PROSLEX: A Novel Dataset for Expert-Annotated Legal Statute Prediction for Indian Judiciary

Anonymous ACL submission

Abstract

Legal Statute Prediction (LSP) involves automatically identifying relevant legal statutes given factual descriptions in legal documents, typically framed as a *multi-label classification task* within natural language processing and information retrieval research. While recent advances have begun incorporating Large Language Models (LLMs) for statute prediction, current approaches primarily focus on accuracy metrics without addressing the critical need for legal reasoning—a fundamental requirement in judicial contexts where decisions must be explainable and justifiable. To address this research gap, we present PROSLEX (**PR**ediction **O**f **S**tatutes and **LE**gal e**X**planation), a comprehensive dataset comprising **1,623 expert-annotated** legal documents from the Indian context. Each document is paired with statute predictions and detailed explanations, **totaling 7,450 explanations**, capturing the underlying legal reasoning. Using this dataset, we systematically evaluate various prompting strategies, including zero-shot, few-shot, chain-of-thought, and tree-of-thoughts approaches, to generate both statute predictions and their corresponding legal rationales. Our evaluation framework measures not only predictive performance but also the coherence and legal validity of generated explanations, positioning PROSLEX as a benchmark for developing explainable AI systems that can support legal practitioners while advancing research in interpretable legal NLP. To ensure reproducibility, we have made our *PROSLEX* dataset and model code available on GitHub[1].

1 Introduction

In Civil Law jurisdictions, written statutory provisions explicitly define `charges` under specific `statutes` (e.g., *theft*, *murder*, *riot*), which must be consulted before determining the applicable charge in a given case[2]. 'Statute' refers to a set of codified legal rules followed by the legal system in the concerned judiciary. In countries such as India, which adhere to the Civil Law system (in conjunction with elements of Common Law) and face the looming pendency of millions of cases[3], automating statute identification emerges as a significant research challenge – not only to support legal professionals but also to aid individuals without legal expertise, for whom seeking professional advice can be costly and time-consuming. Statute prediction makes it easy to find the laws that apply to a case, giving both lawyers and everyday people quick, affordable access to legal guidance.

Statute prediction involves identifying the relevant legal statutes for a given case based on its factual description, making it a practical and impactful application of Natural Language Processing (Paul et al., 2024; Chalkidis et al., 2021). Since a legal case document can be associated with multiple statutes, statute prediction is inherently a *multi-label classification* task.

Explainability in AI-systems, especially in high-risk domains like law is of paramount importance (Liao and Wortman Vaughan, 2024). The EU AI Act Article 13: Transparency and Provision of Information to Deployers [4] states *"***High-risk AI systems** *shall be designed and developed in such a way as to ensure that their operation is* **sufficiently transparent to enable deployers to interpret a system's output** *and use it appropriately. An appropriate type and degree of transparency shall be ensured with a view to achieving compliance with the relevant obligations of the provider and deployer set out in Section 3"* highlighting the non-negotiable necessity of transparency in AI systems.

[1] https://anonymous.4open.science/r/Legal_Statute_Prediction_Explnantion-580A/

[2] https://en.wikipedia.org/wiki/Civil_law_(legal_system)

[3] https://njdg.ecourts.gov.in/njdg_v3/

[4] https://artificialintelligenceact.eu/article/13/

An example document annotated by legal experts.

BHIMANNA v. STATE OF KARNATAKA [2012] INSC 495 (4 September 2012) On 17.11.1999 at about 4.00 p.m., Yenkappa(A-1), along with Bhimanna (A-2) and Suganna (A-3), was returning home with agricultural implements, that is axes and a plough. They attempted to use the land of the deceased as a pathway. The deceased Bheemanna, who was present on his land along with his wife Paddamma (PW.1) and mother, namely, Bheemava, obstructed the accused persons, asking them not to pass through his land. Yenkappa(A-1) then started hurling abuses in filthy language and instigated Bhimanna (A-2) and Suganna (A-3) to assault the deceased.

Statute:**IPC 147**

Thus, Bhimanna (A-2) and Suganna (A-3) began assaulting the deceased with axes over his head and right hand.

Statute:**IPC 302, IPC 147**

Yenkappa (A-1) assaulted the deceased with the wooden part of a plough. Paddamma (PW.1) and Bheemava, mother of the deceased went to save the deceased, but they too, were threatened with assault. Similar threats were hurled when Rangayya (PW.6), nephew of the deceased and his father Hanumappa approached the place of occurrence. The accused persons left the place after assaulting the deceased, throwing away the axes and wooden part of the plough. Rangayya (PW.6) brought a bullock cart as asked by Paddamma (PW.1) from the village and the deceased was then taken to Ramdurga Police Station. Upon the advice of the police, the deceased was taken in a mini lorry, driven by Mahadevappa (PW.10) to Deodurga Hospital, and when they reached there at 8.00 p.m., based on the complaint submitted by Paddamma (PW.1), an FIR was lodged at 8.15 p.m.

Statute:**IPC 302**

Figure 1: An example of annotation by legal experts in the proposed PROSLEX dataset. Here, statutes IPC 147 and IPC 302 are applicable, and the explanation (reason) for the applicability of these statutes is shown as annotations.

