

Legal Judgment Reimagined: PredEx and the Rise of Intelligent AI Interpretation in Indian Courts

Shubham Kumar Nigam^{1*†} Anurag Sharma^{2*} Danush Khanna^{3*}

Noel Shallum⁴ Kripabandhu Ghosh² Arnab Bhattacharya¹

¹ Indian Institute of Technology Kanpur, India ² IISER Kolkata, India

³ Manipal University Jaipur, India ⁴ Symbiosis Law School Pune, India

sknigam@cse.iitk.ac.in, as19ms159@iiserkol.ac.in,

danush.229310455@uj.manipal.edu, kripaghosh@iiserkol.ac.in,

noelshallum@gmail.com, arnabb@cse.iitk.ac.in

Abstract

In the era of Large Language Models (LLMs), predicting judicial outcomes poses significant challenges due to the complexity of legal proceedings and the scarcity of expert-annotated datasets. Addressing this, we introduce **Prediction with Explanation (PredEx)**, the largest expert-annotated dataset for legal judgment prediction and explanation in the Indian context, featuring over 15,000 annotations. This groundbreaking corpus significantly enhances the training and evaluation of AI models in legal analysis, with innovations including the application of instruction tuning to LLMs. This method has markedly improved the predictive accuracy and explanatory depth of these models for legal judgments. We employed various transformer-based models, tailored for both general and Indian legal contexts. Through rigorous lexical, semantic, and expert assessments, our models effectively leverage PredEx to provide precise predictions and meaningful explanations, establishing it as a valuable benchmark for both the legal profession and the NLP community.

1 Introduction

In the evolving landscape of legal technology, the integration of Artificial Intelligence (AI) into the judicial system has emerged as a frontier of immense potential and challenge. The Indian judiciary, characterized by a significant backlog of cases¹, stands to benefit substantially from advancements in AI-assisted legal decision-making. This paper introduces a novel approach to facilitating the legal decision-making process, specifically focusing on the Indian context, in conjunction with explanations for the same. Our work builds upon two foundational studies: (Malik et al., 2021) and (Vats

et al., 2023). Our objective is to develop an advanced system capable of predicting judicial outcomes and providing cogent explanations for these predictions. This system leverages a newly compiled dataset, PredEx, of approximately 15,000 annotated legal documents, considerably larger than those used in previous research, particularly in terms of its volume and depth of annotations. Table 1 compares PredEx with other popularly used corpora for legal judgment prediction, highlighting the uniqueness of our dataset in terms of its size and focus on providing explanations. Unlike previous works that predominantly focused on predicting legal outcomes, PredEx introduces the largest annotated dataset for judgment prediction and explanation in the Indian legal context, addressing a critical gap in legal AI research. This dataset enables us to train and refine sophisticated machine learning models, particularly focusing on instruction tuning, to achieve unprecedented accuracy and relevancy in legal judgment prediction.

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

1. Publication of a New Annotated Dataset (PredEx): We introduce the largest annotated dataset to date for judgment prediction and explanation in the Indian legal context. This dataset surpasses previous efforts in both scope and depth, providing a more robust foundation for training AI models in legal judgment prediction.
2. Exploration of Instruction-Tuning on Large Language Models (LLMs): Our work goes beyond the traditional methods of fine-tuning conventional transformers. We delve into instruction tuning on LLMs, an approach not extensively explored in previous research, to enhance prediction accuracy.

*These authors contributed equally to this work

†Corresponding author

¹<https://www.nytimes.com/2024/01/13/world/asia/india-judicial-backlog.html>

Corpus	Language	Jurisdiction	# of Cases	# of Human annotated Docs	Avg # of Tokens	Annotated LJP Subtasks (# of labels w.r.t Subtask)	Additional Annotation
FCCR (Sulea et al., 2017)	French	France	126,865	0	-	Court Decision (6 and 8 w.r.t. two setups)	date of the court ruling and law area
CAIL (Xiao et al., 2018)	Chinese	China	2,676,075	0	-	Law Article (183) Charge (202) Prison Term (integer value)	the defendant and penalty of money
ECHR (Chalkidis et al., 2019)	English	Europe	11,478	0	2421	Violation (2) Law Article (66)	case importance
ECHR (Chalkidis et al., 2021)	English	Europe	11,000	50 (fact paragraphs)	-	Alleged Law Article (40) Violation (2) Law Article (40)	paragraph-level rationale
SJP (Niklaus et al., 2021)	German French Italian	Switzerland	49,883 (German) 31,094 (French) 4,292 (Italian)	200 (German) (Court Decision)	850	Court Decision (2)	publication year legal area canton of origin
ILDC (Malik et al., 2021)	English	India	34,816	56 (Court Decision and Explanation)	3231	Court Decision (2)	sentence-level explanation
HLDC (Kapoor et al., 2022)	Hindi	India	340,280	0	764	Bail Prediction (2)	extractive summarization
BCD (Lage-Freitas et al., 2022)	Portuguese	Brazil	4,043	0	119	Court Decision (3) decision's unanimity status	unanimity label
(Our dataset) PredEx 2024	English	India	15,222	15,222	4,504	Court Decision (2) Explanation for Decision	expert ratings of generated responses for 50 PredEx and 54 ILDC experts

Table 1: Comparative Overview of Widely Used Legal Judgment Prediction Datasets. Entries marked with ‘-’ denote unknown or unavailable information.

3. Expert Evaluation and Validation: We employ a rigorous evaluation process, utilizing a Likert score scale to assess the efficacy of our system. This evaluation, conducted on a sample of 50 documents, provides critical insights into the performance of our AI models compared to human expert standards.

Our research aims to provide a comprehensive and sophisticated AI-based system for legal judgment prediction and explanation, specifically tailored for the Indian judiciary. This system is not only a technological advancement but also a step towards addressing the pressing challenge of case backlog in India. We believe our contributions will not only enhance the efficiency and transparency of the legal process but also pave the way for further research and development in AI-assisted legal technology. For the sake of reproducibility, we have made the PredEx dataset and the code for our prediction and explanation models accessible via a GitHub link². Additionally, for convenience, we have uploaded the data³ and models⁴ to Huggingface.

2 Related Work

The field of Legal Natural Language Processing (NLP) has witnessed significant advancements, with researchers exploring a variety of complex tasks within the legal domain. A prominent area of

focus has been Legal Judgment Prediction (LJP), where the goal is to predict the outcomes of legal cases based on their facts and contexts. Seminal works in this area include the contributions of (Zhong et al., 2020), (Malik et al., 2021), (Aletras et al., 2016), (Chen et al., 2019) (Long et al., 2019), (Xu et al., 2020) (Yang et al., 2019a), and (Chalkidis et al., 2019). These studies have laid the groundwork for understanding the nuances involved in automating legal decision-making processes.

Another key area of research has been the application of Large Language Models (LLMs) in the legal field. The versatility of models such as GPT, BLOOM, FLAN-T5, and LLaMA has been demonstrated in various studies, including those by (Vats et al., 2023) (Blair-Stanek et al., 2023) and (Katz et al., 2023), highlighting their potential in tasks ranging from statutory reasoning to judgment prediction. However, challenges remain in terms of the acceptability and reliability of LLMs in high-stakes legal contexts. The LegalEval (Modi et al., 2023) workshop further exemplifies the diversity and complexity of legal NLP research, especially on legal judgment prediction and explanation.

