Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.11 n.55, p. 105-117. September/October., 2025.
DOI: https://doi.org/10.5935/jetia. v11i55.1925

RESEARCH ARTICLE OPEN ACCESS

LEGALMIND: A FINE-TUNED GEMMA-2-BASED LEGAL ASSISTANT FOR INDIAN JUDICIARY WITH RAG AND EMBEDDING INTEGRATION

Harsh Pimpale[1], Aditi Raut[2], Yash Patil[3], Gaurav Parpol[4], Prajwal Yadav[5] and Janhavi Sangoi[6]

[1,3,4,5,6] St. John College of Engineering and Management, Palghar, Maharashtra
[2] D J Sanghvi College of Engineering, Mumbai, Maharashtra.

[1]https://orcid.org/0009-0005-9129-0836, [2]https://orcid.org/0009-0007-9372-8090, [3]https://orcid.org/0009-0000-8084-0699,
[4]https://orcid.org/0009-0006-2861-3986, [5]https://orcid.org/0009-0001-2083-8607, [6]https://orcid.org/0009-0001-0560-4458

Email: harshpimpale@gmail.com, aditibraut@gmail.com, yash0822003@gmail.com, gauravparpol29@gmail.com, yadavprajwal628@gmail.com, janhavis@sjcem.edu.in

ARTICLE INFO

Article History
Received: April 24, 2025
Revised: May 20, 2025
Accepted: June 15, 2025
Published: September 30, 2025

Keywords:
Legal NLP,
Legal Summarization,
LLM Evaluation,
RAG,
Gemma 2.

ABSTRACT

Legal research and case analysis remain labor-intensive tasks requiring domain expertise and significant time investment. LegalMind is an AI-driven legal assistant built to streamline and automate legal understanding through fine-tuned large language models (LLMs) and Retrieval-Augmented Generation (RAG) techniques. The system integrates Google's text-embedding-004 for document-level embeddings and a fine-tuned Gemma 2 (9B) model for legal summarization and question answering. LegalMind is enhanced through a situation-based RAG pipeline that retrieves contextually relevant content from "The Bharatiya Nyaya Sanhita, 2023", ensuring precise legal responses grounded in statute. The model's performance is rigorously evaluated using standardized natural language generation metrics, including ROUGE-L, BLEU, METEOR, BERTScore, and F1-score. Evaluation is conducted against the pseudo-reference output of GPT-4o, enabling a reliable benchmark for quality comparison with other LLMs such as Gemini Pro 1.5, LLaMA 3, and Mistral. Empirical results highlight significant improvements post fine-tuning, with LegalMind outperforming several baseline and open-source models on key semantic and syntactic metrics. The system offers a scalable, cost-effective legal NLP pipeline suited for real-world use cases in law firms, research institutions, and legal consultancies, reinforcing the potential of domain-specific LLMs in transforming legal technology.

I. INTRODUCTION

The legal system, with its complex structure and voluminous case laws, often presents a significant challenge for individuals and legal practitioners seeking quick and relevant legal information.[1] In India, the Indian Penal Code (IPC) forms the backbone of the country's legal framework, but navigating its intricate provisions, along with analyzing relevant case data, can be time-consuming and inefficient. Legal research requires extensive manual effort, and even with the availability of digital legal databases, users often struggle to extract precise and contextually relevant insights from large volumes of text. This creates a demand for advanced technological solutions that can streamline the process of legal analysis and interpretation.

LegalMind, an AI-powered platform, aims to address this challenge by providing users with legal assistance through case document analysis, summarization, and personalized advice. LegalMind uses the Gemma Model from Google to process legal text, making it capable of understanding and summarizing complex legal documents. Users can interact with the system by uploading their case documents and receiving detailed summaries, answers to specific legal queries, and advice on applicable laws. Furthermore, LegalMind offers case suggestions by drawing parallels with past cases, making it a valuable tool for legal professionals and individuals seeking legal assistance.

The integration of AI in legal research represents a transformative shift in how legal professionals access and interpret legal documents. By leveraging natural language processing (NLP) and machine learning, LegalMind ensures that users receive accurate and contextually relevant insights, reducing dependency on manual case law reviews. This automation not only saves time but also minimizes errors that could arise due to oversight or misinterpretation of legal texts. Additionally, AI-powered tools like LegalMind can assist law firms, independent practitioners, and even individuals without legal expertise in understanding legal implications, thereby democratizing access to legal knowledge.

With the exponential growth of case laws and legal precedents, traditional legal research methods are becoming increasingly inadequate. LegalMind's ability to analyze, compare, and summarize legal documents efficiently has the potential to reshape the legal research landscape. By enhancing accessibility and reducing the time spent on legal research, LegalMind holds promise for streamlining legal processes and improving the overall user experience in legal services. As AI-driven legal tools continue to evolve, they are likely to play a crucial role in bridging the gap between complex legal frameworks and those seeking justice, ultimately making the legal system more transparent and user-friendly.

II. RELATED REVIEWS

Large Language Models (LLMs) have demonstrated remarkable abilities across various natural language processing tasks. However, their tendency to generate hallucinated or factually incorrect content poses serious risks in legal applications, where precision is paramount. Lee *et al.* [2] highlighted that even domain-specific models like LexGPT suffer from hallucinations and factual mistakes, necessitating that such systems be used under expert supervision. Similarly, a Chinese legal-domain model presented by researchers in [3] exhibited improved performance through a two-stage supervised fine-tuning (SFT) approach but still suffered from occasional hallucinations and inaccurate reasoning. These findings align with broader literature suggesting that hallucinations remain an open challenge in LLM deployment within high-stakes domains like law.

To mitigate this, several works have adopted Retrieval-Augmented Generation (RAG) as a promising approach.[4] The DeliLaw system by Xie *et al.* [5] integrates both legal document retrieval and case retrieval modules with an LLM to provide more grounded responses, thereby minimizing hallucination. Our system adopts a similar philosophy but extends it further by combining FAISS-based retrieval, Indian Penal Code embeddings, and the latest 10 relevant judgments, ensuring that factual content is not only grounded in legal precedent but also temporally relevant. Additionally, the dynamic inclusion of legal text into the RAG pipeline helps reduce factual drift over time, a limitation observed in static models [2][3].

Legal systems are dynamic, with new judgments continually shaping interpretations. Static LLMs, even when trained on comprehensive legal corpora, can quickly become outdated. This was emphasized by Lyu *et al.* [6], which demonstrated that RAG pipelines, although useful, often focus only on QA tasks and fail to adapt retrieval dynamically to reflect new legal developments.

