NYAYAANUMANA and INLEGALLLAMA: The Largest Indian Legal Judgment Prediction Dataset and Specialized Language Model for Enhanced Decision Analysis

Anonymous COLING 2025 submission

Abstract

The integration of AI in legal judgment prediction (LJP) has the potential to transform the legal landscape, particularly in jurisdictions like India, where the legal system is burdened by a significant backlog of cases. This paper introduces `NyayaAnumana`, the largest and most diverse corpus of Indian legal cases compiled for LJP, encompassing a total of 702,945 preprocessed cases. `NyayaAnumana`, which combines the Hindi words "Nyay" (judgment) and "Anuman" (prediction or inference), includes a wide range of cases from the Supreme Court, High Courts, Tribunal Courts, District Courts, and Daily Orders, providing unparalleled diversity and coverage. Our dataset surpasses existing datasets like PredEx and ILDC, offering a comprehensive foundation for advanced AI research in the legal domain. In addition to the dataset, we present `INLegalLlama`, a domain-specific generative LLM tailored to the intricacies of the Indian legal system. It is developed through a two-phase training approach: injecting legal knowledge and enhancing reasoning capabilities. This method allows the model to achieve a deep understanding of legal contexts. Our experiments demonstrate that incorporating diverse court data significantly boosts model accuracy, achieving approximately 90% F1 score in prediction tasks. `INLegalLlama` not only improves prediction accuracy but also offers comprehensible explanations, addressing the need for explainability in AI-assisted legal decisions. These contributions advance both the technological and practical aspects of LJP, highlighting the importance of diverse datasets in developing effective AI solutions for the legal field.

1 Introduction

The integration of AI in legal judgment prediction (LJP) has the potential to revolutionize the legal landscape, offering both challenges and opportunities. In India, where the legal system faces a significant backlog of millions of pending cases, the application of AI in LJP can be crucial for enhancing efficiency and accessibility. However, the complexity and diversity of legal cases present significant challenges in developing effective AI models. To address these challenges, we introduce `NyayaAnumana`, the largest and most diverse corpus of Indian legal cases compiled for LJP, covering various levels of the judiciary. `NyayaAnumana` is derived from the Hindi words "Nyay," meaning judgment, and "Anuman," meaning prediction or inference. This name reflects the core focus of the dataset on legal judgments and their corresponding predictions, emphasizing its role in facilitating AI-driven insights within the legal domain. It stands out when compared to several popular corpora used in legal judgment prediction, surpassing them in terms of the number of cases, diversity of court levels, and comprehensive coverage of Indian legal proceedings. This richness and variety offer a unique opportunity to explore and predict legal judgments with greater accuracy and nuance than ever before. Table 1 provides a comparative overview of `NyayaAnumana` with other notable datasets.

Corpus	Language	Jurisdiction	# of Cases	Avg # of Tokens	# of labels w.r.t Subtask
FCCR (Sulea et al., 2017)	French	France	126,865	-	Court Decision(6 & 8)
CAIL (Xiao et al., 2018)	Chinese	China	2,676,075	-	Law Article (183) Charge (202)
ECHR (Chalkidis et al., 2019)	English	Europe	11,478	2421	Violation (2) Law Article (66)
ECHR (Chalkidis et al., 2021)	English	Europe	11,000	-	Alleged Law Article (40) Violation (2) Law Article (40)
SJP (Niklaus et al., 2021)	German French Italian	Switzerland	49,883 31,094 4,292	850	Court Decision (2)
ILDC (Malik et al., 2021a)	English	India	34,816	3231	Court Decision (2)
HLDC (Kapoor et al., 2022)	Hindi	India	340,280	764	Bail Prediction (2)
BCD (Lage-Freitas et al., 2022)	Portuguese	Brazil	4,043	119	Court Decision (3)
PredEx (Nigam et al., 2024)	English	India	15,222	4,504	Court Decision (2)
(Our dataset) `NyayaAnumana`	English	India	702,945	2,061.17	Court Decision (2 & 3)

Table 1: Comparative overview of widely used legal judgment prediction datasets. Entries marked with '-' denote unknown or unavailable information.

We develop `INLegalLlama`, a domain-specific generative LLM, tailored to the intricacies of the Indian legal domain. `INLegalLlama` is trained not only to enhance the accuracy of legal judgment predictions but also to provide understandable explanations for these decisions. This dual approach caters to the needs of legal experts who seek not just accuracy but also rationale in AI-assisted decisions.

Our work is distinguished by several key contributions that mark significant advancements in the field of legal AI:

1. Largest Indian Legal Corpus for Judgment Prediction: We introduce `NyayaAnumana`, the most extensive legal corpus in India for LJP, encompassing a wide range of courts and orders, ensuring diversity and comprehensive coverage in the dataset.

2. Generative Large Language Model for Prediction and Explanation: Addressing the crucial need for explainability, we develop `INLegalLlama`, a generative LLM that not only predicts outcomes but also provides comprehensible explanations, enhancing the trustworthiness and utility of AI in legal decision-making.

3. Domain-Specific Transformer-Based Classifiers: We fine-tuned several transformer-based models on `NyayaAnumana` to enhance prediction accuracy for the Indian legal domain. This includes testing model performance across different courts and hierarchical data to evaluate the impact of data inclusion.

4. Evaluation on Temporal Data: We tested the model performance on temporal data, assessing its effectiveness on future or unseen data to ensure robustness and generalization capabilities over time.

5. Expert Evaluation and Validation: We conducted a thorough evaluation using a Likert score scale to assess the effectiveness of our system. This evaluation utilized the PredEx test dataset (Nigam et al., 2024) and ILDC_expert (Malik et al., 2021a), offering crucial insights into how our AI models perform compared to human expert benchmarks.

Our research aspires to deliver a sophisticated AI-driven system for legal judgment prediction and explanation, specifically designed for the Indian judicial system. This initiative not only represents a technological leap but also aims to tackle the critical issue of case backlog in India. We anticipate that our work will enhance the efficiency and transparency of the legal process and stimulate further research and development in the realm of AI-assisted legal technologies. To ensure reproducibility, we have made the `NyayaAnumana` dataset and the code for our `INLegalLlama` models accessible via an anonymous link[1].

2 Related Work

The field of Legal Judgment Prediction (LJP) has advanced significantly, driven by various research initiatives addressing the complexities of forecasting legal case outcomes. Traditionally, this task has been the domain of legal professionals, but the development of LJP systems offers potential benefits for both practitioners and the public, particularly in light of overwhelming caseloads in many jurisdictions.

Foundational studies, such as those by (Aletras et al., 2016), (Chalkidis et al., 2019), and (Feng et al., 2021), have laid the groundwork for LJP by outlining its methodologies and emphasizing the importance of explainability in AI-generated predictions. The availability of benchmark datasets, including CAIL2018 (Xiao et al., 2018; Zhong et al., 2020), ECHR-CASES (Chalkidis et al., 2019; Aletras et al., 2016; Medvedeva et al., 2020), and others, has propelled research forward, inspiring models like TopJudge and MLCP-NLN. Despite progress, a gap remains between machine and human performance, and many datasets lack sufficient academic attention.

In the Indian context, notable contributions include PredEx (Nigam et al., 2024), (Huang et al., 2023) and ILDC (Malik et al., 2021b), which highlight the integration of AI in legal judgments and the need for transparent explanations. Additionally, Tiwari et al. (Tiwari et al., 2024) present an AI assistant for legal functions, illustrating AI's potential to enhance legal processes. The role of LLMs in legal contexts has been explored in several studies, balancing their strengths and limitations.

Cross-jurisdictional research, such as that by (Zhao et al., 2018), demonstrates the adaptability of LJP models across different legal systems and languages. Multilingual considerations are also addressed, with studies focusing on linguistic diversity in legal frameworks, such as (Niklaus

[1] `https://anonymous.4open.science/r/Nyayanuman_InLegalLlama-A7DE`

Figure 1: Illustration of the CJPE Task Framework.

et al., 2021) and (Kapoor et al., 2022), which introduces a corpus for Hindi legal documents. Innovative methodologies, including event extraction and multi-stage learning processes, have been proposed in works like (Feng et al., 2022) and (Ma et al., 2021). The geographical scope of LJP research is expanding, as seen in studies exploring the applicability of LJP models in various legal systems.

In summary, the LJP field is dynamic and multifaceted, encompassing both practical applications and theoretical inquiries into AI explainability. Our work aims to contribute to this evolving landscape by developing a comprehensive English legal dataset for India and investigating advanced models capable of processing complete legal judgments.

3 Task Description

Our research focuses on advancing the Court Judgment Prediction and Explanation (CJPE) task, which encompasses two primary components: Prediction and Explanation. These components are executed sequentially to address the crucial needs of predicting legal judgments and providing justifications for these predictions. Figure 1 provides a visual overview of the CJPE framework utilized in our study.

Prediction Task: The core objective of the CJPE task is to predict the outcome of a legal case based on the case proceedings. Unlike previous studies that primarily focused on binary classification (acceptance or rejection), our study also classifies cases with partially accepted outcomes. Given a document D the task is to predict the decision $y \in \{0, 1, 2\}$, where '0' denotes the rejection of all appeals by the appellant, '1' represents the acceptance of all appeals, and '2' indicates partial acceptance of the appeals.

Explanation Task: The second component of the CJPE task involves providing an explanation for the predicted decision by the model. To address the need for explainability, we developed `INLegalLlama`, a generative LLM. Initially, the model was trained on case text to incorporate legal knowledge. Following this, we fine-tuned the model in a supervised manner to enhance its capabilities in both prediction and explanation, thereby increasing the reliability and usefulness of AI in legal decision-making.

4 Dataset

In our research, we introduce "`NyayaAnumana`," a significant advancement in the field of Judgment Prediction in India, particularly within the Indian judiciary. `NyayaAnumana` is the largest and most diverse corpus of Indian legal cases ever compiled, covering judgments from the Supreme Court, High Courts, Tribunal courts, District courts, and Daily orders. This comprehensive dataset offers unparalleled diversity and coverage, addressing existing gaps and providing a rich foundation for advanced legal AI research.

4.1 Dataset Compilation

The dataset compilation process involved gathering a substantial corpus of 2,282,137 Indian court case proceedings up to April 2024. We utilized the IndianKanoon website[2], a well-regarded legal search engine, to collect these documents. This source is widely recognized for its comprehensive database of Indian legal documents, making it an invaluable resource for our dataset.

4.2 Data Statistics

The `NyayaAnumana` dataset exhibits extensive data statistics, vital for understanding the dataset's scope and characteristics. The overall dataset is divided into 'multi' and 'single' categories based on the nature of the decisions. Table 2 and Table 3 provide detailed statistics, including the number of documents, the average number of tokens, and the distribution of clear and partial acceptance decisions across the dataset. This dataset is also analyzed on a court-wise basis, providing insights into the unique characteristics of cases from different

[2] https://indiankanoon.org/

Metric	Train multi	Train single	Validation	Test
SCI				
#Documents	35942	20712	2960	5919
Avg #words	2738.06	2733.54	2638.20	2731.47
Acceptance(%)	59.37	50.79	50.74	50.87
SCI + HCs				
#Documents	484725	283457	40495	80988
Avg #words	2167.78	2115.01	2096.44	2106.51
Acceptance(%)	55.75	51.94	48.22	48.25
SCI + HCs + Tribunals				
#Documents	648356	368639	52663	105326
Avg #words	2080.85	1998.84	2003.11	1988.50
Acceptance(%)	58.27	49.48	49.49	49.36
SCI + HCs + Tribunals + Daily Orders and District Courts				
#Documents	702945	401412	57345	114690
Avg #words	2061.17	1985.37	1986.90	1975.74
Acceptance(%)	57.98	49.30	49.34	49.04

Table 2: Data statistics across different courts for evaluating model performance on binary classification (rejection (0) and clear acceptance (1)).

