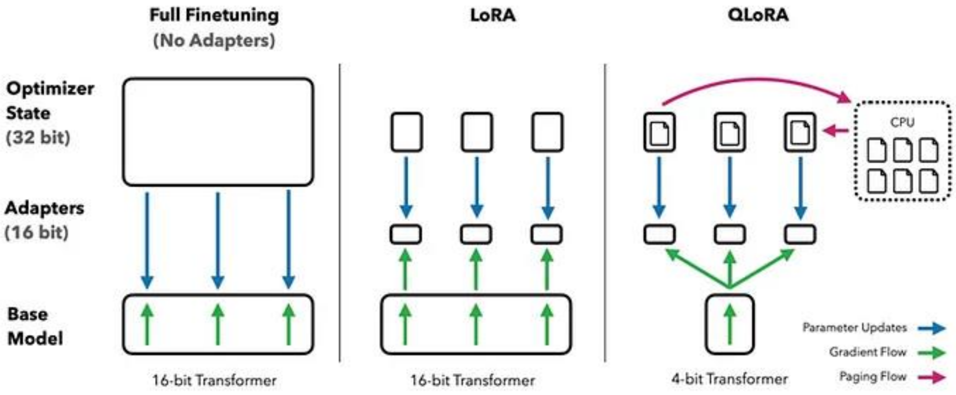


Figure 8 Different Methods of Finetuning a LLM

4.5.3 Quantized Low-Rank Adaptation (QLORA):

Quantized Low-Rank Adaptation (QLORA) is an extension or variation of Low-Rank Adaptation (LORA) incorporating quantization techniques into the adaptation process. Quantization involves mapping continuous values to a finite set of discrete values, which can reduce computational complexity and memory requirements.

In the context of domain adaptation, QLORA aims to adapt a model from a high-dimensional source domain to a low-dimensional target domain while leveraging the low-rank structure in the feature space. However, QLORA further incorporates quantization into the adaptation process, which can offer benefits such as:

- **Reduced Computational Complexity**: By quantizing feature representations, the computational complexity of adaptation algorithms can be reduced, making them more efficient, especially in scenarios with limited computational resources.
- **Improved Generalization**: Quantization can help generalize the adaptation process by reducing overfitting, especially when dealing with limited data in the target domain.
- **Memory Efficiency**: Quantization often reduces memory requirements, benefiting resource-constrained environments like mobile devices or embedded systems.
- **Robustness to Noise**: Quantization can make the adaptation process more robust to noise in the data, as it tends to smooth out the representations.

Quantized Low-Rank Adaptation (QLORA) integrates quantization techniques into the domain adaptation process, aiming to improve efficiency, generalization, and robustness while adapting a model from a high-dimensional source domain to a low-dimensional target domain.

