Rethinking Legal Judgement Prediction in a Realistic Scenario in the Era of Large Language Models

Shubham Kumar Nigam[1] **Aniket Deroy**[2] **Subhankar Maity**[2] **Arnab Bhattacharya**[1]
[1]Indian Institute of Technology Kanpur (IIT-K)
[2]Indian Institute of Technology Kharagpur (IIT-Kgp)
{sknigam, arnabb}@cse.iitk.ac.in
{roydanik18, subhankar.ai}@kgpian.iitkgp.ac.in

Abstract

This study investigates judgment prediction in a realistic scenario within the context of Indian judgments, utilizing a range of transformer-based models, including InLegalBERT, BERT, and XLNet, alongside LLMs such as Llama-2 and GPT-3.5 Turbo. In this realistic scenario, we simulate how judgments are predicted at the point when a case is presented for a decision in court, using only the information available at that time, such as the facts of the case, statutes, precedents, and arguments. This approach mimics real-world conditions, where decisions must be made without the benefit of hindsight, unlike retrospective analyses often found in previous studies. For transformer models, we experiment with hierarchical transformers and the summarization of judgment facts to optimize input for these models. Our experiments with LLMs reveal that GPT-3.5 Turbo excels in realistic scenarios, demonstrating robust performance in judgment prediction. Furthermore, incorporating additional legal information, such as statutes and precedents, significantly improves the outcome of the prediction task. The LLMs also provide explanations for their predictions. To evaluate the quality of these predictions and explanations, we introduce two human evaluation metrics: *Clarity* and *Linking*. Our findings from both automatic and human evaluations indicate that, despite advancements in LLMs, they are yet to achieve expert-level performance in judgment prediction and explanation tasks.

1 Introduction

Predicting case outcomes based on judge-summarized narratives is an important task. Unlike previous studies (Malik et al., 2021; Nigam et al., 2024) and (Vats et al., 2023), we aim to simulate realistic scenarios where legal judgment prediction systems are used to predict and explain judgments as cases arrive on the bench for adjudication. Our approach focuses on the core factual components of the case—specifically, the events that led to the case being filed, which serve as the basis for judgment prediction. These facts are the foundation of legal arguments and provide the context needed for making judicial decisions. In contrast to previous works that have included the entire case text (including proceedings), our focus on facts mirrors real-world conditions, where judges rely primarily on the case facts when delivering judgments.

In addition to the facts of the case, we incorporate additional legal information such as statutes, precedents, and arguments. Statutes represent codified legal principles, while precedents provide case-specific rulings that help guide decision-making. Together, these legal frameworks offer a structured basis upon which judges rely when formulating their rulings. By extracting and integrating these elements into our models, we aim to enhance both the prediction and explanation tasks by grounding the analysis in actual legal texts and the governing principles that are applied in real cases.

We explore the efficacy of various transformer-based models investigate the impact of summarizing legal judgments (Deroy et al., 2021; Deroy and Maity, 2023; Nigam et al., 2023a; Deroy et al., 2024b) using techniques (Deroy et al., 2023, 2024c,a; Nigam and Deroy, 2023) such as BERT-Sum (Liu, 2019), CaseSummarizer (Polsley et al., 2016), LetSum (Farzindar, 2004), and SummaRuN-Ner (Nallapati et al., 2017). Our findings suggest that leveraging summarized information yields decent results in judgment prediction.

To further enhance the quality of prediction, we introduce hierarchical transformer models that utilize the entirety of judgment facts, demonstrating superior performance compared to traditional summarization methods. Additionally, our examination of LLMs, including Llama-2 (13b & 70b) (Touvron et al., 2023) and GPT-3.5 Turbo (Brown, 2020), highlights the exceptional performance of GPT-3.5 Turbo in the context of Indian legal judgment pre-

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 61–80

diction. We find that augmenting our models with additional legal information, such as statutes, precedents, and arguments, significantly improves the quality of both tasks.

