TathyaNyaya and FactLegalLlama: Advancing Factual Judgment Prediction and Explanation in the Indian Legal Context

Anonymous ACL submission

Abstract

In the legal domain, Fact-based Judgment Prediction and Explanation (FJPE) aims to predict judicial outcomes and generate grounded explanations using only factual information, mirroring early-phase legal reasoning. Motivated by the overwhelming case backlog in the Indian judiciary, we introduce TathyaNyaya, the first large-scale, expert-annotated dataset for FJPE in the Indian context. Covering judgments from the Supreme Court and multiple High Courts, the dataset comprises four complementary components, NyayaFacts, NyayaScrape, NyayaSimplify, and NyayaFilter, that facilitate diverse factual modeling strategies. Alongside, we present FactLegalLlama, an instruction-tuned LLaMa-3-8B model fine-tuned to generate faithful, fact-grounded explanations. While FactLegalLlama trails transformer baselines in raw prediction accuracy, it excels in generating interpretable explanations, as validated by both automatic metrics and legal expert evaluation. Our findings show that fact-only inputs and preprocessing techniques like text simplification and fact filtering can improve both interpretability and predictive performance. Together, TathyaNyaya and FactLegalLlama establish a robust foundation for realistic, transparent, and trustworthy AI applications in the Indian legal system.

1 Introduction

The integration of AI technologies into the legal domain holds immense potential for improving the efficiency, accessibility, and transparency of judicial processes, particularly in countries like India, where case backlogs severely burden the courts. As of recent estimates, over 50 million cases are pending across various courts in India (National Judicial Data Grid, 2024), resulting in delays that can stretch into decades. In this context, early-phase legal decision support, i.e., prediction based solely on factual information available at the beginning of a case, has emerged as a highly relevant research goal.

Among the emerging solutions, Fact-based Judgment Prediction and Explanation (FJPE) offers a promising direction. FJPE aims to predict judicial outcomes and provide rationales using only the factual elements of a case, without relying on arguments, precedents, or judicial reasoning. This mirrors real-world scenarios where stakeholders, judges, lawyers, or litigants, must assess case strength based on initial facts to decide whether to proceed, allocate resources, or pursue alternative legal remedies. Furthermore, factual records are often the most reliably documented and readily available components in early legal proceedings, especially in resource-constrained environments.

While previous studies have attempted fact-centric modeling by summarizing multiple legal components or relying on automatically extracted facts (Nigam et al., 2024b; Nigam and Deroy, 2024), these approaches often lack reliable ground truth and blur the boundaries between pure factual inputs and broader legal discourse. Moreover, such works typically reference the full case context, statutes, or reasoning, placing them closer to the domain of Court Judgment Prediction and Explanation (CJPE), which includes post-filing evidence and legal argumentation. In contrast, FJPE distinctly isolates factual segments to simulate the setting of early-phase legal reasoning, where preliminary decisions may be formed even before formal hearings begin.

To advance this direction, we introduce TathyaNyaya, the first large-scale, expertly annotated dataset explicitly designed for FJPE in the Indian legal context. The term combines the Hindi words “Tathya” (fact) and “Nyaya” (justice), underscoring its foundation in factual legal analysis. Unlike prior datasets, TathyaNyaya does not rely on heuristics or summarization techniques; instead, it offers cleanly annotated factual inputs aligned with judicial outcomes and explanations, allowing for reproducible, interpretable, and practical early-stage prediction models.

TathyaNyaya comprises judgments from the Supreme Court of India (SCI) and various High Courts and is organized into four components: NyayaFacts, NyayaScrape, NyayaSimplify, and NyayaFilter. These components support a wide range of fact-centric tasks, from expert annotations and simplified factual paraphrasing to fact vs. non-fact segmentation.

Complementing the dataset, we introduce FactLegalLlama, an instruction-tuned version of LLaMa-3-8B, fine-tuned on TathyaNyaya to perform FJPE tasks. While transformer-based models are strong in predictive performance, FactLegalLlama demonstrates the ability to generate faithful and interpretable factual explanations, thus bridging predictive modeling with legal reasoning.

Our key contributions are:

- TathyaNyaya *Dataset:* We introduce the first extensively annotated, purely fact-centric dataset for judgment prediction and explanation in the Indian legal domain, structured into four components tailored for factual segmentation, simplification, and retrieval.
- *Early-Phase Legal Reasoning:* We focus on realistic and societally impactful early-phase decision-making settings where predictions are made using only the facts, reflecting constraints and needs of India's overburdened judiciary.
- FactLegalLlama *for Explanation:* We propose FactLegalLlama, an instruction-tuned LLaMa-3-8B model designed to generate faithful and fact-grounded explanations for judicial outcomes. It excels in producing coherent and semantically aligned rationales.

To foster transparency and reproducibility, we make both the dataset and model code available[1].

2 Related Work

Legal Judgment Prediction (LJP) has evolved significantly in recent years, propelled by the increasing demand for automation in judicial decision-making. Foundational works such as Aletras et al. (2016); Chalkidis et al. (2019); Feng et al. (2021) introduced outcome prediction techniques using textual court records and inspired benchmark datasets like CAIL2018 (Xiao et al., 2018) and ECHR-CASES (Chalkidis et al., 2019). These works laid the groundwork for deep learning-based models (e.g., TopJudge, MLCP-NLN) that integrate prediction with interpretability.

In the Indian legal context, substantial efforts have emerged to address LJP and adjacent legal NLP tasks. Datasets like ILDC (Malik et al., 2021), CJPE (Nigam et al., 2022), and PredEx (Nigam et al., 2024a) enabled outcome prediction and explanation generation using full-text court judgments. Beyond LJP, research has expanded into legal question answering (AILQA) (Nigam et al., 2023), rhetorical role segmentation (Ghosh and Wyner, 2019; Malik et al., 2022), legal case retrieval (Nigam et al., 2022; Santosh et al., 2025), and document drafting and summarization (Patil et al., 2024). These efforts demonstrate broader engagement with the Indian legal ecosystem, beyond judgment prediction alone.

Fact-based LJP has gained attention as a more realistic and interpretable formulation of traditional LJP. Recent works such as Nigam et al. (2024b) and Nigam and Deroy (2024) highlight the potential of grounding predictions solely on case facts, mirroring how judges might approach decisions during early case stages. However, those works rely on summarization or heuristics rather than expert-annotated factual inputs, which our work addresses directly by introducing NyayaFacts.

Cross-jurisdictional and multilingual LJP research has expanded LJP's applicability across diverse legal systems. Zhao et al. (2018) proposed LJP architectures transferable to different jurisdictions. SwissJudgmentPrediction (Niklaus et al., 2021) and HLDC (Kapoor et al., 2022) introduced multilingual and Hindi legal corpora, respectively, advancing LJP under diverse linguistic and procedural conditions. Moreover, rhetorical and structural representations of legal texts, such as through rhetorical role classification (Marino et al., 2023; Santosh et al., 2024) and event-based modeling (Feng et al., 2022), have improved model understanding and prediction consistency.

3 Task Description

Our work centers on predicting and explaining legal judgments from the Supreme Court of India (SCI) and various High Court cases using a newly introduced annotated dataset, TathyaNyaya. This dataset is the largest of its kind for factual judgment prediction and explanation in the Indian legal

[1] Anonymous GitHub Link

domain. Unlike prior approaches relying on full case texts, `TathyaNyaya` emphasizes factual information alone, reflecting more realistic conditions for automated legal decision-making.

We divide `TathyaNyaya` documents into 2 sets:

- **Single:** These documents either contain a single petition (and thus a single judgment) or multiple petitions where all decisions are identical.
- **Multi:** These documents contain multiple appeals with different outcomes. For simplicity, we convert all *partially accepted* cases into *accepted* (label 1), preserving the binary classification setup. Thus, both `single` and `multi` datasets support binary classification.

The Fact-based Judgment Prediction and Explanation (FJPE) task consists of two subtasks:

Task A: Judgment Prediction: This is a binary classification problem. Given the factual information of a legal case, the goal is to predict whether the judgment favors the appellant or not. A label of "1" denotes acceptance (including partially accepted cases), and "0" denotes complete rejection.

Task B: Rationale Explanation: This subtask involves generating a textual explanation for the predicted decision. The rationale should be grounded in the provided factual information and reflect the reasoning that supports the outcome.

