Srinivas Report.pdf

by Srinivas 1

Submission date: 03-Mar-2025 08:10AM (UTC+0530)

Submission ID: 2603365353

File name: Srinivas_Report.pdf (610.33K)

Word count: 4063

Character count: 23072

ABSTRACT

The Indian legal system is massive, with lakhs of cases still pending and a never-ending stream of legal documents. For lawyers, judges, or even regular people, finding the right case laws, understanding long court judgments, or checking if the legal news is accurate can be a real headache. And with all the fake news floating around misinterpretations of rulings, misleading updates, and viral misinformation it's hard to trust what we see online. This makes it even harder for people to get the correct legal information, which leads to confusion and poor decisions.

That's where NyayaMitra comes in. It's like your AI-powered legal assistant, designed to make legal research simpler and more reliable. Whether you need to analyze a case, find related precedents, summarize long legal documents, or even spot fake legal news, NyayaMitra does it all. Using smart AI models like Phi-3 Mini and some cool search and machine learning techniques, it helps users find the right legal information quickly without wasting hours on paperwork. Made especially for the Indian legal system, NyayaMitra is for everyone lawyers, students, researchers, or even common people who need help understanding legal stuff. The aim is simple: make legal knowledge easier to access, fight fake news, and help people make better, informed decisions without the usual legal drama.

Keywords: AI in Legal Research, Case Law Retrieval, Fake Legal News Detection, Legal Document Summarization, RAG-Based Search, Indian Legal System

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AI	Artificial Intelligence
NLP	Natural Language Processing
RAG	Retrieval-Augmented Generation
NER	Named Entity Recognition
GPU	Graphics Processing Unit
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
FAISS	Facebook AI Similarity Search
LSTM	Long Short-Term Memory (Neural Network)
mBERT	Multilingual BERT
T5	Text-To-Text Transfer Transformer
BART	Bidirectional and Auto-Regressive Transformer
ROUGE	Recall-Oriented Understudy for Gisting Evaluation

NOTATION

- **P@5 (Precision@5)**: Fraction of relevant case precedents retrieved among the top 5 results.
- **R@5 (Recall@5)**: Fraction of relevant case precedents correctly retrieved within the top 5 results.
- **ROUGE-1**: Overlap of unigrams (single words) between the generated summary and the reference summary.
- **ROUGE-2**: Overlap of bigrams (two consecutive words) between the generated and reference summaries.
- **F1-Score**: Harmonic mean of precision and recall, used for evaluating classification tasks.
- **Accuracy**: Fraction of correctly classified cases/news articles among total samples.

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

The Indian legal system is massive, with more than **5 crore pending cases** across various courts. Lawyers, judges, and even common citizens struggle to find relevant case precedents, go through lengthy legal documents, and verify whether legal news is genuine. The way legal research happens in India is still largely **manual and time-consuming**, making it hard for professionals to access the right information when they need it.

I realized how big this problem is when I was following a Supreme Court case about the arrest of a political leader. During the proceedings, the defending lawyer, who had decades of experience quoted an old case to support his argument. But the judge, who was relatively less experienced, had to ask for an explanation and, in some cases, even postponed the hearing to verify the reference. This delay could have been avoided if the judge had an AI-powered tool that could instantly pull up and summarize relevant case precedents.

That incident made me think:

- What if there was a tool that could **instantly fetch past case precedents** when cited in court?
- What if long legal documents could be **summarized in seconds** instead of taking hours to read?
- What if there was a way to **quickly verify the authenticity of legal references**, avoiding unnecessary delays?

Another major issue is **fake legal news**. These days, it's common to see misleading legal updates on WhatsApp, YouTube, and social media—false Supreme Court rulings, misinterpretations of laws, and fabricated judgments. Many people believe these without verifying, and sometimes even professionals get misled.

This is where AI can play a game-changing role. With advancements in **Natural Language Processing (NLP)** and **Machine Learning**, AI models can automate legal research, summarize lengthy judgments, retrieve relevant case laws, and even detect fake legal news.

