# NyayaRAG: Realistic Legal Judgment Prediction with RAG under the Indian Common Law System

**Anonymous ACL submission**

## Abstract

Legal Judgment Prediction (LJP) has emerged as a key area in AI for law, aiming to automate judicial outcome forecasting and enhance interpretability in legal reasoning. While previous approaches in the Indian context have relied on internal case content such as facts, issues, and reasoning, they often overlook a core element of common law systems, reliance on statutory provisions and judicial precedents. In this work, we propose `NyayaRAG`, a Retrieval-Augmented Generation (RAG) framework that simulates realistic courtroom scenarios by providing models with factual case descriptions, relevant legal statutes, and semantically retrieved prior cases. NyayaRAG evaluates the effectiveness of these combined inputs in predicting court decisions and generating legal explanations using a domain-specific pipeline tailored to the Indian legal system. We assess performance across various input configurations using both standard lexical and semantic metrics as well as LLM-based evaluators such as G-Eval. Our results show that augmenting factual inputs with structured legal knowledge significantly improves both predictive accuracy and explanation quality.

## 1 Introduction

The application of artificial intelligence (AI) in legal judgment prediction (LJP) has the potential to transform legal systems by improving efficiency, transparency, and access to justice. This is particularly crucial for India, where millions of cases remain pending in courts, and decision-making is inherently dependent on factual narratives, statutory interpretation, and judicial precedent. India follows a common law system, where prior decisions (precedents) and statutory provisions play a central role in influencing legal outcomes. However, most existing AI-based LJP systems do not adequately replicate this fundamental feature of judicial reasoning.

Previous studies such as Malik et al. (2021); Nigam et al. (2024b, 2025a) have focused on predicting legal outcomes using the current case document, including sections like facts, arguments, issues, reasoning, and decision. More recent efforts have narrowed the scope to factual inputs alone (Nigam et al., 2024a, 2025b), yet these systems still operate in a vacuum, without considering how courts naturally rely on applicable laws and prior rulings. In reality, judges rarely decide in isolation; instead, they actively refer to relevant precedent and statutory law. To bridge this gap, we propose a framework that more closely mirrors actual courtroom conditions by explicitly incorporating external legal knowledge during inference.

Moreover, in critical domains like finance, medicine, and law, decisions must be grounded in verifiable information. Experts in these domains cannot rely on opaque, black-box inferences, and they require systems that ensure factual consistency. Hallucinations, common in large generative models, can have severe consequences in legal decision-making. By retrieving and conditioning model responses on grounded sources such as applicable laws and precedent cases, Retrieval-Augmented Generation (RAG) offers a principled approach to mitigate hallucination and promote trustworthy outputs. Furthermore, RAG frameworks like ours can be flexibly integrated into existing legal systems without requiring the retraining of core models or the sharing of private or sensitive case data. This enhances user trust while allowing the legal community to benefit from AI without sacrificing transparency or data confidentiality.

We introduce **`NyayaRAG`**, a Retrieval-Augmented Generation (RAG) framework for realistic legal judgment prediction and explanation in the Indian common law system. The term "NyayaRAG" is derived from two components: "**Nyaya**" meaning "Justice" and "**RAG**" referring to Retrieval-Augmented Generation. Together, the name re-

flects our vision to build a justice-aware generation system that emulates the reasoning process followed by Indian courts, using facts, statutes, and precedents.

Unlike prior models that operate purely on internal case content, NyayaRAG simulates real-world judicial decision-making by providing the model with: (i) the summarized factual background of the current case, (ii) relevant statutory provisions, (iii) top-$k$ semantically retrieved previous similar judgments. This structure emulates how judges deliberate on new cases, consulting both textual statutes and prior judicial opinions. Through this design, we evaluate how Retrieval-Augmented Generation can help reduce hallucinations, promote faithfulness, and yield legally coherent predictions and explanations.

Our contributions are as follows:

1. *A Realistic RAG Framework for Indian Courts:* We present NyayaRAG, a novel framework that emulates Indian common law decision-making by incorporating not only facts but also retrieved legal statutes and precedents.
2. *Retrieval-Augmented Pipelines with Structured Inputs:* We construct modular pipelines representing different combinations of factual, statutory, and precedent-based inputs to understand their individual and combined contributions to model performance.
3. *Simulating Common Law Reasoning with LLMs:* We show that LLMs guided by RAG and factual grounding can produce legally faithful explanations aligned with how real-world decisions are made under common law reasoning.

Our work moves beyond fact-only or self-contained models by replicating a more faithful legal reasoning pipeline aligned with Indian jurisprudence. We hope that NyayaRAG opens new directions for building interpretable, retrieval-aware AI systems in legal settings, particularly in resource-constrained yet precedent-driven judicial systems like India's. For the sake of reproducibility, we have made our dataset, code, and RAG-based pipeline implementation via an anonymous repository[1].

## 2 Related Work

Recent advancements in natural language processing (NLP) and large language models (LLMs) have significantly improved the performance of question answering (QA) and legal decision support systems. Transformer-based architectures such as BERT (Devlin et al., 2018), GPT (Radford et al., 2019), and their instruction-tuned successors have led to robust capabilities in knowledge-intensive and multi-hop reasoning tasks. The integration of external information via Retrieval-Augmented Generation (RAG) has emerged as a particularly effective approach for enhancing generation fidelity and reducing hallucinations (Han et al., 2024; Hei et al., 2024).

Within the legal domain, Legal Judgment Prediction (LJP) has seen significant progress, with models trained to infer outcomes based on factual and procedural components of court cases (Strickson and De La Iglesia, 2020; Xu et al., 2020; Feng et al., 2023). In the Indian legal context, the ILDC corpus (Malik et al., 2021) and its extended variants (Nigam et al., 2024b; Nigam and Deroy, 2023) have enabled the development of supervised and instruction-tuned models for both judgment prediction and explanation. The emergence of domain-specific datasets and architectures has allowed LJP systems to move from simple binary classification to more complex reasoning tasks aligned with real judicial behavior (Vats et al., 2023).