Our research utilizes advanced Large Language Models and a comprehensive dataset to create a system that predicts and explains judicial outcomes, enhancing legal text processing and transparency. This work supports legal practitioners and the public, especially in complex systems like India’s, and sets the stage for future AI advancements in legal technology.

²<https://github.com/ShubhamKumarNigam/PredEx>

³huggingface.co/collections/L-NLProc/predex-models

⁴huggingface.co/collections/L-NLProc/predex-datasets

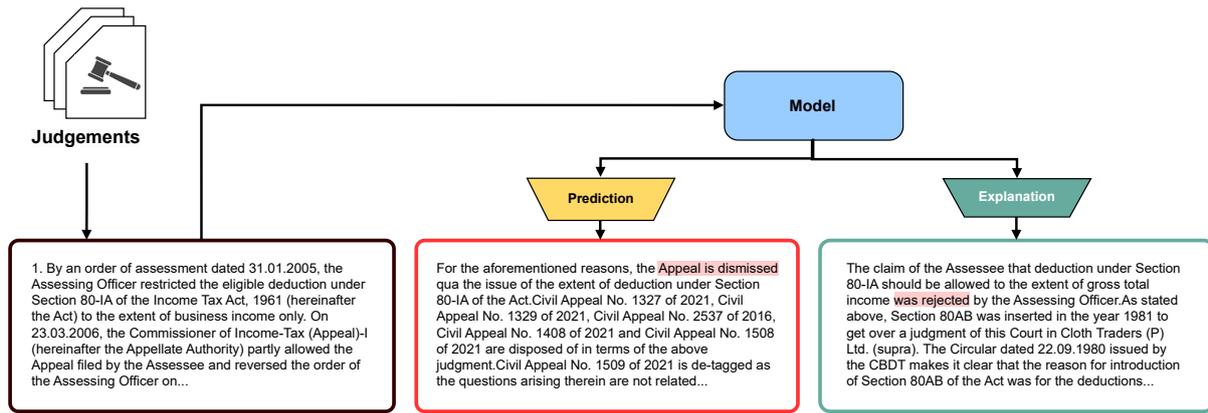


Figure 1: Illustration of the CJPE Task Framework.

3 Task Description

Our research project aims to advance the Court Judgment Prediction and Explanation (CJPE) task, incorporating insights and methodologies from both (Malik et al., 2021) and (Vats et al., 2023). The CJPE task involves two key sub-tasks: Prediction and Explanation. These tasks are performed sequentially, addressing the critical need to predict legal judgments and provide explanations for these predictions. To provide a visual representation of our task framework, Figure 1 illustrates the overall process of Court Judgment Prediction and Explanation (CJPE) as employed in our study. This figure includes the sequential steps of prediction and explanation. To demonstrate an Indian case structure in Table 8 and how explanations are derived from judicial judgments, refer to Figure 2 in the Appendix. For context, you can view the original case text on which this annotation is based.⁵ This example showcases the detailed process by which our annotators have identified and extracted these critical parts, reflecting the essence of judicial reasoning in each case. Similarly, an Example of an Indian Case Structure Table 8.

Prediction Task: The core of the CJPE task is to predict the outcome of a legal case based on the case proceedings. Given a document D that includes the case proceedings from the Supreme Court of India (SCI), the task is to predict the decision $y \in \{0, 1\}$, where ‘1’ signifies the acceptance of the appeal or petition by the appellant or petitioner, and ‘0’ indicates its rejection.

Explanation Task: The second part of the CJPE task involves explaining the predicted decision. Our approach is two-fold, integrating methodolo-

gies from both referenced papers:

1. **Identifying Key Sentences (ILDC for CJPE approach):** Similar to the approach in (Malik et al., 2021), we focus on identifying and highlighting key sentences or segments within the case proceedings that significantly contributed to the predicted outcome. This method relies on extracting specific parts of the text that are directly related to the decision, providing an evidence-based explanation.

2. **Generating Abstract Reasoning (LLMs approach):** Drawing from the approach in (Vats et al., 2023), we attempt to generate more abstract reasoning for the prediction. This involves providing zero and few-shot examples to the LLMs to guide them in generating explanations that are not just tied to specific text excerpts but also encompass broader reasoning and legal principles.

Additionally, we introduce a novel aspect to this task by training the LLMs specifically for both prediction and explanation. This training is tailored to enable the models to understand and process legal texts more effectively, improving their capability to predict outcomes and generate relevant explanations.

4 Dataset

In our research, we introduce “PredEx”, significantly differentiating itself from existing datasets in Legal Natural Language Processing (L-NLP), particularly in the context of the Indian judiciary. This dataset is designed to address the limitations of previous datasets, which primarily focused on prediction tasks and offered limited annotations for explanations.

⁵<https://indiankanoon.org/doc/97694707/>

4.1 Dataset Compilation

In the Data Compilation process, we initially gathered a substantial corpus of about 20,000 court judgments randomly from the Supreme Court of India and various High Courts, utilizing the IndianKanoon website⁶, a legal search engine widely recognized for its comprehensive database of Indian legal documents. The corpus underwent a meticulous annotation process, where our team of legal experts focused on annotating explanations for the judgments. These annotations involved identifying and highlighting key sentences or segments within the case proceedings that significantly influenced the predicted outcomes, as well as providing reasoning for the judgments. Through this process, the original corpus was distilled to approximately 16,000 case files, each richly annotated with expert legal explanations.

Our approach in compiling the PredEx dataset was to randomly select cases, ensuring a broad representation across various types of judgments and legal decisions. This method was deliberately chosen to avoid bias towards any specific temporal aspect or domain. By adopting a random selection process, we aimed to capture the diverse nature of legal cases in the Indian judiciary system. This diversity was crucial to cover various aspects of law and legal decision-making, thereby enhancing the representativeness and applicability of our dataset for training AI models in legal judgment prediction and explanation.

Subsequent to the annotation phase, we undertook a preprocessing step to refine the dataset further. This preprocessing involved the removal of cases that were either too brief or where the final decision segments were challenging to discern. Such preprocessing is crucial for ensuring the quality and consistency of the data, particularly for training robust and reliable AI models; otherwise, it could introduce noise or bias into the model training. As a result of this preprocessing, the total number of case files in our dataset was reduced to 15,222 and is further divided into training and testing sets. We adopted an 80-20 split ratio for this purpose, ensuring a substantial volume of data for model training while still retaining a robust set for testing. Specifically, the training set consists of 12,178 documents, and the test set comprises 3,044 documents.