To build explainable automatic statute prediction systems, test beds are necessary, and in Indian law, apart from initial attempts like (Vats et al., 2023), there has been no large-scale court facts (of cases) dataset containing *legal expert annotated* explanations for the applicability of statute(s). In this paper, we present PROSLEX (**PR**ediction **O**f **S**tatutes and **LE**gal e**X**planation), a dataset comprising **1,623** legal expert–annotated case facts paired with their relevant statutes for explainable statute prediction. Figure 1 illustrates an example annotation, where legal experts highlighted a text span and associated it with the applicable statute. Afterward, **a total of 7,450 spans** were annotated to represent the legal reasoning for the statutes.

The creation of PROSLEX involved two main stages and was carried out by legal experts from a reputed Indian legal institute. In the first stage, experts *identified statutes* based on their legal and social importance. In the second stage, they *selected documents* corresponding to these statutes and annotated the relevant text spans. Subsequently, a senior legal expert resolved annotations with low inter-annotator agreement (IAA) scores. Section 3 provides a detailed description of the PROSLEX creation process. The resulting dataset was then used for our empirical study, which is elaborately described in Section 4. Furthermore, Figure 2 depicts the complete pipeline, from dataset creation to its application in downstream tasks using various in-context learning techniques, including *few-shot*, *chain-of-thought*, and *tree-of-thoughts*.

Our research delivers a comprehensive AI-based system for legal Statute prediction and explanation tailored to the Indian judiciary.

2 Related Work

Legal charge identification or statute prediction has significantly evolved, leveraging advances in machine learning and deep learning methodologies. Early research primarily relied on traditional machine learning methods. For instance, Liu et al. (2015) utilized support vector machines (SVMs) with textual features from legal documents to predict associated statutes.

Motivated by the success of deep learning in NLP tasks, researchers introduced neural network architectures into charge prediction tasks (Chen et al., 2019; Li et al., 2018; Shen et al., 2018; Undavia et al., 2018; Wei et al., 2018; Yang et al., 2019b). Luo et al. (2017) proposed a hierarchical attention network that predicted charges by selecting the most relevant law articles. Similarly, Wei and Lin (2019) developed a knowledge-aware multi-label charge prediction method using external knowledge from legal texts to improve prediction accuracy.

On the other side, to address the challenges of distinguishing confusing charges, Li et al. (2019) developed an element-driven attentive neural net-

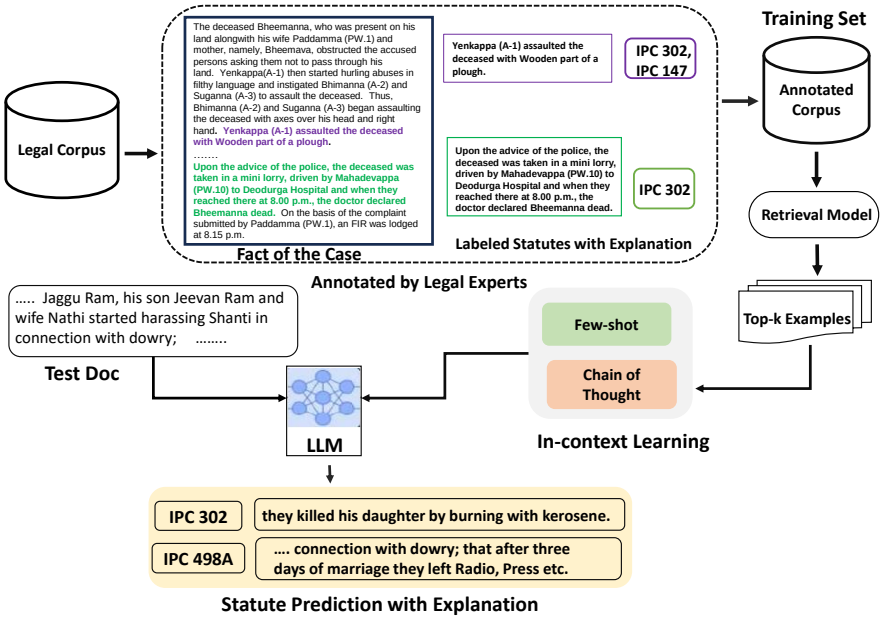

Figure 2: This workflow illustrates a legal prediction framework to identify applicable Indian Penal Code (IPC) statutes for legal cases with the reasoning. The diagram shows a workflow where a legal corpus is annotated by legal experts to create a training set of case facts paired with relevant statutes and explanations. The left side displays an example case involving Bheemanna's assault and death, with highlighted text mapped to IPC sections 302 (*murder*) and 147 (*rioting*). The system uses a retrieval model to find similar cases, then employs in-context learning techniques like few-shot examples and chain-of-thought reasoning to make predictions on unlabeled cases. The bottom portion demonstrates the model analyzing a test document about dowry harassment and correctly predicting applicable statutes (IPC 302 for a *kerosene burning death* and IPC 498A for *dowry harassment*).

work model leveraging legal constitutive elements as discriminative features. Similarly, Yang et al. (2019a) integrated word collocation features using attention mechanisms, and Hu et al. (2018) constructed discriminative attributes to enhance differentiation between confusing charges. More recently, Xu et al. (2020) developed an attention mechanism focusing on subtle differences between confusing legal articles, aiming to improve charge prediction accuracy further. Knowledge-enhanced models have shown effectiveness, particularly for confusing charges. Bi et al. (2024) presented a dual-graph interaction framework that integrates external legal knowledge bases, significantly improving performance in distinguishing among confusing charges. Moreover, Chen et al. (2024) introduced a defendant-aware label representation method tailored for multi-defendant scenarios, addressing complexities in real-world legal cases. Several works emphasized understanding semantic relations between charges and textual descriptions to enhance prediction accuracy (Xu et al., 2020; Yue et al., 2021; Li et al., 2023; Kang et al., 2021).