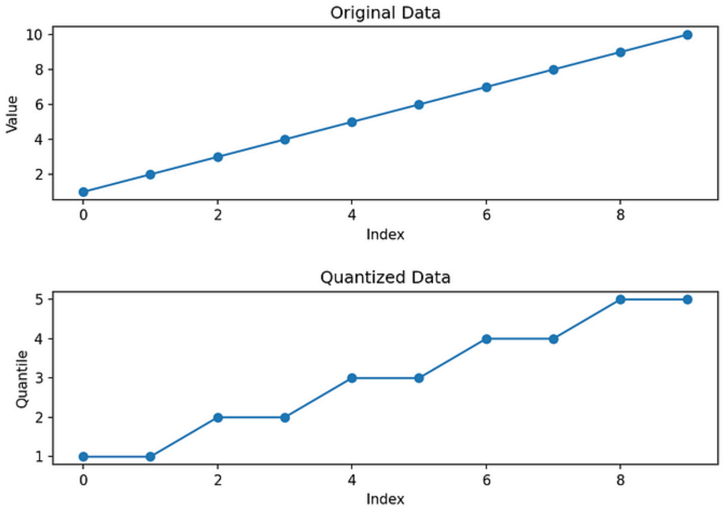
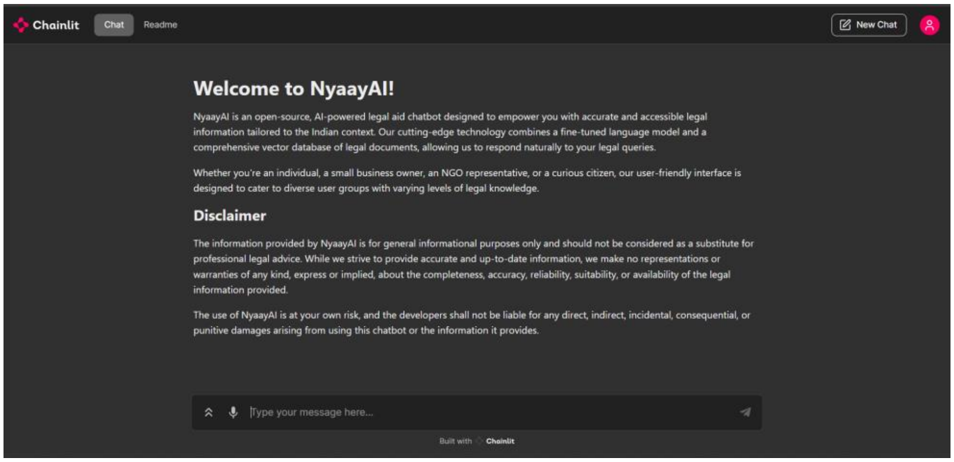


Figure 9 Comparative Analysis of Original and Quantized Data

Overall, integrating PEFT, LORA, and QLORA optimization techniques in the project significantly contributes to the chatbot's performance and effectiveness in providing legal assistance and information retrieval. These algorithms work together to improve the model's ability to understand complex legal text, generate coherent responses, and adapt to diverse user queries, ultimately enhancing the user experience and satisfaction with the chatbot's services.

5. Results and Discussions

The initial interface of the NyaayAI chatbot aimed to provide a user-friendly experience for individuals with varying levels of legal knowledge. While the initial interface offered essential functional expertise, there was room for improvement in the depth and clarity of the chatbot's responses and the overall user experience.



Welcome to NyaayAI!

NyaayAI is an open-source, AI-powered legal aid chatbot designed to empower you with accurate and accessible legal information tailored to the Indian context. Our cutting-edge technology combines a fine-tuned language model and a comprehensive vector database of legal documents, allowing us to respond naturally to your legal queries.

Whether you're an individual, a small business owner, an NGO representative, or a curious citizen, our user-friendly interface is designed to cater to diverse user groups with varying levels of legal knowledge.

Disclaimer

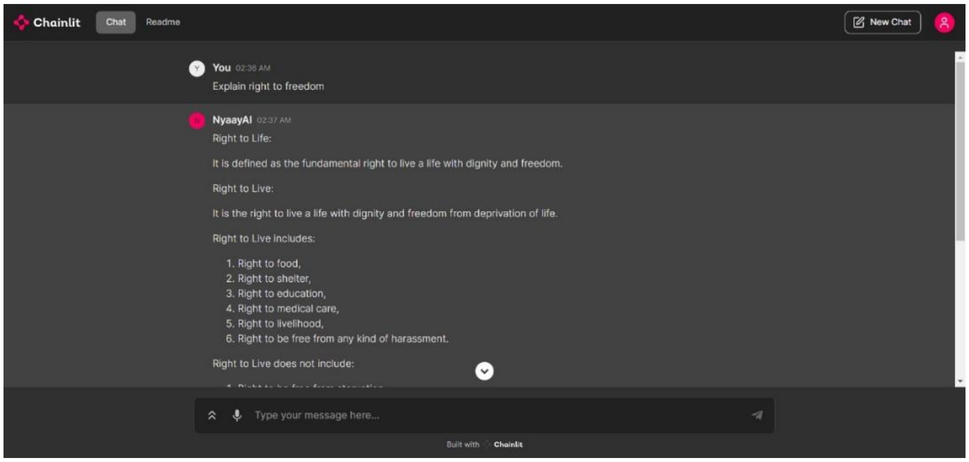
The information provided by NyaayAI is for general informational purposes only and should not be considered as a substitute for professional legal advice. While we strive to provide accurate and up-to-date information, we make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the legal information provided.