That's why we built **NyayaMitra**, an AI-powered legal assistant designed to make legal information more **accessible, accurate and efficient**. Our goal is simple to help lawyers, judges, and even common citizens get reliable legal information without unnecessary delays or confusion.

1.2 Scope of the Project

NyayaMitra is designed specifically for the **Indian legal system** and focuses on four key areas:

1. **Legal Chatbot** : An AI-powered assistant that answers queries related to Indian laws and legal cases.
2. **Fake News Detection in Legal Matters** : Identifying and flagging false or misleading legal news.
3. **Legal Document Summarization** : Summarizing the long judgments in order to provide meaningful insights
4. **Prior Case Retrieval (Precedent Search)** : Uses the Retrieval-Augmented Generation (RAG) to find the previous relevant cases.

The model will be trained on **Indian court judgments**, to adapt for the Indian Legal Domain.

1.3 Challenges in Indian Legal Research

As the Technology advances but, legal research in India faces several hurdles:

1. **Complex and Unstructured Legal Texts** : Indian Court judgments are lengthy, and difficult to analyze.
2. **Lack of Structured Legal Databases** : When we compare with Western legal systems, India does not have a organized legal database, making case retrieval inefficient.
3. **Difficulty in Finding Relevant Precedents** : Indian courts follow the principle that past judgments shape future rulings. However, manually searching for the most relevant precedent is time consuming.
4. **Legal Misinformation** : Misleading interpretations of Supreme Court and High Court judgments, viral WhatsApp forwards, and fake legal news create confusion among the public and legal professionals.

1.4 Why AI is Needed in Indian Legal Research

To overcome these challenges, **Artificial Intelligence (AI) and Natural Language Processing (NLP)** can play a transformative role:

- **AI models fine-tuned for legal texts** (like Phi-3 Mini) can extract essential insights from judgments and legal documents.
- **Retrieval-Augmented Generation (RAG)** can improve **case precedent retrieval**, making legal research faster and more precise.
- **Fake news detection models** trained on legal datasets can help **identify misinformation in legal news**.
- **Advanced text summarization models** (using BART, T5, or GPT-based architectures) can simplify complex legal documents, making them easier to understand.

By integrating these AI-driven techniques, **NyayaMitra** aims to bridge the gap between legal data and actionable insights, making legal research faster, more accurate, and more accessible for everyone in the Indian legal system.

CHAPTER 2

LITERATURE REVIEW

Artificial Intelligence (AI) and Natural Language Processing (NLP) are transforming legal research and judicial decision-making worldwide. In India, where the legal system is vast and complex, AI can play a major role in automating case retrieval, summarizing lengthy legal documents, and detecting fake legal news. This section explores existing research on AI applications in the legal domain, Retrieval-Augmented Generation (RAG)-based case search, legal document summarization, and fake legal news detection.

2.1 AI in Legal Research and Case Analysis

Legal professionals often spend a lot of time searching for old case laws, reading complicated legal texts, and drafting arguments based on past judgments. Traditional legal research is quite **manual and time consuming**, usually relying on **keyword based searches in online databases**. But the problem is, these searches don't always give the most relevant results.

Recently, there have been some great advancements in **Legal Natural Language Processing (NLP)** models, which have really made it easier to find case laws. Models like **CaseLawBERT** and **ECHR-BERT**, which have been trained on legal data, have set new standards for understanding legal texts. However, these models are mostly trained on **Western legal datasets**, so they aren't as good when it comes to understanding **Indian laws and court judgments**.

A study by **Chalkidis et al. (2021)** [1] showed that training AI models on **domain specific legal data** can really improve how well they classify and retrieve legal texts. Since **Indian case law** is huge and complex, an AI system that's specifically trained on **Indian Supreme Court** and **High Court judgments** could make legal research way faster and more accurate. **NyayaMitra** is trying to do this by fine-tuning **Phi-3 Mini** on **Indian legal texts**, improving both the speed of case retrieval and the understanding of context.