Within the Indian legal context, the Artificial Intelligence for Legal Assistance (AILA) track at FIRE 2019 was instrumental in emphasizing legal information retrieval and statute identification (Bhattacharya et al., 2019a). Subsequently, Vats et al. (2023) investigated the strengths and limitations of large language models (e.g., GPT, LLama, Gemini) for statute prediction, demonstrating both substantial promise and critical shortcomings. Further, comprehensive state-of-the-art surveys by Pawar et al. (2023) and Paul et al. (2024) on publicly available datasets underscored notable progress while also identifying persistent challenges in Indian legal statute prediction.

Despite these advances, a notable gap persists in *generating clear explanations alongside charge predictions*, especially within the in-context learning framework. In response to this limitation, we introduce PROSLEX, a novel dataset created to enable in-depth analysis of charge identification with accompanying explanatory context, focusing on Indian legal cases.

Corpus	Language	Jurisdiction	# of Cases	# Lables	Explnanation
ECtHR-B (Chalkidis et al., 2021)	English	Europe	10,000	10	✗
ILSI (Paul et al., 2024)	English	India	66,090	Law Article (100)	✗
LeSICiN (Paul et al., 2022)	English	India	100,000	IPC Statutes (multiple)	✗
FIRE AILA 2019 (Bhattacharya et al., 2019b)	English	India	100	197	✗
(Vats et al., 2023)	English	India	45	80	✓
PROSLEX (Ours)	English	India	1623	7	✓

Table 1: A Comparative Overview of Prominent Legal Statute Prediction Datasets.

3 Dataset Creation

In this section, we provide a detailed account of the creation process of our dataset, **PROSLEX** (**PR**ediction **O**f **S**tatutes and **LE**gal e**X**planation), which comprises a subset of 33,546 Supreme Court of India[5] judgements from 1950 to 2016, covering 444 Acts and 3,428 Sections. The dataset creation was supported by a **Government-funded initiative, involving two postgraduate legal experts per statute and a senior faculty member from a reputed National Law University**. Legal experts initiate the annotation process by first selecting *relevant statutes* and then choosing the *corresponding documents* for each statute.

3.1 Statute selection

We concentrate on crime-related documents, which account for 30% of the Supreme Court collection. In the Indian judiciary, 512 sections related to criminal law, collectively known as the Indian Penal Code (IPC). The selection of statutes for annotation was guided by the following three criteria:

1. **Legal and social significance:** Those statutes have high importance from a legal perspective and the socio perspective. For example, Indian Penal Code 302 represents the sentence for murder and is thus of high relevance.

2. **Sufficient documents:** Those statutes have a substantial representative number of documents for the empirical study.

3. **Distinctiveness of legal provisions:** The nature of statutes needs to be distinct. For instance, IPC 302 indicates a sentence (punishment) for culpable homicide amounting to murder, IPC 300 – culpable homicide, and IPC 304 – culpable homicide but not amounting to murder. Hence, to select the distinctive statute, legal experts only select those documents, which is associated with IPC 302 as it concerns the sentence for the most serious crime within *homicide*.[6]

Statute overview. Based on the aforementioned criteria, the legal experts identified **seven statutes** that meet our requirements, which are described below:

1. **Indian Penal Code,** 1860_147: Whoever is guilty of *rioting*, shall be punished with imprisonment of either description for a term which may extend to two years, or with fine, or with both.

2. **Indian Penal Code,** 1894_201: Whoever, knowing or having reason to believe that an offence has been committed, *causes any evidence of the commission of that offence to disappear*.

3. **Indian Penal Code,** 1860_302: Whoever commits *murder* shall be punished with death, or imprisonment for life, and shall also be liable to fine.

4. **Indian Penal Code,** 1860_376: Deals with the *punishment for rape*.

5. **Indian Penal Code,** 1860_420: Whoever cheats and thereby dishonestly induces the person deceived to deliver any property to any person.

6. **Indian Penal Code,** 1860_498A: An act of *cruelty* for the purposes of Section 498A corresponded to willful conduct of such a nature that it could cause danger to the life, limb, or health of the woman.

7. **Indian Penal Code,** 1860_506: Whoever commits the *offence of criminal intimidation* shall be punished with imprisonment of either description for a term which may extend to two years, or with fine, or with both.

[5] https://www.sci.gov.in/

[6] https://devgan.in/ipc/chapter_16.php

Dataset		Indian Penal Code (IPC) Statutes						
Type	Statistics	IPC 147	IPC 201	IPC 302	IPC 376	IPC 420	IPC 498A	IPC 506
Train	#Documents	214	214	193	105	172	118	116
Train	# Avg. tokens	1052	1584	617	745	496	270	488
Validation	#Documents	30	30	27	15	24	17	16
Validation	# Avg. tokens	1100	1540	583	750	487	274	498
Test	#Documents	62	63	57	30	51	35	34
Test	# Avg. tokens	1043	1534	600	723	415	323	502

Table 2: **Statistics of the PROSLEX Dataset**

3.2 Document selection and annotation

Manually annotating all 33,546 documents is both time-consuming and labor-intensive, and expensive. As a solution, we opted to select a representative *subset of collection* for annotation. In this phase, legal experts focus on selecting a substantial amount of *relevant documents* of the aforementioned statutes. Considering the **budget constraints** of the Government-funded project, the selection was restricted to at most 250 cases per section. To optimally utilize the available resources, we prioritized including a greater number of cases per statute—ensuring sufficient data points for each class—over covering a larger number of statutes with fewer cases. The annotation process involves two key tasks: i) *document selection* and ii) *annotating them*, which is carried out by a team of two legal experts per section, and is described below.