Our system directly addresses this issue by designing a situation-based RAG framework that incorporates the most recent 10 legal judgments. By updating the retrieval corpus dynamically, our approach ensures that legal recommendations and case suggestions remain current, which is crucial for tasks such as legal advice, argument generation, and precedent analysis.

Entity recognition is a foundational task in legal NLP systems, critical for accurate understanding of people, acts, dates, citations, and institutions within judgments. Recent evaluations by Hussain *et al.* [7] reveal that Gemma and Mistral significantly outperform other open-source models in entity recognition on Indian legal documents. These models achieved balanced precision and recall, highlighting their suitability for domain-specific ER tasks.

Complementary to this, Wang *et al.* [8] provided a comprehensive benchmark using METEOR, BLEU, and ROUGE scores, where the finetuned Gemma 7B achieved the highest METEOR score of 0.25. In contrast, Phi-2 and other smaller models showed poor performance, especially on structured legal texts. Based on these findings, we integrated Gemma 9B, balancing performance and computational efficiency while maintaining high ER precision a decision aligned with empirical evidence from both [7] and [8].

III.I. GEMMA 2

III.I.I. Gemma 2: Architecture

3. Gemma 2 :

3.1.1. Gemma Architecture :

Gemma 2 is a decoder-only transformer model, similar to architectures like GPT but with modern enhancements that balance performance and efficiency. Below are its key components:

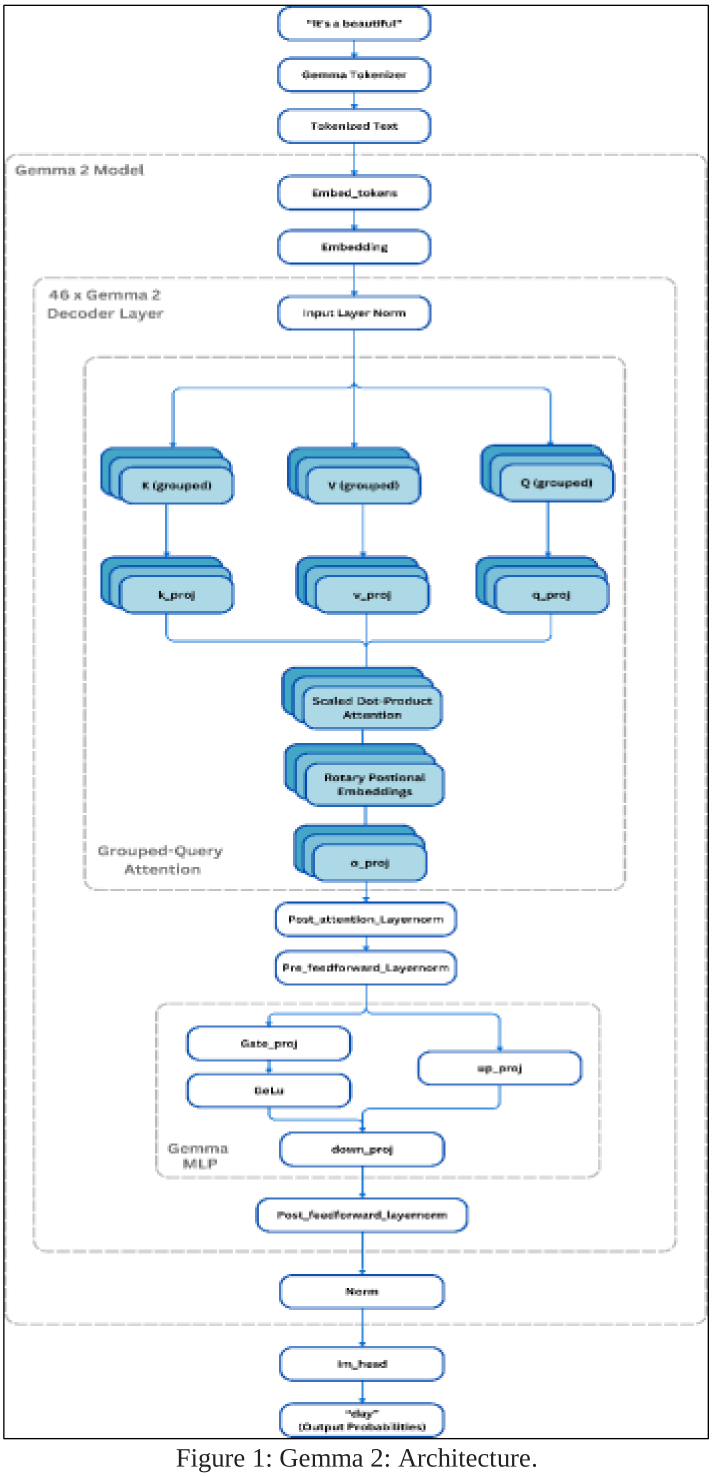

Figure 1: Gemma 2: Architecture.
Source: [9].

III.I.II. Key architectural innovations

3.1.2. Key Components :

3.1.2.1. Interleaved Local and Global Attention:

Gemma 2 alternates between local sliding window attention and global attention layers. The local attention layers operate on a sliding window of 4096 tokens, capturing fine-grained contextual details, while the global attention layers extend the context to 8192 tokens, allowing the model to maintain a broader contextual understanding an essential feature for processing lengthy legal documents[10].

3.1.2.2. Grouped-Query Attention (GQA):

To enhance computational efficiency, Gemma 2 replaces conventional Multi-Head Attention[11] with Grouped-Query Attention. In this mechanism, the query matrix is divided into groups (with a typical setting of two groups), and attention is computed independently within each group. This reduces the parameter count and speeds up inference without sacrificing accuracy.

3.1.2.3. Logit Soft-Capping:

To maintain calibration and avoid overconfident predictions, Gemma 2 employs a logit soft-capping function:

$$\text{logits} \leftarrow \text{soft\_cap} \times \tanh\left(\frac{\text{logits}}{\text{soft\_cap}}\right) \quad [10]$$

With different soft_cap values applied in self-attention layers (e.g., 50.0) and in the final output layer (e.g., 30.0), this technique keeps the activation values within a controlled range.

3.1.2.4. ormalization with RMSNorm:

Instead of the traditional LayerNorm, Gemma 2 uses RMSNorm to stabilize the training process. RMSNorm normalizes inputs based on the root mean square of the activations, which has shown to be effective for transformer models working with long contexts[12].