Figure 2 in the Appendix illustrates the overall FJPE pipeline, outlining the stages from fact input to prediction and explanation generation using the `FactLegalLlama` model.

4 Dataset

In this research, we introduce `TathyaNyaya`, a comprehensive dataset explicitly designed for Fact-based Judgment Prediction and Explanation (FJPE) in the Indian legal domain. This dataset consists of four distinct components: (1) `NyayaFacts`: expert-annotated data that serves as the gold standard for prediction and explanation tasks, (2) `NyayaScrape`: automated fact-extracted data obtained through machine-driven processes, (3) `NyayaSimplify`: a user-friendly dataset created by paraphrasing complex legal language, and (4) `NyayaFilter`: a binary fact vs. non-fact classification dataset designed to streamline the retrieval of relevant factual information. Together, these components form the largest and most diverse factual dataset in the Indian judiciary, enabling the development and evaluation of advanced AI models for transparent and interpretable judgment prediction and explanation. By

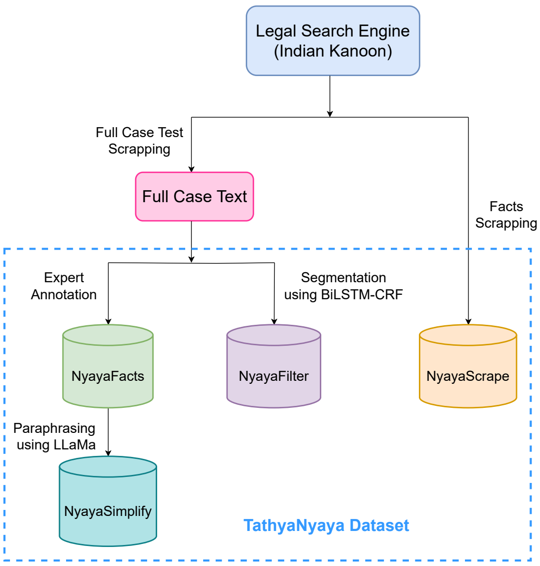

Figure 1: A high-level illustration of the `TathyaNyaya` dataset creation pipeline, showcasing the development process and interconnections of its four components.

Metric	Train (Multi)	Train (Single)	Validation	Test
NyayaFacts				
# Documents	13,629	8,216	1,197	2,389
Avg # Words	855	853	828	865
Acceptance (%)	55.20	47.66	47.45	47.72
NyayaScrape				
# Documents	8,993	3,828	548	1,095
Avg # Words	405	404	412	405
Acceptance (%)	65.77	61.44	59.85	60.55

Table 1: Statistics for `NyayaFacts` and `NyayaScrape` datasets from the `TathyaNyaya` corpus.

focusing exclusively on factual data, `TathyaNyaya` addresses a critical gap in the field, paving the way for more robust and realistic AI-driven solutions tailored to the Indian legal context.

Figure 1 illustrates the `TathyaNyaya` dataset creation pipeline. It provides a high-level overview of how each component which is derived, from expert-curated facts and machine-driven extraction, to fact segmentation and paraphrasing. This end-to-end pipeline ensures that the final dataset captures both breadth and depth in factual legal information, supporting the FJPE task.

4.1 Dataset Compilation and Statistics

The compilation process involved collecting approximately 16,000 judgments from the Supreme Court of India (SCI) and various High Courts through IndianKanoon[2], a widely used legal search engine known for its comprehensive repository of

[2] https://indiankanoon.org/

Indian legal documents. These judgments were then categorized into the following components:

4.1.1 `NyayaFacts`

`NyayaFacts` comprises a subset of SCI and High Court judgments carefully annotated by legal experts. These annotations highlight key factual segments that significantly influence judicial outcomes, serving as high-quality ground truth for both judgment prediction and rationale explanation. After refining and preprocessing, this subset serves as the gold standard for evaluating prediction and explanation tasks.

In particular, the validation and test data were derived from the `NyayaFacts Single` subset to maintain consistency during evaluation, while the training data include both single and multi-case judgments, offering a broad learning landscape. Table 1 provides comprehensive statistics. `NyayaFacts` thus provides a high-quality benchmark for both judgment prediction and explanation tasks.

4.1.2 `NyayaScrape`

`NyayaScrape` comprises judgments sourced from the Indiankanoon website, where cases are automatically segmented into various categories such as facts, issues, conclusions, and assessments of how the courts have treated certain elements (e.g., "Negatively Viewed by Court," "Relied by Party," "Accepted by Court"). Although these segments aim to provide structured insights, the labels are not entirely reliable. They are generated by automated tools rather than human legal experts, resulting in potential inconsistencies and may introduce noise. Moreover, not all judgments contain every type of label, further complicating the data's uniformity.

Despite these limitations, `NyayaScrape` offers valuable machine-derived factual extractions that enable us to compare expert-driven annotations with automated processes. This comparison helps assess the reliability, quality, and shortcomings of model-based fact identification and segmentation. Document-level statistics and comparisons against `NyayaFacts` are provided in Table 1.

4.1.3 `NyayaSimplify`

`NyayaSimplify` aims to enhance model performance and interpretability by transforming complex legal texts into simplified, paraphrased versions. Since most LLMs are pre-trained on general-purpose corpora and not on legal-specific jargon, they often struggle with the dense and domain-specific language found in court judgments. To address this, we paraphrased the `NyayaFacts` test data using the instruction-tuned LLaMA-3-70B-Instruct model. This transformation preserves the factual and legal integrity of the original content while expressing it in more accessible, human-readable language.

Metric	Train	Validation	Test
Facts			
# Documents	13,629	1,197	2,389
# Sentences	3,62,658	30,561	56,240
Avg # Words	29.00	29.00	34.00
Avg # Facts/Document (%)	23.6	23.03	22.7
Overall Facts (%)	19.16	19.09	18.46
Non-Facts			
# Documents	13,629	1,197	2,389
# Sentences	15,29,998	1,29,543	2,48,433
Avg # Words	28.00	28.00	30.00
Avg # Non-Facts/Document (%)	76.4	76.97	77.3
Overall Non-Facts (%)	80.84	80.91	81.54

Table 2: Comparison of factual vs. non-factual statistics used during BiLSTM-CRF classifier training for the `NyayaFilter` dataset.

The resulting dataset allows us to evaluate whether simplifying legal language helps general-purpose models better understand and reason about legal facts. While most dataset statistics remain consistent with `NyayaFacts`, the average word count is notably reduced, indicating a successful simplification. Our findings suggest that simplification improves both the accuracy and interpretability of models on FJPE tasks. Prompt template used for paraphrasing is included in Appendix Table 8.

4.1.4 `NyayaFilter`

`NyayaFilter` addresses the challenges of manual annotation by employing a BiLSTM-CRF model to classify sentences as either factual (1) or non-factual (0). This binary classification replaces the traditional multi-label approach, simplifying the task while maintaining a focus on essential factual information. The model was trained on `NyayaFacts Single` data, with validation and testing on the corresponding splits. This approach achieved approximately 90% accuracy in separating factual statements, as shown in Table 2. This dataset streamlines the retrieval process for FJPE tasks and enables scalable fact extraction.

4.2 Annotation Methodology and Quality Assurance

4.2.1 Expert Participation

The annotation process for `NyayaFacts` was carried out by a team of 10 legal experts, comprising advanced third- and fourth-year law students from premier Indian law colleges. These individuals were chosen based on their academic standing, legal reasoning skills, and familiarity with judicial processes, ensuring that the annotations reflected high-quality and domain-relevant insights.

4.2.2 Timeline and Workload Distribution

The annotation process was conducted over an extended period (April 1, 2022, to October 30, 2023), reflecting the complexity and precision required to analyze diverse legal texts. Each annotator was assigned approximately 30 judgment documents per week, a volume that balanced efficiency with attention to detail. This measured pace allowed the annotators to thoroughly examine the factual segments without compromising quality.

4.2.3 Annotation Protocol

The annotators were tasked with identifying and extracting specific judgment segments that contained factual information, without personal interpretation or summarization. This approach preserved the authenticity of the annotations, ensuring that they faithfully represented the judicial reasoning within each document.