Let the section is S_k and the two legal experts are L_1 and L_2.

1. L_1 and L_2 individually selected cases (same cases annotated by both the annotators) in which *section S_k was upheld by the ruling court*. Thus, the cases in which section S_k was initially applied but ultimately repealed are struck off from the list of cases to be annotated.

2. The legal experts L_1 and L_2, extract *facts of the case* from the original case document.

3. The legal experts L_1 and L_2, perform annotation over facts of the cases. Figure 1 displays legal experts labeling a *span of text* along with applicable statutes.

4. To measure *inter-annotator agreement* between two annotated spans, we compute text overlap using ROUGE-L (Lin, 2004), where the score ranges from 0 (no overlap) to 1 (perfect overlap). A higher score indicates stronger consistency between annotators. For cases with an overlap score < 0.75, the agreement was deemed insufficient, and the senior legal expert finalized the annotation to establish the gold standard. When the score was ≥ 0.75, the agreement was considered satisfactory, and one of the two annotations was retained as the gold standard.

The process of *fact extraction*, *labeling span of text*, *solving inter-annotator disagreement*, and providing additional cases if required, are repeated until for section S_k we have 250 relevant and annotated cases. Although it was not feasible to obtain 250 cases for every statute, our dataset PROSLEX nevertheless offers 1623 annotated documents with 7450 explanations, based on Indian Supreme Court judgments (as shown in Table 1). In Table 2, we showcase overall statistics of our dataset PROSLEX. Later, we employ PROSLEX for various empirical studies.

4 Experimental Setup

In the previous section, we described the entire data set creation process. For the empirical study, we divide our dataset into three parts such as training, validation, and testing. For training, validation, and testing purpose, we utilize 70%, 10% , and 20% of the dataset.

Language Model Based. In this approach, we utilized several language models, including Legal-BERT (Chalkidis et al., 2020), InLegalBert (Paul et al., 2023), Longformer(Beltagy et al., 2020), RoBERTa (Liu et al., 2019), and BERT-based (Large) (Devlin et al., 2019) as baselines for multilabel classification. For training the model, we used a batch size of 32, the Adam optimizer, and a learning rate of 5e-6. The training was conducted over 20 epochs on the Legal Statute Prediction with Explanation (LSPE) dataset. The remaining hyperparameters were set to their default values as provided by the HuggingFace library.

Given a fact of a legal case document as enclosed within angular brackets, enlist the likely stautes that apply for this segment.

Following are the tags along with their descriptions.
1. "[Statute]": "[Description]"
2. "[Statute]": "[Description]"
...

Following is a list of example text for each :

Example 1:
Case_proceeding: "[Text]"
Prediction: "[Statute],[Statute]"
Explanation: "[Reasoning],[Reasoning]"

Example 2:
Case_proceeding: "[Text]"
Prediction: "[Statute]"
Explanation: "[Reasoning]"

Instruction: Learn from the examples provided. Avoid generating fabricated or invalid tags.

Input segment: <[INPUT]>

Table 3: The structure of the prompt used in our ICL experiments. [X] represents a variable that is to be substituted for its value.

Large Language Model Based. By employing LLMs in prediction, we considered two approaches: one involving identifying **legal statutes only**, and the other focusing on **statute identification with explanation**. For this prediction task, we leveraged several models, including Llama-3.1-70B[7], deepseek-r1-distill-llama-70b[8], claude-sonnet-4-20250514[9], GPT-4[10]. The temperature parameter was set to 0 for Claude and GPT, and 0.01 for Llama and DeepSeek. Additionally, we standardized the top-p value to 0.95 across all experiments.

4.1 Zero-shot

In this approach, we rely exclusively on the *label descriptions* (as detailed in Section 3.1) for all labels, without providing any example context, as shown in Table 3.

4.2 Exemplar selection for few-shot

In this approach, we adopt the *K-nearest neighbor selection* method to choose an exemplar for few-shot learning. We compute semantic similarity between the *statute description with annotated span of the corresponding statutes for each document from the training set*, and then select examples for each statute based on the highest score.

Let $D_{\text{train}} = \{d_1, d_2, \ldots, d_n\}$ denote the set of training documents, and $S = \{s_1, s_2, \ldots, s_m\}$ represent the set of statutes. For each document d_i, let a_{ij} be the annotated span corresponding to statute s_j, and let t_j denote the textual description of s_j, as illustrated in Figure 2. We compute the semantic similarity score between the statute description t_j and the annotated span a_{ij} using a similarity function $\sigma(\cdot, \cdot)$, such as cosine similarity in an embedding space, comparing their [CLS] embeddings obtained from InLegalBERT.

The pair (s_{j^*}, a_{i^*j}) is selected as the representative example for statute s_j. For the empirical study, we conduct experiments using $K \in \{2,4,6\}$ examples per statute as few-shot exemplars.