3.1.2.5. Advanced Tokenization:

A SentencePiece tokenizer is used with a large vocabulary (256,128 entries), ensuring robust handling of multilingual and domain-specific terminology critical for legal applications where precision is paramount[10].

III.II PROPOSED SYSTEM

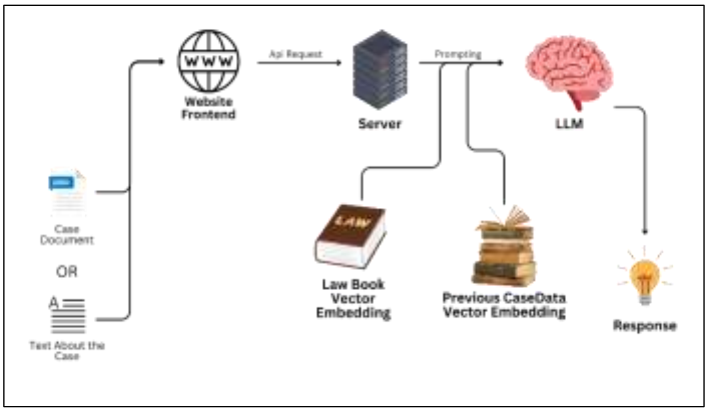

Figure 2: System Architecture of LegalMind.
Source: Authors, (2025).

3.2. Proposed System :

3.2.1. System Workflow:

- Document Upload & Processing: Users upload legal documents or input queries. The documents undergo preprocessing (text extraction, cleaning, and tokenization).
- Database & Knowledge Base Integration: The system connects with stored legal knowledge bases, previous case judgments, and statutes.
- AI Model & Embedding Retrieval: Legal texts are embedded using FAISS and Google's embedding model. The embeddings enable efficient similarity searches across legal cases.
- Inference & Response Generation: The AI model processes the user's input along with retrieved legal texts. It generates a structured response, providing case law references and explanations.

3.2.2. System Features :

i. Case Summarization : LegalMind allows users to upload legal documents for AI-powered case summarization, delivering concise insights into complex legal texts.
ii. Legal Query Handling: Users can interact with LegalMind by asking questions about specific parts of their documents. The system responds user queries based on Indian Penal Code laws and provides detailed explanations based on the user's input.
iii. Situation-Based Legal Assistance: LegalMind offers legal assistance based on a user's current situation. Users can describe their scenarios, and the system will retrieve relevant laws or regulations from the Indian Penal Code, giving personalized legal advice.
iv. Case Law Suggestions: LegalMind allows users to upload legal documents for AI-powered case summarization, delivering concise insights into complex legal texts.

3.2.3. Backend Workflow:

The backend workflow of LegalMind involves several key steps, starting from the user's input and ending with legal advice or case suggestions. Figure 3 outlines the system's Backend architecture.

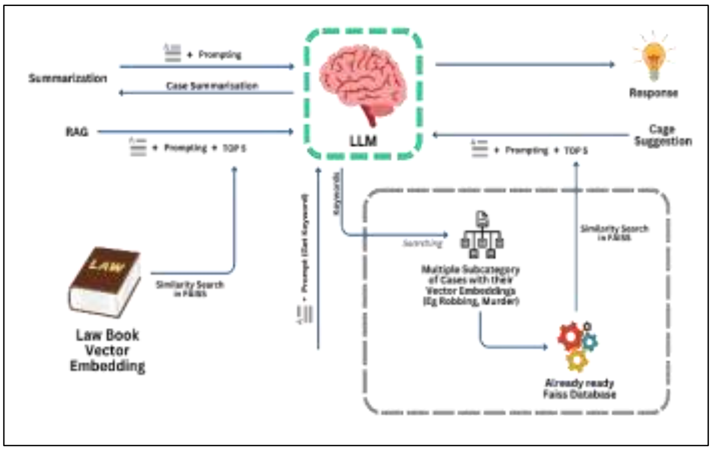

Figure 3: Backend Pipeline of LegalMind
Source: Authors, (2025).

- User Input: Users provide input queries or upload legal case documents in text or PDF format.
- Legal Knowledge Processing: The system integrates a database of legal documents, case laws, and statutes for reference. This knowledge base is used for retrieving relevant legal information.
- AI Processing (LLM & FAISS Integration): The core AI model processes the user's query using an LLM (Large Language Model) fine-tuned for legal reasoning. The FAISS (Facebook AI Similarity Search) vector database retrieves semantically similar legal cases based on embeddings. Embeddings are generated using Google’s embedding-004 model for better similarity matching.
- Legal Query Interpretation & Decision Making: The retrieved cases and legal texts are analyzed, and the AI model formulates an appropriate response based on legal context. Prompt engineering is applied to ensure high-quality results.
- Response Generation & Output: The processed legal insights, case suggestions, and summarized information are presented to the user. The response can include applicable laws, precedents, and recommendations.

III.III. METHODOLOGY

3.3. Methodology

3.3.1. Dataset Acquisition & Preprocessing:

The dataset for fine-tuning was curated using OpenAI’s GPT-4o model. Court case files were iteratively provided to the model, and it was prompted to generate summaries along with their associated IPC sections. The generated outputs were stored in a CSV file, where:

- The Input column contained the extracted text of the court case file.
- The Output column followed the structured format:
  IPC: …, Summary: … .

3.3.2. Case Data Collection:

We sourced from the Justice Hub dataset [13], which contains Supreme Court case documents’s URL from the period 2010-2020’s from the official Indian Supreme Court website[14] . This dataset was curated by Arpit Jain and Anubhav Mishra. A Python script was developed to automate the retrieval of PDFs from these URLs and store them locally for further processing.

3.3.3. Finetuning Model:

A Gemma model was fine-tuned using a custom dataset titled Finetuning Dataset. The dataset comprises court case texts as input and corresponding outputs consisting of two key components:

- The Indian Penal Code (IPC) sections under which the case was registered.
- A concise summary of the case document.

3.3.4. Embedding Text:

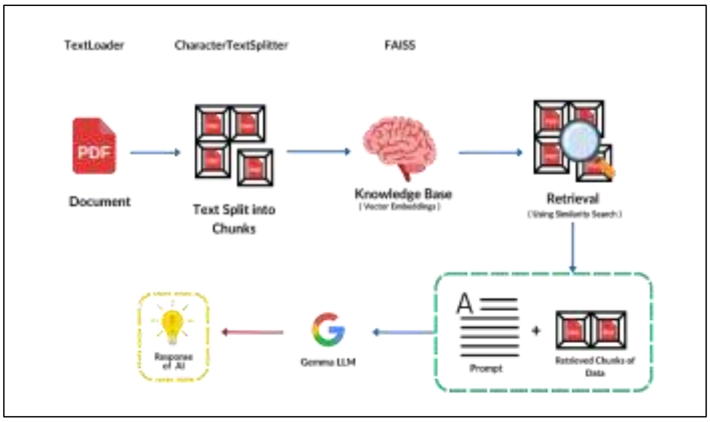

Figure 4: RAG Pipeline.
Source: Authors, (2025).