Metric	Train	Validation	Test
SCI			
#Documents	26823	3833	7665
Avg #words	2776.91	2779.23	2783.02
Clear acceptance(%)	41.95	41.22	41.44
Partial acceptance (%)	17.19	18.03	17.59
SCI + HCs			
#Documents	340972	48711	97421
Avg #words	2172.90	2166.32	2180.39
Clear acceptance(%)	54.31	54.46	54.56
Partial acceptance (%)	1.38	1.29	1.34
SCI + HCs + Tribunals			
#Documents	455514	65074	130147
Avg #words	2085.27	2081.69	2088.82
Clear acceptance(%)	57.23	57.55	57.10
Partial acceptance (%)	1.03	0.96	1.03
SCI + HCs + Tribunals + Daily Orders and District Courts			
#Documents	493726	70533	141065
Avg #words	2060.80	2068.05	2081.41
Clear acceptance(%)	57.02	56.83	57.14
Partial acceptance (%)	0.94	0.97	0.95

Table 3: Data statistics across different courts for evaluating model performance on ternary classification (rejection (0), clear acceptance (1), and partially accepted cases (2)).

courts. This breakdown helps in understanding the diversity within the dataset and the varying complexities associated with different court levels.

Additionally, Figure 9 in the Appendix shows the distribution of cases in different courts in percentage.

4.2.1 Injecting Legal Knowledge

To address the deficiency of legal knowledge in the Vanilla-LLaMA model, we employed a continued pretraining (CPT) approach using a comprehensive Indian legal corpus. This involved utilizing preprocessed data comprising 38,321 cases from the Supreme Court of India (SCI) and a randomly selected 100,000 cases from various High Courts. This extensive training corpus was essential for embedding domain-specific legal knowledge into the model. Additionally, the validation dataset consisted of 12,239 documents sourced from both SCI and High Courts, ensuring that the model was rigorously tested and fine-tuned for the nuances of the Indian legal system. This approach aims to enhance the model's understanding and applicability within the Indian legal framework, thereby improving its predictive capabilities and relevance in legal tasks.

4.2.2 Learning Reasoning Skills

To equip the model with the necessary reasoning capabilities for solving prediction and explanation problems. We conducted supervised fine-tuning (SFT) using selected data from downstream tasks on the PredEx training dataset (Nigam et al., 2024), which consists of 12,178 cases accompanied by corresponding case decisions and explanations annotated by legal experts. By SFT on this dataset, we aimed to enhance its reasoning skills and ability to comprehend and apply legal principles effectively. This targeted fine-tuning approach helps bridge the gap between the model's general knowledge and the specific requirements of legal reasoning tasks, ultimately improving its performance in real-world legal scenarios.

4.2.3 Prediction Task

For prediction, we split the `NyayaAnumana` single dataset into training, validation, and test sets, 70%, 10%, and 20% of the dataset, respectively. A key component of our research involved comparing the performance of models trained on `NyayaAnumana` with those trained on the ILDC 2021 (Malik et al., 2021a) test dataset. This comparison is crucial for benchmarking our models and understanding their efficacy compared to established datasets in the field. By testing against ILDC 2021, we aim to evaluate the improvements in prediction accuracy and model robustness that `NyayaAnumana` offers, showcasing its contribution to the evolving landscape of legal AI in India. We also tested the model performance on temporal data, assessing its effectiveness on future or unseen data from 2020 to

Metric	SCI	HCs	Tribunals	Daily Orders and District Courts
#Documents	1812	29216	19034	17002
Avg #words	4469.79	2671.46	2926.84	1306.02
Acceptance(%)	67.77	53.18	47.42	62.40

Table 4: Statistics of temporal test data (2020-April 2024) across different courts for evaluating model performance.

April 2024 to ensure robustness and generalization capabilities over time, as detailed in Table 4.

4.2.4 Prediction with Explanation Task

For this task, we used the PredEx 2024 test dataset (Nigam et al., 2024), which includes 3,044 robust and balanced cases. The balanced nature of the test set is particularly important for maintaining the validity of our experiments and ensuring the reliability and generalizability of our model's performance.

Overall, `NyayaAnumana` stands as a pioneering dataset in the realm of CJPE, offering unprecedented depth, diversity, and scope for exploring legal AI in the Indian context.

5 Model Training: INLegalLlama

5.1 Injecting Legal Knowledge

To address the limitations of legal knowledge inherent in the Base Llama model, we adopted a continued pretraining (CPT) strategy utilizing a comprehensive Indian legal corpus. For this purpose, we selected the Llama-2 7B architecture (Touvron et al., 2023), which features a substantial context length of 2K, allowing for effective handling of legal texts. This choice facilitates a direct comparison with previous state-of-the-art results on the PredEx dataset (Nigam et al., 2024). This approach significantly enhances the model's understanding and relevance within the Indian legal framework, thereby improving its predictive capabilities and application in legal tasks.

5.2 Learning Reasoning Skills

To further develop the model's reasoning skills, particularly for legal prediction and explanation tasks, we conducted Supervised Fine-Tuning (SFT) using data from specific downstream tasks. This dataset includes case decisions along with their corresponding explanations, all annotated by legal experts. The fine-tuning process was crucial for enhancing the model's ability to understand and apply legal principles effectively, bridging the gap between general knowledge and the specialized

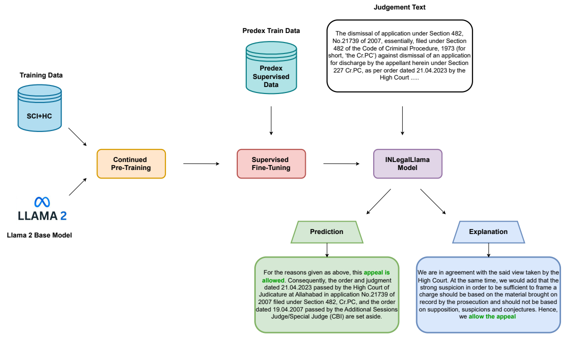

Figure 2: INLegalLlama Flow Diagram.

requirements of legal reasoning tasks. This focused fine-tuning was pivotal in improving the model's performance in real-world legal scenarios.

However, the fine-tuning of such models typically demands substantial computational resources and extensive training data. Given the constraints of limited computational power and the specific nature of our legal task-related dataset, we prioritized efficient training methods to optimize both computational costs and data usage. We employed parameter-efficient tuning techniques, such as the Low-Rank Adaptation (LoRA) method (Hu et al., 2021), to fine-tune the Llama-2 7B model. This approach enabled us to maximize the utility of available data and minimize the need for extensive computational resources, ensuring a cost-effective yet powerful fine-tuning process for developing `INLegalLlama`.

5.3 Methodology

5.3.1 Judgment Prediction

5.3.2 Language Model based

In our approach, we utilized several language models, including InLegalBERT, InCaseLaw (Paul et al., 2023), and XLNet (large) (Yang et al., 2019), as baselines for binary and ternary classification. Due to the length constraints of complete judgments, which exceed the token capacity of these models, we adopted a chunking strategy. Each document was divided into 512-token chunks using a moving window approach with a 100-token overlap to preserve textual context.

5.3.3 Large Language Model based

To utilize LLMs in prediction, we employed two strategies: one involving only prediction instructions and the other prediction with explanation instructions. We used two methods to get predictions from `INLegalLlama` after CPT and CPT followed by SFT. We followed the prompts and instruction-tuning approaches published by (Vats et al., 2023) and (Nigam et al., 2024) in a few-shot setup, and used the PredEx training data for instruction-tuning.

5.4 Judgment Prediction with Explanation

For this task, we employed the same LLMs with settings similar to the Judgment Prediction task, but with modified instructions focusing on both prediction and explanation.

5.5 Prompts Used

For both task inferences we utilized prompts from (Vats et al., 2023). These prompts, which include a case description and a gold standard prediction label, guide the LLM to generate judicial decisions. The details of these prompts can be found in Table 12 in the Appendix. In addition, for instruction tuning, we adopted prompts from (Nigam et al., 2024) for prediction tasks, as listed in Table 13 in the Appendix. The relevant details and examples of the generated predictions and explanations are also available in the referenced tables in the Appendix.

5.6 Instruction-Set

For both judgment prediction and explanation, we used 16 instruction sets correspondingly published by (Nigam et al., 2024). For a comprehensive view of all instruction sets which was randomly given to the model for tuning, we have included the full list in Table 14 in the Appendix of this paper.

6 Evaluation Metrics

In this study, we employed a comprehensive set of evaluation metrics to assess the performance of our models on the NyayaAnumana judgment prediction and PredEx explanation test datasets. 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. **Expert Evaluation:** Human evaluation was a critical component of our assessment framework. Legal experts reviewed the explanations generated by the models, rating them on a 1–5 Likert scale based on criteria such as accuracy, relevance, and completeness. A rating of 1 indicates that the information is irrelevant, while a rating of 5 signifies that the explanation is superior to the expert's own explanation. The full description of the rating scores can be found in Appendix B, which is adapted from (Nigam et al., 2024) and is defined accordingly.

7 Results and Analysis

7.1 Judgment Prediction

Our experiments, as detailed in Table 5 which is trained on `NyayaAnumana` single, and 16 which is trained on `NyayaAnumana` multi reveal interesting insights into the performance of various models on the Judgment prediction results on the binary task across different court cases and temporal test cases, with models trained on SCI + HCs + Tribunal + Daily Orders and District Court data from `NyayaAnumana` test data.

Interestingly, our findings indicate that contrary to previous research, larger models like XLNet did not consistently outperform smaller models. Instead, models specifically trained on Indian legal data, such as InLegalBERT and InCaseLaw, performed comparably and, in some instances, even surpassed XLNet large. This suggests that the inclusion of domain-specific data significantly enhances model performance. The previous state-of-the-art results hovered around 79% accuracy using XLNet with hierarchical BiGRU, while our best models achieved approximately 90% accuracy, highlighting a substantial improvement.