In terms of balancing the test set, special atten-

⁶<https://indiankanoon.org/>

	Train	Test
No. of documents	12,178	3,044
Average no. of tokens	4,586	4,422
Minimum no. of tokens	176	184
Maximum no. of tokens	117,733	83,657
Acceptance percentage	53.44%	50.00%

Table 2: PredEx Statistics.

tion was given to ensure fairness and representativeness in model evaluation. We carefully curated the test set to include a diverse range of case outcomes, such as different types of judgments and legal decisions. This diversity was not just in terms of the nature of cases but also in terms of the outcomes - for instance, balancing cases where appeals were accepted versus those that were dismissed. Such a balanced composition is crucial in avoiding biases towards any particular type of judgment and ensures that our AI models are tested against a wide spectrum of legal scenarios. This balanced nature of the test set is particularly important for maintaining the validity of our experiments and for ensuring the reliability and generalizability of our model's performance. These carefully processed and curated case files now form the core of our PredEx dataset, offering a rich resource for the Court Judgment Prediction and Explanation (CJPE) task. Detailed statistics of the final dataset, post-preprocessing, are presented in the following Table 2.

4.2 Annotation Process

4.2.1 Expert Involvement

We engaged a team of 10 legal experts, primarily law students in their 3rd and 4th years, from various Indian law colleges. These experts were selected based on their academic standing and understanding of legal processes, ensuring high-quality annotations.

4.2.2 Annotation Timeline

The annotation process spanned from April 1, 2022, to October 30, 2023. This extensive period allowed for meticulous and thorough annotation, considering the complexity and detail required in legal document analysis.

4.2.3 Work Allocation

In our annotation process, each student was assigned around 30 judgment documents weekly,

striking a balance between efficiency and the need for thorough, accurate annotations. This workload allocation enabled students to devote adequate time to each document, fostering precise and insightful annotations.

4.2.4 Role of Student Annotators

The role of our student annotators was to meticulously identify and extract specific segments from the judgments that were pivotal to the judge’s reasoning, rather than interpreting or analyzing these segments with their own legal reasoning. Their task was to pinpoint these key sections accurately, ensuring that the extracts faithfully represented the judicial reasoning as stated in the case documents. This extractive approach was critical to maintain the integrity and authenticity of the annotations, allowing the dataset to accurately reflect the content of the original legal texts without the introduction of subjective interpretations by the annotators.

4.2.5 Annotation Quality Control Mechanism

To guarantee the accuracy and consistency of the annotations, we implemented a comprehensive quality control system. Initially, each document was reviewed by a single annotator to ensure a consistent interpretation of the judicial content. Recognizing the complexity of legal texts and the potential for subjective interpretation, we established several layers of review to enhance the reliability of our annotations:

- **Senior Expert Review:** Any disagreements or uncertainties in annotations were promptly escalated to a specialized review panel led by senior legal experts. These experts not only provided additional scrutiny but also mediated discrepancies among the initial annotations. Their extensive experience in legal practice and education enabled them to provide decisive and informed resolutions to any contentious or ambiguous annotations.
- **Regular Training and Meetings:** To further ensure consistency across annotations, regular training sessions and review meetings were conducted. These sessions served to align annotators on the legal framework and annotation criteria, reducing variability and enhancing the uniformity of the annotation process. Training included detailed discussions on identifying key legal arguments and the rationale within the judgments, which are

critical for both the prediction and explanation aspects of our dataset.

This rigorous quality control process has ensured that our dataset meets high standards of reliability and validity. The annotations not only reflect the factual content of the legal decisions but also the detailed reasoning behind these judgments, making our dataset a robust resource for training and evaluating AI models in legal judgment prediction and explanation.

4.2.6 Focus on Prediction and Explanations

Diverging from previous datasets that primarily concentrate on the task of prediction, our PredEx dataset spans both prediction and explanations. The annotations in our dataset serve a dual purpose. Firstly, they identify the outcomes of the cases, fulfilling the prediction aspect. More importantly, they go a step further by providing detailed explanations behind these outcomes. These explanations elucidate the rationale or the legal reasoning that underpins the judgments. This dual emphasis on prediction and explanations fills a significant void in existing legal datasets. Typically, in other datasets, the aspect of explanation is either absent or not explored in depth. By contrast, PredEx enriches the field of legal AI with comprehensive annotations that shed light not just on what the judicial decisions are, but crucially, why these decisions were made. This focus on explanations is particularly vital, as it contributes to a more transparent and understandable AI-driven legal decision-making process.

4.2.7 Largest Explainable Dataset

As a result of this extensive and detailed annotation process, we are releasing what is arguably the largest annotated dataset for legal judgment prediction and explanation in the Indian context. The size and comprehensiveness of this dataset set it apart from existing datasets in the field.

Our dataset represents a significant advancement in legal NLP, particularly for research and applications pertaining to the Indian judiciary. By providing a large-scale, richly annotated dataset that encompasses both prediction and explanation, we aim to facilitate more nuanced and sophisticated AI models capable of understanding and interpreting legal texts in a manner akin to human legal experts. This dataset is not only a resource for advancing AI technology in the legal domain but also a step

towards enhancing transparency and accountability in AI-assisted legal decision-making.

5 Methodology

This section outlines the methodology employed in our research for the tasks of Judgment Prediction and Judgment Prediction with Explanation.

5.1 Judgment Prediction

5.1.1 Language Model based

In our approach, we utilized several language models including InCaseLaw, InLegalBERT (Paul et al., 2023), XLNet (large) (Yang et al., 2019b), and Roberta (large) (Liu et al., 2019) as baselines for binary 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. 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 5 epochs on the PredEx train dataset. The remaining hyperparameters were set to their default values as provided by the HuggingFace library.

5.1.2 Large Language Model based

For utilizing LLMs in prediction, we employed two strategies: one involving prediction instructions only, and the other combining prediction with explanation instructions. Various models like Zephyr (Tunstall et al., 2023), Gemini 1.0 Pro (Team et al., 2023), Llama-2-7B (Touvron et al., 2023), and Llama-2-7B with instruction-tuning were used. We followed the prompts and instruction-tuning approaches published by (Vats et al., 2023) in a few-shot setup, and used the PredEx training data for instruction-tuning.

5.1.3 Prompts Used

For inference, we utilized prompts published by (Vats et al., 2023) and employed Template 2 in a zero and few-shot setup exclusively for prediction tasks, as detailed in Table 9 in the Appendix. These prompts provide a case description alongside a gold standard prediction label, directing the LLM to generate the judicial decision. For instruction tuning, we employed our custom prompts for prediction tasks, which are listed in Table 10 in the Appendix of this paper.

5.1.4 Instruction-Set

We developed 16 instruction sets using ChatGPT4 (DALL-E), validated by legal experts and then used for PredEx training data for instruction tuning. Given the token limit of 4096 in LLMs, we selected the last 1000 words from each document to fit within this constraint. This choice is supported by findings from (Malik et al., 2021) who achieved optimal results using the last 512 tokens of judgments. The input comprised the case proceedings and case decision and a random selection of instructions, with the output being the case outcome prediction. For a comprehensive understanding of our methodology and the full range of instructions used, we have included the complete list of all 16 instruction sets in Table 11 located in the Appendix of this paper.

5.2 Judgment Prediction with Explanation

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

5.2.1 Prompts used

In our approach, we also adopted prompts from (Vats et al., 2023), specifically utilizing Template 1 for the combined task of judgment prediction and explanation. This template, detailed in Table 9 in the Appendix, expands on the prediction-only format of Template 2 by including an explanation component. Here also, for instruction tuning, we employed our custom prompts for prediction with explanation tasks, which are listed in Table 10 in the Appendix of this paper. In this setup, the LLM is instructed not only to predict the outcome of a case but also to articulate the reasoning behind the decision. The precise formatting and examples of how predictions and explanations are structured and solicited from the LLM can be viewed in the referenced table.