4.3 Chain-of-thought

To elicit a series of intermediate reasoning steps while predicting the statutes, we employ the chain-of-thought approach (Wei et al., 2022). Yu et al. (2022) employed zero-shot Chain-of-Thought (CoT) prompting for statute prediction, generating explanations prior to prediction; however, this approach underperformed compared to methods that did not utilize explanations. In contrast, Vats et al. (2023) applied few-shot CoT prompting (Wei et al., 2022) on a dataset of 45 documents. In our work, we conducted experiments on a significantly larger dataset (**1623** documents). Table 5 and Table 4 present the comparative performance of the models in terms of both prediction and explanation.

4.4 Tree-of-thoughts

The Tree-of-Thoughts (ToT) framework has been proposed to improve prompting for complex tasks that demand exploration and look-ahead reasoning (Yao et al., 2023). Unlike Chain-of-Thought (CoT) prompting, ToT organizes intermediate reasoning steps—referred to as thoughts—into a tree structure. Each thought constitutes a coherent language sequence that incrementally progresses toward the final solution. This tree-based organization enables language models to reason deliberately by evaluating how well different thoughts contribute to solving the task. Furthermore, ToT combines the model's capability to generate and assess thoughts with search strategies such as breadth-first and depth-first search. In this study, we employ a breadth-first search (BFS) strategy with the follow-

[7] https://ai.meta.com/blog/meta-llama-3-1/
[8] https://www.deepseek.com/en
[9] https://www.anthropic.com/claude/sonnet
[10] https://openai.com/

	Models	Indian Penal Code (Statutes)							Macro F1 score
		IPC 147	IPC 201	IPC 302	IPC 376	IPC 420	IPC 498A	IPC 506	
Statute Prediction Only									
LM Based	InLegalBERT	0.76	0.80	0.83	0.87	0.94	0.91	0.45	**0.82**
	LegalBERT	0.68	0.72	0.84	0.90	0.93	0.91	0.37	0.80
	BERT-base	0.60	0.55	0.68	0.81	0.80	0.86	0.06	0.67
	Longformer	0.57	0.59	0.73	0.68	0.83	0.86	0.02	0.67
	RoBERTa	0.64	0.41	0.64	0.75	0.86	0.85	0.10	0.64
Zero-shot	DeepSeek	0.68	0.65	0.75	0.83	0.75	0.73	0.26	**0.64**
	Llama-3.1-70b	0.47	0.27	0.66	0.47	0.58	0.55	0.25	0.47
	Claude	0.60	0.45	0.71	0.66	0.73	0.71	0.39	0.61
	GPT-4	0.55	0.37	0.68	0.59	0.69	0.66	0.35	0.56
Few-shot	DeepSeek	0.87	0.72	0.75	0.85	0.82	0.80	0.43	**0.71**
	Llama-3.1-70b	0.66	0.43	0.72	0.83	0.83	0.78	0.37	0.64
	Claude	0.62	0.52	0.74	0.74	0.71	0.74	0.41	0.64
	GPT-4	0.61	0.51	0.73	0.73	0.78	0.77	0.46	0.66
TOT	DeepSeek	0.21	0.08	0.12	0.34	0.05	0.21	0.24	0.17
	Llama-3.1-70b	0.25	0.01	0.09	0.30	0.01	0.24	0.06	0.13
	Claude	0.32	0.02	0.08	0.74	0.11	0.02	0.03	0.18
	GPT-4	0.05	0.26	0.13	0.36	0.41	0.01	0.27	**0.21**
Statute Prediction with Explanation									
COT	DeepSeek	0.65	0.66	0.74	0.91	0.78	0.88	0.40	0.73
	Llama-3.1-70b	0.50	0.44	0.71	0.71	0.56	0.74	0.33	0.57
	Claude	0.49	0.58	0.66	0.86	0.82	0.82	0.40	0.66
	GPT-4	0.50	0.61	0.73	0.93	0.93	0.86	0.54	**0.75**

Table 4: Performance comparison of different models (e.g., DeepSeek, Llama-3.2, Claude, GPT-4) across sections of the Indian Penal Code. This evaluation presents the F1-scores of various statutes using both LM-based and LLM-based approaches. Among the LM-based models, InlegalBERT consistently outperforms others. Additionally, a key finding is that LLMs achieve the highest performance in statement prediction with explanation when using the Chain-of-Thought (CoT) setup within In-Context Learning (ICL) techniques.

ing constraints: a maximum search depth of d_{max} = 7, at most c_{max} = 3 child nodes per state, a limit of n_{max} = 10 per level and a confidence threshold of 0.4 to determine whether a path is expanded further.

5 Results and Analysis

In this section, we discuss the insights derived from our empirical observations. Table 4 shows the F1-score for each statute, followed by the macro-F1 score, while Table 5 compares LLM-generated explanations with expert annotations.

5.1 Insights from Statute Prediction

Specifically, Table 4 provides a comparative analysis of the performance of different models across the aforementioned sections of the Indian Penal Code (IPC), using both traditional language-based models and Large Language Models (LLMs). For the LLM-based systems, we further explore *in-context learning* (ICL), *Chain-of-Thought* (CoT), and *Tree-of-Thoughts* (ToT) strategies. InLegalBERT emerges as the best-performing model under the "Statute Prediction Only" setting, achieving the highest overall accuracy of 0.82, notably outperforming other language models like LegalBERT and Longformer. Among *zero-shot* setups, DeepSeek performs best with 0.64 macro F1-score, though still significantly behind the fine-tuned LM-based models. In the *few-shot* setting, DeepSeek again leads with 0.71 macro F1-score, demonstrating better generalization with limited examples, while GPT-4 and Claude show more moderate performance.