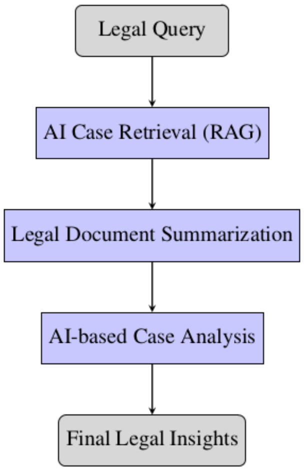


Figure 2.1: AI-Powered Legal Research Workflow

2.2 Retrieval Augmented Generation (RAG) for Case Precedent Search

One of the most important tasks in legal research is finding past cases (also known as precedents) to back up legal arguments. But the problem is, most legal search engines today rely on Boolean searches and keywords, which often lead to results that are either irrelevant or incomplete.

The Retrieval Augmented Generation (RAG) system is gaining attention for its ability to improve legal research and case retrieval. RAG consists of two major modules:

A retriever that searches for relevant legal documents. A generator that gets the key information from the legal documents and provides a summary. Studies have shown that RAG is more effective than traditional search methods, especially in law and technology.

There is also other useful approach that is **vector based legal search**, explored by **Wu et al. (2022)** [2]. in their research they found that search engines using vector databases like FAISS and Milvus will provide more accurate and relevant results. so Similarly our project the NyayaMitra also plans to use Milvus to store and retrieve legal cases efficiently. This will make it easier for lawyers, judges etc to find past cases based on legal reasoning and context.

2.3 Legal Document Summarization

Legal documents are usually long and difficult to understand because they use legal terminology which was not known to common people.so reading the leagal documents will take lot of time. AI can help by creating short and simple summaries, making the information easier to understand. This saves time and helps people quickly grasp the main points without reading everything.

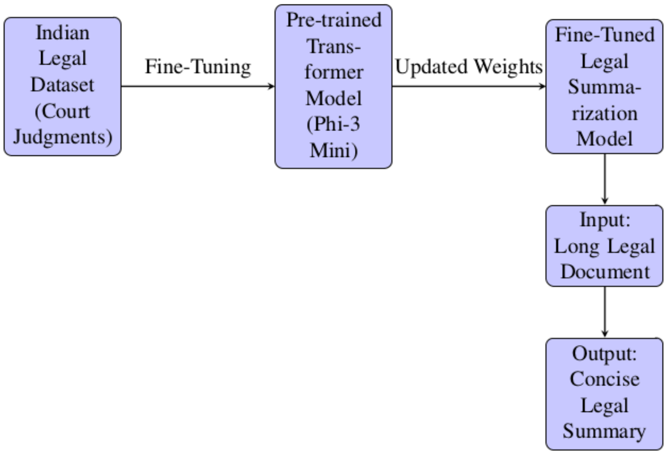

Indian Legal Dataset (Court Judgments)
Fine-Tuning
Pre-trained Transformer Model (Phi-3 Mini)
Updated Weights
Fine-Tuned Legal Summarization Model
Input: Long Legal Document
Output: Concise Legal Summary

Figure 2.2: Fine-Tuning a Transformer for Legal Document Summarization

There are two main types of legal summarization methods:

- **Extractive Summarization**: This method picks out the most important sentences directly from the document (e.g., TextRank, LexRank).
- **Abstractive Summarization**: This method creates a shorter summary in natural language (e.g., BART, T5, PEGASUS).

A study by Shen et al. (2021) [3] found that transformer-based models like BART and T5 work really well for summarizing legal texts when they're trained on legal data. The only issue is that these models need large labeled datasets, which are hard to find for Indian legal research.

To solve this problem, **NyayaMitra** will fine-tune an open-source model, **Phi-3 Mini**, on Indian legal documents. This will help the model create better summaries of court judgments while also understanding the legal context.

2.4 Fake News Detection in the Legal Domain

Fake legal news is a serious issue as it can mislead the public, influence legal opinions, and even spark protests based on misinformation. Most fake news detection models today rely on supervised classification techniques using LSTMs, transformers, and knowledge graphs.