In addition to focusing on the accuracy of legal judgment prediction, it is equally important to assess the quality of the explanations provided by the models. For this reason, we introduce two novel human evaluation metrics: *Clarity* and *Linking*. *Clarity* refers to how well the predictions and explanations are structured and whether they convey the reasoning in a clear and understandable manner. This is critical in the legal domain, where complex legal concepts must be communicated effectively. *Linking*, on the other hand, evaluates the logical consistency between the explanation and the final judgment. It assesses whether the explanation effectively ties back to the outcome and supports the predicted decision. These metrics are vital because, while models may produce accurate predictions, their explanations often lack coherence or fail to justify the decision meaningfully. By incorporating these metrics, we aim to ensure that models provide not only accurate outcomes but also transparent and interpretable explanations that can be trusted by legal professionals.

The key contributions of this study are:

1. We focus on evaluating the performance of several transformer-based models and hierarchical transformer models, specifically on factual data, to mirror real-world conditions in judgment prediction. This approach contrasts with previous works that utilized full case texts.
2. We utilize LLMs to assess their capabilities in legal judgment prediction and explanation tasks.
3. We define two human evaluation metrics, *Clarity* and *Linking*, to assess the quality of LLM-generated judgment predictions and explanations, providing a comprehensive assessment of the overall task performance.

To ensure reproducibility, both the code and dataset have been made publicly available via our repository[1]. Additionally, for convenience, we have uploaded the data[2] and models[3] to Huggingface.

[1] https://github.com/ShubhamKumarNigam/Realistic_LJP

[2] huggingface.co/collections/L-NLProc/Realistic_LJP-models

[3] huggingface.co/collections/L-NLProc/Realistic_LJP-datasets

2 Related Work

The field of Legal Judgment Prediction (LJP) has seen significant advancements, driven by the need to automate legal case outcome forecasting and alleviate the burden of overwhelming caseloads. Early works by (Aletras et al., 2016), (Chalkidis et al., 2019), and (Feng et al., 2021) laid the foundation for LJP, emphasizing the importance of explainability in AI predictions. Benchmark datasets such as CAIL2018 (Xiao et al., 2018), ECHR-CASES (Chalkidis et al., 2019), and others have spurred research in this area, inspiring models like TopJudge and MLCP-NLN. However, there remains a gap between machine and human performance.

In the Indian context, datasets like ILDC (Malik et al., 2021), PredEx (Nigam et al., 2024) and (Nigam et al., 2022; Malik et al., 2022; Nigam et al., 2023b) have highlighted the growing role of AI in legal judgments, with an emphasis on explainability. Research in LJP with LLMs, such as (Vats et al., 2023) and (Nigam et al., 2024), has experimented with models like GPT-3.5 Turbo and Llama-2 on Indian legal datasets. Other studies, such as (Masala et al., 2021) on Romanian legal texts and (Hwang et al., 2022) on Korean legal language, have demonstrated LJP's adaptability across legal systems.

Cross-jurisdictional work, including (Zhao et al., 2018), showcases LJP's applicability in different legal frameworks, with research expanding to multilingual considerations, as seen in (Niklaus et al., 2021) and (Kapoor et al., 2022) for Hindi legal documents. Recent innovations, such as event extraction and multi-stage learning (Feng et al., 2022), continue to push the boundaries of LJP research.

3 Task Definition

This study focuses on Supreme Court of India (SCI) judgments, and the Court Judgment Prediction with Explanation task consists of two subtasks:

Task A: Judgment Prediction: This subtask is framed as a binary classification problem specific to SCI cases. Given a segment of the legal judgment as input, the goal is to predict whether the decision favors or is against the appellant. The prediction is represented by binary labels: {1, 0}, where 1 indicates that the appeal is accepted (i.e., if any part of the appeal is accepted, the decision is considered in favor of the appellant). Although some cases might involve multiple heads of appeal, where an appellant might win on some grounds and lose on

others, for the purposes of this task, the outcome is simplified to a binary decision. Cases with mixed outcomes are excluded or reduced to this binary format for prediction.

Task B: Rationale Explanation: This subtask involves generating a coherent explanation or rationale that justifies the predicted decision, based on the provided segment of the judgment. The explanation seeks to clarify the reasoning behind the predicted outcome.