5.2.2 Instruction-Set

For judgment prediction with explanation, we created 16 instruction sets using ChatGPT4 (DALL-E), also validated by legal experts. This time, the input included case proceedings, decisions, and reasoning, with randomly chosen instructions, and the output being the case outcome prediction with explanation. For a comprehensive view of all 16 instruction sets, we have included the full list in Table 11 in the Appendix of this paper.

6 Evaluation Metrics

In our study, We report Macro Precision, Macro Recall, Macro F1, and Accuracy on the PredEx judgment prediction test dataset and employ a multifaceted approach to evaluate the performance of our models on the PredEx judgment explanation test dataset. Our evaluation metrics encompass both quantitative and qualitative methods, ensuring a thorough assessment of the model’s capabilities in both prediction and explanation tasks.

1. **Lexical Based Evaluation:** We utilized lexical similarity metrics such as Rouge scores (Rouge-1, Rouge-2, and Rouge-L) (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005). These metrics assess the similarity between the generated explanations and the reference texts based on word overlap and order, providing an insight into the lexical accuracy of the model outputs.
2. **Semantic Similarity Based Method:** To capture the semantic essence of the generated explanation, we employed BERTScore (Zhang et al., 2020), which measures the semantic similarity between the generated and ground truth explanations. Additionally, we used BLANC (Vasilyev et al., 2020) to estimate the quality of generated explanations in the absence of a gold standard, offering a perspective on the model’s ability to generate semantically rich and contextually relevant text.
3. **Expert Evaluation:** Human evaluation played a crucial role in our assessment. Legal experts reviewed the explanations generated by the models and rated them on a 1–5 Likert scale based on their accuracy, relevance, and completeness. The criteria for the rating scale were as follows:
 1. The explanation is entirely incorrect or fails to provide any relevant information.
 2. The model’s response is irrelevant or shows misunderstanding of the case judgment.
 3. The explanation is partially accurate but misses critical details.
 4. The response is comparable and relevant to the ground truth.

	Models	Macro Precision	Macro Recall	Macro F1	Accuracy
Prediction only					
LM Based	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
LLM Based	Zephyr	0.5347	0.5295	0.5119	0.5309
	Gemini pro	0.5976	0.5803	0.5610	0.5808
	Llama-2-7B	0.5732	0.5723	0.5713	0.5726
	Llama-2-7B Instruction-tuning on prediction task	0.5186	0.5177	0.5117	0.5177
	Llama-2-7B Instruction-tuning on prediction with explanation task	0.5195	0.5185	0.5127	0.5190
	Prediction with explanation on PredEx				
LLM Based	Gemini pro	0.5184	0.5154	0.4908	0.5081
	Llama-2-7B	0.5087	0.5017	0.3772	0.5025
	Llama-2-7B Instruction-tuning on prediction with explanation task	0.5254	0.5215	0.5031	0.5224
	Prediction with explanation on ILDC expert				
LLM Based	Llama-2-7B	0.3125	0.4259	0.3236	0.4259
	Llama-2-7B Instruction-tuning on prediction with explanation task	0.5750	0.5741	0.5728	0.5741

Table 3: Judgement prediction results. The best results are shown in bold.

5. The explanation is completely accurate, relevant, and potentially superior to the expert’s explanation.

7 Results and Analysis

7.1 Judgment Prediction

Our experiments, as detailed in Table 3, reveal interesting insights into the performance of various models on the PredEx test data. Notably, Roberta emerges as the top performer, outstripping even the Large Language Models (LLMs). This suggests that traditional language models might be more adept at analyzing and predicting outcomes in legal documents compared to generative-based models. Even among the generative models, the few-shot Llama-2-7B model surpassed the fine-tuned Zephyr model, which is surprising given Zephyr’s supervised fine-tuning (SFT) approach and its reinforcement learning training on general corpora. It appears that the Llama-2-7B models, both instruction-tuned for prediction and prediction with explanation tasks, show promising results in this domain.

7.2 Judgment Prediction with Explanation

The results, as shown in Table 4, provide valuable insights into the performance of machine-generated explanations versus expert explanations

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								
Gemini pro	0.3099	0.2428	0.2593	0.0826	0.1870	0.6329	0.1715	2.24
Llama-2-7B	0.3211	0.1886	0.2109	0.0599	0.1760	0.6191	0.1507	3.06
Llama-2-7B Instruction-tuning on prediction with explanation task	0.4972	0.4321	0.4399	0.2531	0.3630	0.6909	0.2844	2.84
Prediction with explanation on ILDC expert (Vats et al., 2023; Malik et al., 2021)								
GPT 3.5 turbo (Reproduced)	0.5383	0.4267	0.4541	0.2842	0.4685	0.7273	0.3394	3.6*
Llama-2-7B	0.4526	0.2454	0.2957	0.1485	0.3440	0.6464	0.2212	3.65
Llama-2-7B Instruction-tuning on prediction with explanation task	0.4939	0.3805	0.3969	0.2918	0.5075	0.6891	0.3636	3.30

Table 4: Explanation performance comparison of various model combinations for judgment prediction across different evaluation metrics, with the highest score in each metric in bold. Entries marked with * denote normalized value.

across a range of models. These assessments include lexical-based, semantic, and expert evaluations on the PredEx test data. To augment our evaluation process, we also incorporated a comparison with the instruction-tuned models on the 54 ILDC_expert (Malik et al., 2021) dataset. This dataset, to our knowledge, represents the largest collection of legal expert-annotated data available for Indian cases, offering a valuable benchmark for assessing the performance of our models. This multi-faceted evaluation offers a comprehensive understanding of the models’ capabilities in generating explanations.

Given the expense and time required to obtain legal expert annotations, we carefully sampled 50 cases from our dataset for Likert score evaluations by legal experts. This sampling strategy was chosen to provide a representative and manageable subset of cases for in-depth expert analysis, while also considering the practical constraints associated with expert-driven evaluations.

7.3 Lexical Based Evaluation

In the lexical-based evaluation, the performance of LLMs in generating explanations shows that verbatim matches are not at a satisfactory level. However, it’s important to note that these metrics, while valuable, do not fully encapsulate the models’ proficiency in analyzing cases, predicting outcomes, and generating reasoning. Thus, we turn to Semantic Similarity-Based Evaluation and Expert Score Evaluation for a more thorough assessment.

7.4 Semantic Evaluation

Semantic evaluation, particularly the BERTScore, indicates better alignment of the explanations with the gold standard, suggesting a good semantic understanding in the generated explanations. The Llama-2-7B model with instruction-tuning for prediction and explanation tasks excels in semantic similarity. Nevertheless, lower scores in open-source models point to challenges in accurately generating case analysis, predictions, and reasoning. It’s crucial to recognize that generative models may exhibit hallucination issues, not entirely captured by this metric, necessitating manual evaluation by legal experts for a more complete assessment.