Table 6 showcases the binary prediction results across various LLM-based models. Our findings

Test Data	Macro Precision	Macro Recall	Macro F1	Accuracy
InLegalBert				
ILDC	0.7209	0.7169	0.7189	0.7172
SCI (2019)	0.8261	0.8255	0.8258	0.8258
SCI (2020-24)	0.8515	0.8588	0.8552	0.8720
SCI+HCs (2019)	0.8739	0.8735	0.8737	0.8739
HCs (2020-24)	**0.8940**	**0.8943**	**0.8942**	0.8945
SCI+HCs+Tribunal (2019)	0.8637	0.8634	0.8635	0.8635
Tribunal (2020-24)	0.8308	0.8249	0.8278	0.8277
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8722	0.8718	0.8720	0.8720
Daily_orders (2020-24)	0.8897	0.8869	0.8883	**0.8955**
InCaseLaw				
ILDC	0.7347	0.7335	0.7341	0.7337
SCI (2019)	0.8271	0.8272	0.8271	0.8272
SCI (2020-24)	0.8449	0.8579	0.8513	0.8670
SCI+HCs (2019)	0.8585	0.8570	0.8578	0.8579
HCs (2020-24)	**0.8891**	**0.8898**	**0.8895**	0.8898
SCI+HCs+Tribunal (2019)	0.8544	0.8521	0.8532	0.8526
Tribunal (2020-24)	0.8160	0.7939	0.8048	0.8001
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8573	0.8553	0.8563	0.8559
Daily_orders (2020-24)	0.8868	0.8795	0.8831	**0.8911**
XLNet Large				
ILDC	0.6851	0.6850	0.6851	0.6849
SCI (2019)	0.8150	0.8137	0.8143	0.8142
SCI (2020-24)	0.8507	0.8562	0.8535	0.8709
SCI+HCs (2019)	0.8590	0.8585	0.8588	0.8590
HCs (2020-24)	**0.8848**	0.8863	0.8856	**0.8851**
SCI+HCs+Tribunal (2019)	0.8580	0.8576	0.8578	0.8577
Tribunal (2020-24)	0.8180	0.8053	0.8116	0.8098
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8660	0.8655	0.8657	0.8657
Daily_orders (2020-24)	0.8832	**0.8904**	**0.8868**	0.8924

Table 5: Judgment prediction results on the binary task across different court cases and temporal test cases, with models trained on SCI + HCs + Tribunal + Daily Orders and District Court data from NyayaAnumana single split data. The best results are highlighted in bold.

Models	Macro Precision	Macro Recall	Macro F1	Accuracy
Using LMs on PredEx				
InLegalBert	0.7546	0.7526	0.7536	0.7526
InCaseLaw	0.7421	0.7395	0.7408	0.7395
XLNet Large	0.7736	0.7707	0.7722	0.7707
RoBerta Large	**0.7831**	**0.7822**	**0.7827**	**0.7822**
Using LLMs on PredEx				
Zephyr	0.5347	0.5295	0.5119	0.5309
Gemini pro	0.5976	0.5803	0.5610	0.5808
Llama-2	0.5732	0.5723	0.5713	0.5726
Llama-2 SFT	0.5186	0.5177	0.5177	0.5177
Llama-2 CPT Prediction only	0.4684	0.4967	0.3490	0.4910
INLegalLlama CPT+SFT Prediction only	0.5439	0.5204	0.4462	0.5204
Llama-2 CPT Prediction + Explanation	0.4554	0.4965	0.3512	0.5086
INLegalLlama CPT+SFT Prediction + Explanation	**0.7623**	**0.7605**	**0.7601**	**0.7605**
Using LLMs on ILDC expert				
Zephyr	0.4011	0.3704	0.3524	0.5556
Gemini pro	0.6467	0.5741	0.5139	0.5741
Llama-2	0.3125	0.4259	0.3236	0.4259
Llama-2 SFT	0.5750	0.5741	0.5728	0.5741
Llama-2 CPT Prediction only	0.3125	0.4545	0.3704	0.5882
INLegalLlama CPT+SFT Prediction only	0.5926	0.5926	0.5926	0.5926
Llama-2 CPT Prediction + Explanation	0.2692	0.5000	0.3500	0.5385
INLegalLlama CPT+SFT Prediction + Explanation	**0.7301**	**0.7223**	**0.7198**	**0.7223**

Table 6: Judgment prediction results using different LLMs on PredEx and ILDC datasets. The best results are highlighted in bold.

indicate that while injecting legal knowledge and enhancing reasoning skills on the prediction task did not significantly boost performance, the combination of both on tasks integrating both prediction and explanation yielded more promising results. This improvement is likely attributed to the incorporation of a chain of thought (CoT) process, which aids in making more accurate judgment predictions by facilitating deeper reasoning.

Moreover, our comparison with different transformer-based models, including our INLegalLlama model, reveals that it is competitive with classifier-based models. This highlights the importance of infusing the model with domain-specific legal knowledge and further refining it through reasoning tasks. Table 7 showcases the results for the ternary judgment prediction task on SCI court cases. Similar trends were observed in the binary prediction tasks, with models trained on the Indian legal corpus showing marked improvements in performance.

7.2 Judgment Prediction with Explanation

The results, as presented in Table 8, offer valuable insights into the comparative performance of machine-generated explanations against those provided by legal experts across various models. These evaluations cover lexical-based, semantic, and expert assessment metrics, specifically using the 50 test cases from the PredEx (Nigam et al., 2024) and 54 ILDC expert (Malik et al., 2021a), comparisons with the INLegalLlama model with different settings. This comprehensive evaluation framework allows us to thoroughly assess the models' abilities to generate accurate and contextually relevant explanations.

Additionally, we experimented with the Aalap (Tiwari et al., 2024) model, which is instruction-tuned on various Indian legal tasks, but it underperforms in this task. This may be due to its lack of focus on generating explanations alongside predictions, a complex requirement that might not have been sufficiently addressed during training. In contrast, comparisons with the INLegalLlama model under different settings demonstrate our approach's effectiveness in improving the explainability and accuracy of AI-generated legal judgments.

Models	Metric	Overall	Class 0	Class 1	Class 2
InLegalBert	Macro Precision	0.6376	0.73	0.67	0.52
	Macro Recall	0.5903	0.77	0.81	0.19
	Macro F1	0.5868	0.75	0.73	0.28
InCaseLaw	Macro Precision	0.6260	0.72	0.67	0.49
	Macro Recall	0.5812	0.78	0.80	0.17
	Macro F1	0.5746	0.75	0.73	0.25
XLNet	Macro Precision	**0.6443**	0.75	0.70	0.48
	Macro Recall	**0.6142**	0.80	0.82	0.22
	Macro F1	**0.6125**	0.77	0.76	0.31

Table 7: Judgment prediction results on the ternary task on SCI court cases. The best results are highlighted in bold.

7.2.1 Lexical-Based Evaluation

The performance of LLMs in generating explanations reveals that verbatim matches to reference texts are not consistently high. However, it is important to recognize that these metrics, although useful, do not fully capture the models' capabilities in analyzing legal cases, predicting outcomes, and generating reasoning. Therefore, we also employed Semantic Similarity-Based Evaluation and Expert Score Evaluation to provide a more comprehensive assessment of the models' performance.

7.2.2 Semantic Evaluation

The semantic evaluation, particularly utilizing BERTScore, demonstrates better alignment of the generated explanations with the gold standard, indicating a strong semantic understanding of the explanations produced. `INLegalLlama` shows superior performance in terms of semantic similarity. It is important to note that generative models may occasionally produce hallucinated content, which these metrics do not fully capture, emphasizing the need for manual review by legal experts to ensure comprehensive assessment.

7.2.3 Expert Evaluation

Assessing the performance of generative models in the task of CJPE requires the insight of domain-specific experts. The expert evaluation, summarized in Table 8, indicates that our `INLegalLlama` model, performs exceptionally well, although it sometimes generates truncated or repetitive content. Despite these minor drawbacks, the instruction-tuned model produces fewer non-factual responses and delivers a higher overall quality of explanations compared to other pre-trained models. Notably, models equipped with carefully designed prompts for explanation generation showed improved performance and did not suffer from hallucination issues. The expert ratings, detailed in Table 28, further emphasize the effectiveness of our

Models	Lexical Based Evaluation					Semantic Evaluation		Expert Evaluation
	Rouge-1	Rouge-2	Rouge-L	BLEU	METEOR	BERTScore	BLANC	Rating Score
	Prediction with explanation on PredEx (Nigam et al., 2024)							
Gemini pro	0.3099	0.2428	0.2593	0.0826	0.1870	0.6329	0.1715	2.24
Aalap	0.2711	0.1001	0.1703	0.0324	0.1528	0.5541	0.0742	2.46
Llama-2	0.3211	0.1886	0.2109	0.0599	0.1760	0.6191	0.1507	3.06
Llama-2 SFT	0.4972	0.4321	**0.4399**	0.2531	0.3630	**0.6909**	0.2844	2.84
Llama-2 CPT	0.3355	0.1549	0.2287	0.0898	0.2326	0.5834	0.1118	3.26
INLegalLlama CPT+SFT	**0.5076**	**0.4338**	0.4379	**0.2555**	**0.3643**	0.6825	**0.2927**	**3.54**
	Prediction with explanation on ILDC expert (Vats et al., 2023; Malik et al., 2021a)							
GPT 3.5 turbo	**0.5383**	**0.4267**	**0.4541**	0.2842	0.4685	**0.7273**	0.3394	3.6
Aalap	0.2991	0.0948	0.1808	0.0491	0.2564	0.5379	0.0944	2.3
Llama-2	0.4526	0.2454	0.2957	0.1485	0.3440	0.6464	0.2212	3.65
Llama-2 SFT	0.4939	0.3805	0.3969	**0.2918**	0.5075	0.6891	0.3636	3.30
Llama-2 CPT	0.3083	0.2211	0.2550	0.1418	0.3681	0.5929	0.2572	3.41
INLegalLlama CPT+SFT	0.5088	0.4026	0.4229	0.2820	**0.5412**	0.6758	**0.4072**	**3.67**

Table 8: Explanation performance comparison of various model combinations for explanation across different evaluation metrics, with the highest score in each metric in bold.

instruction-tuned model, which in some instances even surpasses the quality of explanations provided by human legal experts, achieving an impressive rating score of 4. This highlights the potential of generative models, particularly those enhanced by our approach, in delivering accurate and contextually relevant legal explanations.

8 Conclusions and Future Work

In this study, we presented `NyayaAnumana`, the largest and most diverse dataset of Indian legal cases, alongside `INLegalLlama`, a specialized language model fine-tuned for legal judgment prediction and explanation. Our findings demonstrate that domain-specific models, particularly those enhanced with legal data, significantly outperform generic large language models in both accuracy and the quality of explanations provided. Notably, we achieved very good accuracy in the prediction task after including data from all court levels, underscoring the value of comprehensive datasets.

Future work will focus on expanding the dataset to include judgments in regional languages, better reflecting India's linguistic diversity. We plan to explore larger and more advanced models, potentially using more efficient quantization techniques and enhanced hardware resources, to better handle complex legal documents. Refining our fine-tuning methodologies by incorporating a broader range of legal documents, such as statutes and contracts, will further enrich the model's knowledge base.

By addressing these challenges and expanding the scope of our research, we aim to enhance the performance and reliability of AI models in the legal domain, contributing to more efficient and accurate legal decision-making processes.

Limitations

Our study faced several significant limitations that influenced both our approach and the findings. One of the primary constraints was the use of a 4-bit quantized model due to resource constraints, which restricted our ability to leverage larger parametric models, such as those with 70B or 40B parameters. The token limitation and high subscription charges for paid cloud services further exacerbated this issue, limiting our capacity to perform inference and fine-tuning on more advanced models. This limitation likely restricted the full exploration of these models' capabilities, potentially affecting the depth and quality of the insights and performance metrics we could achieve.

Additionally, the resource-intensive nature of obtaining legal expert annotations presented another challenge. The high costs and significant time required for acquiring these annotations made it impractical to obtain expert evaluations for the entire PredEx test dataset. As a result, we use the same 50 random documents as used in (Nigam et al., 2024) for expert review and Likert score evaluations. While necessary, this approach potentially limits the breadth and depth of our expert-based evaluation, as it does not encompass the entire dataset.