Parallel to these developments, there has been a sharp rise in interest in RAG techniques for legal NLP. Several benchmark and system-level contributions have explored how retrieval-enhanced generation can be leveraged to assist legal professionals, improve legal QA systems, and support document analysis. Notably, LegalBench-RAG (Pipitone and Alami, 2024) introduced a benchmark suite for evaluating RAG in the legal domain. Survey papers like (Hindi et al., 2025) provide comprehensive overviews of techniques aimed at improving RAG performance, factual grounding, and interpretability in legal settings.

Several system-level contributions have demonstrated the power of RAG in specialized applications. Graph-RAG for Legal Norms (de Martim, 2025) and Bridging Legal Knowledge and AI (Barron et al., 2025) proposed methods to integrate structured legal knowledge such as statutes and normative hierarchies into the retrieval pipeline. Similarly, CBR-RAG (Wiratunga et al., 2024) applied case-based reasoning to leverage historical decisions, showing strong gains in legal question answering. HyPA-RAG (Kalra et al., 2024) explored hybrid parameter-adaptive retrieval to dynamically adjust context based on query specificity.

[1] https://anonymous.4open.science/r/RAGLegal

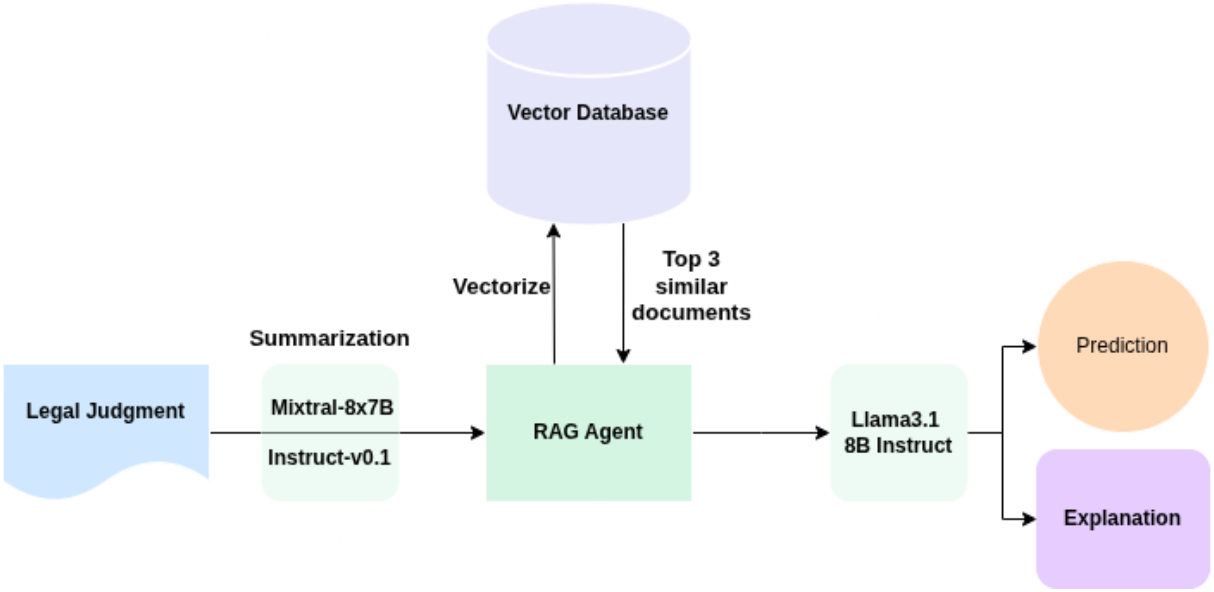


Figure 1: Illustration of our Legal Judgment Prediction framework using RAG. The input legal judgment is first summarized; a RAG agent retrieves top-3 relevant documents from a vector database; and an instruction-tuned LLM (e.g., LLaMA-3.1 8B Instruct) generates the final prediction and explanation.

Further domain-specific applications include AI-powered legal assistants like Legal Query RAG (Wahidur et al., 2025) and RAG-based solutions for dispute resolution in housing law (Rafat, 2024). Optimizing Legal Information Access (Amato et al., 2024) showcased federated RAG architectures for secure document retrieval, and Augmenting Legal Judgment Prediction with Contrastive Case Relations (Liu et al., 2022) illustrated the benefits of encoding contrastive precedents for predictive reasoning.

## 3 Task Description

India's judicial system operates within the common law framework, where judges deliberate cases based on three fundamental pillars: (i) the factual context of the case, (ii) applicable statutory provisions, and (iii) relevant judicial precedents. Our task is designed to simulate such realistic legal decision-making by leveraging Retrieval-Augmented Generation (RAG), enabling models to access external legal knowledge during inference.

Figure 1 illustrates our Legal Judgment Prediction (LJP) pipeline enhanced with RAG. The pipeline begins with a full legal judgment document, which undergoes summarization to reduce its length and retain essential factual meaning. This is necessary because legal judgments tend to be long, and appending retrieved knowledge further increases the input size. Given limited model capacity and computational resources, we employ a summarization step (using `Mixtral-8x7B-Instruct-v0.1`) to create a condensed representation of both the input case and the retrieved legal context.

**Prediction Task:** Based on the summarized factual description $D$ and the retrieved top-$k$ (e.g., $k = 3$) similar legal documents (statutes or precedents), the model predicts the likely court judgment. The prediction label $y \in \{0, 1\}$ indicates whether the appeal is fully rejected (0) or fully/partially accepted (1). This binary framing captures the most common forms of judicial decisions in Indian appellate courts.

**Explanation Task:** Alongside the decision, the model is also required to generate an explanation that justifies its output. This explanation should logically incorporate the facts, cited statutes, and relevant precedents retrieved during the RAG process. This step emulates how judges provide reasoned opinions in written judgments.