The workflow of the system, as illustrated in Figure 1, captures the entire process—from extracting facts and additional legal information (such as statutes, precedents, and lower court rulings) to feeding this data into transformer models, hierarchical transformers, and LLMs. The diagram visually represents the pipeline of both tasks, highlighting how the prediction and explanation processes interact to form a comprehensive legal judgment prediction system.

4 Dataset

We utilize the ILDC-multi dataset, as described by (Malik et al., 2021), which comprises a total of 34,816 legal judgments from the Supreme Court of India, collected from 1947 to April 2020 via the Indian Kanoon website[4]. This dataset is divided into three subsets: training, validation, and test which contains 32,305, 994, and 1,517 judgments correspondingly. It is specifically designed to support the tasks of Court Judgment Prediction and Explanation (CJPE), with a portion of the legal judgment serving as input for both prediction and explanation processes. Additionally, a subset of this corpus is annotated with gold-standard explanations provided by legal experts, enhancing its utility for developing automated systems that predict and explain judicial outcomes.

5 Methodology

5.1 Extraction of Facts and Additional Information from Judgments

To extract relevant sentences from legal judgments, we employ a Hierarchical BiLSTM-CRF classifier, focusing on different rhetorical roles as identified by (Ghosh and Wyner, 2019). To create a realistic scenario for our model, we utilize the factual and additional contextual information such as statutes and precedents of the judgments as input for transformer models and LLMs.

[4] https://indiankanoon.org/

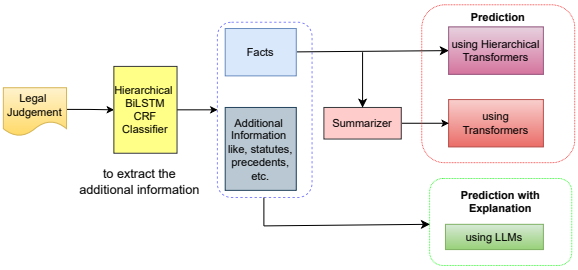

Figure 1: Workflow for Legal Judgment Prediction with explanation.

5.2 Transformer and Hierarchical Transformer Models

The extracted facts undergo summarization using various techniques, including CaseSummarizer (Polsley et al., 2016), BertSum (Liu, 2019), SummaRuNNer (Nallapati et al., 2017), and LetSum (Farzindar, 2004), to ensure they fit within the input constraints of transformer models. Given that models like XLNET-large (Yang et al., 2019), BERT (Devlin et al., 2018), and InLegalBERT (Paul et al., 2022) can process a maximum input length of 512 tokens, we summarize the facts accordingly. Additionally, we utilize hierarchical transformer models that allow us to input the entire set of facts without the need for summarization. This approach facilitates the handling of comprehensive legal information during the prediction task, which is a binary classification problem.

5.3 Prediction with Explanation using LLMs

For the explanation task, we leverage LLMs such as Llama-2 (70b & 13b) (Touvron et al., 2023) and GPT-3.5 Turbo (Brown, 2020), employing a prompting strategy. Given that the combined input and response length for these models is 4096 tokens, we segment the inputs into chunks of 2048 words. This segmentation allows us to generate judgment predictions, as one token corresponds to approximately three-quarters of a word, translating to about 750 words for 1000 tokens[5]. We then aggregate the outputs from multiple chunks using a majority voting mechanism to determine the final judgment; in the event of a tie, the judgment is considered in favor of the appellant. For inputs shorter than 2048 words, we directly input the entire text into the LLM without requiring majority voting. We explore two prompting techniques:

Normal Prompting: The prompt states, "You are asked to be a judge of a legal case and pro-

[5] what-are-tokens-and-how-to-count-them

vide a judgment of the following legal judgment: <Legal judgment>."

Chain-of-Thought Prompting (CoT): Following the chain-of-thought approach proposed by (Wei et al., 2022), the prompt is modified to include, "Think Step by Step."

We investigate six variations for each model input including sentences from:

V1: Only facts.
V2: V1 + statutes, and precedents.
V3: V2 + rulings by lower courts.
V4: V3 + arguments.
V1+CoT: Similar to V1, but incorporates the CoT prompt, "Think Step by Step."
V4+CoT: Similar to V4, but includes the CoT.