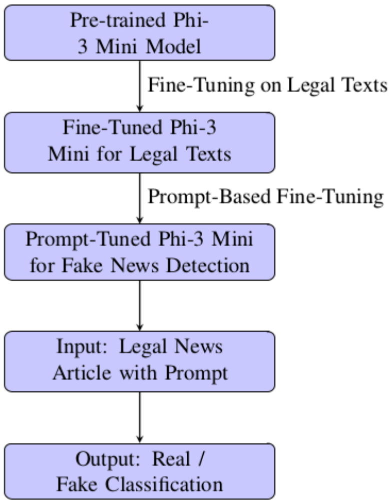

Pre-trained Phi-3 Mini Model
Fine-Tuning on Legal Texts
Fine-Tuned Phi-3 Mini for Legal Texts
Prompt-Based Fine-Tuning
Prompt-Tuned Phi-3 Mini for Fake News Detection
Input: Legal News Article with Prompt
Output: Real / Fake Classification

Figure 2.3: Direct Prompt Fine-Tuning for Fake Legal News Detection

A study by Zhou & Zafarani (2020) [4] found that graph-based models are particularly effective in detecting misinformation by analyzing how fake news spreads on social media. However, most of these models focus on general news, and there is very little research on detecting fake legal news.

Recent research has also looked into prompt engineering for misinformation detection. Studies have shown that fine-tuned transformers like GPT and BERT can detect fake news through zero-shot and few-shot learning techniques.

NyayaMitra will use prompt-based fine-tuning to classify fake and misleading legal news articles, enhancing accuracy in identifying misinformation specifically within the legal domain.

2.5 Summary and Research Gaps

There has been significant progress in AI-driven legal research, case retrieval, summarization, and fake news detection. However, most studies have focused on Western legal systems, leaving a major gap in AI applications for Indian law.

Research Gap	**NyayaMitra's Approach**
No AI models trained on Indian legal data	Fine-tuning Phi-3 Mini on Indian Supreme Court & High Court judgments
Inefficient legal case search	RAG-based retrieval with Milvus vector database
Limited work on fake legal news detection	Prompt-based fine-tuning for misinformation detection
Complex legal documents hard to summarize	Transformer-based summarization fine-tuned on Indian legal texts

Table 2.1: Research Gaps and NyayaMitra's Solutions

By addressing these gaps, NyayaMitra aims to transform legal research in India, making legal information faster, more accurate, and accessible to everyone.

CHAPTER 3

Work Done

3.1 Data Collection and Preprocessing

3.1.1 Data Collection through Web Scraping

Data collection is a crucial step in building a robust legal chatbot. In this project, legal judgments and court orders were gathered through web scraping. Since these documents are primarily available in PDF format, extracting structured data required specialized tools.

Court	Number of Cases Collected
Supreme Court of India [5]	27,517
Calcutta High Court [6]	17,652
Delhi High Court [7]	20,637
Madras High Court [8]	101,063
Patna High Court [9]	33,978
Total legal documents collected	**200,847**

Table 3.1: Collected Case Judgments from Indian Courts

3.1.2 Data Preprocessing

These are the major steps followed for preprocessing:

- **Separating Judgments and Orders**: The scraped data contained both judgments and orders, which needed to be classified appropriately. Different courts follow varied styles of writing judgments, prioritizing the use of multiple classification techniques for proper categorization.
- **Extracting Data from PDFs**: Various tools were evaluated for extracting text from PDFs, including `pymupdf` and `pdfplumber`. After thorough comparison, `pymupdf` proved to be the most effective.
- **Comparison of Extraction Tools**: Below is a comparison graph between `pymupdf` and `pdfplumber`.
- **Converting Extracted Data into CSV Files**: Once the text was extracted, it was converted into structured CSV files for further processing.