By structuring the LJP task in this way, summarizing long documents and integrating retrieval-based augmentation, we study the effectiveness of RAG agents in producing judgments that are both faithful to legal reasoning and grounded in precedent and statute. The overall framework allows us to approximate a real-world decision-making environment within Indian courtrooms.

| Dataset | #Documents | Avg. Length | Max | Min |
|---|---|---|---|---|
| **SCI (Full)** | 56,387 | 3,495 | 401,985 | 14 |
| **Summarized Single** | 4,962 | 302 | 875 | 1 |
| **Summarized Multi** | 4,930 | 300 | 879 | 1 |
| **Sections** | 29,858 | 257 | 27,553 | 9 |

Table 1: Statistics of the dataset across various processed categories.

# 4 Dataset

Our dataset is designed to simulate realistic court decision-making in the Indian legal context, incorporating facts, statutes, and precedent, essential elements under the common law framework. This dataset enables exploration of Legal Judgment Prediction (LJP) in a Retrieval-Augmented Generation (RAG) setup.

## 4.1 Dataset Compilation

We curated a large-scale dataset consisting of 56,387 Supreme Court of India (SCI) case documents up to April 2024, sourced from IndianKanoon[2], a trusted legal search engine. The website provides structural tags for various judgment components (e.g., facts, issues, arguments), which allowed for clean and structured scraping. These documents serve as the foundation for our summarization, retrieval, and reasoning experiments.

## 4.2 Dataset Composition

The corpus supports multiple downstream pipelines, each focusing on specific judgment elements or legal context. Table 1 presents key statistics across different configurations, and an example breakdown is shown in the Appendix Table 6.

### 4.2.1 Case Text

Each judgment includes complete narrative content such as factual background, party arguments, legal issues, reasoning, and verdict. Due to length constraints exceeding model context windows, we summarized these documents using `Mixtral-8x7B-Instruct-v0.1`(Jiang et al., 2024), which supports up to 32k tokens. The summarization preserved critical legal elements through carefully designed prompts (see Table 2).

### 4.2.2 Precedents

From each judgment, cited precedents were extracted using metadata tags provided by IndianKanoon. These citations represent explicit legal reasoning and are retained for use during inference to replicate how courts consider prior judgments.

### 4.2.3 Statutes

Statutory references were also programmatically extracted, including citations to laws like the Indian Penal Code and the Constitution of India. Where statute sections exceeded length limits, they were summarized using the same LLM pipeline. Only statutes directly cited in the respective cases were retained, ensuring relevance.

### 4.2.4 Previous Similar Cases

To simulate implicit precedent-based reasoning, we employed semantic similarity retrieval to identify relevant previous cases beyond explicit citations:

- **Corpus Vectorization:** All 56,387 documents were embedded into dense vector representations using the `all-MiniLM-L6-v2` sentence transformer.
- **Target Encoding:** The 5,000 selected training samples were vectorized similarly.
- **Top-$k$ Retrieval:** Using `ChromaDB`, we retrieved the top-3 most semantically similar cases for each document based on cosine similarity.
- **Augmentation:** Retrieved cases were appended to the factual input to form the "`casetext + previous similar cases`" input during model inference.

This retrieval step enriches context with precedents that are semantically close, even if not cited, enhancing the legal realism of our setup.

### 4.2.5 Facts

We separately extracted the factual portions of all 56,387 judgments. These include background information, chronological events, and party narratives, excluding legal reasoning. These fact-only subsets were used to simulate realistic courtroom scenarios where judges primarily rely on facts, relevant law, and precedent for decision-making.

Overall, our dataset is uniquely structured to test legal decision-making under realistic constraints, aligning with the Indian legal system's reliance on factual narratives, statutory frameworks, and prior rulings.

# 5 Methodology

To simulate realistic judgment prediction and evaluate the role of Retrieval-Augmented Generation (RAG) in enhancing legal decision-making, we design a modular experimental setup. This setup ex-

[2] https://indiankanoon.org/

| Summarization Prompt |
| --- |
| The text is regarding a court judgment for a specific case. Summarize it into 1000 tokens but more than 700 tokens. The summarization should highlight the Facts, Issues, Statutes, Ratio of the decision, Ruling by Present Court (Decision), and a Conclusion. |

Table 2: Instruction prompt used with `Mixtral-8x7B-Instruct-v0.1` for summarizing legal judgments.

plores how different types of legal information, such as factual summaries, statutes, and precedents, affect model performance on the dual tasks of prediction and explanation.

### 5.1 Pipeline Construction

To systematically evaluate the impact of legal knowledge sources, we constructed multiple input pipelines using combinations of the dataset components described in Section 4. Each pipeline configuration represents a distinct input scenario reflecting different degrees of legal context and retrieval augmentation. These pipelines are as follows:

- **CaseText Only:** Includes only the summarized version of the full case judgment, which contains factual background, legal arguments, and reasoning.
- **CaseText + Statutes:** Appends summarized statutory references cited in the judgment to the case text, simulating scenarios where relevant laws are explicitly considered.
- **CaseText + Precedents:** Incorporates prior cited judgments mentioned in the original case, representing explicitly relied-upon precedents.
- **CaseText + Previous Similar Cases:** Adds top-3 semantically similar past judgments (retrieved via ChromaDB using `all-MiniLM-L6-v2` embeddings), allowing the model to learn from precedents not explicitly cited.
- **CaseText + Statutes + Precedents:** A comprehensive legal input pipeline combining the full judgment summary, statutes, and cited prior judgments.
- **Facts Only:** A minimal pipeline containing only the factual summary, excluding all legal reasoning and verdicts. This setup evaluates whether a model can infer judgments from facts alone.
- **Facts + Statutes + Precedents:** Combines factual input with statutory and precedent context to simulate realistic courtroom conditions where judges rely on facts, applicable law, and relevant past cases.

This modular design enables granular control over input features and facilitates direct comparison of how each knowledge source contributes to judgment prediction and explanation generation.

### 5.2 Prompt Design

To ensure consistency and interpretability across all pipelines, we used fixed instruction prompts with minor variations depending on the available contextual inputs (e.g., facts only vs. facts + law + precedent). These prompts guide the model in producing both binary predictions and natural language explanations. Prompts were structured to reflect real judicial inquiry formats, aligning with the instruction-following capabilities of modern LLMs. Full prompt templates are listed in Appendix Table 7, along with prediction examples.