4.2.4 Quality Control Framework

To maintain annotation consistency and reliability, a multi-layered quality control mechanism was implemented:

- **Initial Review:** Each case was initially annotated by a single expert. This ensured efficiency while maintaining focus on factual segments. Subsequently, the annotations underwent multiple validation layers.
- **Senior Expert Validation:** Discrepancies or ambiguous annotations were escalated to a review panel comprising senior legal practitioners, who provided final judgments on contentious segments, enhancing the reliability of the final annotations.
- **Training and Alignment Meetings:** Regular training sessions and coordination meetings were conducted to align all annotators on annotation protocols, legal conventions, and factual identification criteria. These interactive forums helped minimize subjectivity, solidify common standards, and maintain uniform annotation quality throughout the project's duration.
- **Inter-Annotator Agreement (IAA) Evaluation:** To evaluate the stability and reproducibility of our quality control framework, we conducted a re-annotation study on a randomly selected subset of documents using the same annotators and protocols. We computed IAA scores across multiple metrics, including Intraclass Correlation Coefficient (ICC) (Koo and Li, 2016), Krippendorff's Alpha (Krippendorff, 2011), and Pearson Correlation Coefficient (Cohen et al., 2009). The results demonstrate high agreement among annotators, suggesting strong consistency and reliability of the annotation process. Full IAA results and agreement tables are provided in Appendix B.

5 Methodology

In this section, we present our overall methodology for extracting factual segments from legal judgments, training our custom model `FactLegalLlama` for FJPE, and finally addressing both the prediction-only and prediction-with-explanation tasks. We also detail the prompts we used and instruction-tuning strategies employed to refine our model's outputs.

5.1 Fact Extraction from Full Judgments

To prepare the dataset for Fact-based Judgment Prediction and Explanation (FJPE), we first extracted the factual statements from full-text legal judgments. We adopted a streamlined binary classification approach by fine-tuning a BiLSTM-CRF model (Ghosh and Wyner, 2019), a previous state-of-the-art (SoTA) model for semantic segmentation of legal documents. Instead of using the original multi-class rhetorical role framework, which distinguishes between roles such as issue, statute, precedent, and argument, we simplified the task by treating all non-factual segments as a single class labeled "non-facts."

This transformation into a binary classification problem enabled the model to focus solely on identifying factual segments critical to judgment prediction. Training was conducted using the `NyayaFacts` multi, which provided expert-annotated labels for factual and non-factual segments. By isolating the facts, we laid the groundwork for developing AI models capable of making decisions and generating explanations based solely on factual data. This preprocessing ensured that the subsequent models trained on the dataset remained focused on the most relevant and actionable information in legal cases.

5.2 Training FactLegalLlama

The FactLegalLlama model, based on the LLaMa-3-8B architecture, was fine-tuned specifically for the FJPE task using NyayaFacts. The training process involved instruction-tuning with a diverse set of 16 templates designed to guide the model in judgment prediction and explanation tasks. We utilized low-rank adaptation (LoRA) to optimize model training on limited computational resources. Training parameters, such as quantization to 4-bit precision and gradient accumulation, ensured efficient usage of resources while maintaining model performance.

To further enhance its capabilities, FactLegalLlama was fine-tuned with both prediction-only and prediction-with-explanation tasks, enabling it to handle a wide range of factual judgment scenarios. The fine-tuning process emphasized the use of simplified prompts to ensure clarity and relevance in the generated outputs.

5.3 Fact-Based Judgment Prediction

5.3.1 Language Model-Based Approach

For baseline comparisons, we utilized transformer-based models like InLegalBERT (Paul et al., 2023), and XLNet Large (Yang et al., 2019) for binary classification. Due to the token length constraints of these models, we adopted a chunking strategy by dividing documents into 512-token segments with a 100-token overlap to preserve context. Chunk-level predictions were aggregated to generate final case-level predictions.

5.3.2 Large Language Model-based Approach

We utilized FactLegalLlama, our instruction-tuned LLaMa-3-8B model (Dubey et al., 2024), for judgment prediction-only instructions, where the model predicts judicial outcomes solely based on the factual inputs. The training data from TathyaNyaya was used to train the factual prediction context, emphasizing precision.

5.4 Prediction with Explanation (FJPE)

For the combined task of prediction and explanation, we employed FactLegalLlama with modified instruction prompts. Instructions guided the model to first predict the outcome and then generate a rationale grounded in the provided factual data.

5.5 Prompts Used

Prompts for both prediction and explanation tasks were carefully designed the prompts. For prediction-only tasks, the prompts instructed the model to output a binary decision. For prediction-with-explanation tasks, the prompts included directives to explain the reasoning behind the prediction. These templates are detailed in Table 7 in the Appendix.

5.6 Instruction Sets

The fine-tuning process for FactLegalLlama involved using a diverse set of 16 instruction templates for judgment prediction and explanation. These templates ensured the model could generalize effectively across a wide range of cases and factual scenarios. The complete list of instruction sets used for tuning is in Table 9 in the Appendix.

6 Evaluation Metrics

To rigorously assess the performance of our models on judgment prediction and factual explanations in the TathyaNyaya test dataset, we employed a suite of evaluation metrics. For judgment prediction, we report Macro Precision, Macro Recall, Macro F1, and Accuracy. For evaluating the quality of explanations, both quantitative and qualitative methods were applied.

1. **Lexical-Based Evaluation:** We used traditional lexical similarity metrics, including ROUGE-1/2/L (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005). These metrics measure word overlap and sequence alignment between generated explanations and reference texts, providing a quantitative measure of the accuracy of lexical content.
2. **Semantic Similarity Evaluation:** To assess the semantic alignment of the generated explanations, we applied BERTScore (Zhang et al., 2020), which evaluates semantic similarity between the generated text and reference explanations. Additionally, BLANC (Vasilyev et al., 2020) was utilized to estimate the contextual relevance and coherence of the generated text in the absence of a gold-standard reference.
3. **Expert Evaluation:** To further validate the interpretability and legal coherence of the generated explanations, we conducted a small-scale expert evaluation. Legal experts rated the model-generated explanations on a 1–10 Likert scale based on three key criteria: factual accuracy, legal relevance, and completeness. A score of 1 indicates the explanation is irrelevant or misleading, while 10 denotes that the explanation

Model	Macro Precision	Macro Recall	Macro F1	Accuracy	Training Data
Results on NyayaFacts Test Data					
InLegalBert	0.5934	0.5936	0.5935	0.5932	NyayaFacts Single
XLNet_Large	**0.6064**	**0.6040**	**0.6052**	**0.6061**	
FactLegalLlama	0.5416	0.5312	0.5036	0.5386	
InLegalBert	0.6001	0.5836	0.5917	0.5740	NyayaFacts Multi
XLNet_Large	**0.6145**	**0.5965**	**0.6054**	**0.5908**	
FactLegalLlama	0.5390	0.5368	0.5318	0.5401	
InLegalBert	0.5480	0.5192	0.5332	0.5082	NyayaScrape Single
XLNet_Large	**0.5807**	**0.5781**	**0.5794**	**0.5756**	
FactLegalLlama	0.5139	0.5122	0.4922	0.5042	
InLegalBert	0.5735	0.5269	0.5492	0.5157	NyayaScrape Multi
XLNet_Large	**0.5935**	**0.5878**	**0.5906**	**0.5842**	
FactLegalLlama	0.4951	0.4966	0.4516	0.4884	
Results on NyayaScrape Test Data					
InLegalBert	0.6718	0.5748	0.6195	0.6521	NyayaScrape Single
XLNet_Large	**0.6754**	**0.6394**	**0.6569**	**0.6849**	
FactLegalLlama	0.5574	0.5372	0.5191	0.6045	
InLegalBert	0.7976	0.7268	0.7606	0.7717	NyayaScrape Multi
XLNet_Large	**0.8098**	**0.7781**	**0.7936**	**0.8055**	
FactLegalLlama	0.5439	0.5317	0.5177	0.5877	
InLegalBert	**0.6237**	0.5243	0.5697	**0.6183**	NyayaFacts Single
XLNet_Large	0.5433	0.5282	0.5357	0.5918	
FactLegalLlama	0.5832	**0.5868**	**0.5792**	0.5840	
InLegalBert	**0.6784**	0.5027	0.5775	0.6073	NyayaFacts Multi
XLNet_Large	0.6124	0.5129	0.5583	0.6119	
FactLegalLlama	0.6541	**0.6583**	**0.6552**	**0.6651**	

Table 3: Performance metrics of models evaluated on NyayaFacts and NyayaScrape test data. Each block shows results obtained by training on either NyayaFacts or NyayaScrape data (single or multi variants), then testing on corresponding subsets. The best scores in each section are highlighted in bold.

is highly accurate and legally insightful.