Variations V1 and V2 simulate realistic scenarios where only essential elements, such as facts, statutes, and precedents, are provided to the LLM. These components mirror how judges typically approach cases by relying on the factual context and legal frameworks. V3 accounts for cases where a lower court has previously ruled on the matter, adding another layer of realism by simulating situations where an appeal is being heard. V4 enhances the prediction process by including arguments from legal counsel, simulating the complexity of real courtroom proceedings.

Prompting strategies engage both Task A (prediction) and Task B (explanation), thereby facilitating a comprehensive approach to judgment prediction and rationale generation.

6 Evaluation of Model Performance

6.1 Automatic Evaluation

Table 1 summarizes the performance of judgment predictions made by different LLMs through prompting. The results demonstrate that relying solely on factual information leads to lower performance scores. However, incorporating additional legal case-specific information, such as statutes, precedents, rulings from lower courts, and arguments, significantly enhances the quality of predictions. Among the evaluated models, GPT-3.5 Turbo demonstrates the best overall performance.

Table 2 provides further insights into the performance of various hierarchical transformer models and other transformer architectures. The results show that hierarchical transformer models outperform traditional summarization methods. Notably, models specifically pre-trained on Indian legal data, such as InlegalBERT, exhibit superior performance compared to those trained on generic datasets like BERT. The results indicate that LLMs have yet to reach the performance level of legal experts, who demonstrate a 94% agreement rate, as noted by (Malik et al., 2021).

Metric	V1	V2	V3	V4	V1+CoT	V4+CoT
Llama-2-13b						
Precision	0.6443	0.6839	0.6941	0.6997	0.6821	**0.7221**
Recall	0.6292	0.6246	0.6228	0.6416	0.6319	**0.6824**
F1-score	0.6365	0.6528	0.6445	0.6693	0.6560	**0.7016**
Llama-2-70b						
Precision	0.7011	0.7344	0.7416	**0.7518**	0.7322	0.7416
Recall	0.6644	0.6851	0.7147	0.6952	0.6817	**0.7234**
F1-score	0.6822	0.7088	0.7278	0.7223	0.7059	**0.7323**
GPT-3.5 Turbo						
Precision	0.7016	0.7014	0.7411	0.7609	0.7261	**0.7687**
Recall	0.6894	0.6914	0.6949	**0.7155**	0.6847	0.7132
F1-score	0.6953	0.6962	0.7172	0.7374	0.7047	**0.7398**

Table 1: Performance Metrics for the Judgment Prediction Task on the ILDC-multi dataset using different LLMs across various input configurations (V1, V2, V3, V4, V1+CoT, V4+CoT), utilizing both normal prompting and CoT prompting. Bold values indicate the highest score for each metric and model.

6.2 Expert Evaluation

For the expert evaluation, we selected 25 explanations generated by the GPT-3.5 Turbo model, corresponding to different judgments, and enlisted three legal experts to assess these outputs. Each expert rated the explanations on a scale of 1 to 5 based on two criteria: (i) Clarity, the quality and coherence of the rationale behind the legal judgment, and (ii) Linking, the degree to which the explanation is logically connected to the final outcome of the judgment.

To ensure consistency and reliability in the evaluation, the experts were provided with clear guidelines. They were first instructed to familiarize themselves with both the legal judgments and the model-generated outputs to ensure informed assessments. For each explanation, they evaluated:

Clarity: This criterion focuses on how well the rationale is presented. A clear explanation should have a logical flow, use appropriate terminology, and be easily understood by both legal professionals and laypeople. The experts were asked to consider whether the explanation was coherent and if the reasoning behind the judgment was easy to follow.