The use of NyaayAI is at your own risk, and the developers shall not be liable for any direct, indirect, incidental, consequential, or punitive damages arising from using this chatbot or the information it provides.

Figure 10 Welcome Screen for Interface

5.1 Pre-Optimization

The figure below illustrates the response generated by the chatbot before any optimization or fine-tuning processes were implemented. Hallucinations and inaccuracies characterize the unoptimized chatbot's response.



You 02:36 AM

Explain right to freedom

NyaayAI 02:37 AM

Right to Life:

It is defined as the fundamental right to live a life with dignity and freedom.

Right to Live:

It is the right to live a life with dignity and freedom from deprivation of life.

Right to Live includes:

1. Right to food,
2. Right to shelter,
3. Right to education,
4. Right to medical care,
5. Right to livelihood,
6. Right to be free from any kind of harassment.

Right to Live does not include:

Figure 11 Before Optimization

5.2 Optimization Process

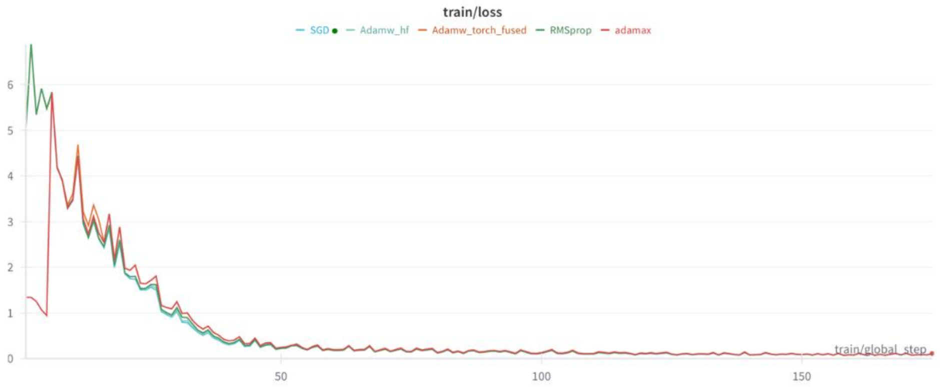


Figure 12 Training Loss of all Optimizers

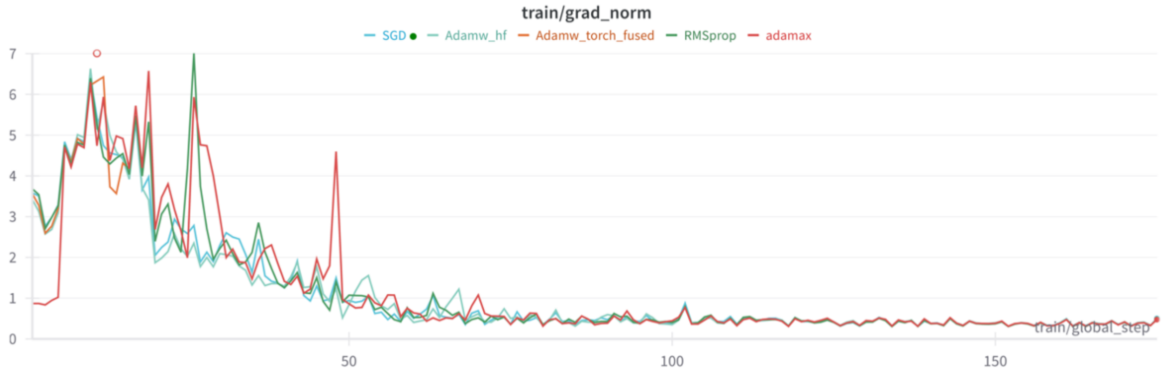


Figure 13 Gradient Normalization

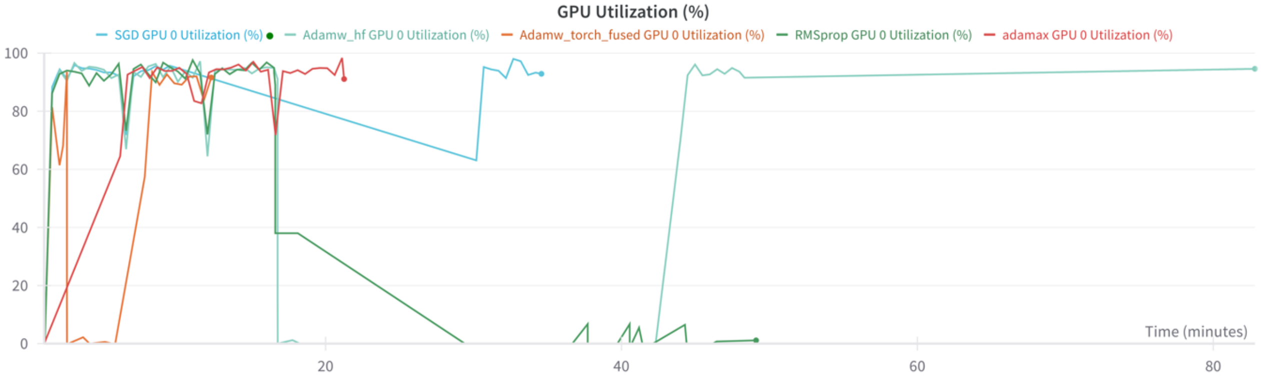


Figure 14 GPU utilization (%)