7 Results and Analysis

In this section, we present and interpret the performance of our models across various datasets and experimental settings. We focus first on raw judgment prediction results using `NyayaFacts` and `NyayaScrape` data, then on the performance improvements or trade-offs observed in the `NyayaFilter` and `NyayaSimplify` settings. Finally, we analyze the explanation quality generated by `FactLegalLlama` using both lexical and semantic metrics.

7.1 Performance on NyayaFacts and NyayaScrape

We begin by examining model performances on the `NyayaFacts` and `NyayaScrape` test sets, as reported in Table 3. Each model (InLegalBERT, XLNet_Large, and FactLegalLlama) was evaluated under different training configurations, including Single and Multi.

Language Model-Based Baselines: Across both `NyayaFacts` and `NyayaScrape` test sets, XLNet_Large consistently outperforms InLegalBERT on macro Precision, Recall, F1, and Accuracy metrics. For instance, when trained on `NyayaFacts Single`, XLNet_Large surpasses InLegalBERT's macro F1 and Accuracy. This trend persists in most training and testing configurations, highlighting XLNet_Large's robust capability for factual judgment prediction in the given domain.

FactLegalLlama's Prediction-Only Performance: `FactLegalLlama`, while instruction-tuned for outcome prediction, lags behind the transformer-based baselines in raw prediction performance. For example, when trained on `NyayaFacts Single` and tested on `NyayaFacts`,

Model	Macro Precision	Macro Recall	Macro F1	Accuracy	Training Data
Results on NyayaFilter Test Data					
InLegalBert	**0.5870**	**0.5857**	**0.5864**	**0.5885**	NyayaFacts Single
XLNet_Large	0.5805	0.5775	0.5790	0.5818	
InLegalBert	0.5886	0.5560	0.5719	0.5421	NyayaFacts Multi
XLNet_Large	**0.5977**	**0.5874**	**0.5925**	**0.5797**	
InLegalBert	0.5342	0.5180	0.5260	0.5023	NyayaScrape Single
XLNet_Large	**0.5577**	**0.5509**	**0.5543**	**0.5429**	
InLegalBert	**0.5789**	**0.5409**	**0.5592**	**0.5249**	NyayaScrape Multi
XLNet_Large	0.5581	0.5364	0.5470	0.5224	
Results on NyayaSimplify Test Data					
InLegalBert	**0.6199**	**0.6197**	**0.6198**	0.6167	NyayaFacts Single
XLNet_Large	0.6179	0.6169	0.6174	**0.6200**	
InLegalBert	**0.6222**	0.5986	**0.6102**	0.5839	NyayaFacts Multi
XLNet_Large	0.6160	**0.6002**	0.6080	**0.5878**	
InLegalBert	0.5760	0.5311	0.5526	0.5061	NyayaScrape Single
XLNet_Large	**0.5864**	**0.5845**	**0.5854**	**0.5789**	
InLegalBert	0.5659	0.5215	0.5428	0.4950	NyayaScrape Multi
XLNet_Large	**0.5978**	**0.5891**	**0.5934**	**0.5789**	

Table 4: Model performance on NyayaFilter and NyayaSimplify test datasets. For NyayaFilter, results illustrate how automatically retrieved factual data affects performance when models are trained on NyayaFacts or NyayaScrape datasets. For NyayaSimplify, results show the impact of paraphrasing complex legal texts into simpler language. Bolded scores indicate the best performance in each section.

Training Data	Testing Data	Lexical Based Evaluation					Semantic Evaluation	
		R1	R2	RL	BLEU	METEOR	BERT Score	BLANC
No Training	NyayaFacts	0.28	0.10	0.14	0.04	0.17	0.53	0.08
No Training	NyayaScrape	0.19	0.08	0.13	0.04	0.18	0.48	0.09
NyayaFacts Single	NyayaFacts	0.32	0.11	0.19	0.04	0.18	0.58	0.10
NyayaFacts Multi	NyayaFacts	**0.34**	**0.11**	**0.20**	**0.05**	**0.21**	**0.58**	**0.10**
NyayaScrape Single	NyayaScrape	0.12	0.05	0.09	0.02	0.10	0.39	0.06
NyayaScrape Multi	NyayaScrape	0.17	0.08	0.13	0.03	0.13	0.45	0.08
NyayaSimplify	NyayaSimplify	0.28	0.08	0.18	0.02	0.17	0.56	0.07

Table 5: Performance of FactLegalLlama on the FJPE task. The base model is LLaMa-3-8B. "No Training" indicates results from the unmodified (vanilla) model. Other rows show improvements after fine-tuning with different subsets of the `TathyaNyaya` data. Bolded values represent the best performance within a given evaluation scenario.

it obtains a macro F1 of 0.5036 compared to XLNet_Large's 0.6052. A similar gap is observed across other splits. Although `FactLegalLlama` underperforms in direct classification metrics, its strength lies in generating explanations, as discussed later.

Single vs. Multi Cases: Both baselines and `FactLegalLlama` exhibit more stable performance on the Single subsets compared to the Multi subsets. The complexity introduced by multiple petitions with varying outcomes in the Multi cases reduces overall accuracy and F1 scores, emphasizing the challenge of fact-based judgment prediction in more intricate legal scenarios.

7.2 Impact of Fact Retrieval (NyayaFilter) and Text Simplification (NyayaSimplify)

Table 4 reports model performances on the `NyayaFilter` and `NyayaSimplify` test datasets. These results highlight how the preprocessing choices affect model accuracy on automatic fact retrieval and paraphrasing complex legal texts.

NyayaFilter Results: When comparing NyayaFilter results to the original `NyayaFacts` and NyayaScrape sets, we see that while performance can fluctuate, some models benefit from training on data where fact and non-fact segments are clearly distinguished. For example, on the NyayaFilter test set derived from NyayaFacts Single, InLegalBERT attains a macro F1 of 0.5864, maintaining competitive performance. XLNet_Large, although not always the top performer here, still sustains a strong baseline. These findings suggest that automatically retrieved factual subsets can be used without severely degrading model performance.

NyayaSimplify Results: Paraphrasing complex legal language into simpler text (the NyayaSimplify scenario) generally helps models retain or slightly improve performance. For instance, with NyayaFacts Single, InLegalBERT reaches a macro F1 of 0.6198 and XLNet_Large hits an Accuracy of 0.6200 on the simplified data, both representing small yet noteworthy improvements compared to their performance on the original complex texts. This trend indicates that reducing linguistic complexity can aid models in understanding and classifying factual statements more accurately.

7.3 Quality of Explanations from FactLegalLlama

Table 5 presents the evaluation of FactLegalLlama on the explanation generation task, measured through both lexical and semantic metrics. We compare a "No Training" scenario (using the LLaMa-3-8B model) with fine-tuned versions on different subsets of TathyaNyaya data.

Fine-tuning Benefits: Fine-tuning LLaMa-3-8B (`FactLegalLlama`) on factual data substantially improves its explanation quality. For `NyayaFacts`, training on the Multi subset yields the strongest results, with Rouge-1 at 0.34 and a BERTScore of 0.58, outperforming both the "No Training" scenario and the Single subset training. This suggests that exposure to more complex, multi-petition cases helps the model generate richer, more contextually sensitive explanations.

Domain-Specific Fine-tuning: The contrast between "No Training" and the various training configurations highlights the necessity of domain-specific adaptation. Without fine-tuning, the model's explanations remain weak and less aligned with factual inputs, as indicated by lower Rouge and BLEU scores. After training with `NyayaFacts Multi`, the model better captures the underlying legal rationale, producing explanations that align more closely with reference annotations.