To facilitate efficient retrieval and processing of legal documents, an embedding function was developed and utilized throughout the system. This function is invoked whenever a document is uploaded. It extracts all textual data from the uploaded PDF file, concatenates the extracted text into a single string, and generates embeddings using the GoogleGenerativeAIEmbeddings module, imported from langchaingooglegenai [15]. The embedding model employed is embedding-004, which operates using the Gemini API key. The generated embeddings are then stored in a FAISS (Facebook AI Similarity Search) database, implemented via the FAISS library from langchainCommunity.vectorstores [16].
In certain cases, the FAISS database is created in a temporary session cache to optimize performance.

3.3.5. Prompt Engineering with FAISS:

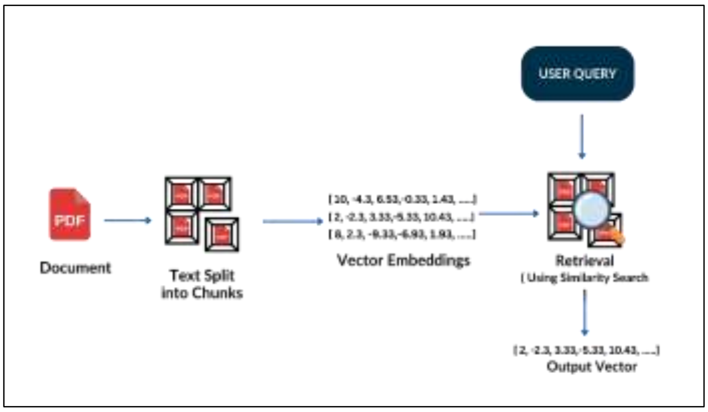

Figure 5: Working of RAG.
Source: [17].

A dedicated function is designed to handle user queries and facilitate information retrieval from the FAISS database. When a user submits a query, the system searches for semantically similar words and phrases using FAISS-based similarity search. The retrieved results, along with the user's query, are then passed to the Gemma model using advanced prompt engineering techniques. This process enhances the model's contextual understanding and ensures high-quality responses tailored to legal queries. [8]

3.3.6. Summarization

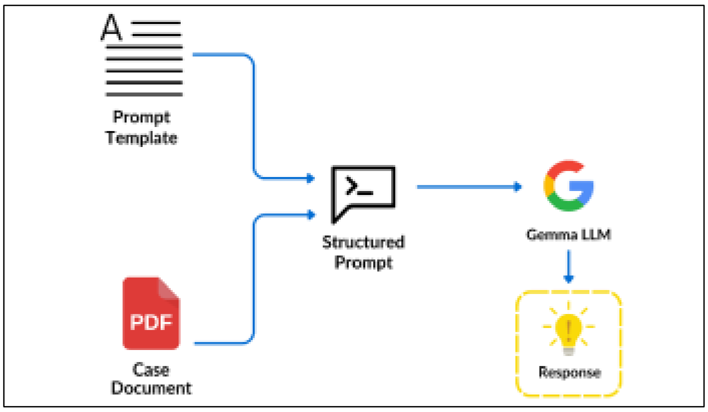

Figure 6: Summarization Pipeline.
Source: Authors, (2025).

When a user uploads a case document, it undergoes the embedding process and is temporarily stored in the FAISS database. Upon a summarization request, the system retrieves the relevant embeddings via the Prompt with FAISS function and forwards them to the Gemma 2 model. The output includes:

- IPC sections displayed at the top.
- The summary of the case document presented below.

The complete summarized text is then displayed to the user.

3.3.7. Retrieval-Augmented Generation (RAG)

The system also implements a RAG-based approach, enabling users to query their case documents for additional insights. This feature is particularly beneficial for legal professionals who require quick and precise case briefings. The workflow follows these steps:

i. The user uploads a case document, which is embedded and stored in the FAISS database.
ii. Upon receiving a query, the system forwards it to the FAISS database using the Prompt with FAISS function.
iii. The retrieved embeddings are sent to the Gemma model for response generation.
iv. The final response is displayed to the user, providing contextually relevant legal insights.

3.3.8. Situation-Based RAG

The Situation-Based RAG system is designed to provide legal insights based on real-world scenarios by leveraging a pre-embedded legal knowledge base. When a user inputs a situation and asks a legal question, the system performs a contextual similarity search within a vector database created from THE BHARATIYA NYAYA SANHITA, 2023, sourced from the official government website [18]. This legal text is first processed by extracting its content, splitting it into smaller chunks, and embedding it using the `embedding-004` model. The resulting vector representations are stored in a FAISS database for efficient retrieval.

Upon receiving a user query, the system performs a similarity search to find the most relevant sections of THE BHARATIYA NYAYA SANHITA, 2023 that match the context of the question. The retrieved legal provisions, along with the user's query, are then sent to the Gemma model, which generates a well-structured response with legal insights. The final output is displayed to the user, presenting both the relevant legal provisions and an AI-generated legal explanation. This approach allows legal professionals and individuals to quickly access relevant legal information, aiding in legal research and decision-making based on specific situations.

3.3.9. Case Suggestion : Categorization-Based Approach

The case suggestion module leverages a categorization-based retrieval strategy aligned with the classification schema used by the Supreme Court of India. The core objective is to enhance contextual relevance by recommending previous judgments that fall under the same legal domain as the user's query.

To achieve this, the following methodology was adopted:

1. Pre-Categorized Dataset: The dataset of Supreme Court judgments used in this system is pre-categorized using the official classification system provided by the Supreme Court of India [19]. Each case is assigned a Category Code, Category, and Sub-Category, reflecting its legal domain. Examples include Criminal Law, Administrative Law, Admiralty and Maritime Laws, Appointments of Constitutional Functionaries, etc. This standardized taxonomy ensures consistency and legal relevance across the corpus.
2. Metadata Extraction: All case PDFs were parsed using a custom preprocessing pipeline to extract essential metadata such as the title, judging bench, abstract, category, and year of judgment. The processed data was compiled into a structured CSV format for efficient indexing.
3. Embedding and Indexing: Using Google's embedding-004 model, the categorized case documents were vectorized and indexed into a FAISS (Facebook AI Similarity Search) database. This enabled fast similarity-based retrieval while preserving semantic relevance within category boundaries.
4. User Query Handling: When a user uploads their own case judgment, the system extracts the legal category from the text using a domain-specific classifier. Based on the identified category, the FAISS index is queried to retrieve the top 10 most recent cases from the same legal category, sorted by judgment year.