7.5 Expert Evaluation

Evaluating generative models in the legal judgment prediction task with explanation requires domain-specific expertise. The expert evaluation, detailed in Table 5, shows that the Llama-2-7B model with instruction-tuning performs notably well, although it sometimes produces truncated or repetitive responses. Despite these limitations, the instruction-tuned model demonstrates fewer non-factual responses and better overall explanation quality compared to other pre-trained models. Interestingly, models with well-designed prompts for explanation generation displayed enhanced performance without instances of hallucination.

The expert ratings, as reflected in Table 5, further underscore the efficacy of our instruction-

Generative Models	Rating Score				
	1	2	3	4	5
	PredEx				
Llama-2-7B	2	11	22	12	3
Llama-2-7B Instruction-tuned	5	13	18	13	1
	ILDC expert				
Llama-2-7B	0	9	22	21	2
Llama-2-7B Instruction-tuned	2	3	16	24	9

Table 5: Distribution of Expert Rating Scores for Generative Models on PredEx and ILDC Expert Data.

tuned model, which even surpasses the quality of explanations provided by legal professionals (achieving a rating score of 4). This underlines the potential of generative models, particularly those leveraging our instruction-tuning approach, in generating accurate and relevant legal explanations. The average expert rating scores, presented in Table 4, corroborate the superiority of our generative models over other approaches.

7.6 Hallucination

We address the issue of hallucinations in model-generated text, a common challenge in using large language models for generating legal judgments. Hallucinations refer to instances where the model generates false or irrelevant information that is not supported by the input data. To combat this, we have implemented a fine-tuning strategy that significantly reduces these errors. A detailed comparative analysis in the Appendix B showcases these strategies and their effectiveness. This analysis demonstrates how fine-tuning and instruction-tuning specifically tailored to the legal domain can mitigate hallucinations, providing clearer, more accurate, and legally coherent outputs.

8 Conclusions and Future Work

We introduced PredEx, the largest dataset for legal judgment prediction and explanation in this context, marking a significant advancement over previous datasets. Our research explored instruction tuning on Large Language Models (LLMs), showing promise in improving prediction accuracy and explanatory depth.

Looking ahead, our focus will be on training Indian Legal domain-specific Large Language Models. This approach will ensure that the models are

ingrained with domain-specific knowledge, crucial for tasks like legal judgment prediction with explanations. Furthermore, we plan to undertake Supervised Fine-Tuning (SFT) on various downstream tasks, including the judgment prediction with explanation task. Another key objective will be to incorporate contextual understanding into the models to mitigate issues like hallucinated responses, a common challenge with generative models.

The question remains as we advance in this field: How ready is the State-of-the-Art to aid in explainable judgment prediction? Our future efforts aim to answer this question by refining the capabilities of AI in legal applications, making a significant contribution to the evolving field of AI-assisted legal judgment prediction and explanation. The ultimate goal is to develop AI tools that can not only alleviate the backlog in the Indian judiciary but also deliver justice efficiently and transparently. To further enhance the accuracy and reliability of our system, we plan to implement a Reinforcement Learning from Human Feedback (RLHF) pipeline. This pipeline aims to refine the model’s predictions and explanations based on human feedback, ensuring that the outputs align more closely with expert legal understanding and reasoning. The inclusion of RLHF represents a significant advancement in developing AI systems for legal judgment prediction and explanation, as it allows for continuous improvement and adaptation based on real-world feedback and interactions.

Limitations

Our study faced several significant limitations that impacted our approach and findings. A primary constraint was the token limitation and high subscription charges for paid cloud services, which restricted our ability to perform inference and fine-tuning on larger parametric models, particularly those with 70B or 40B parameters. This limitation likely curtailed our exploration of the full capabilities of these advanced models, which could have provided deeper insights or enhanced performance.

Another critical limitation was the resource-intensive nature of obtaining legal expert annotations. Due to the high costs and extensive time required for this process, it was not feasible for us to obtain expert evaluations for the entire PredEx test dataset. Consequently, we opted to sample 50 random documents for expert review and Likert score evaluations. While necessary, this ap-

proach potentially limits the breadth and depth of our expert-based evaluation, as it does not encompass the entire dataset.

In terms of the effectiveness of Large Language Models (LLMs) in the legal domain, our findings suggest that while these models are proficient in conversational contexts, their applicability in logic or knowledge-intensive tasks like legal judgment prediction and explanation is less convincing. Analyzing lengthy legal documents and generating predictions with explanations poses a significant challenge for generative-based models. This is particularly true in cases where the models need to process and understand complex legal reasoning and contexts.

Furthermore, the performance of the open-source baseline model, which was intended to jointly predict and generate explanations, did not meet our expectations. This underperformance could be attributed to the token limitations imposed during our study. By only using the last 1000 tokens of documents for fine-tuning, there is a possibility that the model did not fully grasp the entire context of the cases. Moreover, our fine-tuned models frequently produced truncated responses due to the 512-token limit set for generation. This limitation may have hindered the models' ability to generate comprehensive and nuanced explanations.

Lastly, the pre-trained models used in our study inherently lacked detailed knowledge specific to Indian legal cases. Even after undergoing tuning processes, these models struggled to generate explanations that paralleled the depth and specificity of human-like legal reasoning. This shortfall highlights the challenge of adapting general AI models to specialized domains such as law, where domain-specific knowledge and reasoning are crucial.

These limitations underscore the challenges in applying LLMs to complex and specialized tasks like legal judgment prediction and explanation. They also highlight the necessity for continued research and development efforts aimed at enhancing the capabilities of AI models in interpreting and understanding legal documents and contexts.

Ethics Statement

Ethical conduct was a cornerstone in our research, especially considering the sensitive nature of the data and the methodologies involved. In collecting and annotating the PredEx dataset, we ensured that the law students involved in the annotation process

were treated fairly and compensated appropriately. Their consent was obtained for all participation, and while they made significant contributions to the dataset, they are not listed as authors of this paper. This distinction is made to acknowledge their contribution while also maintaining the academic integrity of the publication process.

Significantly, the senior legal expert who played a pivotal role in mentoring the annotation process, as well as providing guidance on the Likert rating system and evaluating the generated explanations for both the PredEx and ILDC datasets, is credited as one of the authors of this paper. This inclusion reflects the expert's substantial intellectual contribution to the research, in line with ethical norms and authorship guidelines in academic publishing.

Moreover, for the computational resources used in this study, we adhered to ethical standards by duly paying the subscription fees for Google Colab Pro. This payment ensured legitimate access to the necessary paid cloud services, which were instrumental in the development and testing of our AI models. We believe in supporting the services and platforms that enable research like ours, and this includes the responsible financial support of technology providers.

In summary, our approach to ethics encompassed not only the respectful and fair treatment of all individuals involved but also the adherence to legal and financial obligations. This comprehensive ethical stance underscores our commitment to conducting research that is not only innovative and impactful but also responsible and respectful of all parties involved.