These are some of the insights gained from the training process of the LLM model based on PEFT, LORA, and QLORA:

- Adamax: An extension of the Adam optimizer utilizes the infinity norm for computing the weights. It tends to perform well in many cases, but in this scenario, it resulted in a relatively higher training loss than other optimizers.

- SGD: A classic optimizer that updates the weights with the gradient of the loss function. Despite its simplicity, it performs reasonably well. In this case, it performed slightly better than Adamax.

- AdamW: A variant of Adam with weight decay. It is known to perform well with large-scale models. The usage of the Hugging Face library suggests integration with pre-trained models. The low training loss indicates that this optimizer configuration worked well for this task.

- AdamW with Torch Fused: Similar to the AdamW with Hugging Face, this variant also utilizes weight decay. The slightly higher loss compared to AdamW with the HF library suggests that different implementations or configurations may yield different results.

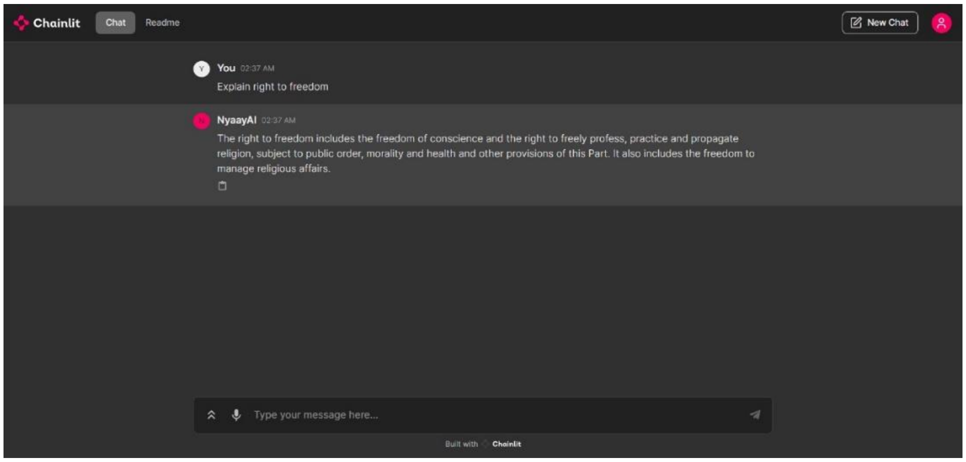
- RMSprop: An optimizer that adapts the learning rate based on the magnitude of recent gradients. It typically performs well in scenarios where the gradients vary widely. In this case, it performed similarly to AdamW variants.