The applicability of Large Language Models (LLMs) in the legal domain, particularly for tasks involving legal judgment prediction and explanation, remains uncertain based on our findings. While LLMs demonstrate proficiency in conversational contexts, their performance in tasks requiring complex logic or specialized knowledge, such as legal reasoning, is less convincing. Analyzing lengthy legal documents and generating predictions along with explanations proved to be challenging for generative-based models. This challenge is particularly evident when the models must process and understand intricate legal reasoning and contexts.

Lastly, the dataset used in this study comprised only English-language judgments, excluding other regional languages such as Hindi and Bengali. This limitation underscores the need for more inclusive datasets that represent the linguistic diversity present in legal documents across different jurisdictions.

These limitations highlight the complexities and challenges inherent in applying LLMs to specialized tasks like legal judgment prediction and explanation. They also underscore the need for ongoing research and development to enhance AI models' capabilities in interpreting and understanding legal documents and contexts comprehensively.

Ethics Statement

In conducting this research, we placed a strong emphasis on ethical considerations, particularly due to the sensitive nature of legal data and the methodologies employed. The `NyayaAnumana` dataset, used extensively in this study, was sourced from publicly accessible legal search engines, ensuring compliance with data privacy and usage regulations. We have taken steps to remove any meta-information such as judge names, case titles, and case IDs to protect the privacy and confidentiality of individuals involved.

Furthermore, the computational resources utilized in this study were obtained through ethical and legitimate means. We subscribed to Google Colab Pro and other necessary cloud services, ensuring that all resources used for model training and testing were accessed legally. This financial support not only facilitated our research but also contributed to the sustainability of these services.

In addition to adhering to legal and ethical guidelines in data handling and resource usage, we are committed to transparency and reproducibility in our research. The `NyayaAnumana` dataset and the code for our models, including `INLegalLlama`, for now, have been made available via an anonymous link to promote open science and enable other researchers to replicate and build upon our work.

Finally, we acknowledge the potential societal impact of deploying AI in legal settings. Our models are designed to assist, not replace, human judgment, and we stress the importance of human oversight in any AI-assisted legal decision-making process. We remain committed to ongoing ethical scrutiny as we advance this research field.

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.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

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

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

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*.

André Lage-Freitas, Héctor Allende-Cid, Orivaldo Santana, and Lívia Oliveira-Lage. 2022. Predicting brazilian court decisions. *PeerJ Computer Science*, 8:e904.

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.

Luyao Ma, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Wei Ye, Changlong Sun, and Shikun Zhang. 2021. Legal judgment prediction with multi-stage case representation learning in the real court setting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 993–1002, New York, NY, USA. Association for Computing Machinery.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021a. 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.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021b. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4046–4062. Association for Computational Linguistics.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28:237–266.

Shubham Kumar Nigam, Anurag Sharma, Danush Khanna, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2024. Legal judgment reimagined: Predex and the rise of intelligent ai interpretation in indian courts. *Preprint*, arXiv:2406.04136.

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.

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*.

Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef Van Genabith. 2017. Exploring the use of text classification in the legal domain. *arXiv preprint arXiv:1710.09306*.

Aman Tiwari, Prathamesh Kalamkar, Atreyo Banerjee, Saurabh Karn, Varun Hemachandran, and Smita Gupta. 2024. Aalap: Ai assistant for legal & paralegal functions in india. *arXiv preprint arXiv:2402.01758*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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 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.

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.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

A Ablation Study

In our ablation study, we investigated the impact of various court-level training data configurations on the performance of our models in the binary classification judgment prediction task. We observed consistent trends across multiple experiments, particularly when analyzing the results from the full NyayaAnumana dataset in relation to subsets that excluded specific court cases. The performance metrics are detailed in Appendix Tables 15, 17, 18, 19, 20, 21, 22, and 23.

Our findings indicate that models trained on High Court cases exhibited the best performance, likely due to the substantial representation of High Court data in the training set. However, when these models were evaluated on the ILDC dataset, a noticeable drop in performance was observed. This suggests that while the models excel in familiar contexts, they struggle to generalize to datasets with different characteristics or distributions.

We also conducted experiments on the ternary judgment prediction task, incorporating additional court cases, with results presented in Appendix Tables 24, 25, and 26. These experiments further reinforced the significance of including diverse court data and highlighted the benefits of utilizing large volumes of training data.

Our findings emphasize the importance of a diverse and comprehensive dataset. Including a broad spectrum of court cases not only enhances the model's understanding of various judicial contexts but also significantly improves its performance metrics. In particular, the models trained on the most extensive datasets achieved an F1 score of around 90%, underscoring the critical role of data diversity and volume in achieving high accuracy.

Overall, the ablation study illustrates that the diversity and volume of training data play a crucial role in enhancing model performance, particularly in the context of legal judgment prediction tasks. Future work should continue to explore the impact of various data configurations to further optimize model accuracy and generalization capabilities.

B Rating Score Description

The evaluation of the explanations generated by the models was conducted using a 1–5 Likert scale, where each score reflects the quality and relevance of the provided explanation. The criteria for rating are as follows:

[1]: The explanation is entirely incorrect or fails to provide any relevant information. This score indicates that the response does not address the case judgment in any meaningful way.

[2]: The response is irrelevant or demonstrates a misunderstanding of the case judgment. A rating of 2 suggests that while some effort was made to respond, the explanation does not accurately reflect the case details.

[3]: The explanation is partially accurate but lacks critical details. This score indicates that the response contains some correct information, but it is insufficient for a complete understanding of the case judgment.

[4]: The response is generally accurate and relevant, comparable to the ground truth. A rating of 4 signifies that the explanation aligns well with the expected outcomes and provides a solid understanding of the case.

[5]: The explanation is fully accurate, relevant, and potentially superior to the expert's explanation. This highest rating reflects an exceptional response that not only meets the criteria of accuracy and relevance but also offers insights that exceed standard expert evaluations.

C Experimental Setup and Hyper-parameters

C.1 Transformers Training Hyper-parameters

For model training, we used a batch size of 16, the Adam optimizer (Kingma and Ba, 2014), and a learning rate of 2e-6. The training was conducted over 3 epochs on the NyayaAnumana train dataset. The remaining hyperparameters were set to their default values as provided by the HuggingFace library.

C.2 INLegalLlama Training Procedure

The fine-tuning of the INLegalLlama model was conducted using the Llama 2 7B model architecture, with the model loaded in Bfloat16 precision. The training was done in Google Colab Pro, utilizing a single A100 GPU with 40GB of memory. Given the constraints of limited computational resources, we carefully selected parameters to fully utilize the available compute power. This setup enabled us to develop a highly capable model within a reasonable time frame of 48 hours, incurring a cost of approximately $59. During the training process, the maximum token length was set at 2096. We employed the Low-Rank Adaptation (LoRA) technique, initializing the LoRA rank at 16 and setting the alpha parameter to 64, with a dropout rate of 0.1. This configuration was applied to all layers of the model, aiming to achieve performance comparable to a fully fine-tuned model. The integration of flash-attention 2 significantly improved the training speed.

The optimization process utilized a Paged Adam 32-bit optimizer with a learning rate of 1e-4, alongside a "cosine" learning rate scheduler. The gradient accumulation steps were set to 4, and the warm-up ratio was established at 0.05. We employed DeepSpeed Stage 3 optimization, with a per-device batch size of 4. The model was trained for a total of 3000 steps, which corresponds to approximately 0.347 epochs.

The entire training process was meticulously tracked using Weights & Biases (Biewald, 2020), allowing for detailed monitoring and analysis.The progression of training and evaluation losses during the fine-tuning process is illustrated in Appendix Figures 3, 4, 5, 6, 7 and 8. These figures highlight the model's loss trends during the Continued Pre-training (CPT) and Supervised Fine-Tuning (SFT) phases, for both prediction-only and prediction-with-explanation tasks. Notably, the CPT phase shows a consistent reduction in training and evaluation losses, indicating the model's ability to learn effectively from the legal corpus. Similarly, the SFT phase results display refined loss curves, especially in the prediction-with-explanation task, suggesting the model's enhanced capability to generate comprehensive and accurate legal explanations.

D Hallucination

In our study, we address the issue of hallucinations in model-generated text, which is a prevalent challenge when using large language models for generating legal judgments. Hallucinations occur when the model produces information that is false or irrelevant, not supported by the input data. A sample of hallucination has been provided in Appendix Table 9. To tackle this issue, we employed a specialized fine-tuning strategy aimed at significantly reducing such errors. A detailed comparative analysis provided in Appendix D.1 highlights the effectiveness of these strategies. This analysis illustrates how fine-tuning and instruction-tuning, specifically tailored to the legal domain, can help minimize hallucinations, resulting in outputs that are clearer, more accurate, and legally coherent.

D.1 CPT Llama-2 hallucinations

In the subsection, we conduct a thorough comparison between ground truth and fine-tuned models to demonstrate some samples where the CPT model showed signs of hallucination. Table 9 in the Appendix presents an extensive analysis of the model, illustrating the performance of the LLama-2 pre-trained model on legal judgment prediction with explanation tasks in our PredEx dataset. In the given table, this can be observed that the model did not only produce incorrect judgment but also delivered the wrong explanation, if not repeating the sentences and printing a random repetitive set of numbers. Some examples in the table show the model hallucinating by repeating a statement irrelevant to the case information. This comparison specifically focuses on how instruction-tuning can mitigate the issues of inaccuracies often found in the responses generated by pre-trained Large Language Models (LLMs).

D.2 CPT Llama-2 vs INLegalLlama (CPT+SFT)

In the subsection, we conduct a detailed comparison between Llama-2-7B CPT and INLegalLlama (CPT+SFT) to demonstrate some samples where the fine-tuned model INLegalLlama performed better than Llama-2 CPT in all aspects including prediction and explanation. Table 10 in the Appendix presents an extensive analysis of the model, illustrating the performance of the LLama-2 pre-trained model against the INLegalLlama model on legal judgment prediction with explanation tasks in our PredEx dataset. In the given table, it can be observed that the performance of the Llama-2 CPT looks good at once but when we refer to the explanation, we realize that the explanation given for the prediction does not follow the context of the legal judgment. On the other hand, INLegalLlama maintains the contextual information and retains a lot of information from the legal judgment.As a result, INLegalLlama outperforms the Llama-2 CPT with better context information and reasonable facts inclusion.

E Preprocessing

The preprocessing of the NyayaAnumana dataset involved several critical steps to ensure the data's quality and relevance. Given the unstructured nature of the documents, which varied in format and size, we faced challenges such as spelling errors

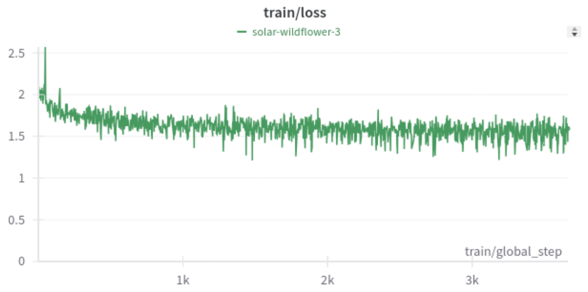

Figure 3: Training loss progression for the INLegalLlama model using a 4-bit quantized Llama-2 architecture with 4096 maximum sequence length during the Continued Pretraining (CPT) phase.

Figure 4: Evaluation loss progression for the INLegalLlama model using a 4-bit quantized Llama-2 architecture with 4096 maximum sequence length during the Continued Pretraining (CPT) phase.