### 5.3 Inference Setup

We use the `LLaMA-3.1 8B Instruct` model for all experiments in a few-shot prompting setup. Each input sequence, composed according to one of the pipeline templates, is paired with a relevant prompt. The model is required to output:

- A binary judgment prediction: 0 (appeal rejected) or 1 (appeal fully/partially accepted)
- A justification: a coherent natural language explanation based on legal facts, statutes, and precedent

The model is explicitly instructed to reason with the provided information and emulate judicial writing. Retrieved knowledge (via RAG) is included in-context to enhance legal reasoning while minimizing hallucinations.

This experimental design allows us to evaluate the effectiveness of legal retrieval and summarization under realistic judicial decision-making constraints in the Indian common law setting.

## 6 Experimental Setup and Hyper-parameters

### 6.1 Summarization Hyper-parameters

To condense lengthy Indian Supreme Court judgments into structured and model-friendly inputs, we employed `Mixtral-8x7B-Instruct-v0.1`, a mixture-of-experts, instruction-tuned language model developed by Mistral AI. The summarization was conducted in a zero-shot setting using

tailored legal prompts that extracted key elements such as facts, statutes, precedents, reasoning, and the final ruling.

The model was accessed via the HuggingFace Transformers interface and run on an NVIDIA A100 GPU with 80GB VRAM. Inputs were truncated to a maximum of 27,000 tokens to comply with the model's context window. The output length was constrained to between 700 and 1,000 tokens to ensure consistency and legal completeness. A low decoding temperature of 0.2 was used to encourage determinism and factual alignment. These summaries served as inputs to the Retrieval-Augmented Generation (RAG) pipelines used for downstream judgment prediction and explanation.

### 6.2 Judgment Prediction Hyper-parameters

For the legal judgment prediction task, we used the `LLaMA 3-8B Instruct` model, which supports high-quality reasoning in instruction-following settings. The model was applied in a few-shot prompting setup without any task-specific fine-tuning. Input prompts consisted of structured summaries (produced by Mixtral) along with retrieved statutes and prior similar cases. These inputs followed a consistent legal instruction format to guide the model's prediction and explanation generation.

Inference was performed using the PyTorch backend with HuggingFace Transformers on an NVIDIA A100 GPU (80GB). The model was loaded using `device_map="auto"` for automatic device allocation. We used deterministic generation parameters (temperature = 0.2, top-p = 0.9) and controlled output format to ensure faithful and interpretable outputs. Each output consisted of a binary prediction (`0` for appeal rejected, `1` for appeal accepted/partially accepted) followed by a free-text legal explanation. No supervised fine-tuning was used, which allows our framework to be easily adapted to different legal datasets without retraining.

## 7 Evaluation Metrics

To evaluate the effectiveness of our Retrieval-Augmented Legal Judgment Prediction framework, we adopt a comprehensive set of metrics covering both classification accuracy and explanation quality. The evaluation is conducted on two fronts: the judgment prediction task and the explanation generation task. These metrics are selected to ensure a holistic assessment of model performance in the legal domain. We report Macro Precision, Macro Recall, Macro F1, and Accuracy for judgment prediction, and we use both quantitative and qualitative methods to evaluate the quality of explanations generated by the model.

1. **Lexical-based Evaluation:** We utilized standard lexical similarity metrics, including Rouge scores (Rouge-1, Rouge-2, and Rouge-L) (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005). These metrics measure the overlap and order of words between the generated explanations and the reference texts, providing a quantitative assessment of the lexical accuracy of the model outputs.
2. **Semantic Similarity-based Evaluation:** To capture the semantic quality of the generated explanations, we employed BERTScore (Zhang et al., 2020), which measures the semantic similarity between the generated text and the reference explanations. Additionally, we used BLANC (Vasilyev et al., 2020), a metric that estimates the quality of generated text without a gold standard, to evaluate the model's ability to produce semantically meaningful and contextually relevant explanations.
3. **LLM-based Evaluation (LLM-as-a-Judge):** To complement traditional metrics, we incorporate an automatic evaluation strategy that uses large language models themselves as evaluators, commonly referred to as *LLM-as-a-Judge*. This evaluation is crucial for assessing structured argumentation and legal correctness in a format aligned with expert judicial reasoning. We adopt G-Eval (Liu et al., 2023), a GPT-4-based evaluation framework tailored for natural language generation tasks. G-Eval leverages chain-of-thought prompting and structured scoring to assess explanations along three key criteria: *factual accuracy*, *completeness & coverage*, and *clarity & coherence*. Each generated legal explanation is scored on a scale from 1 to 10 based on how well it aligns with the expected content and a reference document. The exact prompt format used for evaluation is shown in Appendix Table 8. For our experiments, we use the GPT-4o-mini model to generate reliable scores without manual intervention. This setup provides an interpretable, unified judgment metric that captures legal soundness, completeness of reasoning, and logical coherence, beyond what traditional similarity-based metrics can offer.

Together, these metrics provide a robust and mul-

| Pipeline Name | Partition | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| CaseText Only | Single | 62.273 | 33.498 | 30.883 | 29.450 |
| | Multi | 53.103 | 25.258 | 23.946 | 20.808 |
| CaseText + Statutes | Single | **67.067** | 45.288 | 44.547 | 44.318 |
| | Multi | 60.358 | **64.221** | **64.039** | 60.351 |
| CaseText + Precedents | Single | 61.733 | 41.919 | 41.349 | 40.806 |
| | Multi | 57.532 | 61.340 | 61.187 | 57.525 |
| CaseText + Previous Similar Cases | Single | 57.532 | **61.341** | **61.187** | **57.525** |
| | Multi | 61.733 | 41.919 | 41.349 | 57.525 |
| CaseText + Statutes + Precedents | Single | 64.705 | 43.495 | 42.976 | 42.775 |
| | Multi | **65.864** | 63.942 | 63.986 | **63.963** |
| CaseFacts Only | Single | 51.125 | 51.355 | 51.298 | 50.677 |
| | Multi | 53.713 | 51.177 | 51.182 | 51.180 |
| Facts + Statutes + Precedents | Single | 50.576 | 33.573 | 33.556 | 33.24 |
| | Multi | 52.574 | 52.009 | 52.009 | 52.009 |

Table 3: Performance of Various Pipelines on Binary and Multi-label Legal Judgment Prediction. The best result has been marked in bold.

tidimensional view of the model's capabilities, not only in predicting judicial outcomes but also in generating coherent, contextually grounded, and legally meaningful explanations.