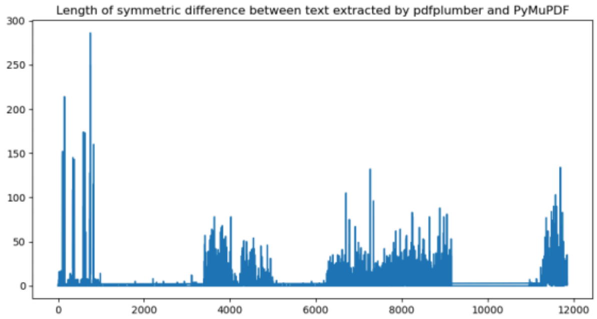


Figure 3.1: Comparison of `pymupdf` and `pdfplumber` extraction accuracy.

3.2 Model Development

The initial step in model development is to choose an appropriate model for the problem statement.

3.2.1 Choosing an Open-Source Development Model

To train the chatbot, an open-source model called Phi-3 Mini was selected. This model has 4 billion parameters and allows for customization and adaptation to new data, enabling it to be claimed as a proprietary model after fine-tuning.

3.2.2 Legal Chatbot

The legal chatbot was developed and fine-tuned using multiple approaches:

- **Fine-Tuning Techniques**: Different fine-tuning methods were explored, including full fine-tuning, parameter-efficient fine-tuning (PEFT), LoRA (Low-Rank Adaptation), and QLoRA (Quantized LoRA). Since GPU resources were available, LoRA was chosen as the preferred method.
- **Training Approach**: In LoRA, only the attention layers were fine-tuned. This approach effectively adds an extra layer to the pre-existing ones without disrupting the model's core architecture.
- **Training Dataset**: The model was trained using approximately 200,000 legal judgments, with 10% of the dataset reserved for evaluation.

- **Training Time and Epochs**: The training process spanned approximately 30 hours and was conducted over three epochs.
- **Performance Metrics**: Several graphs were plotted to assess model performance, including:
  - Training loss over epochs
  - Mean token accuracy vs. epochs
  - Loss vs. learning rate

Below are the graphs showcasing these performance metrics:

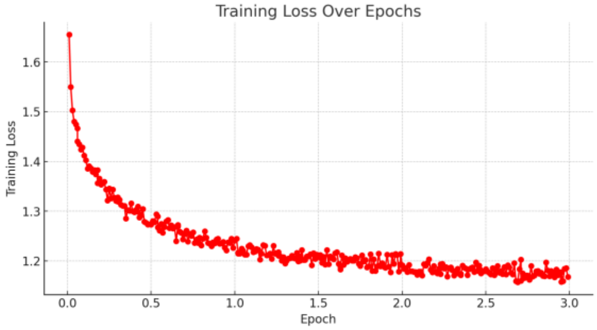


Figure 3.2: Training Loss Over Epochs

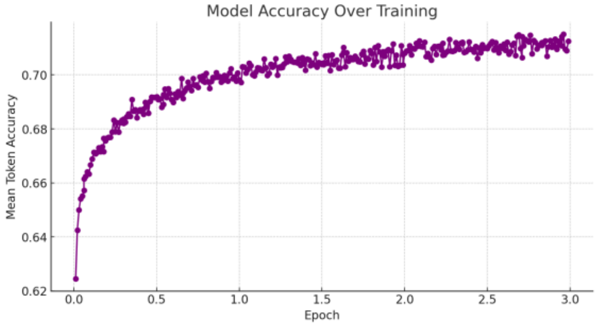


Figure 3.3: Mean Token Accuracy vs Epochs

This trained model was subsequently integrated into the legal chatbot system. The working prototype of it:

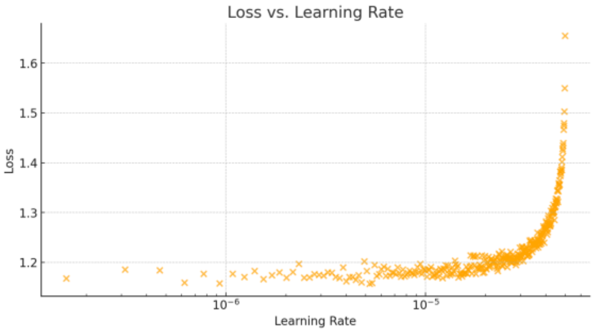
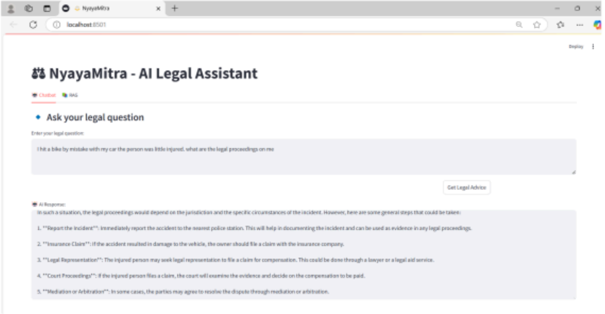


Figure 3.4: Loss vs Learning Rate



NyayaMitra - AI Legal Assistant

Ask your legal question

Enter your legal question:

I hit a bike by mistake with my car the person was little injured. what are the legal proceedings on me

AI Response:

In such a situation, the legal proceedings would depend on the jurisdiction and the specific circumstances of the incident. However, here are some general steps that could be taken:

1. **Report the Incident**: Immediately report the accident to the nearest police station. This will help in documenting the incident and can be used as evidence in any legal proceedings.

2. **Insurance Claim**: If the accident resulted in damage to the vehicle, the owner should file a claim with the insurance company.

3. **Legal Representation**: The injured person may seek legal representation to file a claim for compensation. This could be done through a lawyer or a legal aid service.

4. **Court Proceedings**: If the injured person files a claim, the court will examine the evidence and decide on the compensation to be paid.

5. **Mediation or Arbitration**: In some cases, the parties may agree to resolve the dispute through mediation or arbitration.

Figure 3.5: Working prototype of chatbot

3.3 Retrieval-Augmented Generation (RAG) for Case Retrieval

To improve **legal case retrieval**, the **RAG framework** was implemented.

- **Fine-Tuning for Retrieval:**
  - The model was **trained on 50,000 legal judgments** using **Phi-3 Mini** to adapt it to Indian legal terminology and case structures.
- **Embedding and Vector Storage:**
  - All legal documents were converted into **vector embeddings** using Phi-3 Mini.
  - These embeddings were stored in **Milvus (vector database)** for efficient retrieval.
- **Retrieval Mechanism:**

1. **Query Input:** A user enters a legal query.
2. **Embedding Conversion:** The query is converted into a vector representation.
3. **Vector Matching:** The system searches the Milvus database for the most relevant case precedents.
4. **Final Output:** The retrieved documents are passed through the fine-tuned model for contextual refinement and answer generation.

- **Why only this type of RAG?**
  - Inspired by the research paper **"Searching for Best Practices in Retrieval-Augmented Generation"**[10].
  - RAG improves accuracy and contextual understanding compared to traditional keyword-based search methods.

3.4 Evaluation Results

Metric	Score
Perplexity (PPL)	5.32
BLEU Score	47.8
ROUGE-1	52.4
ROUGE-2	39.2
Mean Token Accuracy	87.3%
Precision@5	80.4%
Recall@5	84.5%
Hit Rate@5	92.4%

Table 3.2: Evaluation Metrics for the Legal Chatbot and Retrieval System

3.5 Testing and Evaluation

Testing included human evaluation of chatbot accuracy, retrieval precision for case law search benchmarks.

Component	Evaluation Metric	Performance
Legal Chatbot	Response Accuracy (Human Evaluation) Legal Relevance	85% High
RAG-Based Case Retrieval	Precision@5 Recall@5	80.4% 84.5%