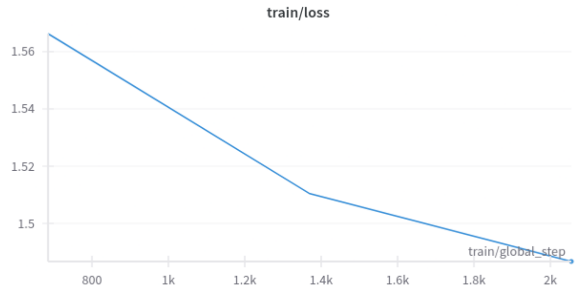

Figure 5: Training loss progression for the INLegalLlama model during the Supervised Fine-Tuning (SFT) phase, focused solely on prediction tasks, using the Llama-2 architecture at checkpoint 3000.

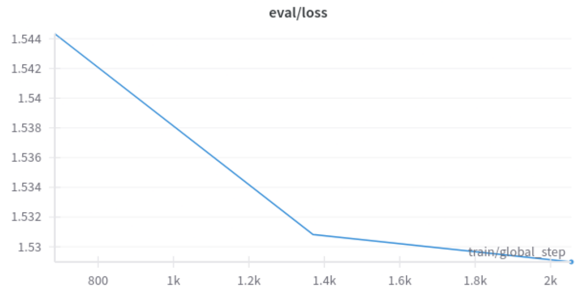

Figure 6: Evaluation loss progression for the INLegalLlama model during the Supervised Fine-Tuning (SFT) phase, focused solely on prediction tasks, using the Llama-2 architecture at checkpoint 3000.

and inconsistencies. To mitigate these, we used regular expressions to remove noisy text and meta-information, including case numbers, titles, judge names, petitioners, respondents, and dates. We identified key sections using specific terms like 'ORDER,' 'JUDGMENT,' and 'JUDGEMENT' to isolate the essential content and filter out irrelevant details.

We further refined the dataset by removing cases without clear outcomes or insufficient information. To maintain consistency and manageability, we excluded extremely short cases (less than 50 words) and excessively long ones (more than 32,000 words). Additionally, tokens with characters repeated more than twice consecutively were removed to clean up text errors. This meticulous refinement process reduced the number of cases to 1,125,604, retaining only the most relevant cases, as detailed in Table 27, which compares the number of cases before and after preprocessing, categorized by court type.

E.1 Label Making

After filtering, the documents were labeled using keywords indicative of positive outcomes like "approved," "allowed," and "granted," or negative outcomes such as "rejected," "disapproved," and "dismissed." This labeling process helped classify the cases into likely acceptance or rejection categories. We categorized the cases into two groups: single-labeled cases, where all appeals had the same outcome or only a single appeal was filed, marked as 0 (rejection) or 1 (acceptance), and multi-labeled judgments, where at least one appeal was accepted, indicating partial acceptance, marked as 2.

To ensure accurate labeling, we focused on the last 750 words of each document, typically where decisions are summarized. Special attention was given to a context window around key terms like "appeal," "petition," or "case" to accurately determine the judgment nature. For example, phrases like "Appeals Allowed" or "The appeal is granted" indicated a positive outcome, while "The appeal has no proper evidence and hence we reject it" indicated rejection. We also looked for indicators like "partly" to identify multi-labeled judgments for cases with partial approvals. In instances where the judgment was phrased negatively (e.g., "No appeal is allowed"), we used a label-flipping strategy if negation words like "no" or "not" were found

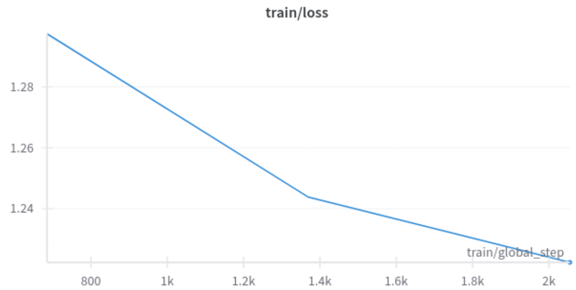

Figure 7: Training loss progression for the INLegalLlama model during the Supervised Fine-Tuning (SFT) phase, focused on both prediction and explanation tasks, using the model setup for 1000 input tokens and 500 output tokens.

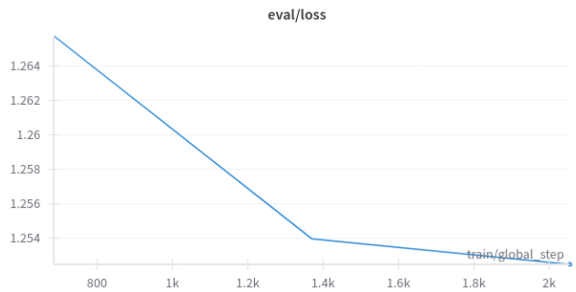

Figure 8: Evaluation loss progression for the INLegalLlama model during the Supervised Fine-Tuning (SFT) phase, focusing on both prediction and explanation tasks, using the model setup for 1000 input tokens and 500 output tokens.

close to key terms, ensuring the labels accurately reflected the judgment's intent.

This meticulous approach to labeling, focusing on the judgment's context and nuanced expression, enhances the reliability of our dataset and prepares it for effective training of judgment prediction models.

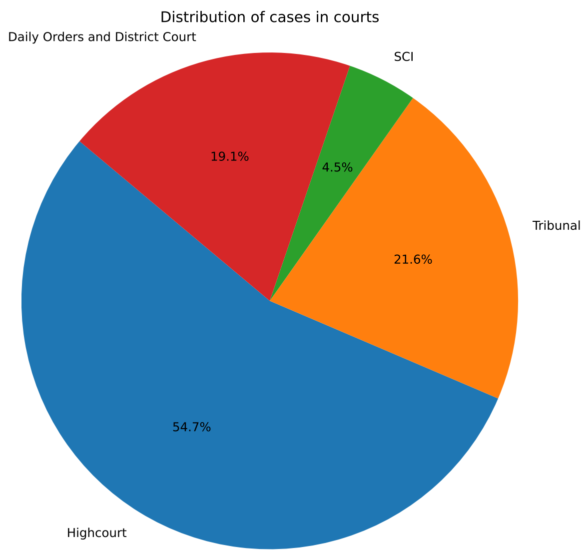

Figure 9: Distribution of cases in different courts (percentage).

Model	Decision and Explanation
Ground Truth	Decision: Rejected. Explanation: IT first requires to be noticed that the above two agreements dated November 1, 1966 and may 13, 1967, are agreements made otherwise than in the course of conciliation proceedings and are of the nature referred to in(1) of S. 18 Admittedly, these agreements were not signed on behalf of the workmen who were were not members of the 2nd respondent Union. .... As we are unable to accept that contention, the third contention alsowe have held that these agreements have ceased to be in operation and did not require to be terminated, this contention must fail. We have already recorded reasons in that connection in the previous discussion.
Llama2-7B CPT	Decision: No clear decision generated. Explanation: 1.0 The order is upheld. 1.1 The order is upheld. 1.2 The order is upheld. 1.3 The order is upheld. 1.4 The order is upheld. 1.5 The order is upheld. 1.6 The order is upheld. 1.7 The order is upheld. 1.8 The order is upheld. 1.9 The order is upheld. 1.10 The order is upheld. .... 1.73 The order is upheld. 1.74 The order is upheld. 1.75 The order is upheld. 1.76 The order is upheld. 1.77 The order is upheld. 1.78 The order is upheld. 1.79 The order is upheld. 1.80 The order is upheld. 1.81 The order is upheld. 1.82 The order is upheld. 1.83 The order is upheld. 1.
Ground Truth	Decision: Rejected Explanation: For proper appreciation of the contentions advanced by the parties before us, it is necessary to refer to some of the provisions of the Act. Section 160(1)(i) of the Act prov ides that in respect of the income of a non-resident specified in sub-section (1) of section 9 of the Act, the agent of the non-resident, including a person who is treated as an agent under section 163 is a representative assessee .... The above cable shows that the London Solicitors had sought information about the suits in Calcutta t o enable them to engage Mr. Blanco White to plead on behalf of the German Corporation in the said suits. On December 23, 1969, the London Solicitors wrote a letter to the appellants in which it was stated that the copies of certain documents sent by the appellants had been handed over to Mr. Blanco White in addition to copies of certain other documents which they themselves had handed over to him.
Llama2-7B CPT	Decision: Rejected Explanation: The appellants are a partnership firm and the partners are Mr. and Mr. D.K.Jain. Mr. D.K.Jain is a non-resident Indian.Income from the partnership firm is subject to tax in India under the provisions of the Income-tax Act, 1961....................The firm has no in Germany.The firm has no in Germany.The firm has no in Germany.The firm has no in Germany.The firm has no in Germany.The firm has no in Germany.The firm has no in Germany.The firm has no in Germany.The firm has no in Germany.The firm has no in Germany.The firm has no in Germany.The

Table 9: We observe 'reject' judgments in the response generated by Llama-7B CPT with varying explanations, conclusions, and repeating responses.

Model	Decision and Explanation	Explanation of error
Ground Truth	Case decision: Accepted Explanation: In our opinion, the High Court should have examined these questions and especially the question as to whether the had been able to substantiate the claim that it was in possession of the entire underground rights in respect of Ac.covered by the agreement of 1934. The agreement of 1934 bound the Government and the Hindu undivided family: the was not a party to it. Even if the Hindu undivided family continued in possession of some of the mines, seams and veins after October 1, 1947, it is a moot question whether such possession meant possession of the seams, mines, and veins which were not actually being worked. A further question would arise as to how and what right the company could lawfully claim in respect of the colliery by virtue of the deed of assignment in 1950. Without an adjudication of these questions the High Court was not justified in directing the State and other authorities to put the company in possession of the area described as a virgin area. Whether the case is a fit for examination on affidavits will be for the High Court to consider.	
Llama-2-7B CPT	Case decision: Accepted Explanation: 1.The petitioner was in actual physical of the area which was taken over by the State Government on February 28, 1970. The High Court was not justified in directing the State and other authorities to put the petitioner in the physical area. The State Government had taken over the area for the purpose of the public interest. The High Court should have refrained from granting any relief to the petitioner. 2. The respondent company was not in actual physical of the area which was taken over by the State Government on February 28, 1970. The High Court was not justified in directing the State and other authorities to put the petitioner in physical of the area. The State Government had taken over the area for the purpose of the public interest. The High Court should have refrained from granting any relief to the petitioner. ... 12. The High Court should have examined the questions of whether the petitioner was in actual physical of the area which was taken over by the State Government on February 28, 1970, and whether the State Government had taken over the area for the purpose of the public interest. The High Court should have refrained from granting any relief to the petitioner. 13 The High Court should have examined the questions whether the petitioner was in actual physical of the area which was taken	The vanilla incorrectly explains the explanation of the case text. The reason behind hallucination is lack of contextual understanding of the case information, overemphasis on certain parts of the input, and missing information such as the agreement of 1934 and 1950. One of the reasons could also be the complex legal language that the model could not fully grasp.
INLegalLlama (CPT+SFT) Instruction-tuning for prediction with explanation task	Case decision: Accepted Explanation: Further there was no agreement as regards the royalties to be paid under the renewed lease. 23. With regard to the right claimed under the deed of assignment it was argued that the assignment could take place only when the lease was in force and as the lease of 1917 had expired in 1947 no valid assignment could be taken place if indeed, there was, in fact, any assignment of the colliery as alleged. 24. It was urged that without examining those questions and without coming to the conclusion that the petitioner-company had by their petition disclosed a legal right in them with regard to the colliery or to continue in possession thereof the High Court was not competent to grant them any relief as it had purported to do. It was submitted that at the most there was a term in the agreement of 1934 read with the lease of 1917 which entitled the Hindu undivided family to sue for specific performance of the contract to renew for a further term. The demise in its favor of the veins etc., under the lease of 1917 had come to an end and even if the person in possession of the seams and veins which were being worked could not be dispossessed by an order of the nature disclosed in the letter of February 1970, the petitioner-company had not been able to show any right in it to claim possession of the virgin area, that it to say, the seems, veins, mines , etc., of which they were not in physical possession and which they were not working at the moment, when the impugned order was served on them. 25. In our opinion, the High Court should have examined these questions and especially the question as to whether the petitioner-company had been able to substantiate the claim that it was in possession of the entire underground rights in respect of Ac. 871-49 covered by the agreement of 1934. The agreement of 1934 bound the Government and the Hindu undivided family: the petitioner company was not a party to it. Even if the Hindu undivided family continued in possession of some of the mines	The Finetuned model correctly explains the reasoning behind the case judgment comparatively better. The reasoning contains better context information and is organized logically. This explanation includes specific legal arguments related to the agreements. The explanation also compares the petitioner's claim with actual agreements and classes.