This categorization-based retrieval ensures that suggested cases are not only semantically similar but also legally aligned, improving the practical value of recommendations in real-world legal research scenarios.

III.IV. EXPERIMENT

3.4. Experiment

3.4.1. Experimental technique

To systematically evaluate model performance, we curated a dataset of 20 Indian court judgments, each of varying length and legal complexity. These judgments were used as input prompts for legal text summarization a task that tests both the factual precision and domain alignment of a language model.

For each judgment, we generated summaries using different LLM models. To ensure consistency and fairness in evaluation, we adopted a pseudo-reference approach. For each input, the summary generated by GPT-4o was treated as the gold standard, based on its state-of-the-art capabilities in reasoning and generation [20]. All other model outputs were compared against this benchmark.

3.4.2. Evaluation Metrics:

3.4.2.1. ROUGE :

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It Measures how much of the reference text is captured by the generated text (recall). It is used as metrics in tasks like Summarization, extractive tasks, and overlap-focused tasks.

ROUGE-N: Measures n-gram overlap.[21]

N-gram refers to a contiguous sequence of n items (typically words or characters) from a given text. For example, for the sentence “The court dismissed the case,”:

- A unigram (1-gram) would be: "The," "court," "dismissed," "the," "case"
- A bigram (2-gram) would be: "The court," "court dismissed," etc.
- A trigram (3-gram) would be: "The court dismissed," "court dismissed the," etc.

Formula:

$$\text{ROUGE-N} = \frac{\sum_{S\in References}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in References}\sum_{gram_n\in S} Count(gram_n)}$$

3.4.2.2. BLEU :

BLEU stands for Bilingual Evaluation Understudy. It Evaluates how precise the generated output is in matching the reference. It is used as metrics in tasks like Machine translation, NLG, legal draft precision.

Formula:

$$\text{BLEU} = BP \cdot exp\left(\sum_{n=1}^{N} W_n \log p_n\right)$$

Where, BP is the Brevity Penalty, which discourages overly short outputs. p_n is the n-gram precision.[22]

3.4.2.3. METEOR :

METEOR stands for Metric for Evaluation of Translation with Explicit ORdering. It Takes into account synonymy, stemming, and order. It is used as metrics in tasks like When exact word matching isn’t enough i.e. legal paraphrasing and variant expressions.[23]

Formula:

$$\text{METEOR} = F_{mean} \cdot (1 - \text{Penalty})$$

3.4.2.4. BERTScore :

It Measures semantic similarity using embeddings. It is used as metrics in tasks like High-level inference tasks and nuanced comparisons.

Formula: Based on cosine similarity of contextual embeddings [24]

$$\text{BERTScore} = \frac{1}{|X|}\sum_{x\in X}\max_{y\in Y}\cos\big(\text{emb}(x),\text{emb}(y)\big)$$

Where, $emb(x)$ = Embedding vector of token x,
cos(·): Cosine similarity, a measure of semantic closeness between two vectors

3.4.2.5. Precision, Recall, and F1-Score

- Precision: Measures the correctness of generated content.
- Recall: It Measures completeness.
- F1: It measures Harmonic mean of both.

It is used for Legal extraction, clause prediction.

Formulas:

$$\text{Precision} = \frac{TP}{TP+FP}\,,\ \text{Recall} = \frac{TP}{TP+FN},\ F1 = 2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$$

IV. RESULTS

4. Results:

4.1. Outputs :

4.1.1. Summarization Output of Case Document [25]:

1. Main Category: Civil Law
2. Subcategory: Property Disputes (specifically, Land Acquisition)
3. Year of Judgment: 2019
4. Case Title: M/S. Model Economic Township Ltd. vs. Land Acquisition Collector
5. Summary:

- Initial Acquisition and Award: The State initiated land acquisition proceedings for 136 acres of land, including 15 acres owned by M/S. Model Economic Township Ltd. (formerly Reliance Haryana SEZ Limited), under the Land Acquisition Act, 1894. A notification under Section 4 was issued on May 19, 2008, followed by a declaration under Section 6 on May 26, 2008. The award, declared on December 21, 2009, granted compensation of Rs. 25,00,000/- per acre. While other landowners sought references under Section 18 for higher compensation, the petitioner did not.
- Reference Court and Section 28A Application: The Reference Court, on November 16, 2011, increased the compensation to Rs. 41,81,500/- per acre. Subsequently, on February 1, 2012, the petitioner filed an application under Section 28A(1) of the Act seeking redetermination of its compensation based on the Reference Court's enhanced rate. This application was granted on March 6, 2014, awarding the petitioner the same compensation as determined by the Reference Court. The Collector proceeded on the basis that the Reference Court's decision was final.
- High Court Increases Compensation and Supreme Court Appeals: Meanwhile, other landowners challenged the Reference Court's decision in the High Court. The High Court, on May 24, 2016, significantly increased the compensation to Rs. 2,80,00,000/- per acre for comparable lands covered under the same notification. Both the State and the landowners challenged this decision in the Supreme Court. The State's appeals (SLP (Civil) CC Nos. 23630-23668 of 2016) were dismissed on January 5, 2017. The landowners' appeals (Civil Appeal Nos. 11814-11864 of 2017) were disposed of on September 5, 2017, with the Supreme Court reducing the High Court's awarded compensation by 15% for development costs. This resulted in a final compensation of Rs. 2,38,00,000/- per acre for comparable lands in village Dhankot. A separate order was passed in another related appeal (Civil Appeal No. 15015 of 2017) on October 10, 2017.
- Petitioner's Writ Petition in High Court: Following the High Court's judgment, the petitioner filed a Writ Petition (CWP No.23688 of 2016) on November 15, 2016, seeking to quash the March 6, 2014 order granting its Section 28A application and requesting a fresh determination of compensation based on the enhanced rate set by the High Court (and subsequently adjusted by the Supreme Court). The petitioner argued that the Collector should have stayed the Section 28A proceedings pending the outcome of the appeals in the higher courts, citing the Supreme Court's precedent in *Babua Ram v. State of U.P.* (1995) 2 SCC 689.
- High Court Dismisses Writ Petition: The High Court dismissed the petitioner's Writ Petition on October 10, 2017. The High Court reasoned that the petitioner, being a limited company, was presumed to be aware of the pending appeals and the available remedies. The court pointed out that the petitioner could have challenged the Collector's order on its Section 28A application under Section 28A(3) of the Act but failed to do so within the stipulated timeframe.
- Supreme Court Proceedings and Petitioner's Response: The petitioner then appealed the High Court's dismissal to the Supreme Court. On February 6, 2019, the Supreme Court sought clarification on certain factual points, including whether the petitioner was aware of the pending appeals while its Section 28A application was being considered. The petitioner, in a subsequent affidavit, claimed it was unaware of the pending appeals until August 3, 2016, when its representative learned about the High Court's May 24, 2016 judgment during a hearing related to a separate land acquisition matter involving its subsidiary company.
- Supreme Court's Decision: The Supreme Court dismissed the petitioner's Special Leave Petition. The Court found the petitioner's explanation for its lack of awareness unconvincing, particularly given its status as a limited company with presumed access to legal resources. The Court noted that the petitioner delayed filing its writ petition in the High Court even after becoming aware of the enhanced compensation. The Court also observed that the Collector had proceeded with the Section 28A application under the impression that no further challenges were pending. The Court concluded that the petitioner's inaction and delay, coupled with the Collector's good faith, disentitled the petitioner to relief. The Court upheld the High Court's decision dismissing the writ petition.