References

- Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoțiu-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.
- Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. **Can gpt-3 perform statutory reasoning?**
- 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.
- Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. **Charge-based prison term prediction with deep gating network**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6362–6367, Hong Kong, China. Association for Computational Linguistics.
- 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.
- Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. **Gpt-4 passes the bar exam**.
- 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.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Shangbang Long, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Automatic judgment prediction via legal reading comprehension. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 558–572. Springer.
- Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. **ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.
- Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik Kumar Guha, Sachin Malhan, and Vivek Raghavan. 2023. **SemEval-2023 task 6: LegalEval - understanding legal texts**. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2362–2374, Toronto, Canada. Association for Computational Linguistics.
- 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 Șulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef van Genabith. 2017. Exploring the use of text classification in the legal domain.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai,

- Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- 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*.
- Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Cl  mentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.
- Oleg V. Vasilyev, Vedant Dharnidharka, and John Bohnannon. 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*.
- Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3086–3095, Online. Association for Computational Linguistics.
- Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019a. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI'19*, page 4085–4091. AAAI Press.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. 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.
- Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1250–1257.

A Experimental Setup and Hyper-parameters

Our experimental setup was designed to optimize the performance of instruction fine-tuning on LLMs and to accurately assess their capabilities in legal judgment prediction and explanation tasks. We utilized two cores of NVIDIA A100-PCIE-40GB with 126GB RAM of 32 cores for instruction fine-tuning, ensuring powerful computational resources for processing and model training. In addition to the dedicated hardware, we employed a Google Colab Pro subscription having A100 Hardware accelerator for conducting inference and other experiments. This platform provided us with the necessary flexibility and scalability for our extensive experimentation.

Regarding the model training specifics, we fine-tuned the LLMs for 5 epochs. This duration was chosen to balance between adequately training the models on our PredEx dataset and preventing overfitting. During our experiments, we encountered a common issue with generative models – the tendency to hallucinate and repeat sentences. To address this, we implemented a post-processing step after inference. This step involved selecting the first occurrences of the decision and explanation parts from the model outputs and omitting any subsequent repetitions. This approach helped us refine the output quality, ensuring the results to be coherent and concise.

However, it is important to note that certain LLMs did not yield inference results in some cases. In such instances, we excluded those cases from our evaluation process. This decision was made to maintain the integrity and accuracy of our experimental findings, as including non-inferential results could have skewed our overall assessment of the models' performance.

Overall, our experimental setup was carefully crafted to provide a robust and reliable framework for evaluating the efficacy of instruction-tuned LLMs in the context of legal judgment prediction and explanation.

B Hallucination Examples

B.1 Pre-trained vs Fine-tuned

In the Appendix, we conduct a thorough comparison between pre-trained and fine-tuned models to demonstrate the reduction of hallucinations through our fine-tuning methods. Table 7 in the

Appendix presents this comparative analysis, illustrating the impact of instruction-tuning on legal judgment prediction with explanation tasks in our PredEx dataset. This section 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).

B.2 Zephyr-7b-beta and Gemini pro

For a practical illustration of hallucination issues, Table 6 in the Appendix provides examples from the Zephyr-7b-beta and Gemini Pro models. Notably, Zephyr-7b-beta sometimes produced contradictory outcomes ('accepted' and 'rejected') for the same case with identical explanations. Additionally, an instance was observed where Gemini Pro produced a decision in Chinese, despite the prompt and the rest of the proceedings being in English. These examples highlight the critical need for model training and tuning specific to the dataset and task to ensure the reliability and accuracy of AI-generated legal predictions and explanations.

Model	Decision and Explanation
Ground Truth	<p>Decision: Accepted. Explanation: Insofar as first issue is concerned that need not detain us for long as we find that the treatment given thereto by the High Court is without any blemish.....The High Court has held that it provides for only one situation where the time period can be stretched, viz. where the assessment is held up (after due service of notice) due to a stay order by a court of law. That covers only one eventuality, namely assessment proceedings. The fact that the other periods have not been mentioned leads to the inference that the delay in issuance of notice, or delays in other situations have been intentionally left out, and the consequence of such imperatively framed time period being breached, result in invalidity of the act.</p>
Zephyr-7b-beta	<p>[1, The appeal is likely to be allowed.]</p> <p>Explanation: The High Court has held that the notice issued for revising the assessment list was not valid for the assessment year 1997-98 as it was issued after the expiry of the assessment year. However, the notice was received during the assessment year 1998-99. The Court has clarified that the notice could not have been used to revise the assessment list for the year 1997-98, but it was valid for the year 1998-99. Therefore, the Court is likely to allow the appeal as the assessment list for the year 1998-99 was validly revised.</p> <p>[0, The appeal is not likely to be allowed.]</p> <p>Explanation: The High Court has held that the notice issued for revising the assessment list was not valid for the assessment year 1997-98 as it was issued after the expiry of the assessment year. However, the notice was received during the assessment year 1998-99. The Court has clarified that the notice could not have been used to revise the assessment list for the year 1997-98, but it was validly used to revise the assessment list for the year 1998-99. Therefore, the appeal is not likely to be allowed as the issue at hand is related to the assessment year 1997-98, for which the notice was not valid.</p>
Ground Truth	<p>Decision: Rejected</p> <p>Explanation: At the outset we would like to say that we do not find any substance in the contention of learned counsel for the appellant that there has been any confusion of the issue or that any wrong approach has been adopted by the taxing authorities, the Tribunal or the High Court.....no deduction in respect of interest credited by the assessee to the account of the college over the said sum can be allowedIn the circumstances, in our view, the High Court rightly answered the questions referred to it against the assessee in both the references.</p>
Gemini pro	<p>Decision: Chinese Character</p> <p>Explanation: The High Court held that such a view was clearly sustainable and observed that it is not the requirement under s. 27(iii) that the assessee must further show that the borrowing of the capital was necessary for the business so that if at the time of the borrowing the assessee has sufficient amount of its own the deduction could not be allowed and the High Court further took the view that in deciding whether a claim of interest on borrowing can be allowed the fact that the assessee had ample resources its disposal and need not have borrowed was not a relevant matter for consideration.</p>

Table 6: We observe both ‘accept’ and ‘reject’ judgments in the response generated by Zephyr-7B-beta with the same explanation and opposite conclusion and repetitive responses. On the other hand, Gemini Pro gives responses in the Chinese language instead of English language in a few instances.