Table 10: Comparative analysis of responses generated by Llama-2-7B (CPT) and INLegalLlama (CPT+SFT).

CASE NO:
CIVIL APPEAL NOS. 3088-3089 OF 2020
APPELLANTS:
BHARTI AIRTEL LIMITED AND ANOTHER
RESPONDENT:
VIJAYKUMAR V. IYER AND OTHERS
DATE OF JUDGMENT:
03/01/2024
BENCH:
Dipankar Datta
CASE TEXT:
... The present appeals raise an interesting question on the right to claim set-off in the Corporate Insolvency Resolution Process when the Resolution Professional proceeds in terms of clause (a) to sub-section (2) of Section 25 of the Insolvency and Bankruptcy Code, 20161 to take custody and control of all the assets of the corporate debtor ....The NCLAT vide order dated 17th May 2019 allowed the appeal, inter alia, holding that set-off is violative of the basic principles and protection accorded under any insolvency law. Set-off is antithetical to the objective of the IBC. Reference was made to the non-obstante provisions in the form of Section 238 of the IBC. As moratorium under Section 14(4) applies till the date of completion of the Corporate Insolvency Resolution Process, which is till the resolution plan... ...The judgment of this Court in The Official Liquidator of High Court of Karnataka v. Smt. V. Lakshmikutty53 had applied Section 46 of the Provincial Insolvency Act, 1920 and had accordingly permitted insolvency set-off on interpretation and application of Sections 529 and 530 of the Companies Act, 1956. In that context, it is observed that the English courts, on 52 Career Institute Educational Society v. Om Shree Thakurji Educational Society, 2023 SCC OnLine SC 586.... ... Thereupon the Airtel entities made a payment of Rs.341.80 crores due to the Aircel entities on 10th January 2019. The balance amount of Rs.145.20 crores was set-off by the Airtel entities on the ground that this amount was owned by the Aircel entities to the Airtel entities. According to Airtel entities, Rs.145.20 crores was the adjusted or the net amount payable by the Aircel entities towards operational charges, SMS charges and interconnect usage charges to the Airtel entities ...
JUDGEMENT:
.... Having considered the contentions raised by the appellant Airtel entities in detail, and in light of the provisions of the IBC relating to the Corporate Insolvency Resolution Process, we do not find any merit in the present appeals and the same are dismissed. There will be no order as to costs....

Table 11: 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: Given a Supreme Court of India case proceeding enclosed in angle brackets < >, your task is to predict the decision of the case (with respect to the appelant) and provide an explaination for the decision. **Prediction**: Given a case proceeding, the task is to predict the decision 0 or 1, where the label 1 corresponds to the acceptance of the appeal/petition of the appellant/petitioner and the label 0 corresponds to the rejection of the appeal/petition of the appellant/petitioner, Explanation: The task is to explain how you arrived at the decision by predicting important sentences that lead to the decision. **Context**: Answer in a consistent style as shown in the following two examples: **case_proceeding**: # case_proceeding example 1 **Prediction**: # example 1 prediction **Explanation**: # example 1 explanation **case_proceeding**: # case_proceeding example 2 **Prediction**: # example 2 prediction **Explanation**: # example 2 explanation **Instructions**: Learn from the above two examples and perform the task for the following case proceeding. case_proceeding: <{case_proceeding}> Format your output in list format: [prediction, explanation]”””
Template 2 (prediction only)
prompt = f“““Task: Given a Supreme Court of India case proceeding enclosed in angle brackets < >, your task is to predict the decision of the case (with respect to the appellant). **Prediction**: Given a case proceeding, the task is to predict the decision 0 or 1, where the label 1 corresponds to the acceptance of the appeal/petition of the appellant/petitioner and the label 0 corresponds to the rejection of the appeal/petition of the appellant/petitioner **Context**: Answer in a consistent style as shown in the following two examples: **case_proceeding**: # case_proceeding example 1 **Prediction**: # example 1 prediction **case_proceeding**: # case_proceeding example 2 **Prediction**: # example 2 prediction **Instructions**: Learn from the above two examples and perform the task for the following case proceeding. **case_proceeding**: <{case_proceeding}> Give the output predicted case decision as either 0 or 1.”””

Table 12: Prompts for Judgment Prediction taken from (Vats et al., 2023).

Template 3 (prediction only)
prompt = f“““ ### **Instructions**: Analyze the case proceeding and predict whether the appeal/petition will be rejected (0) or accepted (1). ### **Input**: <{case_proceeding}> ### Response: ”””
Template 4 (prediction with explanation)
prompt = f“““ ### **Instructions**: Analyze the case proceeding and predict whether the appeal/petition will be accepted (1) or rejected (0), and subsequently provide an explanation behind this prediction with important textual evidence from the case. ### **Input**: <{case_proceeding}> ### Response: ”””

Table 13: Prompts for Judgment Prediction used for instruction fine-tuned models taken from (Nigam et al., 2024). Instructions were randomly chosen from Table 14.

Instruction sets for Predicting the Decision	
1	Analyze the case proceeding and predict whether the appeal/petition will be accepted (1) or rejected (0).
2	Based on the information in the case proceeding, determine the likely outcome: acceptance (1) or rejection (0) of the appellant/petitioner's case.
3	Review the case details and predict the decision: will the court accept (1) or deny (0) the appeal/petition?
4	Considering the arguments and evidence in the case proceeding, predict the verdict: is it more likely to be in favor (1) or against (0) the appellant?
5	Examine the details of the case proceeding and forecast if the appeal/petition stands a chance of being upheld (1) or dismissed (0).
6	Assess the case proceedings and provide a prediction: is the court likely to rule in favor of (1) or against (0) the appellant/petitioner?
7	Interpret the case information and speculate on the court's decision: acceptance (1) or rejection (0) of the presented appeal.
8	Given the specifics of the case proceeding, anticipate the court's ruling: will it favor (1) or oppose (0) the appellant's request?
9	Scrutinize the evidence and arguments in the case proceeding to predict the court's decision: will the appeal be granted (1) or denied (0)?
10	Analyze the legal arguments presented and estimate the likelihood of the court accepting (1) or rejecting (0) the petition.
11	From the information provided in the case proceeding, infer whether the court's decision will be positive (1) or negative (0) for the appellant.
12	Evaluate the arguments 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 proceeding and predict the outcome: is the judgment expected to be in support (1) or in denial (0) of the appeal?
14	Using the case data, forecast whether the court is likely to side with (1) or against (0) the appellant/petitioner.
15	Examine the case narrative and anticipate the court's decision: will it result in an approval (1) or disapproval (0) of the appeal?
16	Based on the legal narrative and evidentiary details in the case proceeding, 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 (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 proceeding (1 for acceptance, 0 for rejection) and subsequently provide an explanation based on significant sentences in the proceeding.
4	Evaluate the case proceeding 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 in the case proceeding, and then clarify this prediction by discussing critical sentences from the text.
6	Judge the probable resolution of the case (approval (1) or disapproval (0)), and elaborate on this forecast by extracting and interpreting significant sentences from the proceeding.
7	Forecast the likely verdict of the case (granting (1) or denying (0) the appeal) and then 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)), and then expound on this prediction by highlighting and analyzing key textual elements from the proceeding.
9	Decide if the appeal in the case proceeding is more likely to be successful (1) or unsuccessful (0), and then justify your decision by focusing on essential sentences in the document.
10	Conjecture the end result of the case (acceptance (1) or non-acceptance (0) of the appeal), followed by a detailed explanation using crucial sentences from the case proceeding.
11	Predict whether the case will result in an affirmative (1) or negative (0) decision for the appeal, 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) and then give 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 proceeding, and subsequently give 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), and then dissect the proceeding to provide a detailed explanation using key textual passages.
15	Speculate on the likely judgment (yes (1) or no (0) to the appeal) and then delve into the case proceeding to elucidate your prediction, focusing on critical sentences.
16	Hypothesize the court's verdict (affirmation (1) or negation (0) of the appeal), and then clarify this hypothesis by interpreting significant sentences from the case proceeding.

Table 14: Instruction Sets for Predicting Legal Decisions and Providing Explanations taken from (Nigam et al., 2024)

Test Data	Macro Precision	Macro Recall	Macro F1	Accuracy
InLegalBert				
ILDC	0.7486	0.7464	0.7475	0.7462
SCI (2019)	0.8391	0.8391	0.8391	0.8388
SCI (2020-24)	0.8712	0.8875	0.8793	0.8913
SCI+HCs (2019)	0.8843	0.8831	0.8837	0.8837
HCs (2020-24)	**0.9006**	**0.9018**	**0.9012**	**0.9012**
SCI+HCs+Tribunal (2019)	0.8753	0.8749	0.8751	0.8751
Tribunal (2020-24)	0.8545	0.8527	0.8536	0.8541
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8800	0.8795	0.8797	0.8797
Daily_orders (2020-24)	0.8734	0.8801	0.8768	0.8830
InCaseLaw				
ILDC	0.7090	0.7085	0.7088	0.7086
SCI (2019)	0.8209	0.8209	0.8209	0.8209
SCI (2020-24)	0.8453	0.8659	0.8555	0.8681
SCI+HCs (2019)	0.8564	0.8543	0.8554	0.8554
HCs (2020-24)	**0.8708**	**0.8717**	**0.8712**	**0.8690**
SCI+HCs+Tribunal (2019)	0.8603	0.8593	0.8598	0.8596
Tribunal (2020-24)	0.8451	0.8412	0.8431	0.8433
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8578	0.8565	0.8571	0.8570
Daily_orders (2020-24)	0.8457	0.8609	0.8532	0.8564
XLNet Large				
ILDC	0.7112	0.7044	0.7078	0.7040
SCI (2019)	0.8095	0.8082	0.8089	0.8076
SCI (2020-24)	0.8255	0.8570	0.8410	0.8482
SCI+HCs (2019)	0.8583	0.8533	0.8558	0.8550
HCs (2020-24)	**0.8725**	**0.8724**	**0.8724**	**0.8688**
SCI+HCs+Tribunal (2019)	0.8653	0.8624	0.8639	0.8629
Tribunal (2020-24)	0.8625	0.8570	0.8598	0.8595
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8640	0.8601	0.8620	0.8610
Daily_orders (2020-24)	0.8282	0.8473	0.8376	0.8363

Table 15: Judgment prediction results on the binary task across different court cases and temporal test cases, with models trained on SCI + HCs + Tribunal data from `NyayaAnumana` single-split data. The best results are highlighted in bold.