Another observation is that Large Language Models under the *ToT* mechanism can struggle with accurate statute prediction. When generating multiple reasoning branches, the model may interpret the same statutory language differently across branches, resulting in inconsistent or incorrect predictions. In contrast, *CoT* follows a single reasoning chain. This focused strategy allows LLMs to

Models	Lexical Based Evaluation				Semantic Evaluation	
	Rouge-1	Rouge-2	Rouge-L	BLEU	BELURT	BERTScore
Prediction with explanation on StatandEx						
Llama-3.1-70b	0.17	0.18	0.21	0.25	0.58	0.65
Deepseek	0.10	0.13	0.15	0.21	0.42	0.46
Claude	0.10	0.11	0.14	0.19	0.40	0.43
GPT-4	0.20	0.23	0.31	0. 28	0.63	**0.67**

Table 5: Explanation performance comparison of various model combinations for statute prediction across different evaluation metrics, with the highest score in each metric in **bold**.

concentrate computational resources on one line of reasoning rather than dispersing attention across multiple branches, leading to more accurate statute identification. Consequently, GPT-4 achieves the highest macro F1-score of 0.75, demonstrating strong reasoning capabilities, particularly when explanations are required.

Overall, while domain-specific models like InLegalBERT dominate in direct statute prediction, large foundational models like GPT-4 excel when interpretability and reasoning are involved.

5.2 Insights of LLM Reasoning

To evaluate the LLMs generated rational along with predicted statutes, we employ a comprehensive evaluation strategy that integrates both quantitative and qualitative analyses. This multifaceted approach ensures a robust assessment of our model's performance for the explanation tasks.

1. **Lexical Based Evaluation:** We employed lexical similarity metrics, including ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) (Lin, 2004), and BLEU (Papineni et al., 2002) to evaluate the generated explanations. These metrics measure the overlap and sequence of words between the model outputs and reference texts, offering insight into the lexical accuracy of the explanations.

2. **Semantic Similarity Based Evaluation:** To capture the semantic quality of the generated explanations, we utilized BERTScore (Zhang et al., 2019), BELURT (Sellam et al., 2020), which evaluate the semantic similarity between the generated outputs and the reference explanations. Therefore, these metrics provide deeper insight into the model's ability to produce meaningful and contextually appropriate explanations.

Table 5 presents a comparative analysis of explanation quality across various models—LLaMA 3.1, DeepSeek, Claude, and GPT-4—using both lexical-based (ROUGE-1, ROUGE-2, ROUGE-L, BLEU) and semantic-based (METEOR, BERTScore) evaluation metrics. Overall, GPT-4 consistently outperforms the other models across all metrics, achieving the highest scores in ROUGE-1 (0.20), ROUGE-2 (0.23), ROUGE-L (0.31), BLEU (0.28), BLEURT(0.63), and BERTScore (0.67), indicating that its explanations are both textually aligned and semantically richer compared to other models. LLaMA 3.1 ranks second, showing relatively strong lexical overlap (e.g., ROUGE-2: 0.18) and decent semantic quality (BLEURT: 0.58, BERTScore: 0.65), suggesting it generates explanations that are closer to reference texts than DeepSeek or Claude. DeepSeek and Claude lag behind significantly on all metrics, with Claude performing the lowest semantically (BERTScore: 0.43), indicating less coherent or relevant explanations. This analysis confirms that GPT-4 is the most capable model for generating legally meaningful and well-aligned statute explanations, with LLaMA-3-70b being a strong secondary contender. In Figure 5, we present an error analysis illustrating a case where GPT-4 correctly predicts the statute but provides incorrect reasoning.

6 Conclusion and Future Work

We present PROSLEX, the largest dataset to date for legal statute prediction and explanation, representing a major step forward in this domain. Our research investigates In-Context Learning (ICL) methods such as Few-Shot and Chain-of-Thought (CoT) prompting, which show promising improvements in both prediction accuracy and explanatory depth.

Our future work seeks to address this question by advancing AI capabilities in legal applications. We aim to make a meaningful contribution to the growing field of AI-assisted legal statute prediction and explanation.

Limitations

While this study makes important steps forward in statute prediction for legal documents, there are still a few areas that could benefit from further refinement. We see these not as major limitations, but as valuable opportunities for future work. They don't take away from the contributions of this research, but rather point to ways it can continue to grow and improve.

The PROSLEX dataset, while the largest and most comprehensive resource for Indian legal statutes, is inherently specialized to the Indian judicial system and focuses primarily on the crime domain. This specialization enables deep insights within this specific legal context but may limit the immediate applicability of the models to legal documents from other jurisdictions. To adapt the models for cross-jurisdictional use one could likely require additional techniques. While this does not detract from the contributions of the current work, it does highlight a valuable direction for future research in enhancing the generalizability of legal NLP models across different legal systems.