## 8 Results and Analysis

We conducted extensive evaluations across multiple pipeline configurations to study the impact of different legal information components on both judgment prediction and explanation quality. Tables 3 and 4 summarize the model's performance across these configurations for binary and multi-label settings.

### 8.1 Judgment Prediction Performance

As shown in Table 3, the pipeline combining *CaseText + Statutes* achieved the highest accuracy in the single-label setting, with a notable 67.07% accuracy. This suggests that legal statutes provide substantial contextual cues for the model to infer the likely decision. In contrast, *CaseText Only* achieved 62.27%, highlighting the importance of augmenting case narratives with applicable laws. Interestingly, the *CaseText + Previous Similar Cases* pipeline showed the highest precision, recall, and F1-score in the single-label case, indicating that semantically retrieved precedents, despite not being explicitly cited, help the model align with actual judicial outcomes.

In the multi-label setting, the best accuracy was observed for the *CaseText + Statutes + Precedents* pipeline, with 65.86% accuracy and 63.96% F1-score. This comprehensive context provides the model with structured legal knowledge, improving generalization across different outcome labels. Conversely, the *Facts Only* pipeline performed worst overall, reaffirming that factual narratives alone, without legal context, are insufficient for reliably predicting legal outcomes. The poor performance of the *Facts + Statutes + Precedents* pipeline in the single-label setting suggests that factual sections might lack the interpretive cues that full case texts offer when combined with legal references.

### 8.2 Explanation Generation Quality

Table 4 presents the results of explanation evaluation using both automatic metrics (ROUGE, BLEU, METEOR, BERTScore, BLANC) and the LLM-based evaluation (G-Eval). Across the board, the *CaseText + Statutes* pipeline consistently outperformed others in both single and multi-label setups, achieving top scores in lexical and semantic similarity, as well as in G-Eval. In the single-label case, it obtained a BLEU score of 0.0321, a ROUGE-2 score of 0.0764, and a G-Eval score of 4.21, substantially higher than the *CaseText Only* baseline. This indicates that access to statutory provisions improves not only prediction performance but also the interpretability and factual alignment of explanations, as judged by a strong LLM evaluator.

Interestingly, while the *CaseText + Previous Similar Cases* pipeline attained the highest ROUGE-1 score (0.2744), it lagged behind the *CaseText + Statutes* pipeline in other semantic and coherence-based metrics, including G-Eval. This suggests

| Pipelines | Data Split | ROUGE-1 | ROUGE-2 | ROUGE-L | BLEU | METEOR | BERTScore | BLANC | G-Eval |
|---|---|---|---|---|---|---|---|---|---|
| CaseText Only | Single | 0.2447 | 0.0656 | 0.1582 | 0.0266 | 0.1758 | 0.5158 | 0.0772 | 4.168 |
| | Multi | 0.2481 | 0.0666 | 0.1595 | 0.0261 | 0.1781 | 0.5223 | 0.0807 | 4.001 |
| CaseText + Statutes | Single | 0.2706 | **0.0764** | **0.1671** | **0.0321** | **0.1985** | **0.5279** | **0.0925** | **4.21** |
| | Multi | **0.2665** | **0.0759** | **0.1656** | **0.0336** | **0.2024** | **0.5326** | **0.0921** | **4.10** |
| CaseText + Precedents | Single | 0.2525 | 0.0641 | 0.1585 | 0.0290 | 0.1938 | 0.5143 | 0.0775 | 3.45 |
| | Multi | 0.2664 | 0.0671 | 0.1636 | 0.0292 | 0.2024 | 0.5274 | 0.0854 | 3.413 |
| CaseText + Previous Similar Cases | Single | **0.2744** | 0.0671 | 0.1617 | 0.0280 | 0.1973 | 0.5221 | 0.0818 | 3.722 |
| | Multi | 0.2656 | 0.0693 | 0.1590 | 0.0285 | 0.1922 | 0.5233 | 0.0788 | 3.67 |
| CaseText + Statutes + Precedents | Single | 0.2616 | 0.0685 | 0.1623 | 0.0271 | 0.1885 | 0.5215 | 0.0837 | 4.11 |
| | Multi | 0.2612 | 0.0710 | 0.1641 | 0.0304 | 0.1997 | 0.5269 | 0.0860 | 3.923 |
| CaseFacts Only | Single | 0.2563 | 0.0604 | 0.1573 | 0.0200 | 0.1781 | 0.5199 | 0.0626 | 3.532 |
| | Multi | 0.2458 | 0.0614 | 0.1512 | 0.0189 | 0.1727 | 0.5231 | 0.0770 | 3.742 |
| Facts + Statutes + Precedents | Single | 0.2533 | 0.0542 | 0.1571 | 0.0191 | 0.1849 | 0.5061 | 0.0623 | 2.968 |
| | Multi | 0.2536 | 0.0575 | 0.1538 | 0.0198 | 0.1918 | 0.5170 | 0.0698 | 3.083 |

Table 4: Comparison of Explanation Generation Across Different Legal Context Pipelines. The best result has been marked in bold.

that semantically similar cases enhance stylistic or linguistic similarity but may lack the depth of structured legal justification provided by statutes. The *CaseText + Statutes + Precedents* pipeline also achieved strong overall scores, performing nearly as well as the best configurations, demonstrating the additive benefit of combining both explicit citations and statutory references.

In contrast, pipelines that relied solely on factual narratives, *Facts Only* and *Facts + Statutes + Precedents*, consistently underperformed. These configurations yielded the lowest BLEU, METEOR, and G-Eval scores, reaffirming that facts alone are insufficient for generating legally persuasive or complete explanations. Notably, the *Facts + Statutes + Precedents* pipeline scored as low as 2.968 on G-Eval in the single-label case, emphasizing that interpretive and argumentative components are critical for producing human-aligned legal reasoning.