Table 2 Optimization process summary

Optimizer	**Learning Rate**	**Time Taken (hrs)**	**Train Loss**
Adamax	2e-1	5:20	0.139000
SGD	2e-1	5:10	0.137000
Adamw_hf	2e-4	5:12	**0.076200**
Adamw_torch_fused	2e-4	5:11	0.113600
RMSprop	2e-1	5:15	0.114500

Overall, the choice of optimizer and its hyperparameters significantly impacts the training process and the resulting model performance. In this scenario, AdamW with the Hugging Face library yielded the lowest training loss, indicating its effectiveness for this particular task. However, further experimentation and fine-tuning might be necessary to optimize the model's performance further.

5.3 Post Optimization



You 02:37 AM

Explain right to freedom

NyaayAI 02:37 AM

The right to freedom includes the freedom of conscience and the right to freely profess, practice and propagate religion, subject to public order, morality and health and other provisions of this Part. It also includes the freedom to manage religious affairs.

Figure 15 post-optimization

After the optimization process, which involved fine-tuning the language model, integrating the Retrieval-Augmented Generation (RAG) component, and incorporating user feedback, the content of the NyaayAI chatbot interface was significantly enhanced. The optimized interface content, coupled with the improved performance of the fine-tuned language model and the RAG system, aimed to provide users with accurate, comprehensive, and actionable legal information in a user-friendly and accessible manner.

6. Conclusion

The Legal Assistance Chatbot was developed using RAG (Retrieval-Augmented Generation) Optimization, significantly improving legal assistance technology. This project integrates advanced natural language processing techniques, retrieval-based methods, and generation models to create an advanced interface that provides personalized legal guidance and information.

By leveraging RAG optimization, the chatbot can now understand user queries, retrieve relevant legal documents and cases from databases, and generate relevant and accurate responses. Optimization has increased the efficiency of information retrieval and improved the quality of the responses provided to users.

While the Legal Chatbot has demonstrated promising results in its current state, there is still room for further improvements and expansion. Future iterations of the project could focus on fine-tuning the natural language understanding capabilities, integrating with additional legal databases and APIs, and enhancing the dialogue management system for more seamless interactions.

Overall, this project offers valuable insights into the potential of advanced AI technologies to transform the legal industry, improving access to justice and empowering individuals with more excellent knowledge and understanding of legal matters. It represents a significant step forward in the intersection of artificial intelligence and legal services, with implications for practitioners and the public.

7. Future Work

Within the ever-evolving domain of legal assistance technology, the future of our legal chatbot using RAG Optimization assures more significant improvements in improving and enhancing its efficiency, scalability, and functionality. Building upon the current capabilities of the model, future work for this project involves using relevant, optimized techniques and technologies to advance legal services. The following are some of the noted improvements on the current proposed system:

1. **Efficiency and Scalability:**
    a. Use various techniques to create optimized smaller and faster RAG versions.
    b. Reduce required computational memory and resource utilization by using compression models.
2. **Multi-modal Capabilities:**
    a. Diversify input methods by allowing text, audio, and visual interactions to improve the chatbot's versatility.
    b. Allow users to upload documents or images regarding legal queries, allowing the chatbot to analyze and aid.
3. **Integration with Legal Research Tools:**
    a. Access updated legal precedents, laws, and decrees by linking to legal research platforms.
    b. Enhance the chatbot's knowledge base by linking to legal databases, citation systems, and repositories.
4. **Legal Document Generation and Review:**
    a. Additional functionality to generate legal documents like contracts, agreements, or letters based on defined user specifications.
    b. Work with NLG techniques to draft documents and review compliance and accuracy.

8. References

[1] 'Welcome to NJDG - National Judicial Data Grid'. Accessed: Mar. 24, 2024. [Online]. Available: https://njdg.ecourts.gov.in/njdgnew/?p=main/pend_dashboard

[2] 'The Implications of ChatGPT for Legal Services and Society,' Harvard Law School Center on the Legal Profession. Accessed: Mar. 25, 2024. [Online]. Available: https://clp.law.harvard.edu/article/the-implications-of-chatgpt-for-legal-services-and-society/

[3] R. Singh, S. D. Rao, P. Rai, and A. Singh, 'ACCESS TO LEGAL INFORMATION & RESEARCH IN DIGITAL AGE.'