Ethical Consideration

This research was carried out with a firm commitment to ethical principles. In this work, we have introduced a new legal statute dataset, curated from Indian Supreme Court cases. We would like to state that the data that were used for annotation are already publicly available on the website. We do not collect any personal information (e.g., age, address, or phone number) available in these data, and all data were collected in compliance with legal and privacy regulations. However, person names do exist in these legal case documents, which may raise some privacy concerns.

The legal experts were annotators of the dataset annotation process. The annotators were treated with fairness, and they received appropriate compensation for their contributions in Indian Rupees. They were provided informed consent before participating in the project.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019a. Fire 2019 aila track: artificial intelligence for legal assistance. In *Proceedings of the 11th annual meeting of the forum for information retrieval evaluation*, pages 4–6.

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019b. Overview of the fire 2019 aila track: Artificial intelligence for legal assistance. In *FIRE (working notes)*, pages 1–12.

Sheng Bi, Zafar Ali, Tianxing Wu, and Guilin Qi. 2024. Knowledge-enhanced model with dual-graph interaction for confusing legal charge prediction. *Expert Systems with Applications*, 249:123626.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-based prison term prediction with deep gating network. *arXiv preprint arXiv:1908.11521*.

Yu Chen, You Zhang, Jin Wang, Dan Xu, and Xuejie Zhang. 2024. Learning defendant-aware label representation for multi-defendant charge prediction. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th international conference on computational linguistics*, pages 487–498.

Liangyi Kang, Jie Liu, Lingqiao Liu, and Dan Ye. 2021. Label definitions augmented interaction model for legal charge prediction. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43*, pages 270–283. Springer.

Penghua Li, Fen Zhao, Yuanyuan Li, and Ziqin Zhu. 2018. Law text classification using semi-supervised convolutional neural networks. In *2018 Chinese control and decision conference (CCDC)*, pages 309–313. IEEE.

Shang Li, Boyang Liu, Lin Ye, Hongli Zhang, and Binxing Fang. 2019. Element-aware legal judgment prediction for criminal cases with confusing charges. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 660–667. IEEE.

Wei Li, Shiyu Zhu, and Li Li. 2023. Leveraging knowledge-aware method for charge prediction. In *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1496–1501. IEEE.

Q. Vera Liao and Jennifer Wortman Vaughan. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, (Special Issue 5). Https://hdsr.mitpress.mit.edu/pub/aelql9qy.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yi-Hung Liu, Yen-Liang Chen, and Wu-Liang Ho. 2015. Predicting associated statutes for legal problems. *Information Processing Management*, 51(1):194–211.

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*.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. *arXiv preprint arXiv:1707.09168*.

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.

Shounak Paul, Rajas Bhatt, Pawan Goyal, and Saptarshi Ghosh. 2024. Legal statute identification: A case study using state-of-the-art datasets and methods. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2231–2240.

Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. 2022. Lesicin: A heterogeneous graph-based approach for automatic legal statute identification from indian legal documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11139–11146.

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 the nineteenth international conference on artificial intelligence and law*, pages 187–196.

Sachin Pawar, Basit Ali, Girish K Palshikar, Ramandeep Singh, and Dhirendra Singh. 2023. Extraction and classification of statute facets using few-shot learning. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 197–206.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Yatian Shen, Jun Sun, Xiaopeng Li, Lei Zhang, Yan Li, and Xiajiong Shen. 2018. Legal article-aware end-to-end memory network for charge prediction. In *Proceedings of the 2nd International Conference on Computer Science and Application Engineering*, pages 1–5.

Samir Undavia, Adam Meyers, and John E Ortega. 2018. A comparative study of classifying legal documents with neural networks. In *2018 Federated conference on computer science and information systems (FedCSIS)*, pages 515–522. IEEE.

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

Duan Wei and Li Lin. 2019. An external knowledge enhanced multi-label charge prediction approach with label number learning. *arXiv preprint arXiv:1907.02205*.

Fusheng Wei, Han Qin, Shi Ye, and Haozhen Zhao. 2018. Empirical study of deep learning for text classification in legal document review. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3317–3320. IEEE.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. *arXiv preprint arXiv:2004.02557*.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019a. Legal judgment prediction via multi-perspective bi-feedback network. *arXiv preprint arXiv:1905.03969*.

Ze Yang, Pengfei Wang, Lei Zhang, Linjun Shou, and Wenwen Xu. 2019b. A recurrent attention network for judgment prediction. In *Artificial Neural Networks and Machine Learning–ICANN 2019: Text and*

Time Series: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part IV 28, pages 253–266. Springer.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 973–982.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

A.1 LLM Generated Explanation

Fact Text
"On 23.9.1985 in the evening prosecutrix alongwith her husband Dayaram went to Khajuria. On the way near the field of Mangal, appellants met them and started to abuse prosecutrix. Appellants also started beating husband of \fprosecutrix and took the prosecutrix near the well of Kamal Singh where accused Pooran Singh and Lalliram talked to Chaturbhuj. Dayaram was locked in a room. Then appellants took away the prosecutrix to the upper room of the house and committed rape repeatedly in the night. In the next morning they released prosecutrix and warned her not to report to anybody. Then she brought Dayaram from the room and at that time Latura, Gyarasa, Bharo Singh, Kamal Singh and Harihar reached there. They were informed about the incident. Appellants also snatched a bag from the prosecutrix containing Rs.25/- and identity card of Dayaram. Prosecutrix lodged the report in Police Station Madhogarh out post which is Ex.P1. Investigation was undertaken. Spot map was prepared which is Ex.P2. Prosecutrix was sent for medical examination vide Ex.P6 and she was examined by the lady doctor, Smt. C.P. Jain (PW-11) twice, first time on 25.9.1985 and then on 5.10.1985. Ex.P6 is the report pertaining to medical examination dated 25.9.1985. As per report Ex.P6 she gave opinion that no definite opinion can be given about rape. In the report it was stated that no external injuries were found \fon her body. The trial Court found that the evidence of Latura (PW-3) who is father of PW-2, Bharosa (PW-4) and Puliabai (PW-5) was inconsistent and the defence witness Maya probabilised the defence taken by the trial Court. Doctor also categorically stated that she was not pregnant on the alleged date of occurrence."