Overall, these results highlight the strength of Retrieval-Augmented Generation when paired with structured legal knowledge, especially statutory content. The addition of automatic LLM-based evaluation via G-Eval provided further insights into the factuality and coherence of model-generated legal explanations, going beyond traditional similarity metrics to better approximate human evaluative standards.

## 9 Conclusion and Future Scope

In this paper, we presented NyayaRAG, a Retrieval-Augmented Generation (RAG) framework designed to simulate realistic legal judgment prediction and explanation within the Indian common law system. Unlike prior models that rely solely on the content of the current case, our approach integrates three key components, factual descriptions, relevant statutory provisions, and semantically similar prior cases, closely emulating the way judges reason in real courtroom settings.

Through extensive experiments, we demonstrated that incorporating structured legal knowledge via RAG significantly improves both the predictive performance and the legal soundness of generated explanations. Specifically, pipelines augmented with statutes and precedents yielded higher accuracy and interpretability, as confirmed by both traditional NLP metrics and LLM-based evaluation using G-Eval. Our work underscores the importance of simulating actual legal reasoning processes to develop more faithful and trustworthy AI systems for legal applications.

Future work may explore several promising directions. First, we aim to extend our framework to multi-class or hierarchical verdict structures that better capture real-world legal complexity. Second, while our current retrieval is based on dense semantic similarity, future iterations could integrate symbolic reasoning or graph-based legal knowledge for more structured retrieval. Third, we plan to incorporate temporal dynamics of precedent evolution, enabling the system to weigh older versus newer case law appropriately. Lastly, we envision incorporating human-in-the-loop feedback and expert validation to further align AI predictions with judicial expectations.

By aligning AI with the foundational principles of Indian jurisprudence, NyayaRAG contributes toward the long-term vision of explainable, realistic, and accessible legal AI. We hope this work sparks further research into retrieval-enhanced, court-aligned AI systems in underrepresented legal ecosystems.

## Limitations

While NyayaRAG demonstrates promising results in simulating realistic courtroom decision-making, several limitations remain that open avenues for future improvements.

First, although the use of Retrieval-Augmented Generation (RAG) mitigates hallucinations, it does not eliminate them entirely. There may still be instances where the model generates factually plausible but legally incorrect rationales. This risk is particularly sensitive in high-stakes domains like law, where interpretability and precision are critical.

Second, the explanation outputs are not currently validated by human experts on a large scale. While we employed G-Eval for automatic assessment, legal AI systems benefit from domain expert validation to ensure that generated rationales align with acceptable legal standards. Additionally, the current system is designed for binary and multi-label prediction but does not yet handle full multi-class or hierarchical verdicts.

Third, the system assumes access to accurate, cleaned, and structured legal documents. In practice, real-world court data may be noisy, incomplete, or inconsistently formatted, which can affect retrieval accuracy and downstream generation quality. Moreover, while our summarization process helps reduce input length, it may lead to information loss if key legal details are omitted.

Finally, we currently do not perform explicit fine-tuning of LLMs on Indian legal corpora due to computational constraints. Instead, we rely on instruction-tuned models with domain-aligned prompts. While this makes the system more accessible and privacy-preserving, dedicated legal domain fine-tuning could further improve legal reasoning capabilities.

Despite these limitations, NyayaRAG offers a strong step toward realistic, explainable legal AI systems by aligning with how human judges reason using facts, law, and precedent. Addressing these challenges in future work will further improve robustness, trustworthiness, and adoption in practical legal settings.

## Ethics Statement

This work complies with ethical standards for conducting research in sensitive and high-stakes domains such as law. The legal documents used in this study were sourced from IndianKanoon {https://indiankanoon.org/}, a publicly accessible repository of Indian court judgments. These documents are already in the public domain and do not include sealed cases or sensitive personal information. As such, the data used does not violate privacy norms or confidentiality requirements.

We acknowledge that legal AI systems must be used with caution. The proposed system is intended for academic research and the simulation of realistic legal reasoning processes, not for deployment in real-world legal decision-making. The outputs of the model should not be interpreted as legal advice or judicial determinations. They are designed to support interpretability and experimentation within controlled research environments. We discourage the use of these outputs in actual litigation, policy decisions, or contexts that may affect individuals' legal rights or outcomes without appropriate oversight by legal professionals.

Our system does not involve any human subject experimentation, crowd-sourced annotations, or interaction with individuals. All summarization and evaluation processes were conducted automatically using publicly available models and pre-defined prompts. In evaluating explanation quality, we relied on G-Eval, an automatic evaluation framework based on GPT-4, which requires no human intervention or subjective annotation.

We are aware that legal documents may reflect existing societal biases, and while our system attempts to replicate the reasoning structure used in legal practice, it may also inherit some of these biases. We do not introduce new bias into the model intentionally, but we recognize the need for further work in auditing legal AI systems for fairness, especially with respect to litigant identity, demographic context, and jurisdictional variation.

## References


Flora Amato, Egidia Cirillo, Mattia Fonisto, and Alberto Moccardi. 2024. Optimizing legal information access: Federated search and rag for secure ai-powered legal solutions. In *2024 IEEE International Conference on Big Data (BigData)*, pages 7632–7639. IEEE.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ryan C Barron, Maksim E Eren, Olga M Serafimova, Cynthia Matuszek, and Boian S Alexandrov. 2025. Bridging legal knowledge and ai: Retrieval-augmented generation with vector stores, knowledge graphs, and hierarchical non-negative matrix factorization. *arXiv preprint arXiv:2502.20364*.

Hudson de Martim. 2025. Graph rag for legal norms: A hierarchical and temporal approach. *arXiv preprint arXiv:2505.00039*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Geya Feng, Yongbin Qin, Ruizhang Huang, and Yanping Chen. 2023. Criminal action graph: a semantic representation model of judgement documents for legal charge prediction. *Information Processing & Management*, 60(5):103421.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. Rag-qa arena: Evaluating domain robustness for long-form retrieval augmented question answering. *arXiv preprint arXiv:2407.13998*.