8 Conclusions and Future Work

We introduced `TathyaNyaya`, a fact-focused dataset for judgment prediction and explanation within the Indian legal domain, and `FactLegalLlama`, an instruction-tuned model delivering fact-grounded rationales. By emphasizing factual content rather than full judgments, `TathyaNyaya` aligns more closely with actual legal decision-making scenarios, while `FactLegalLlama` highlights the value of coupling predictive accuracy with transparent explanations. Preprocessing steps such as fact filtering and paraphrasing further enhance model clarity and performance, and domain-specific fine-tuning proves essential for capturing legal subtleties. Future work may extend these findings to other jurisdictions, refine fact extraction techniques, integrate, and interpretability frameworks. These efforts collectively advance transparent, accessible, and reliable AI-assisted judicial processes.

Limitations

This study faced several limitations that influenced both the scope and outcomes of our research. A key constraint was the reliance on a 4-bit quantized model due to resource limitations, which restricted our ability to experiment with larger parametric models, such as 70B or 40B parameter LLMs. Additionally, the high computational costs and token limitations associated with cloud-based services further hindered our capacity to perform extensive inference and fine-tuning. This restricted exploration may have limited the depth of insights and performance metrics achievable with FactLegalLlama.

The model's performance on scrapped datasets was also not fully evaluated due to configuration constraints, leaving gaps in understanding its generalizability to non-annotated factual data. Furthermore, challenges such as hallucinations in generative outputs and maintaining factual consistency in explanations remain unresolved, which can impact the reliability of the model in real-world legal applications.

Lastly, the dataset used in this study comprises only English-language judgments, which limits its applicability in multilingual contexts, especially in jurisdictions where regional languages dominate legal proceedings. This exclusion highlights the need for more inclusive datasets that reflect the linguistic diversity of legal documents in India and beyond.

These limitations underscore the challenges of applying LLMs to specialized legal tasks such as judgment prediction and explanation. They also point to areas requiring further research, including resource optimization, multilingual dataset development, and enhancing the factual consistency and reasoning capabilities of AI models.

Ethics Statement

This research was conducted with a strong commitment to ethical considerations, particularly given the sensitive nature of legal data and the implications of deploying AI in legal contexts. The `TathyaNyaya` dataset, central to this study, was compiled from publicly accessible sources, such as Indian legal search engines, ensuring adherence to data privacy and usage regulations. To further safeguard privacy, we removed identifiable meta-information, including judge names, case titles, and case IDs, from the dataset.

The computational resources used for model training and evaluation were obtained through ethical and legitimate means. These resources were either institutional or subscribed services, ensuring compliance with licensing agreements and financial support for these platforms. By adhering to these practices, we ensured that our research activities aligned with sustainable and lawful resource usage.

Transparency and reproducibility were foundational principles of this study. The `TathyaNyaya` dataset and the code for FactLegalLlama will be made publicly available, enabling researchers to replicate and extend our findings. This open-access approach is intended to foster collaboration within the research community and drive further advancements in AI-assisted legal decision-making.

We recognize the potential societal impact of AI applications in the legal domain, particularly regarding fairness, accountability, and the risk of misuse. Our models are explicitly designed to assist legal professionals rather than replace human judgment, emphasizing the necessity of human oversight in AI-assisted decision-making processes. As we continue this line of research, we remain vigilant in addressing ethical challenges and aligning our efforts with principles of fairness, transparency, and societal benefit.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ computer science*, 2:e93.

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.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. *Association for Computational Linguistics (ACL)*.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yi Feng, Chuanyi Li, Jidong Ge, Bin Luo, and Vincent Ng. 2021. Recommending statutes: A portable method based on neural networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2):1–22.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664, Dublin, Ireland. Association for Computational Linguistics.

Saptarshi Ghosh and Adam Wyner. 2019. Identification of rhetorical roles of sentences in indian legal judgments. In *Legal Knowledge and Information Systems: JURIX 2019: The Thirty-second Annual Conference*, volume 322, page 3. IOS Press.

Arnav Kapoor, Mudit Dhawan, Anmol Goel, Arjun T H, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, and Ashutosh Modi. 2022. HLDC: Hindi legal documents corpus. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3521–3536, Dublin, Ireland. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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.

Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Angshuman Hazarika, Shubham Kumar Nigam, Arnab Bhattacharya, and Ashutosh Modi. 2022. Semantic segmentation of legal documents via rhetorical roles. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 153–171, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Gabriele Marino, Daniele Licari, Praveen Bushipaka, Giovanni Comandé, Tommaso Cucinotta, et al. 2023. Automatic rhetorical roles classification for legal documents using legal-transformeroverbert. In *CEUR WORKSHOP PROCEEDINGS*, volume 3441, pages 28–36. CEUR-WS.

National Judicial Data Grid. 2024. National judicial data grid statistics. `https://www.njdg.ecourts.gov.in/njdgnew/index.php`.

Shubham Nigam, Anurag Sharma, Danush Khanna, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2024a. 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 and virtual meeting. Association for Computational Linguistics.

Shubham Kumar Nigam and Aniket Deroy. 2024. Fact-based court judgment prediction. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, page 78–82, New York, NY, USA. Association for Computing Machinery.

Shubham Kumar Nigam, Aniket Deroy, Subhankar Maity, and Arnab Bhattacharya. 2024b. 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, Navansh Goel, and Arnab Bhattacharya. 2022. nigam@ coliee-22: Legal case retrieval and entailment using cascading of lexical and semantic-based models. In *JSAI International Symposium on Artificial Intelligence*, pages 96–108. Springer.

Shubham Kumar Nigam, Shubham Kumar Mishra, Ayush Kumar Mishra, Noel Shallum, and Arnab Bhattacharya. 2023. Legal question-answering in the indian context: Efficacy, challenges, and potential of modern ai models. *arXiv preprint arXiv:2309.14735*.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. *arXiv preprint arXiv:2110.00806*.

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.

Atharv Patil, Kartik Bapna, and Ayush Shah. 2024. Legal docgen using ai: Your smart doc generator. *International Journal of Novel Research and Development*, 9(5):536–543.

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2023. Pre-trained language models for the legal domain: A case study on indian law. In *Proceedings of 19th International Conference on Artificial Intelligence and Law - ICAIL 2023*.

T. Y. S. S. Santosh, Isaac Misael Olguín Nolasco, and Matthias Grabmair. 2025. Lecopcr: Legal concept-guided prior case retrieval for european court of human rights cases. *Preprint*, arXiv:2501.14114.

TYSS Santosh, Apolline Isaia, Shiyu Hong, and Matthias Grabmair. 2024. Hiculr: Hierarchical curriculum learning for rhetorical role labeling of legal documents. *arXiv preprint arXiv:2409.18647*.

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

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

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.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

A Experimental Setup and Hyper-parameters

In this section, we detail the experimental configurations, training procedures, and hyper-parameters employed to develop and evaluate our models. We first describe the training of transformer-based baseline models for fact-based judgment prediction, then outline the instruction-tuning process used to adapt `FactLegalLlama` for both prediction-only and prediction-with-explanation tasks.

A.1 Transformers Training Hyper-parameters

To establish competitive baselines, we fine-tuned transformer models such as InLegalBERT and XLNet_Large on the `NyayaFacts` dataset. Each model was trained with a batch size of 16 using the AdamW optimizer (Kingma and Ba, 2014) and a learning rate of 2e-6. We ran the training for three epochs, adopting default hyper-parameter settings from the HuggingFace Transformers library. Experiments were carried out on an NVIDIA A100 40GB GPU, ensuring adequate computational resources for handling extensive legal text. This training protocol allowed the models to capture the nuances of fact-based segments and reliably predict judicial outcomes.

A.2 FactLegalLlama Instruction Fine-Tuning

To develop `FactLegalLlama`, we began with the meta-llama/Meta-Llama-3-8B base model. We applied 4-bit quantization to optimize memory usage and introduced Low-Rank Adaptation (LoRA) with a rank of 16 for parameter-efficient fine-tuning. The maximum input sequence length was set to 2,500 tokens, accommodating the substantial factual inputs characteristic of legal documents.

We employed the paged AdamW optimizer in 32-bit precision with a learning rate of 1e-4 and implemented a cosine decay learning rate scheduler for smoother convergence. Mixed-precision training (fp16) and a gradient accumulation of 4 steps were used to further manage GPU memory. We utilized a per-device batch size of 4 and trained the model for three epochs, a process that required approximately 38 hours on an NVIDIA A100 40GB GPU. Under these conditions, the model achieved a training loss of 1.5060 and a validation loss of 1.6745, indicating effective adaptation to the underlying factual patterns in the data.