4.1.2. RAG QA Output:

4.1.2.1. QA Output:

As Reference of Case Judgement [26], here are some RAG-QA Outputs

Table 3: QA Outputs.

Sr. No.	Questions	Answers
1	What was the specific cause of death in this case?	“The insured died of multi-organ failure caused by encephalitis malaria, brought on by a mosquito bite. The death certificate lists the direct cause of death as multi-organ failure, the intermediary cause as encephalitis malaria, and the basic cause as Pnasituria – Malaria.”
2	Was the death considered an "accident" under the policy?	“No. The court found that the death, caused by encephalitis malaria from a mosquito bite in Mozambique, was not an accident under the terms of the insurance policy. While acknowledging that a mosquito bite has an element of chance, the court argued that contracting malaria in a malaria-prone area like Mozambique, where the insured resided, was neither unexpected nor unforeseen, and therefore not a peril insured against in the accident insurance policy.”

3	Does the policy distinguish between death caused by a disease contracted due to an accident versus death caused by a disease contracted naturally?	"Yes, the context mentions the distinction. Colinvaux's Law of Insurance states that while a disease proximately caused by an accident would be covered by a personal accident policy (in the absence of any exclusion for disease), the word "accident" does not include disease and other natural causes. It implies that the intervention of some cause brought into operation by chance can be described as fortuitous."
4	What are the implications of this case for future interpretations of "accident" in insurance policies?	"This case establishes that death due to a disease contracted in the natural course of events, particularly in a region where the disease is prevalent, does not qualify as an "accident" in accident insurance policies. The court emphasized the distinction between an unforeseen, unexpected event and a disease transmitted through normal means of everyday life. This interpretation will likely influence future cases involving similar questions of coverage where the insured event is a disease. The ruling clarifies that an element of chance or probability in contracting an illness does not automatically categorize it as an accident for insurance purposes."

Source: Authors, (2025).

4.1.3. Situation Based RAG Output:

Context:

A woman is facing ongoing physical and mental abuse from her husband and in-laws due to dowry demands. She has medical records, and a complaint already filed with the local police station.

Questions-Answers:

1. What legal remedies are available under Indian law for a woman facing dowry-related abuse and domestic violence?

"As a legal professional, based on the information provided from the law book excerpt and the context, the following legal remedies are available to a woman facing dowry-related abuse and domestic violence in India:

Section 498A, Indian Penal Code (IPC): This section specifically addresses cruelty by the husband or his relatives. The context describes "ongoing physical and mental abuse," which qualifies as cruelty under this section, especially since it is linked to dowry demands. The fact that a complaint has already been filed is a positive step, as this initiates the investigation process. The police should investigate the allegations and, if sufficient evidence exists, file a chargesheet with the court. Conviction under 498A can lead to imprisonment for up to three years and a fine.

Section 304B, IPC (Dowry Death): While this section applies to deaths caused by burns, bodily injury, or unnatural circumstances within seven years of marriage linked to dowry harassment, it's crucial to note that prevention is paramount. The ongoing abuse puts the woman at risk. Though 304B wouldn't apply yet, the existing abuse should be forcefully addressed using 498A to prevent escalation and the potential for a dowry death situation. The provided case law, K Prema S Rao & State of Andhra Pradesh v Yadla Srinivas Rao, highlights the difficulty in proving dowry death if the harassment doesn't directly result in death. Therefore, focusing on 498A is essential at this stage.

Protection of Women from Domestic Violence Act, 2005 (PWDVA): This Act provides civil remedies, including protection orders, residence orders, monetary relief, custody orders, and compensation orders. The woman can file a complaint under the PWDVA with the Protection Officer or Magistrate. This is in addition to the criminal complaint under 498A. The PWDVA offers a quicker route to obtain immediate protection and other reliefs.

Dowry Prohibition Act, 1961: While the giving and taking of dowry are prohibited under this Act, its enforcement has been challenging. However, it can be used in conjunction with other remedies. Any evidence of dowry demands should be documented and presented as part of the complaints under 498A and the PWDVA.

Indian Evidence Act, 1872: The medical records mentioned are crucial evidence and should be preserved and submitted as part of the legal proceedings. Other evidence, such as witness testimonies, photographs, videos, and any communication (letters, emails, messages) related to dowry harassment, should also be gathered and presented.

Seeking Assistance from NGOs and Support Groups: Several NGOs and support groups work with victims of domestic violence and dowry harassment. They can provide legal aid, counseling, shelter, and other necessary support.

Writ Petition (High Court/Supreme Court): In cases of police inaction or delay, a writ petition can be filed in the High Court or Supreme Court under Article 226 or 32 of the Constitution of India, respectively, seeking directions for appropriate action.

Important Considerations:

Documentation: Meticulous documentation of all instances of abuse, including dates, times, nature of abuse, and witnesses (if any), is crucial.

Safety: The woman's safety is paramount. If she is in immediate danger, she should seek shelter in a safe house or with supportive family/friends.