[4] J. Dias *et al.*, 'State of the Art in Artificial Intelligence applied to the Legal Domain.' arXiv, Mar. 10, 2022. Accessed: Mar. 24, 2024. [Online]. Available: http://arxiv.org/abs/2204.07047

[5] R. Sil, A. Roy, B. Bhushan, and A. K. Mazumdar, 'Artificial Intelligence and Machine Learning based Legal Application: The State-of-the-Art and Future Research Trends,' in *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Greater Noida, India: IEEE, Oct. 2019, pp. 57–62. doi: 10.1109/ICCCIS48478.2019.8974479.

[6] D. Charlotin, 'Large Language Models and the Future of Law.' Rochester, NY, Aug. 22, 2023. doi: 10.2139/ssrn.4548258.

[7] A. R. Kandula, M. Tadiparthi, P. Yakkala, S. Pasupuleti, P. Pagolu, and S. M. Chandrika Potharlanka, '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)*, Kanjirapally, India: IEEE, Nov. 2023, pp. 1–6. doi: 10.1109/AICERA/ICIS59538.2023.10420298.

[8] N. Jain and G. Goel, 'An Approach to Get Legal Assistance Using Artificial Intelligence,' in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India: IEEE, Jun. 2020, pp. 768–771. doi: 10.1109/ICRITO48877.2020.9198029.

[9] E. J. Hu *et al.*, 'LoRA: Low-Rank Adaptation of Large Language Models.' arXiv, Oct. 16, 2021. Accessed: Mar. 26, 2024. [Online]. Available: http://arxiv.org/abs/2106.09685

[10] Y. Gao *et al.*, 'Retrieval-Augmented Generation for Large Language Models: A Survey.' arXiv, Jan. 04, 2024. Accessed: Mar. 26, 2024. [Online]. Available: http://arxiv.org/abs/2312.10997

[11] K. Pichai, 'A Retrieval-Augmented Generation Based Large Language Model Benchmarked On a Novel Dataset,' *Journal of Student Research*, vol. 12, Nov. 2023, doi: 10.47611/jsrhs.v12i4.6213.

[12] M. Kulkarni, P. Tangarajan, K. Kim, and A. Trivedi, 'Reinforcement Learning for Optimizing RAG for Domain Chatbots'. arXiv, Jan. 09, 2024. Accessed: Mar. 26, 2024. [Online]. Available: http://arxiv.org/abs/2401.06800

● 10% Overall Similarity

Top sources found in the following databases:

- 9% Internet database
- 6% Publications database
- Crossref database
- Crossref Posted Content database
- 7% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	i-scholar.in Internet	4%
2	medium.com Internet	<1%
3	gabormelli.com Internet	<1%
4	Coventry University on 2023-08-09 Submitted works	<1%
5	kalaharijournals.com Internet	<1%
6	"ICICIS 2023 Schedule", 2023 International Conference on Integration o... Crossref	<1%
7	The British College on 2023-05-31 Submitted works	<1%
8	ifms.edu.br Internet	<1%

9	Kieran Pichai. "A Retrieval-Augmented Generation Based Large Langua... Crossref	<1%
10	theprint.in Internet	<1%
11	timesofindia.indiatimes.com Internet	<1%
12	Jia-na Meng. "Transfer Learning Based on SVD for Spam Filtering", 201... Crossref	<1%
13	globalscientificguild.com Internet	<1%
14	pdfs.semanticscholar.org Internet	<1%
15	Emirates Aviation College, Aerospace & Academic Studies on 2019-0... Submitted works	<1%
16	University of Birmingham on 2023-09-18 Submitted works	<1%
17	export.arxiv.org Internet	<1%
18	coursehero.com Internet	<1%
19	www4.austlii.edu.au Internet	<1%
20	Jooyeup Lee, Wooyong Jung, Seungwon Baek. "In-House Knowledge ... Crossref	<1%

21	University of Lancaster on 2023-05-28 Submitted works	<1%
22	University of Limerick on 2024-03-13 Submitted works	<1%
23	jsr.org Internet	<1%