Model	Decision and Explanation	Explanation of error
Ground Truth	<p>Case decision: Accepted</p> <p>Explanation: Thus, going by the principles governing the matter, propounded by this Court there cannot be any doubt that the rights of the appellant-bank over the pawned sugar had precedence over the claims of the Cane Commissioner and that of the workmen. The High Court was, therefore, in error in passing an interim order to pay parts of the proceeds to the Cane Commissioner and to the Labour Commissioner for disbursal to the cane growers and to the employees. There is no dispute that the sugar was pledged with the appellant bank for securing a loan of the first respondent and the loan had not been repaid. The goods were forcibly taken possession of at the instance of the revenue recovery authority from the custody of the pawnee, the appellant-bank. In view of the fact that the goods were validly pawned to the appellant bank, the rights of the appellant-bank as pawnee cannot be affected by the orders of the Cane Commissioner or the demands made by him or the demands made on behalf of the workmen. Both the Cane Commissioner and the workmen in the absence of a liquidation, stand only as unsecured creditors and their rights cannot prevail over the rights of the pawnee of the goods.¹⁹ We are also of the view that pending the writ appeals, the High Court ought not to have passed such an interim order of consequence especially in the light of the legal principles settled by this Court. The order of the High Court, therefore, cannot be sustained and calls for interference.</p>	
Llama-2-7B	<p>Case decision: Rejected</p> <p>Explanation:</p> <p>The case proceedings provide several significant sentences that indicate the outcome of the case will be rejected. Firstly, the court notes that the rights of a pawnee over the pawned goods have precedence over the claims of a secured creditor, even in the absence of a liquidation (para 15). This suggests that the appellant bank's rights as a pawnee cannot be affected by the orders of the Cane Commissioner or the demands made by him, even though the goods were taken possession of at the instance of the revenue recovery authority.</p> <p>Furthermore, the court observes that the Cane Commissioner and the workmen stand only as unsecured creditors in the absence of a liquidation, and their rights cannot prevail over the rights of the pawnee of the goods (para 16). This implies that the appellant bank's rights as a pawnee are superior to those of the Cane Commissioner and the workmen, and cannot be ignored or set aside in favor of the latter.</p> <p>Finally, the court notes that the High Court's interim order was passed in error, as it failed to recognize the appellant bank's rights as a pawnee (para 18). This suggests that the High Court's decision will be reversed, and the appellant bank's rights will be upheld.</p> <p>Based on these significant sentences in the case proceedings, it can be concluded that the case will be rejected, and the appellant bank's rights as a pawnee will be upheld.</p>	<p>The Vanilla model incorrectly categorises the sugarcane growers and the workmen, represented by the Cane commissioner and the Labour Commissioner respectively, as "secured creditors" when in actuality they are "unsecured creditors" as already established in the case text. This hallucination is all the more apparent as the model correctly identifies the parties as "unsecured" creditors a few lines later. Furthermore, the model incorrectly places references to different paragraphs of the case text, i.e., para 15 and para 16. as the paragraphs in question do not reflect the text as purported by the model in its output.</p>
Llama-2-7B Instruction-tuning on prediction with explanation task	<p>Case decision: Accepted</p> <p>Explanation:</p> <p>18. Thus, going by the principles governing the matter, propounded by this Court there cannot be any doubt that the rights of the appellant-bank over the pawned sugar had precedence over the claims of the Cane Commissioner and that of the workmen. The High Court was, therefore, in error in passing an interim order to pay parts of the proceeds to the Cane Commissioner and to the Labour Commissioner for disbursal to the cane growers and to the employees. There is no dispute that the sugar was pledged with the appellant bank for securing a loan of the first respondent and the loan had not been repaid. The goods were forcibly taken possession of at the instance of the revenue recovery authority from the custody of the pawnee, the appellant-bank. In view of the fact that the goods were validly pawned to the appellant bank, the rights of the appellant-bank as pawnee cannot be affected by the orders of the Cane Commissioner or the demands made by him or the demands made on behalf of the workmen. Both the Cane Commissioner and the workmen in the absence of a liquidation, stand only as unsecured creditors and their rights cannot prevail over the rights of the pawnee of the goods.¹⁹ We are also of the view that pending the writ appeals, the High Court ought not to have passed such an interim order of consequence, especially in the light of the legal principles settled by this Court. The order of the High Court, therefore, cannot be sustained and calls for interference.</p>	<p>The Finetuned model accurately categorises the workmen and the sugarcane growers as "unsecured creditors" and correctly determines that the right of the pawnee (Appellant Bank) will have precedence over their rights to recompensation. Furthermore, the finetuned model also accurately states if the liquidation of the company had been put into motion, the workmen would THEN ONLY be considered "secured creditors" in pari-passu with other secured creditors.</p>

Table 7: Comparative analysis of responses generated by Pretrained Llama-2-7B and Instruction Finetuned Llama-2-7B.

Supreme Court of India
Sandeep Khaitan vs Jsvm Plywood Industries Ltd. on 22 April, 2021
Author: K.M. Joseph
Bench: Uday Umesh Lalit, Indira Banerjee, K.M. Joseph

REPORTABLE

IN THE SUPREME COURT OF INDIA
CRIMINAL APPELLATE JURISDICTION

CRIMINAL APPEAL NO.447 OF 2021
(Arising out of SLP (CRL.) No. 1549 of 2021)

SANDEEP KHAITAN, RESOLUTION PROFESSIONAL
FOR NATIONAL PLYWOOD INDUSTRIES LTD. ... APPELLANT(S)

VERSUS

J SVM PLYWOOD INDUSTRIES LTD. & Anr. ... RESPONDENT(S)

J U D G M E N T

K.M. JOSEPH, J.
Leave granted.

1. The appeal is directed against order dated 04.02.2021 passed by the Hon'ble High Court of Guwahati. In the impugned order, the High Court has allowed an interlocutory application filed by the Respondent No. 1 to allow it to operate its bank account maintained with the ICICI Bank Shubaneswar and to unfreeze the bank account of its creditors over which the lien has been created and the accounts frozen pursuant to the lodging of an FIR by the appellant before us. It was made subject to conditions.

2. An application under Section 7 of the Insolvency and Bankruptcy Code, 2016, hereinafter referred to as the IBC was admitted on 26.08.2019 against one National Plywood Industries Limited (NPIL). The Appellant was appointed as the Interim Resolution Professional. A moratorium also came to be passed by the very same order within the meaning of Section 14 of the IBC. The Appellant came to be appointed as the Resolution Professional by an order dated 08.11.2019. In the meantime, the Respondent No.1 claiming to be an operational creditor lay the claim for the amounts due to it from the Corporate Debtor before the Appellant vide communication dated 22.11.2019. It would appear that the former Managing Director of the Corporate Debtor challenged the order of the NCLT, Guwahati, admitting the application under Section 7. The NCLAT by order dated 24.11.2019 dismissed the appeal interalia holding that the application under Section 7 was not barred by limitation. Civil Appeal No. 9142 of 2019 filed by the former Managing Director of the Corporate Debtor came to be however allowed by this Court by an order dated 20.01.2020. The NCLT was directed to consider the matter in accordance with law. It would appear that on 28.01.2020 interlocutory application 7 of 2020 filed by the former Managing Director of the Corporate Debtor seeking an injunction restraining the Respondents therein from interfering in the operation of the Corporate Debtor and to disperse the cost of the CIRP was disposed of interalia as follows: -

i. "Today the Respondents submitted across the Bar that except ratifying the expenses of the IRP, no major decisions have been taken by the COC in the yesterday's COC meeting. Both the respondents informed that they are conscious about the order passes by the Hon'ble Supreme Court and the legal consequences thereof. ii. In view of the above submissions of the respondents, this Tribunal expects that the respondents would maintain status-quo in respect of the IRP proceedings. As the main company petition was remanded back to the Hon'ble NCLAT for fresh disposal in accordance with law, this Tribunal is of the considered opinion that the petitioner has to approach the Hon'ble NCLAT for any further directions in the above matter and accordingly above application stands disposed of with the above observations. Even otherwise, the order of admission of the company petition has not attained finality and, therefore, no interim orders as prayed for needs to be passed today. iii. In the result, IA No. 07 of 2020 is disposed of with the above observations." Therafter there is order dated 20.03.2020 passed which we will advert to.