A.3 Training Objectives

The instruction-based fine-tuning of `FactLegalLlama` targeted two primary objectives: fact-driven judgment prediction and fact-driven prediction with explanation. By employing a carefully designed set of instructions and incorporating LoRA-based parameter updates, the model learned to generate outcomes and accompanying rationales rooted in the factual segments. This combination of parameter-efficient fine-tuning and instruction-oriented training yielded a model well-suited for practical applications in legal NLP, balancing computational feasibility with interpretability and domain relevance.

A.4 Training Procedure for Hierarchical BiLSTM-CRF Classifier

The Hierarchical BiLSTM-CRF classifier is designed to classify sentences in legal documents into factual and non-factual categories by leveraging the hierarchical structure of the data. The model architecture comprises a word-level BiLSTM coupled with a CRF layer and a sentence-level BiLSTM. The word-level BiLSTM encodes contextual dependencies within sentences, while the CRF ensures coherence in predicted tag sequences. The sentence-level BiLSTM aggregates these representations to capture inter-sentence dependencies, enabling the model to account for both local and global patterns in the data.

Training is conducted using the AdamW optimizer with a learning rate of 2e-6, a batch size of 16, and for five epochs. A CRF-based loss function is used to optimize sequence-level tagging accuracy. During training, metrics such as precision, recall, F1-score, and loss are evaluated on a validation set after each epoch to monitor performance and ensure generalization. The model configuration includes a word embedding size of 100 and a sentence embedding size of 200, with training conducted on an NVIDIA A100 40GB GPU.

To enhance generalization, K-fold cross-validation is employed, where the dataset is split into multiple folds, and the model is trained and validated on different subsets. The average performance across folds provides a robust measure of the model's capability. Checkpoints are saved periodically during training, enabling the model to be restored for inference or further fine-tuning.

B Inter-Annotator Agreement (IAA) for Expert Evaluation

To ensure the reliability of expert-based evaluation for AI-generated legal documents, we conducted an Inter-Annotator Agreement (IAA) analysis using standard agreement metrics. This evaluation quantifies the consistency of expert assessments in scoring factual accuracy and completeness across different models.

B.1 IAA Metrics and Methodology

We employed three widely used agreement metrics:

- **Intraclass Correlation Coefficient (ICC)** (Koo and Li, 2016): Measures the absolute agreement among raters for continuous variables, commonly used for reliability assessment in research.
- **Krippendorff's Alpha** (Krippendorff, 2011): A robust reliability measure applicable to ordinal and interval data, ensuring agreement beyond chance.
- **Pearson Correlation Coefficient** (Cohen et al., 2009): Measures linear correlation between two annotators' scores, assessing the strength of agreement.

Three legal experts independently rated the generated legal documents for *Factual Accuracy* and *Completeness & Comprehensiveness* using a structured rubric. Each model's outputs were rated without knowledge of the generating model to prevent bias.

B.2 Findings and Observations

The IAA scores reveal several important trends. The highest inter-annotator agreement is observed when the Meta-Llama-3-8B model is both trained and tested on the NyayaSimplify dataset, achieving strong scores across all metrics—Fleiss' Kappa (0.3934), Cohen's Kappa (0.4065), ICC (0.8054), and Pearson Correlation (0.8725). This suggests that simplified factual inputs help the model generate outputs that are highly consistent with human judgments. In contrast, models trained on the NyayaScrape dataset perform poorly, with some metrics like Fleiss' Kappa dropping to -0.3151 and Pearson Correlation turning negative, indicating noisy data or inconsistency in capturing factual content. Although Krippendorff's Alpha appears relatively high (e.g., 0.5005) in some NyayaScrape settings, this may reflect ordinal agreement rather than overall reliability. Meanwhile, fine-tuning on NyayaFacts with multi-instance input shows a modest improvement in agreement over single-instance training, suggesting that exposure to more structured context helps the model align better with human annotators. Table 6 presents IAA scores.

Base Model	Training Data	Testing Data	Fleiss' Kappa	Cohen's Kappa	ICC	Krippendorff's Alpha	Pearson Corr.
Meta-Llama-3-8B	No Training	NyayaFacts	0.1456	0.1737	0.373	-0.0025	0.4765
	No Training	NyayaScrape	-0.0524	0.0641	0.373	-0.0025	0.4765
	NyayaFacts Single	NyayaFacts	-0.0738	-0.0488	-0.0699	-0.0028	-0.0697
	NyayaFacts Multi	NyayaFacts	0.1415	0.1936	0.2282	0.0863	0.3411
	NyayaScrape Single	NyayaScrape	-0.146	0.0428	0.0477	0.3556	0.2026
	NyayaScrape Multi	NyayaScrape	-0.3151	-0.0177	-0.0151	0.5005	-0.0929
	NyayaSimplify	NyayaSimplify	0.3934	0.4065	0.8054	-0.0136	0.8725

Table 6: Inter-Annotator Agreement (IAA) metrics for Meta-Llama-3-8B using various training and testing dataset configurations.

Template 1 (prediction only)
prompt = f“““ ### **Instructions**: Given the facts of the case,just predict the outcome as ’1’ for acceptance or ’0’ for rejection. ### **Input**: <{case_facts}> ### Response: ”””
Template 2 (prediction with explanation)
prompt = f“““ ### **Instructions**: Given the facts of the case,first predict the outcome as ’1’ for acceptance or ’0’ for rejection. Then, provide key sentences from the facts or clear reasoning that support your decision. ### **Input**: <{case_facts}> ### Response: ”””

Table 7: Prompts for Factual Judgment Prediction and Explanation used for instruction fine-tuned models. Instructions were selected based on the templates provided in Table 9.

Template 1 (Paraphrasing facts)
prompt = f“““ ### **Instructions**:You are an Indian legal expert with extensive knowledge of legal terms, statutes, and laws. Your task is to explain a legal case to your clients in simple and understandable language. Avoid legal jargon and focus on conveying the meaning of the case in everyday language, making it clear and easy for someone without legal knowledge to understand. While simplifying, ensure that the key points of the case, including the facts, legal claims, and decisions, are clearly communicated without losing any critical information. You should Preserve the key legal terms and references,Clarify complex legal processes,Avoid excessive legal jargon,Be concise but complete,Explain court actions clearly, Provide Only Paraphrased Outcome ### **Input**: Paraphrase the following text:<{case_facts}> ### Response: ”””

Table 8: Prompt for paraphrasing facts to change legal jargons to interpretable terms.