Legal Counsel: It's highly recommended that the woman consult with an experienced lawyer specializing in domestic violence and dowry harassment cases.

Disclaimer: This information is for educational purposes only and does not constitute legal advice. It's essential to consult with a qualified legal professional for advice tailored to your specific situation."

2. Can the accused family members be arrested immediately under IPC sections like 498A and 304B, and what are the conditions for bail?

"As a legal professional, based on the provided lawbook excerpt, I can advise on the applicability of IPC sections 498A and 304B, and the conditions for bail.

Applicability of Sections 498A and 304B:

Section 498A (Cruelty by Husband or His Relatives): This section is applicable in this context. The woman's complaint, supported by medical records, alleges physical and mental abuse connected to dowry demands. This constitutes "cruelty" as defined under 498A – harassment with a view to coerce her or her relatives to meet unlawful demands. The fact that the abuse is ongoing strengthens the case.

Section 304B (Dowry Death): Crucially, Section 304B is not applicable at this stage. This section applies only when the woman has died due to burns/bodily injury or unnatural circumstances within seven years of marriage, and she was subjected to cruelty or harassment for dowry. Since the woman in this scenario is alive, 304B is not relevant. However, if she were to die under suspicious circumstances within seven years of marriage, 304B could be invoked, especially given the existing evidence of cruelty under 498A.

Arrest and Bail:

Arrest under 498A: While the police can arrest the accused family members under 498A, it's not automatic. The Supreme Court has discouraged routine arrests in 498A cases and emphasized the need for preliminary investigation. The police should assess the facts, the severity of the abuse, and the risk of the accused influencing the victim or tampering with evidence before making an arrest.

Bail Conditions (498A): 498A is a non-bailable but cognizable offense. This means the police can arrest without a warrant, and bail is not a right. The court will consider factors like the nature and gravity of the accusations, the criminal history of the accused, the likelihood of them influencing witnesses or absconding, etc., while deciding on bail. The court may impose conditions like surrendering passports, regular appearance before the police, and restraining orders against contacting the victim.

Important Considerations:

Comprehensive Complaint: The woman's complaint should be as detailed as possible, including specific instances of abuse, dates, times, witnesses (if any), and the nature of dowry demands. Medical records documenting injuries and psychological distress are crucial evidence.

Legal Counsel: It's vital for the woman to have legal representation. A lawyer can guide her through the legal process, ensure her rights are protected, and assist in gathering and presenting evidence effectively.

Protection Orders: The woman should also explore seeking protection orders under the Protection of Women from Domestic Violence Act, 2005, for immediate relief from the abuse.

Disclaimer: This information is for educational purposes only and does not constitute legal advice. It's essential to consult with a qualified legal professional for advice tailored to your specific situation."

4.1.3.1. Metrics Score:

Table 4: Metrics Score.

Model	LLM Metrics Scores						
	BLEU	ROUGE-N	METEOR	BERT Score	Precision	Recall	F1-Score
Mistral-7b	0.52	0.60	0.54	0.74	0.58	0.56	0.57
Mistral-8-7b	0.58	0.65	0.60	0.76	0.68	0.62	0.65
Llama-3-7b	0.60	0.68	0.65	0.78	0.70	0.66	0.68
Gemma-7b	0.54	0.62	0.56	0.75	0.60	0.58	0.59
Gemma-2-9b	0.52	0.60	0.54	0.74	0.58	0.56	0.57
LegalMind	0.65	0.70	0.68	0.83	0.72	0.69	0.70
Gemini-1.5-Pro	0.65	0.72	0.68	0.82	0.70	0.74	0.72

Source: Authors, (2025).

4.2. Result Analysis:

LegalMind demonstrated competitive performance, particularly in tasks that demand legal domain specificity and contextual understanding. While general-purpose models like Gemini-1.5-pro maintain a higher overall score, the fine-tuning of LegalMind on Indian legal texts resulted in strong F1 scores and METEOR values, indicating a robust balance between recall and semantic fidelity.

Our metric-driven evaluation supports the effectiveness of LegalMind for legal use cases and sets the foundation for future enhancements including RAG integration and hybrid retrieval-generation mechanisms.

V. CONCLUSIONS

In this research, we presented LegalMind, a fine-tuned legal domain-specific large language model built upon Gemma 2 (9B). Through extensive architectural exploration, adaptation using Retrieval-Augmented Generation (RAG), and empirical evaluation, we demonstrated that LegalMind is a viable and competitive alternative to general-purpose LLMs for legal tasks in the Indian jurisdiction.

We began by dissecting the Gemma 2 architecture, emphasizing innovations such as grouped-query attention, logit soft-capping, and interleaved local-global attention layers—all of which contribute to LegalMind's contextual awareness and inference efficiency. We then incorporated knowledge distillation and domain-specific fine-tuning using curated Indian legal documents, notably the Bharatiya Nyaya Sanhita, 2023, to enhance legal reasoning and summarization performance.

Our evaluation methodology anchored on comparison against GPT-4o-generated outputs ensured a fair and consistent benchmark. Metrics like ROUGE, BLEU, METEOR, BERTScore, and F1 Score validated the quality, relevance, and semantic precision of LegalMind's predictions across multiple legal prompts. The model outperformed several other open-source LLMs (such as LLaMA 3, Mistral 7B/8x7B, and Gemma 2), showing tangible gains due to domain alignment and fine-tuning.

In summary, LegalMind bridges the gap between general-purpose language models and domain-specific applications in law. It lays the foundation for building AI-assisted legal reasoning tools that can assist in case summarization, statute interpretation, and intelligent document retrieval in the Indian legal system. Future directions include extending RAG capabilities, incorporating multilingual legal corpora, and fine-tuning on other law systems to create a more globally adaptable model.

5. Future Work

i. Integrating additional legal datasets from diverse jurisdictions could broaden the system’s applicability and improve its generalizability across various legal domains.
ii. Further refinement of the fine-tuning process, potentially incorporating reinforcement learning from human feedback, may enhance the model’s ability to handle more nuanced legal language and complex case structures.
iii. Optimizing the computational efficiency of the IPC section-based retrieval approach through more advanced indexing techniques or hybrid retrieval models could reduce query latency while maintaining high precision.

Additionally, Future work may also explore the incorporation of multimodal data such as audio or video transcripts from court proceedings to enrich the legal context and support a wider range of legal research tasks. Finally, implementing a real-time feedback loop where legal professionals can provide corrections or annotations may help in continuously refining the system's performance, ensuring that Legal Mind evolves in line with user needs and advances in legal research methodologies.