Test Data	Macro Precision	Macro Recall	Macro F1	Accuracy
InLegalBert				
ILDC	0.7209	0.7169	0.7189	0.7172
SCI (2019)	0.8261	0.8255	0.8258	0.8258
SCI (2020-24)	0.8515	0.8588	0.8552	0.8720
SCI+HCs (2019)	0.8739	0.8735	0.8737	0.8739
HCs (2020-24)	**0.8940**	**0.8943**	**0.8942**	0.8945
SCI+HCs+Tribunal (2019)	0.8637	0.8634	0.8635	0.8635
Tribunal (2020-24)	0.8308	0.8249	0.8278	0.8277
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8722	0.8718	0.8720	0.8720
Daily_orders (2020-24)	0.8897	0.8869	0.8883	**0.8955**
InCaseLaw				
ILDC	0.7347	0.7335	0.7341	0.7337
SCI (2019)	0.8271	0.8272	0.8271	0.8272
SCI (2020-24)	0.8449	0.8579	0.8513	0.8670
SCI+HCs (2019)	0.8585	0.8570	0.8578	0.8579
HCs (2020-24)	**0.8891**	**0.8898**	**0.8895**	0.8898
SCI+HCs+Tribunal (2019)	0.8544	0.8521	0.8532	0.8526
Tribunal (2020-24)	0.8160	0.7939	0.8048	0.8001
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8573	0.8553	0.8563	0.8559
Daily_orders (2020-24)	0.8868	0.8795	0.8831	**0.8911**
XLNet Large				
ILDC	0.6851	0.6850	0.6851	0.6849
SCI (2019)	0.8150	0.8137	0.8143	0.8142
SCI (2020-24)	0.8507	0.8562	0.8535	0.8709
SCI+HCs (2019)	0.8590	0.8585	0.8588	0.8590
HCs (2020-24)	**0.8848**	0.8863	0.8856	**0.8851**
SCI+HCs+Tribunal (2019)	0.8580	0.8576	0.8578	0.8577
Tribunal (2020-24)	0.8180	0.8053	0.8116	0.8098
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8660	0.8655	0.8657	0.8657
Daily_orders (2020-24)	0.8832	**0.8904**	**0.8868**	0.8924

Table 16: Judgment prediction results on the binary task across different court cases and temporal test cases, with models trained on SCI + HCs + Tribunal + Daily Orders and District Court data from `NyayaAnumana` multi-split data. The best results are highlighted in bold.

Test Data	Macro Precision	Macro Recall	Macro F1	Accuracy
InLegalBert				
ILDC	0.7492	0.7351	0.7421	0.7357
SCI (2019)	0.8532	0.8437	0.8484	0.8451
SCI (2020-24)	**0.9102**	0.8798	**0.8947**	**0.9095**
SCI+HCs (2019)	0.8822	0.8814	0.8818	0.8799
HCs (2020-24)	0.8908	**0.8869**	0.8888	0.8891
SCI+HCs+Tribunal (2019)	0.8785	0.8744	0.8764	0.8737
Tribunal (2020-24)	0.8275	0.8194	0.8234	0.8142
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8544	0.8493	0.8519	0.8483
Daily_orders (2020-24)	0.8852	0.8686	0.8768	0.8855
InCaseLaw				
ILDC	0.6856	0.6369	0.6604	0.6381
SCI (2019)	0.7965	0.7744	0.7852	0.7767
SCI (2020-24)	**0.8673**	0.8385	**0.8526**	**0.8742**
SCI+HCs (2019)	0.8211	0.8152	0.8182	0.8124
HCs (2020-24)	0.8501	**0.8446**	0.8474	0.8477
SCI+HCs+Tribunal (2019)	0.8262	0.8164	0.8213	0.8153
Tribunal (2020-24)	0.8234	0.8194	0.8214	0.8154
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8259	0.8174	0.8216	0.8157
Daily_orders (2020-24)	0.8624	0.8379	0.8500	0.8610
XLNet Large				
ILDC	0.7257	0.7107	0.7181	0.7113
SCI (2019)	0.8479	0.8356	0.8417	0.8371
SCI (2020-24)	0.8965	0.8794	0.8878	0.9034
SCI+HCs (2019)	0.8686	0.8668	0.8677	0.8650
HCs (2020-24)	**0.9065**	**0.9023**	**0.9044**	**0.9045**
SCI+HCs+Tribunal (2019)	0.8634	0.8571	0.8602	0.8562
Tribunal (2020-24)	0.8255	0.8198	0.8227	0.8153
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8613	0.8557	0.8585	0.8544
Daily_orders (2020-24)	0.8813	0.8638	0.8725	0.8815

Table 17: Judgment prediction results on the binary task across different court cases and temporal test cases, with models trained on SCI + HCs + Tribunal data from NyayaAnumana multi-split data. The best results are highlighted in bold.

Test Data	Macro Precision	Macro Recall	Macro F1	Accuracy
InLegalBert				
ILDC	0.7544	0.7536	0.7540	0.7535
SCI (2019)	0.8309	0.8308	0.8308	0.8309
SCI (2020-24)	0.8742	0.8807	0.8774	0.8918
SCI+HCs (2019)	0.8691	0.8692	0.8692	0.8693
HCs (2020-24)	**0.8952**	**0.8953**	**0.8952**	**0.8956**
SCI+HCs+Tribunal (2019)	0.8542	0.8538	0.8540	0.8540
Tribunal (2020-24)	0.8086	0.7849	0.7966	0.7914
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.7499	0.7421	0.7460	0.7406
Daily_orders (2020-24)	0.8687	0.8684	0.8685	0.8767
InCaseLaw				
ILDC	0.7513	0.7503	0.7508	0.7502
SCI (2019)	0.8327	0.8328	0.8328	0.8327
SCI (2020-24)	0.8555	0.8710	0.8632	0.8769
SCI+HCs (2019)	0.8660	0.8652	0.8656	0.8658
HCs (2020-24)	**0.8961**	**0.8969**	**0.8965**	**0.8967**
SCI+HCs+Tribunal (2019)	0.8511	0.8494	0.8503	0.8498
Tribunal (2020-24)	0.7923	0.7599	0.7758	0.7679
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8497	0.8479	0.8488	0.8485
Daily_orders (2020-24)	0.8696	0.8721	0.8709	0.8784
XLNet Large				
ILDC	0.7247	0.7245	0.7246	0.7245
SCI (2019)	0.8375	0.8368	0.8372	0.8371
SCI (2020-24)	0.8680	0.8792	0.8736	0.8874
SCI+HCs (2019)	0.8788	0.8792	0.8790	0.8789
HCs (2020-24)	**0.9101**	**0.9099**	**0.9100**	**0.9104**
SCI+HCs+Tribunal (2019)	0.8649	0.8650	0.8649	0.8649
Tribunal (2020-24)	0.8114	0.8008	0.8060	0.8049
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8653	0.8654	0.8654	0.8654
Daily_orders (2020-24)	0.8684	0.8741	0.8713	0.8780

Table 18: Judgment prediction results on the binary task across different court cases and temporal test cases, with models trained on SCI + HCs data from NyayaAnumana single split data. The best results are highlighted in bold.

Test Data	Macro Precision	Macro Recall	Macro F1	Accuracy
InLegalBert				
ILDC	0.7364	0.7078	0.7218	0.7086
SCI (2019)	0.8480	0.8323	0.8401	0.8341
SCI (2020-24)	0.8986	0.8614	0.8796	0.8968
SCI+HCs (2019)	0.8689	0.8653	0.8671	0.8630
HCs (2020-24)	**0.8949**	**0.8881**	**0.8915**	**0.8912**
SCI+HCs+Tribunal (2019)	0.8529	0.8474	0.8501	0.8465
Tribunal (2020-24)	0.8069	0.8072	0.8071	0.8074
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.7672	0.7406	0.7537	0.7381
Daily_orders (2020-24)	0.8779	0.8593	0.8685	0.8779
InCaseLaw				
ILDC	0.6963	0.6675	0.6816	0.6684
SCI (2019)	0.8062	0.7930	0.7995	0.7947
SCI (2020-24)	0.8652	0.8516	0.8584	**0.8780**
SCI+HCs (2019)	0.8315	0.8265	0.8290	0.8239
HCs (2020-24)	**0.8678**	**0.8584**	**0.8631**	0.8624
SCI+HCs+Tribunal (2019)	0.8174	0.8112	0.8143	0.8102
Tribunal (2020-24)	0.7831	0.7828	0.7829	0.7836
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8172	0.8118	0.8145	0.8104
Daily_orders (2020-24)	0.8583	0.8304	0.8441	0.8555
XLNet Large				
ILDC	0.7265	0.7181	0.7223	0.7185
SCI (2019)	0.8412	0.8292	0.8351	0.8307
SCI (2020-24)	**0.9118**	0.8910	0.9013	**0.9150**
SCI+HCs (2019)	0.8645	0.8624	0.8635	0.8605
HCs (2020-24)	0.9060	**0.9014**	**0.9037**	0.9037
SCI+HCs+Tribunal (2019)	0.8505	0.8466	0.8485	0.8459
Tribunal (2020-24)	0.8159	0.8166	0.8162	0.8164
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.8506	0.8471	0.8488	0.8462
Daily_orders (2020-24)	0.8783	0.8606	0.8693	0.8787

Table 19: Judgment prediction results on the binary task across different court cases and temporal test cases, with models trained on SCI + HCs data from NyayaAnumana multi-split data. The best results are highlighted in bold.

Test Data	Macro Precision	Macro Recall	Macro F1	Accuracy
InLegalBert				
ILDC	0.7147	0.7145	0.7146	0.7146
SCI (2019)	0.8212	0.8187	0.8200	0.8194
SCI (2020-24)	**0.8983**	**0.8532**	**0.8752**	**0.8929**
SCI+HCs (2019)	0.7691	0.7664	0.7677	0.7642
HCs (2020-24)	0.7919	0.7585	0.7748	0.7683
SCI+HCs+Tribunal (2019)	0.7514	0.7447	0.7481	0.7437
Tribunal (2020-24)	0.7300	0.6655	0.6963	0.6514
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.7499	0.7421	0.7460	0.7406
Daily_orders (2020-24)	0.7849	0.7389	0.7612	0.7814
InCaseLaw				
ILDC	0.7067	0.7066	0.7066	0.7067
SCI (2019)	0.8093	0.8083	0.8088	0.8088
SCI (2020-24)	**0.8782**	0.8394	0.8584	**0.8791**
SCI+HCs (2019)	0.7596	0.7578	0.7587	0.7559
HCs (2020-24)	0.8678	**0.8584**	**0.8631**	0.8624
SCI+HCs+Tribunal (2019)	0.7420	0.7369	0.7394	0.7359
Tribunal (2020-24)	0.7031	0.6493	0.6751	0.6355
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.7397	0.7340	0.7368	0.7324
Daily_orders (2020-24)	0.7826	0.7356	0.7584	0.7790
XLNet Large				
ILDC	0.7356	0.7342	0.7349	0.7343
SCI (2019)	0.8495	0.8488	0.8492	0.8491
SCI (2020-24)	**0.9185**	**0.8961**	**0.9071**	**0.9200**
SCI+HCs (2019)	0.8064	0.8063	0.8063	0.8052
HCs (2020-24)	0.8342	0.8288	0.8315	0.8320
SCI+HCs+Tribunal (2019)	0.7936	0.7912	0.7924	0.7906
Tribunal (2020-24)	0.7810	0.7629	0.7718	0.7555
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.7916	0.7894	0.7905	0.7887
Daily_orders (2020-24)	0.8224	0.8167	0.8195	0.8319

Table 20: Judgment prediction results on the binary task across different court cases and temporal test cases, with models trained on SCI data from NyayaAnumana single split data. The best results are highlighted in bold.