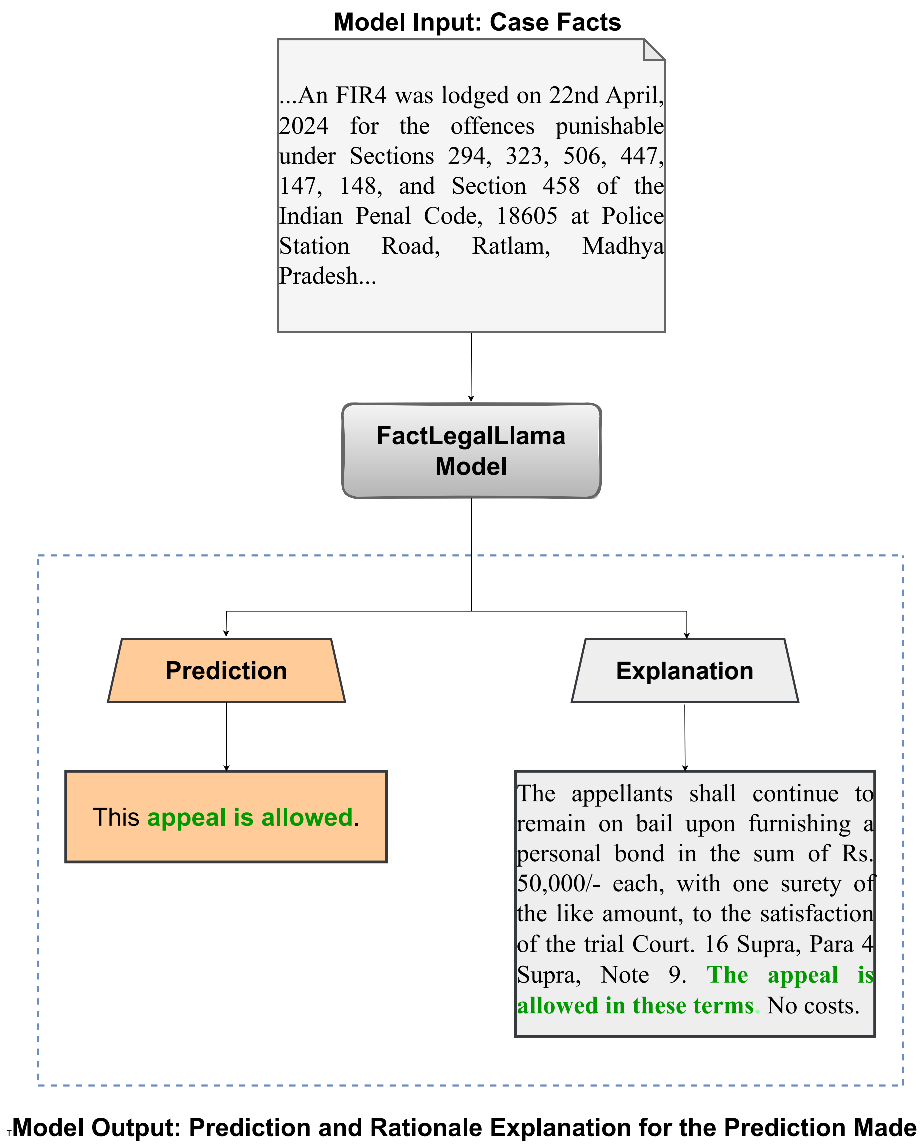

Figure 2: Illustration of the Fact-based Judgment Prediction and Explanation (FJPE) pipeline using the `FactLegalLlama` model.

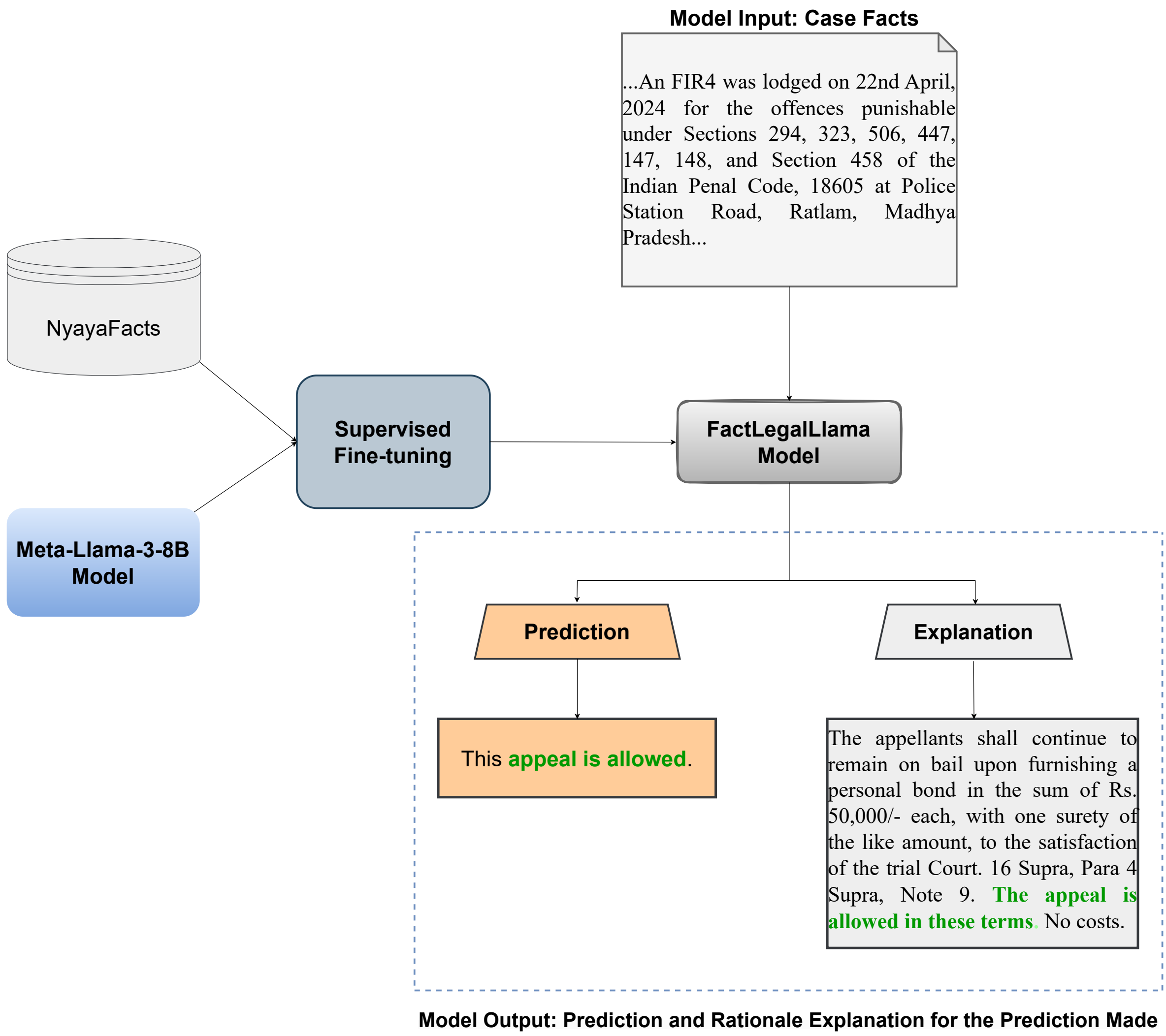

Figure 3: Training dynamics of `FactLegalLlama` for the combined judgment prediction and explanation task. The model learns to produce both the outcome and its underlying rationale directly from factual inputs, guided by instruction-based fine-tuning.

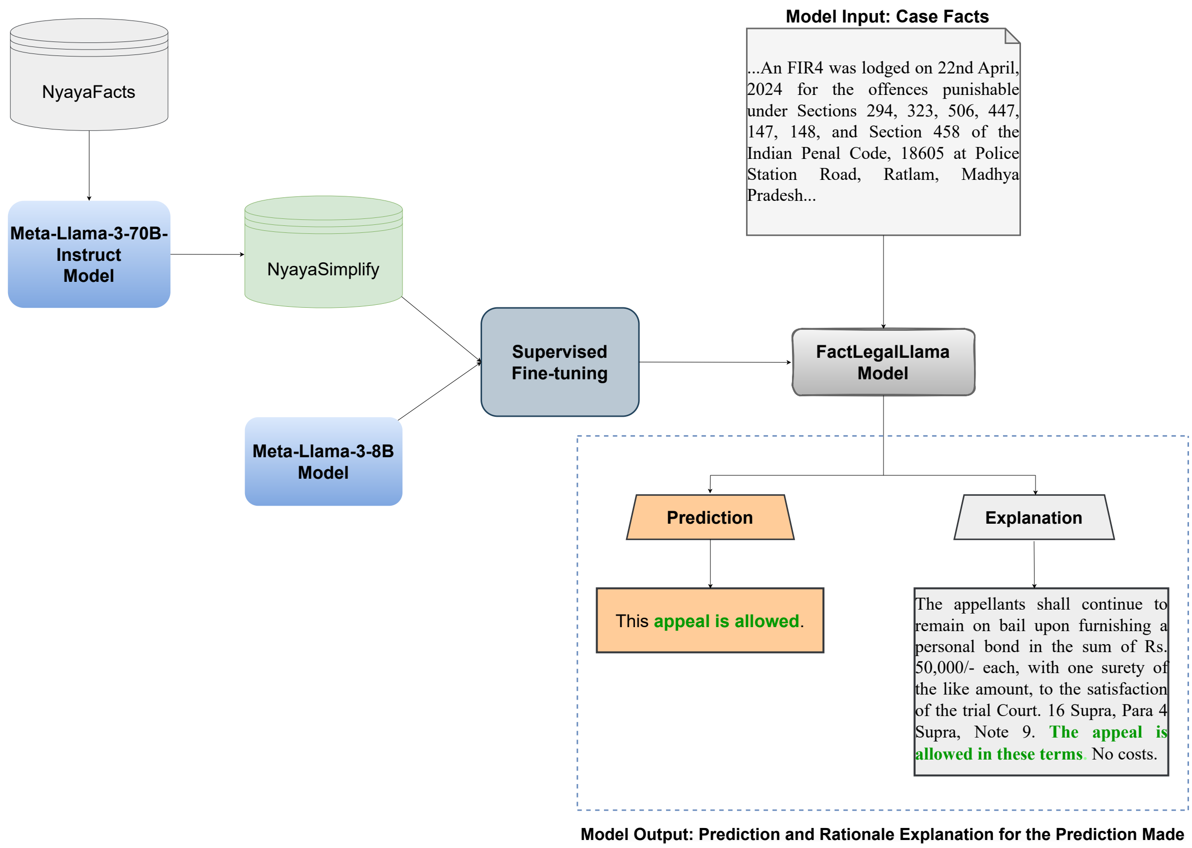

Figure 4: Overview of the simplification and fine-tuning process. First, complex legal facts are paraphrased into simpler language using LLaMA-3-70B, creating the `NyayaSimplify` dataset, followed by supervised fine-tuning (SFT) using LLaMa-3-7B for the FJPE task.