Table 3.3: Testing and Evaluation Metrics

CHAPTER 4

Conclusion and Future Scope

4.1 Conclusion

The NyayaMitra project is set to change the way legal research and case analysis are done in the Indian judicial system by using the power of AI. Over the past seven weeks, a dataset of 200,847 legal case has been gathered from judgments from the Supreme Court and various High Courts through web scraping. Legal texts were then carefully cleaned, pre-processed, and embedded using Phi-3 Mini, which helps the system better understand the cases. A RAG-based case retrieval system built on Milvus has also been developed, boosting the accuracy of finding relevant past cases. So, Phi-3 Mini has been trained to summarize those never-ending legal judgments, making them crisp and easy to understand while still keeping the important points. On top of that, the team has even built a legal chatbot that can answer law-related questions and quickly fetch past cases. Early tests show it's doing a pretty solid job accurate retrieval, good summaries, and smooth chatbot performance. Going ahead, they're planning to tackle legal misinformation by creating an AI system that can detect fake or misleading legal news, ensuring people get only verified and reliable info.

4.2 Future Scope

The next phase of NyayaMitra will be focusing on improvements, testing, and deployment. Our key priorities include:

4.2.1 Enhancing Case Retrieval

- Fine-tuning our search algorithms to improve accuracy and ranking of case precedents.

- Implementing multi-document retrieval, enabling users to access multiple related case precedents instead of just one.

4.2.2 Developing Legal Summarization

- Developing abstractive summarization with reinforcement learning to improve coherence and readability.
- Integrating citation tracking to maintain references to critical legal points in summaries.

4.2.3 Developing a Fake Legal News Detection System

- Building a dataset of real and fake legal news sourced from credible platforms.
- Fine-tuning Phi-3 Mini using prompt-based tuning to detect misinformation.
- Implementing a classification model using BERT and LSTMs to distinguish between genuine and fake legal news.
- Developing a real-time fake news verification API to help users validate the credibility of legal news.

4.2.4 Expanding Chatbot Capabilities

- Improving chatbot responses by enhancing contextual understanding for better legal explanations.
- Making user interactions more intuitive and structured for a smoother experience.

4.2.5 Frontend Development & System Integration

- We will make a simple and easy-to-use website where people can search for legal cases, chat with a legal AI, get case summaries, and check for fake news.
- Everything will be connected smoothly using APIs so that all features work together properly.
- The main goal is to make sure the website is user-friendly and useful for anyone looking for legal information.

4.2.6 Final Testing and Deployment

- Conducting comprehensive testing with legal professionals and students to ensure reliability..
- Deploying NyayaMitra as a fully functional AI-powered legal assistant, ready for use by legal professionals and researchers.

4.2.7 Agentic RAG System

- Introducing an Agentic RAG (Retrieval-Augmented Generation) system to improve the quality and relevance of responses generated by the AI. This system will leverage external knowledge from relevant legal sources and integrate it into the model's decision-making process.
- The Agentic RAG system will be designed to intelligently select the most relevant documents during the retrieval phase, enhancing the quality of generated responses for both the legal chatbot and case retrieval features.
- With the integration of the Agentic RAG system, NyayaMitra will be able to provide even more accurate and contextually aware legal research assistance, making the AI system more dynamic and responsive to complex legal queries.

4.3 Summary

NyayaMitra is well on its way to becoming a powerful AI-driven legal assistant, designed to simplify legal research, enhance case retrieval, improve judgment summarization, and tackle misinformation in the legal space. Over the next few weeks, we'll focus on refining our models, improving the user experience, and ensuring seamless deployment.

Srinivas Report.pdf

ORIGINALITY REPORT

7%	5%	1%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Indian Institute of Information Technology, Design and Manufacturing - Kancheepuram Student Paper	2%
2	idoc.pub Internet Source	1%
3	Submitted to University of Strathclyde Student Paper	1%
4	www.mdpi.com Internet Source	1%
5	Submitted to Liverpool John Moores University Student Paper	1%
6	leanpub.com Internet Source	<1%
7	dspace.mist.ac.bd Internet Source	<1%
8	kiaraygtn130625.articlesblogger.com Internet Source	<1%
9	dev.to Internet Source	<1%
10	fastercapital.com Internet Source	<1%
	www.elmbs.com	

11	Internet Source	<1 %
12	aaltodoc.aalto.fi Internet Source	<1 %
13	ijream.org Internet Source	<1 %
14	lawcirca.com Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		