Zijian Hei, Weiling Wei, Wenjie Ou, Juyi Qiao, Junming Jiao, Zhiqing Zhu, and Guowen Song. 2024. Dr-rag: Applying dynamic document relevance to retrieval-augmented generation for question-answering. *arXiv preprint arXiv:2406.07348*.

Mahd Hindi, Linda Mohammed, Ommama Maaz, and Abdulmalik Alwarafy. 2025. Enhancing the precision and interpretability of retrieval-augmented generation (rag) in legal technology: A survey. *IEEE Access*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *ArXiv*, abs/2401.04088.

Rishi Kalra, Zekun Wu, Ayesha Gulley, Airlie Hilliard, Xin Guan, Adriano Koshiyama, and Philip Treleaven. 2024. Hypa-rag: A hybrid parameter adaptive retrieval-augmented generation system for ai legal and policy applications. *arXiv preprint arXiv:2409.09046*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Dugang Liu, Weihao Du, Lei Li, Weike Pan, and Zhong Ming. 2022. Augmenting legal judgment prediction with contrastive case relations. In *Proceedings of the 29th international conference on computational linguistics*, pages 2658–2667.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Shubham Kumar Nigam, Deepak Patnaik Balaramamahanthi, Shivam Mishra, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025a. NYAYAANUMANA and INLEGALLLAMA: The largest Indian legal judgment prediction dataset and specialized language model for enhanced decision analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11135–11160, Abu Dhabi, UAE. Association for Computational Linguistics.

Shubham Kumar Nigam and Aniket Deroy. 2023. Fact-based court judgment prediction. *arXiv preprint arXiv:2311.13350*.

Shubham Kumar Nigam, Aniket Deroy, Subhankar Maity, and Arnab Bhattacharya. 2024a. Rethinking legal judgement prediction in a realistic scenario in the era of large language models. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 61–80, Miami, FL, USA. Association for Computational Linguistics.

Shubham Kumar Nigam, Balaramamahanthi Deepak Patnaik, Shivam Mishra, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025b. Tathyanyaya and factlegalllama: Advancing factual judgment prediction and explanation in the indian legal context.

Shubham Kumar Nigam, Anurag Sharma, Danush Khanna, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2024b. Legal judgment reimagined: PredEx and the rise of intelligent AI interpretation in Indian courts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4296–4315, Bangkok, Thailand. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Md Irfan Rafat. 2024. Ai-powered legal virtual assistant: Utilizing rag-optimized llm for housing dispute resolution in finland.

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal judgement prediction for uk courts. In *Proceedings of the 3rd International Conference on Information Science and Systems*, pages 204–209.

Oleg V. Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: human-free quality estimation of document summaries. *CoRR*, abs/2002.09836.

Shaurya Vats, Atharva Zope, Somsubhra De, Anurag Sharma, Upal Bhattacharya, Shubham Kumar Nigam, Shouvik Guha, Koustav Rudra, and Kripabandhu Ghosh. 2023. LLMs – the good, the bad or the indispensable?: A use case on legal statute prediction and legal judgment prediction on Indian court cases. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12451–12474, Singapore. Association for Computational Linguistics.

Rahman S. M. Wahidur, Sumin Kim, Haeung Choi, David S. Bhatti, and Heung-No Lee. 2025. Legal query rag. *IEEE Access*, 13:36978–36994.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.

Zhuopeng Xu, Xia Li, Yinlin Li, Zihan Wang, Yujie Fanxu, and Xiaoyan Lai. 2020. Multi-task legal judgement prediction combining a subtask of the seriousness of charges. In *Chinese Computational Linguistics: 19th China National Conference, CCL 2020, Hainan, China, October 30–November 1, 2020, Proceedings 19*, pages 415–429. Springer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

| Pipeline | Decision and Explanation |
|---|---|
| Ground Truth | Decision: Rejected.<br>Explanation:<br>1. The parties to these appeals belong to Chokkura Thaliyadath tarwad, a hindu family governed by Marumakkathayam system of law as modified by the Madras Marumakkathayam Act,1932. They belong to the thavazhi of Cheriyammu Amma. Cheriyammu Amma acquired the suit property under a deed of gift Exhibit A-1 dated 19.6.1905.<br>.....<br>Thus, in both the suits, the essential question that fell for decision, especially in the context of the only contention raised before us by learned Senior Counsel Mr. A.S. Nambiar, appearing on behalf of the appellant, was whether the transaction Exhibit-B-9 could be ignored by the sons of Kunhimalu Amma as a void transaction.", '6. Exhibit-B-9 is termed a deed of release |
| Case Text only | Decision: Rejected.<br>Explanation:<br>The appeal is rejected because the lower courts have already upheld the decision that the transaction Exhibit-B-9 is not void in law and the present court also dismissed the second appeals filed by the plaintiff.<br>.....<br>The plaintiff has failed to present any compelling argument or evidence that would justify overturning the decision of the lower courts. The appeal is therefore rejected. The decision of the present court is well-supported by the law and the facts of the case. The plaintiff has no valid grounds for appeal. |
| Case Text with Statutes | Decision: Rejected<br>Explanation:<br>The appeal is rejected because the present court upholds the decision of the lower courts, and the transaction Exhibit-B-9 is not void in law as it is a release of rights by all the other members of the thavazhi in favor of the only other member and not an assignment of undivided shares.<br>.....<br>The plaintiff's argument that the transaction Exhibit-B-9 is void in law is not supported by the evidence or the applicable laws, and the court's decision is a correct application of the law. The court's decision is also consistent with the principles of the Marumakkathayam Law and the Madras Marumakkathayam Act, 1932. The facts of the case do not support the plaintiff's argument that the |
| Case Text with Precedents | Decision: Rejected<br>Explanation:<br>The appeal is rejected because the transaction Exhibit-B-9 is not void in law as it is a release of rights by all the other members of the thavazhi in favor of the only other member and not an assignment of undivided shares, which is in accordance with the Madras Marumakkathayam Act, 1932, and the ratio of the decision.<br>.....<br>The ratio of the decision is that the transaction Exhibit-B-9 is not void in law as it is a release of rights by all the other members of the thavazhi in favor of the only other member and not an assignment of undivided shares. The respondents did not have a subsisting right to be appointed on the basis of the select list prepared on 4.4.87 as its life had expired. The Appointing Authority was justified in filling up the |
| Case Text with Statutes & Precedents | Decision: Rejected<br>Explanation:<br>The appeal is accepted because the transaction Exhibit-B-9 is not void in law as it is a release of rights by all the other members of the thavazhi in favor of the only other member and not an assignment of undivided shares.<br>.....<br>The Supreme Court's decision in this case is a significant step towards ensuring transparency and accountability in the electoral process, but it is not directly related to the issues in this case. Therefore, the appeal is accepted. The decision in S. G. Jaisinghani vs Union Of India And Ors |