•
•
•

FINDINGS

16. The contours of the jurisdiction under 482 of the Cr.P.C. are far too well settled to require articulation or reiteration. Undoubtedly in this case by 26.08.2019 an application filed under section 7 of the IBC was admitted, the appellant appointed as the interim resolution professional and what is more a moratorium declared. With the declaration of the moratorium the prohibitions as enacted in section 14 came into force. It is clear that the assets of the company would include the amounts lying to the credit in the bank accounts. There cannot be any dispute that well after the order under section 14 was passed, a sum of Rs. 32.50 lakhs has been remitted into the account of Respondent No.1 company. No doubt it is the definite case of the Respondent No.1 that it has had business relations with the Corporate Debtor since more than 15 years and that the amount remitted in its account represented the price of the materials supplied to the Corporate Debtor. Apart from this amount a sum of rupees more than Rs. 20 lakhs is still due. It is to be noticed that though an appeal was filed against the order admitting the petition under Section 7 the same was dismissed by the NCLAT. The appellate order was undoubtedly set aside by this court and the appeal remanded to the NCLT for its consideration. We would think that setting aside the appellate order of the NCLAT by this court and remanding the appeal would not have the effect of setting aside the order admitting the application. Initially, as was noticed by us an order was passed on 28.02.2020. The ambiguity created by the said order was removed by the subsequent order of the Tribunal dated 20.03.2020. In other words, by the order dated 20.03.2020 the NCLT, Guwahati ordered that the appellant was at liberty to act as per law and the words used in the earlier order dated 28.02.2020 relied upon by the Respondent No. 1 were found to be a mere casual observation which did not culminate into any direction. We need not say anything further particularly in view of the fact that there is an FIR and which is pending consideration in the High Court also. It is significant only for us to notice that the Appellant is essentially aggrieved by the transactions representing a sum of Rs. 32.50 lakhs all of which took place after order dated 20.03.2020.

17. It may be true that in the interim order passed by the NCLT Guwahati, the Tribunal had directed the Directors to refund the amount of the Corporate Debtor less any amount paid for supplies. It is also true that the review petition filed by the Appellant is dismissed, essentially based on the limitations on the power of review.

18. The provisions of the IBC contemplate resolution of the insolvency if possible, in the first instance and should if not be possible, the winding up of the Corporate Debtor. The role of the insolvency professional is neatly carved out. From the date of admission of application and the appointment of Interim Resolution Professional, the management of the affairs of the Corporate Debtor is to vest in the Interim Resolution Professional. With such appointment, the powers of the Board of Directors or the partners of the Corporate Debtor as the case may be are to stand suspended. Section 17 further declares that the powers of the Board of Directors or partners are to be exercised by the Interim Resolution Professional. The financial institutions are to act on the instructions of the Interim Resolution Professional. Section 14 is emphatic, subject to the provisions of sub section (2) and (3). The impact of the moratorium includes prohibition of transferring, encumbering, alienating or disposing of by the Corporate Debtor of any of its assets.

•
•
•

i. The Respondent No.1 is **allowed** to operate its account subject to it first remitting into the account of the Corporate Debtor, the amount of Rs 32.50 lakhs which stood paid to it by the management of the Corporate Debtor. The assets of the Corporate Debtor shall be managed strictly in terms of the provisions of the IBC. The Appellant as RP will bear in mind the provision of Section 14 (2A) and the object of IBC. We however make it clear that our order shall not be taken as our pronouncement on the issues arising from the FIR including the petition pending under Section 482 of the Cr.P.C.

ii. We also make it clear that the judgment will not stand in the way of the Respondent No.1 pursuing its claim with regard to its entitlement to a sum of Rs.32.50 lakhs and any other sum from the Corporate Debtor or any other person in the appropriate forum and in accordance with law. There will be no order as to costs.

•
•
•

Input to model

Ratio of decision/Explanation

To be predicted by model

Decision (Not given as input)

Figure 2: Annotated Example of Judicial Reasoning Extraction.

CASE NO:
Appeal (civil) 3499-3500 of 2007
PETITIONER:
CENTRAL BANK OF INDIA
RESPONDENT:
SIRIGUPPA SUGARS & CHEMICALS LTD. & ORS
DATE OF JUDGMENT:
07/08/2007
BENCH:
TARUN CHATTERJEE & P.K. BALASUBRAMANYAN
CASE TEXT:
<p>...These appeals challenge the interim order passed by the Division Bench of the High Court in a pending writ appeal, directing disbursement of certain amounts realised on sale of stocks of sugar, owned by the first respondent company held under pledge by the appellant–bank. The Labour Commissioner had passed an order under Section 33(c) of the Industrial Disputes Act against the first respondent company in respect of the dues to the workmen. The same was challenged by the first respondent in the writ petition as also by others...</p> <p>...In Giles vs. Grover (1832 (131) ER 563 : 9 Bing 128) it has been held that the Crown has no precedence over a pledgee of goods. In Bank of Bihar vs. State of Bihar (supra) the principle has been recognised by this Court holding that the rights of the pawnee who has parted with...</p> <p>...There is no difference between the common law of England and the law with regard to pledge as codified. Under Section 172 a pledge is a bailment of the goods as security for payment of a debt or performance of a promise. Section 173 entitles a pawnee to retain the goods pledged as security for payment of a debt and under Section 175 he is entitled to receive from the pawner any extraordinary expenses he incurs for the preservation of the goods pledged with him...</p> <p>...In State of M.P. vs. Jaura Sugar Mills Ltd. And others (supra) dealing with the Madhya Pradesh Sugar Cane (Regulation and Supply) Act, it was only held that the Cane Commissioner having power to compel the cane growers to supply cane to the factory, has incidental power and is duty bound to ensure payment of the price of the sugarcane supplied by the sugarcane growers...</p>
JUDGEMENT:
<p>...We, therefore, allow these appeals and set aside the impugned order of the High Court, directing payment out of parts of the sale proceeds to the Labour Commissioner and to the Cane Commissioner. We hold that the appellant as the pawnee, is entitled to the amount in satisfaction of its debt to secure which, the goods had been pawned and to appropriate the sale proceeds towards the debt due and only if there is surplus...</p>

Table 8: Example of Indian Case Structure. Sections referenced are highlighted in blue, previous judgments cited are in magenta, and the final decision is indicated in green.

<p>Template 1 (prediction + explanation)</p> <p>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) and provide an explanation 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 given two examples and perform the task for the following case proceeding. case_proceeding: <{case_proceeding}> Format your output in list format: [prediction, explanation]””””</p>
<p>Template 2 (prediction only)</p> <p>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 given 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.””””</p>

Table 9: Prompts for Judgment Prediction taken from (Vats et al., 2023).

<p>Template 3 (prediction only)</p> <p>prompt = f““““ ### Instructions: Analyze the case proceeding and predict whether the appeal/petition will be rejected (0) or accepted (1). ### Input: <{case_proceeding}> ### Response: ””””</p>
<p>Template 4 (prediction with explanation)</p> <p>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: ””””</p>

Table 10: Prompts for Judgment Prediction used for instruction fine-tuned models. Instructions were randomly chosen from Table 11.

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 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 11: Instruction Sets for Predicting Legal Decisions and Providing Explanations.