VI. AUTHOR’S CONTRIBUTION

Conceptualization: Harsh Pimpale, Aditi Raut, Yash Patil, Gaurav Parpol, Prajwal Yadav and Janhavi Sangoi.
Methodology: Harsh Pimpale, Aditi Raut, Yash Patil, Gaurav Parpol, Prajwal Yadav and Janhavi Sangoi.
Investigation: Harsh Pimpale, Aditi Raut, Yash Patil, Gaurav Parpol, Prajwal Yadav and Janhavi Sangoi.
Discussion of results: Harsh Pimpale, Aditi Raut, Yash Patil, Gaurav Parpol, Prajwal Yadav and Janhavi Sangoi.
Writing – Original Draft: Harsh Pimpale, Aditi Raut, Yash Patil, Gaurav Parpol, Prajwal Yadav and Janhavi Sangoi.
Writing – Review and Editing: Harsh Pimpale, Aditi Raut, Yash Patil, Gaurav Parpol, Prajwal Yadav and Janhavi Sangoi.
Resources: Harsh Pimpale, Aditi Raut, Yash Patil, Gaurav Parpol, Prajwal Yadav and Janhavi Sangoi.
Supervision: Harsh Pimpale, Aditi Raut, Yash Patil, Gaurav Parpol, Prajwal Yadav and Janhavi Sangoi.
Approval of the final text: Harsh Pimpale, Aditi Raut, Yash Patil, Gaurav Parpol, Prajwal Yadav and Janhavi Sangoi.

VII. REFERENCES

[1] M. Medvedeva, M. Vols, and M. Wieling, "Legal information retrieval systems: State-of-the-art and open issues," Artificial Intelligence and Law, vol. 30, no. 2, pp. 181–206, Jun. 2022.

[2] J.-S. Lee, "LexGPT 0.1: Pre-trained GPT-J models with Pile of Law," arXiv preprint arXiv:2306.05431, Jun. 2023.

[3] Y. Fei et al., "InternLM-Law: An Open-Sourced Chinese Legal Large Language Model," in Proc. 31st Int. Conf. Comput. Linguistics (COLING), Abu Dhabi, UAE, 2025, pp. 9376–9392.

[4] A. C. P. L. de Oliveira et al., "On the legal implications of Large Language Model answers: A prompt engineering approach and a view beyond by exploiting Knowledge Graphs," Artificial Intelligence and Law, vol. 32, no. 1, pp. 1–27, Mar. 2024.

[5] N. Xie et al., "DeliLaw: A Chinese Legal Counselling System Based on a Large Language Model," arXiv preprint arXiv:2408.00357, Aug. 2024.

[6] Y. Lyu et al., "CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models," arXiv preprint arXiv:2401.17043, Jan. 2024.

[7] A. S. Hussain and A. Thomas, "Large Language Models for Judicial Entity Extraction: A Comparative Study," arXiv preprint arXiv:2407.05786, Jul. 2024.

[8] M. Macias, "Finetuning and Improving Prediction Results of LLMs Using Synthetic Data," Bachelor's thesis, Metropolia University of Applied Sciences, 2024.

[9] Gemma Team, "Gemma Explained: New in Gemma 2," Google Developers Blog, [Online]. Available: https://developers.googleblog.com/en/gemma-explained-new-in-gemma-2/. [Accessed: Jan. 10, 2025].

[10] Gemma Team, "Gemma 2: Improving Open Language Models at a Practical Size," arXiv preprint arXiv:2408.00118, Aug. 2024.

[11] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[12] B. Zhang and R. Sennrich, "Root Mean Square Layer Normalization," arXiv preprint arXiv:1910.07467, Oct. 2019.

[13] Justice Hub, "Supreme Court Cases 2010–2020 Dataset," [Online]. Available: https://justicehub.in/dataset/supreme-court-cases-2010-2020. [Accessed: Sept. 20, 2024].

[14] Supreme Court of India, "Official Website," [Online]. Available: https://main.sci.gov.in/. [Accessed: Apr. 20, 2025].

[15] LangChain, "Google Generative AI Integration," [Online]. Available: https://python.langchain.com/api_reference/google_genai/. [Accessed: Dec. 20, 2024].

[16] LangChain, "FAISS Vector Store Integration," [Online]. Available: https://python.langchain.com/docs/integrations/vectorstores/faiss/. [Accessed: Dec. 21, 2024].

[17] M. Shaik and M. Y. Bin Ghazali, “LLM Based News Research Tool Using LangChain with Enhancing Similarity Search and Token Limit,” International Journal for Research in Applied Science and Engineering Technology (IJRASET), vol. 12, no. 1, pp. 366–372, Jan. 2024. [Online]. Available: https://www.ijraset.com/research-paper/llm-based-news-research-tool-using-langchain

[18] Ministry of Home Affairs, The Bharatiya Nyaya Sanhita, 2023 (English Version), Apr. 2024. [Online]. Available: https://www.mha.gov.in/sites/default/files/250883_english_01042024.pdf
[Accessed: Jan. 10, 2025].

[19] Supreme Court of India, Case Categories, [Online]. Available: https://www.sci.gov.in/case-category/
[Accessed: Jan. 1, 2025].

[20] OpenAI, "ChatGPT," [Online]. Available: https://chatgpt.com/. [Accessed: Jan. 20, 2025].

[21] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in Proc. Workshop Text Summarization Branches Out (WAS), Barcelona, Spain, 2004, pp. 74–81.

[22] K. Papineni et al., "BLEU: a Method for Automatic Evaluation of Machine Translation," in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL), Philadelphia, PA, USA, 2002, pp. 311–318.

[23] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Machine Translation Summarization, Ann Arbor, MI, USA, 2005, pp. 65–72.

[24] T. Zhang et al., "BERTScore: Evaluating Text Generation with BERT," in Proc. Int. Conf. Learn. Representations (ICLR), Addis Ababa, Ethiopia, 2020.

[25] Supreme Court of India, “Judgment in Criminal Appeal No. 1902 of 2012,” [Online]. Available: https://main.sci.gov.in/supremecourt/2012/1902/1902_2012_Judgement_12-Feb-2019.pdf. [Accessed: Oct. 10, 2024].

[26] Supreme Court of India, “Judgment in Criminal Appeal No. 686 of 2018,” [Online]. Available: https://main.sci.gov.in/supremecourt/2018/686/686_2018_Judgement_26-Feb-2019.pdf. [Accessed: Oct. 10, 2024].