Table 5: We observe 'reject' judgments in the response generated by LLaMa-3x8B with varying supporting information.

| |
|---|
| **CASE NO:** |
| CIVIL APPEAL NO. . . . . OF 2014<br>(Arising out of SLP (C) No.22035 of 2013) |
| **APPELLANTS:** |
| Usha Bharti |
| **RESPONDENT:** |
| State Of U.P. & Ors |
| **DATE OF JUDGMENT:** |
| 28/03/2014 |
| **BENCH:** |
| Fakkir Mohamed Ibrahim Kalifulla |
| **CASE TEXT:** |
| ... The earlier judgment of the High Court in the writ petition clearly merged with the judgment of the High Court dismissing the review petition. Therefore, it was necessary only, in the peculiar facts of this case, to challenge only the judgment of the High Court in the review petition. It....<br><br>...These Rules can be amended by the High Court or the Supreme Court but Section 114 can only be amended by the Parliament. He points out that Section 121 and 122, which permits the High Court to make their own rules on theprocedure to be followed in the High Court as well as in...<br><br>...The principle of Ejusdem Generis should not be applied for interpreting these provisions. Learned senior counsel relied on Board of Cricket Control (supra). He relied on Paragraphs 89, 90 and 91. learned senior counsel also relied on S. Nagaraj & Ors. Vs. State of Karnataka & Anr .[13] He submits finally that all these judgments show that justice is above all. Therefore, no...<br><br>... We are unable to accept the submission of Mr. Bhushan that the provisions contained in Section 28 of the Act cannot be sustained in the eyes of law as it fails to satisfy the twin test of reasonable classification and rational nexus with the object sought to be achieved. In support of this submission, Mr. Bhushan has relied on the judgment of this Court in D.S. Nakara vs. Union of India[16]. We... |
| **JUDGEMENT:** |
| .... When the order dated 19th February, 2013 was passed, the issue with regard to reservation was also not canvassed. But now that the issue had been raised, we thought it appropriate to examine the issue to put an end to the litigation between the parties.<br><br>In view of the above, the appeal is accordingly dismissed..... |

Table 6: Example of Indian Case Structure. Sections referenced are highlighted in blue, previous judgments cited are in magenta, and the final decision is indicated in red.

| **Template 1 (prediction + explanation)** |
| --- |
| **prompt** = f“““Task: Your task is to evaluate whether the appeal should be accepted (1) or rejected (0) based on the case proceedings provided below..<br>**Prediction**: You are a legal expert tasked with making a judgment about whether an appeal should be accepted or rejected based on the provided summary of the (case/facts) along with (Precedents/statutes/both) depending on the pipeline. Your task is to evaluate whether the appeal should be accepted (1) or rejected (0) based on the case proceedings provided below.<br>**case_proceeding**: # case_proceeding example 1<br>**Prediction**: # example 1 prediction<br>**Explanation**: # example 1 explanation<br>**case_proceeding**: # case_proceeding example 2<br>**Prediction**: # example 2 prediction<br>**Explanation**: # example 2 explanation<br>**Instructions**: L### Now, evaluate the following case:<br>Case proceedings: summarized_text<br>Provide your judgment by strictly following this format:<br>##PREDICTION: [Insert your prediction here]<br>##EXPLANATION: [Insert your reasoning here that led you to your prediction.]<br>Strictly do not include anything outside this format. Strictly follow the provided format. Do not generate placeholders like [Insert your prediction here]. Just provide the final judgment and explanation. Do not hallucinate/repeat the same sentence again and again””” |

Table 7: Prompts for Judgment Prediction.

**Instructions**:
You are an expert in legal text evaluation. You will be given:
A document description that specifies the intended content of a generated legal explanation.
An actual legal explanation that serves as the reference. A generated legal explanation that needs to be evaluated. Your task is to assess how well the generated explanation aligns with the given description while using the actual document as a reference for correctness.

**Evaluation Criteria (Unified Score: 1-10)**
Your evaluation should be based on the following factors:
*Factual Accuracy (50%)* – Does the generated document correctly represent the key legal facts, reasoning, and outcomes from the original document, as expected from the description?
*Completeness & Coverage (30%)* – Does it include all crucial legal arguments, case details, and necessary context that the description implies?
*Clarity & Coherence (20%)* – Is the document well-structured, logically presented, and legally sound?

**Scoring Scale:**
1-3 $\rightarrow$ Highly inaccurate, major omissions or distortions, poorly structured.
4-6 $\rightarrow$ Somewhat accurate but incomplete, missing key legal reasoning or context.
7-9 $\rightarrow$ Mostly accurate, well-structured, with minor omissions or inconsistencies.
10 $\rightarrow$ Fully aligned with the description, factually accurate, complete, and coherent.

**Input Format:**
Document Description:
{{doc_des}}

**Original Legal Document (Reference):**
{{Actual_Document}}

**Generated Legal Document (To Be Evaluated):**
{{Generated_Document}}

**Output Format:**
Strictly provide only a single integer score (1-10) as the response, with no explanations, comments, or additional text.

Table 8: The prompt is utilized to obtain scores from the G-Eval automatic evaluation methodology. We employed the GPT-4o-mini model to evaluate the quality of the generated text based on the provided prompt/input description, alongside the actual document as a reference.