Test Data	Macro Precision	Macro Recall	Macro F1	Accuracy
InLegalBert				
ILDC	0.7216	0.6866	0.7037	0.6875
SCI (2019)	0.8370	0.8230	0.8299	0.8246
SCI (2020-24)	**0.9107**	**0.8546**	**0.8818**	**0.8979**
SCI+HCs (2019)	0.7781	0.7623	0.7701	0.7577
HCs (2020-24)	0.8051	0.7709	0.7876	0.7806
SCI+HCs+Tribunal (2019)	0.7678	0.7427	0.7550	0.7407
Tribunal (2020-24)	0.7354	0.6537	0.6921	0.6380
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.7672	0.7406	0.7537	0.7381
Daily_orders (2020-24)	0.8241	0.7509	0.7858	0.8003
InCaseLaw				
ILDC	0.7134	0.6881	0.7005	0.6889
SCI (2019)	0.8385	0.8265	0.8325	0.8280
SCI (2020-24)	**0.9042**	0.8535	**0.8781**	**0.8951**
SCI+HCs (2019)	0.7696	0.7546	0.7620	0.7501
HCs (2020-24)	0.8678	**0.8584**	0.8631	0.8624
SCI+HCs+Tribunal (2019)	0.7631	0.7391	0.7509	0.7371
Tribunal (2020-24)	0.7272	0.6485	0.6856	0.6328
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.7607	0.7367	0.7485	0.7337
Daily_orders (2020-24)	0.8140	0.7386	0.7744	0.7903
XLNet Large				
ILDC	0.7229	0.7020	0.7123	0.7027
SCI (2019)	0.8851	0.8776	0.8813	0.8787
SCI (2020-24)	**0.9187**	**0.8787**	**0.8982**	**0.9123**
SCI+HCs (2019)	0.8026	0.7921	0.7973	0.7884
HCs (2020-24)	0.8282	0.8097	0.8189	0.8162
SCI+HCs+Tribunal (2019)	0.7913	0.7726	0.7818	0.7710
Tribunal (2020-24)	0.7748	0.7402	0.7571	0.7303
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.7901	0.7717	0.7808	0.7697
Daily_orders (2020-24)	0.8300	0.7825	0.8056	0.8201

Table 21: Judgment prediction results on the binary task across different court cases and temporal test cases, with models trained on SCI data from NyayaAnumana multi-split data. The best results are highlighted in bold.

Test Data	Macro Precision	Macro Recall	Macro F1	Accuracy
InLegalBert				
ILDC	0.7447	0.7437	0.7442	0.7436
SCI (2019)	0.7559	0.7516	0.7538	0.7527
SCI (2020-24)	**0.8259**	**0.7305**	**0.7753**	**0.8113**
SCI+HCs (2019)	0.6810	0.6657	0.6733	0.6604
HCs (2020-24)	0.6766	0.6155	0.6446	0.6338
SCI+HCs+Tribunal (2019)	0.6562	0.6384	0.6472	0.6362
Tribunal (2020-24)	0.6446	0.5566	0.5974	0.5363
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.6793	0.6722	0.6757	0.6737
Daily_orders (2020-24)	0.6865	0.6033	0.6422	0.6829
InCaseLaw				
ILDC	0.7451	0.7380	0.7415	0.7376
SCI (2019)	0.7367	0.7347	0.7357	0.7354
SCI (2020-24)	**0.7876**	**0.6971**	**0.7396**	**0.7848**
SCI+HCs (2019)	0.6573	0.6460	0.6516	0.6411
HCs (2020-24)	0.6306	0.5864	0.6077	0.6046
SCI+HCs+Tribunal (2019)	0.6341	0.6188	0.6264	0.6167
Tribunal (2020-24)	0.6362	0.5480	0.5888	0.5271
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.6333	0.6158	0.6244	0.6123
Daily_orders (2020-24)	0.6648	0.5959	0.6285	0.6736
XLNet Large				
ILDC	0.7429	0.7361	0.7395	0.7357
SCI (2019)	0.7577	0.7575	0.7576	0.7577
SCI (2020-24)	**0.7963**	**0.7855**	**0.7909**	**0.8201**
SCI+HCs (2019)	0.7114	0.7114	0.7114	0.7118
HCs (2020-24)	0.6982	0.6988	0.6985	0.6990
SCI+HCs+Tribunal (2019)	0.6897	0.6897	0.6897	0.6897
Tribunal (2020-24)	0.6613	0.6515	0.6564	0.6446
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.6895	0.6895	0.6895	0.6894
Daily_orders (2020-24)	0.7028	0.7064	0.7046	0.7199

Table 22: Judgment prediction results on the binary task across different court cases and temporal test cases, with models trained on data from ILDC single data. The best results are highlighted in bold.

Test Data	Macro Precision	Macro Recall	Macro F1	Accuracy
InLegalBert				
ILDC	0.7706	0.7650	0.7678	0.7647
SCI (2019)	0.7799	0.7735	0.7767	0.7721
SCI (2020-24)	**0.7868**	**0.8231**	**0.8045**	**0.8046**
SCI+HCs (2019)	0.7035	0.6907	0.6971	0.6949
HCs (2020-24)	0.6551	0.6446	0.6498	0.6355
SCI+HCs+Tribunal (2019)	0.6840	0.6770	0.6805	0.6782
Tribunal (2020-24)	0.6463	0.6464	0.6463	0.6447
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.6544	0.6337	0.6439	0.6308
Daily_orders (2020-24)	0.5982	0.5918	0.5950	0.5428
InCaseLaw				
ILDC	0.7592	0.7461	0.7526	0.7456
SCI (2019)	**0.7691**	0.7622	0.7656	0.7608
SCI (2020-24)	0.7687	**0.8068**	**0.7873**	**0.7765**
SCI+HCs (2019)	0.6908	0.6822	0.6865	0.6857
HCs (2020-24)	0.6367	0.6271	0.6319	0.6180
SCI+HCs+Tribunal (2019)	0.6691	0.6651	0.6671	0.6661
Tribunal (2020-24)	0.6242	0.6241	0.6242	0.6221
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.6667	0.6628	0.6647	0.6642
Daily_orders (2020-24)	0.5842	0.5818	0.5830	0.5399
XLNet Large				
ILDC	0.7873	0.7840	0.7856	0.7838
SCI (2019)	**0.8104**	**0.7959**	**0.8031**	**0.7939**
SCI (2020-24)	0.7483	0.7839	0.7657	0.7417
SCI+HCs (2019)	0.7224	0.7023	0.7122	0.7074
HCs (2020-24)	0.6918	0.6677	0.6795	0.6556
SCI+HCs+Tribunal (2019)	0.7100	0.6949	0.7023	0.6965
Tribunal (2020-24)	0.6704	0.6559	0.6631	0.6631
SCI+HCs+Tribunal+ DailyOrders+DistrictCourts (2019)	0.7096	0.6942	0.7018	0.6963
Daily_orders (2020-24)	0.6358	0.6247	0.6302	0.5709

Table 23: Judgment prediction results on the binary task across different court cases and temporal test cases, with models trained on data from ILDC multi data. The best results are highlighted in bold.

Models	Metric	Overall	Class 0	Class 1	Class 2
InLegalBert	Macro Precision	0.6950	0.81	0.84	0.44
	Macro Recall	**0.5883**	0.82	0.85	0.10
	Macro F1	**0.6062**	0.81	0.85	0.16
InCaseLaw	Macro Precision	**0.6984**	0.80	0.83	0.46
	Macro Recall	0.5608	0.81	0.84	0.03
	Macro F1	0.5653	0.81	0.84	0.05
XLNet	Macro Precision	0.6853	0.80	0.84	0.42
	Macro Recall	0.5800	0.81	0.84	0.08
	Macro F1	0.5952	0.81	0.84	0.14

Table 24: Judgment prediction results on the ternary task on SCI + HCs court cases. The best results are highlighted in bold.

Models	Metric	Overall	Class 0	Class 1	Class 2
InLegalBert	Macro Precision	**0.5432**	0.80	0.83	0.00
	Macro Recall	**0.5453**	0.77	0.87	0.00
	Macro F1	**0.5440**	0.79	0.85	0.00
InCaseLaw	Macro Precision	0.4957	0.72	0.77	0.00
	Macro Recall	0.4981	0.69	0.81	0.00
	Macro F1	0.4966	0.70	0.79	0.00
XLNet	Macro Precision	0.5376	0.79	0.82	0.00
	Macro Recall	0.5411	0.77	0.85	0.00
	Macro F1	0.5392	0.78	0.84	0.00

Table 25: Judgment prediction results on the ternary task on SCI + HCs + Tribunals court cases. The best results are highlighted in bold.

Models	Metric	Overall	Class 0	Class 1	Class 2
InLegalBert	Macro Precision	**0.5401**	0.77	0.85	0.00
	Macro Recall	**0.5476**	0.82	0.83	0.00
	Macro F1	**0.5436**	0.79	0.84	0.00
InCaseLaw	Macro Precision	0.4516	0.63	0.73	0.00
	Macro Recall	0.4564	0.64	0.72	0.00
	Macro F1	0.4540	0.64	0.73	0.00
XLNet	Macro Precision	0.5362	0.76	0.85	0.00
	Macro Recall	0.5446	0.82	0.81	0.00
	Macro F1	0.5399	0.79	0.83	0.00

Table 26: Judgment prediction results on the ternary task on SCI + HCs + Tribunals + Daily Orders + District Court cases. The best results are highlighted in bold.

Court-wise	Raw files	FIles After Preprocessing	FIles After Lebeling
Supreme Court	55928	54831	40562
High Court	1324373	977849	482295
Tribunal Court	477397	318681	186671
Daily Orders and District Courts	424439	312480	92771
Total	2282137	1663841	802299

Table 27: Number of cases before and after preprocessing, by court type

Generative Models	Rating Score 1	2	3	4	5
	PredEx				
Llama-2-7B	2	11	22	12	3
Llama-2 SFT	5	13	18	13	1
Llama-2 CPT	2	2	27	19	0
INLegalLlama CPT+SFT	0	0	23	27	0
	ILDC expert				
Llama-2-7B	0	9	22	21	2
Llama-2 SFT	2	3	16	24	9
Llama-2 CPT	1	3	25	23	2
INLegalLlama CPT+SFT	0	0	22	28	4

Table 28: Distribution of Expert Rating Scores for Generative Models on PredEx and ILDC Expert Data.