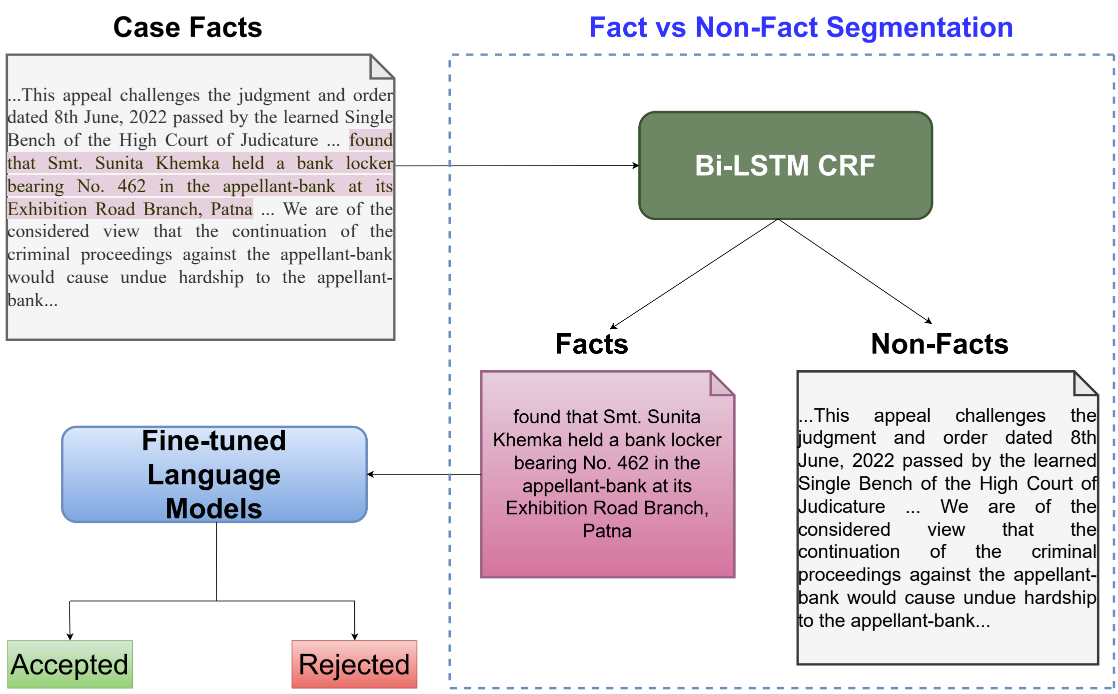

Figure 5: The Fact vs. Non-Fact segmentation framework employing a BiLSTM-CRF model. This segmentation step separates factual statements from non-factual content in legal judgments, creating the `NyayaFilter` dataset. The refined dataset is subsequently used for downstream judgment prediction and explanation tasks.

Instruction sets for Predicting the Decision	
1	Analyze the facts presented in the case and predict whether the outcome will be favorable (1) or unfavorable (0).
2	Based on the facts provided, determine the likely outcome: favorable (1) or unfavorable (0) for the appellant/petitioner
3	Review the facts of the case and predict the decision: will the court rule in favor (1) or against (0) the appellant/petitioner?
4	Considering the facts and evidence in the case, predict the verdict: is it more likely to be in favor (1) or against (0) the appellant?
5	Examine the facts of the case and forecast whether the appeal/petition is likely to be upheld (1) or dismissed (0).
6	Assess the facts of the case and provide a prediction: is the court likely to rule in favor of (1) or against (0) the appellant/petitioner?
7	Interpret the facts of the case and speculate on the court's decision: will the appeal be accepted (1) or rejected (0) based on the provided information?
8	Given the specifics of the case facts, anticipate the court's ruling: will it favor (1) or oppose (0) the appellant's request?
9	Scrutinize the facts and arguments presented in the case to predict the court's decision: will the appeal be granted (1) or denied (0)?
10	Analyze the facts presented and estimate the likelihood of the court accepting (1) or rejecting (0) the petition.
11	From the facts provided in the case, infer whether the court's decision will be favorable (1) or unfavorable (0) for the appellant.
12	Evaluate the facts and evidence in the case and predict the verdict: is an acceptance (1) or rejection (0) of the appeal more probable?
13	Delve into the case facts and predict the outcome: is the judgment expected to be in support (1) or in denial (0) of the appeal?
14	Using the case facts, forecast whether the court is likely to side with (1) or against (0) the appellant /petitioner.
15	Examine the case facts and anticipate the court's decision: will it result in an approval (1) or disapproval (0) of the appeal?
16	Based on the facts and evidence in the case, predict the court's stance: favorable (1) or unfavorable (0) to the appellant.
Instruction sets for Integrated Approach for Prediction and Explanation	
1	First, predict whether the appeal in case proceeding will be accepted (1) or not (0), and then explain the decision by identifying crucial sentences from the document.
2	Determine the likely decision of the case facts (acceptance (1) or rejection (0)) and follow up with an explanation highlighting key sentences that support this prediction.
3	Predict the outcome of the case based on the facts provided (acceptance (1) or rejection (0)) and explain your reasoning by extracting key sentences that justify the decision.
4	Evaluate the case facts to forecast the court's decision (1 for yes, 0 for no), and elucidate the reasoning behind this prediction with important textual evidence from the case.
5	Ascertain if the court will uphold (1) or dismiss (0) the appeal based on the case facts, and then clarify this prediction by discussing the critical sentences that support the decision.
6	Judge the probable resolution of the case based on the facts (approval (1) or disapproval (0)), and elaborate on this forecast by extracting and interpreting significant sentences from the case facts.
7	Forecast the likely verdict of the case (granting (1) or denying (0) the appeal) based on the facts, and rationalize your prediction by pinpointing and explaining pivotal sentences in the case document.
8	Assess the case to predict the court's ruling (favorably (1) or unfavorably (0)) based on the facts, and expound on this prediction by highlighting and analyzing key textual elements from the case facts.
9	Assess the case to predict the court's ruling (favorably (1) or unfavorably (0)) based on the facts, and expound on this prediction by highlighting and analyzing key textual elements from the case facts.
10	Conjecture the end result of the case (acceptance (1) or non-acceptance (0) of the appeal) based on the facts, followed by a detailed explanation using crucial sentences from the case facts.
11	Predict whether the case will result in an affirmative (1) or negative (0) decision for the appeal based on the facts, and then provide a thorough explanation using key sentences to support your prediction.
12	Estimate the outcome of the case (positive (1) or negative (0) for the appellant) based on the facts, and then provide a reasoned explanation by examining important sentences within the case documentation.
13	Project the court's decision (favor (1) or against (0) the appeal) based on the case facts, and subsequently provide an in-depth explanation by analyzing relevant sentences from the document.
14	Make a prediction on the court's ruling (acceptance (1) or rejection (0) of the petition) based on the case facts, and then dissect the case to provide a detailed explanation using key textual passages.
15	Speculate on the likely judgment (yes (1) or no (0) to the appeal) based on the case facts, and then delve into the case to elucidate your prediction, focusing on critical sentences.
16	Hypothesize the court's verdict (affirmation (1) or negation (0) of the appeal) based on the case facts, and then clarify this hypothesis by interpreting significant sentences from the case.

Table 9: Instruction sets for Prediction and Explanation using factual data from case proceedings.