Applicable statutes
IPC 506: Whoever commits the offence of criminal intimidation shall be punished with imprisonment of either description for a term which may extend to two years, or with fine, or with both IPC 376: Whoever, except in the cases provided for in sub-section (2), commits rape, shall be punished with rigorous imprisonment of either description for a term which shall not be less than ten years, but which may extend to imprisonment for life, and shall also be liable to fine.

Figure 3: Example case from the statute prediction and explanation train dataset. Two students independently did the annotations, and a senior faculty member of law resolved disagreements, if any.

Annotated explanation	Generated explanation
"From the evidence of the doctor (P.W.7), it appears that Jhalar Mandal had only the injury on his head, which is said to be the cause of his death.": ["IPC 147"], "From the evidence on record, it is clear that at the time when the second incident took place which led to the death of the deceased even according to the prosecution, none of the accused persons was motivated by any particular desire to attack Jhalar Mandal.": ["IPC 147"], "The incident in question occurred on the spur of the moment without there being any intention of causing death or of causing such injury as they knew was likely to cause death, and was an act arising out of the enmity they had with the nephew of the deceased and aggravated by the unwanted questioning by the deceased.": ["IPC 147"], "From the evidence it is clear that the act of the appellant cannot be construed as an act other than causing grievous hurt.": ["IPC 147"], "In this background, we agree with the learned counsel for the appellants that in the absence of any motive, intention or knowledge as to their act which led to the death of the deceased the appellants can only be held guilty for an offence punishable under section 326 read with 34 Indian Penal Code, 1860 since there is material to show that these 2 appellants did wield their lathis out of which at least one blow, if not both, struck the head of the deceased causing him grievous injury which ultimately led to his death.": ["IPC 147"]	Jhalar Mandal had only the injury on his head, which is said to be the cause of his death.:["IPC 302"],The appellant herein and 4 other accused persons formed themselves into an unlawful assembly armed with lathis and Tangi on 1980 and went to the field of Ugan Mandal PW-4.: ["IPC 147"]
"they killed his daughter by burning with kerosene".: ["IPC 302"], "cremated her body at 00 a.m.": ["IPC 201"]	that immediately after the marriage, Jaggu Ram, his son Jeevan Ram and wife Nathi started harassing Shanti in connection with dowry.:["IPC 498A"],they killed his daughter by burning with kerosene and cremated her body.:["IPC 201"]
"while deceased Baburao and his son Sharanappa were proceeding from their house to their field carrying food for their father who was staying in the field and when they were passing through the field of Suleman, the appellants along with three others assaulted Baburao.": ["IPC 302"], "2 gave a blow with an axe on the neck of the deceased as a result of which he fell down and soon thereafter died.": ["IPC 302"]}	Appellant no. 2 gave a blow with an axe on the neck of the deceased as a result of which he fell down and soon thereafter died.: ["IPC 302"]

Figure 4: High-quality explanations with statute predictions generated by LLMs. The explanations produced by LLMs closely align with those annotated by legal experts.

An example of Misaligned Reasoning Despite Correct Statute Prediction

... The appellant was bewildered and motionless for some time. The appellant apprehended that Dr. Bhattacharya would create a situation which may adversely affect appellant's health. There was heated exchange of words which resulted in commotion. **There was scuffle on the arrival of outsiders and two sons of Parul Jana out of hospital premises.** The appellant immediately contacted the General Manager (Projects) and requested him to help to control the situation. When the General Manager reached the hospital, the appellant explained the situation to him. The General Manager also met Dr. Bhattacharya to get true and correct facts as to how the incident had happened. The General Manager then advised the appellant to go back. Immediately, the appellant left the hospital. In the entire incident, asserted the appellant, save and except accompanying sons of Parul Jana, he did nothing. He was not involved in the incident in any manner whatsoever. It was the Chief Medical Officer, who alone was responsible for the entire unfortunate situation. He also inflicted injuries on two sons of Parul Jana. Dr. Bhattacharya, however, cooked up a false case against the appellant alleging that the appellant had assaulted and injured him. On 6th May, 1999, i.e. on the next day, the Chief Medical Officer, **Dr. Bhattacharya reported to the management that at the late night hours of 5th May, 1999, the appellant led by a bunch of hooligans had visited the hospital, assaulted him,** i.e. Dr. Bhattacharya and abused and threatened other officers....

Figure 5: The red-colored text span represents the portion labeled by the GPT-4 model in the *Chain-of-Thought* setup, where the predicted statute is *Indian Penal Code 147*. The blue-colored span corresponds to the annotation provided by legal experts, with the same predicted statute, *Indian Penal Code 147*. Thus, although GPT-4 correctly identified the statute, its reasoning was incorrect.