Linking: This metric captures how well the explanation ties back to the final outcome. A strong

Metric	HT	CS	SR	BS	LS
			XLNET-large		
Precision	0.6424	0.6313	**0.6478**	0.6227	0.5778
Recall	**0.6036**	0.5713	0.5472	0.5683	0.5602
F1-score	**0.6223**	0.5998	0.5993	0.5942	0.5689
			InlegalBERT		
Precision	0.6534	0.6415	0.6338	**0.6604**	0.6010
Recall	**0.6202**	0.5673	0.5613	0.5885	0.5532
F1-score	**0.6363**	0.6022	0.5954	0.6223	0.5761
			BERT		
Precision	**0.6039**	0.5557	0.5589	0.5592	0.5475
Recall	**0.5838**	0.5540	0.5589	0.5589	0.5457
F1-score	**0.5936**	0.5548	0.5589	0.5590	0.5466

Table 2: Comparative Performance of Transformer Models on the Judgment Prediction Task on the ILDC-multi Dataset. These models are with fact summarization techniques such as CaseSummarizer (CS), SummaRuNNer (SR), BertSum (BS), and LetSum (LS), as well as Hierarchical Transformer (HT) models using the complete facts. Bold values indicate the highest score for each metric and model.

linking score indicates that the rationale clearly leads to the conclusion of the judgment, without any gaps or inconsistencies. The experts were tasked with identifying whether the explanation logically and explicitly supports the final decision. The evaluators used the following rating scales:

- **For Clarity:**
  [1]: Very Poor (Unclear rationale)
  [2]: Poor (Some clarity but weak rationale)
  [3]: Fair (Moderately clear rationale)
  [4]: Good (Clear rationale)
  [5]: Excellent (Very clear rationale)

- **For Linking:**
  [1]: Very Poor (Unclear and disconnected explanation)
  [2]: Poor (Weak linkage between explanation and judgment)
  [3]: Fair (Moderate linking, some gaps)
  [4]: Good (Clear linkage to the judgment)
  [5]: Excellent (Strong and coherent linking)

These ratings, calculated as the average scores for each criterion across the three experts, are presented in Table 3. To ensure objectivity and ethical standards, the experts were instructed to maintain impartiality and avoid conflicts of interest throughout the evaluation process.

The results indicate that Variation 4 with chain-of-thought prompting (V4+CoT) achieved the highest scores for both clarity and linking, demonstrating its effectiveness in producing coherent and well-connected explanations. The average Fleiss' Kappa scores for Clarity and Linking were 0.64 and 0.70, respectively, indicating substantial agreement among the evaluators.

Metric	V1	V2	V3	V4	V1+ CoT	V4+ CoT
Clarity	3.13	3.20	3.33	3.47	3.20	**3.73**
Linking	3.66	3.80	3.87	4.00	3.73	**4.27**

Table 3: Expert Evaluation Results for the Explanation Task Using GPT-3.5 Turbo. Bold values indicate the highest scores for each metric.

The combination of automatic and human evaluations offers a comprehensive assessment of the models' performance, revealing areas for improvement and confirming the efficacy of specific prompting techniques—such as chain-of-thought (CoT) in enhancing the quality of legal judgment prediction and explanation.

7 Conclusions

In this study, we explored the effectiveness of various LLMs and transformer architectures in the task of judgment prediction and explanation using the ILDC-multi dataset. Our results demonstrate that incorporating additional case-specific information significantly enhances the prediction accuracy compared to using only factual information. The results also highlight the superiority of hierarchical transformer models over traditional summarization techniques, suggesting that a comprehensive approach to input data yields better predictive outcomes. Despite the promising results, our evaluations reveal that automated metrics still fall short of matching the performance levels of human legal experts, who demonstrate a high degree of agreement in judgment assessments. This gap underscores the need for further refinement of LLMs and transformer models to improve their interpretability and reliability in legal contexts.

Limitations

This study is focused solely on Supreme Court of India (SCI) judgments, which may limit the generalizability of the models to other courts or jurisdictions. Legal systems in different countries, or even lower courts within the same system, may have distinct structures, procedures, and nuances that are

not captured in this study.

Additionally, the judgment prediction task is simplified as a binary classification problem. In real-world cases, particularly in multi-issue appeals, an appellant may win on some points and lose on others. This complexity is not fully addressed here, as our model reduces the outcome to a binary decision, which may overlook the nuances of cases with multiple heads of appeal.

While we incorporate facts, statutes, precedents, and arguments to simulate a realistic scenario, this approach still does not capture the full range of judicial reasoning. Judges often rely on implicit legal reasoning, judicial discretion, and a wider array of contextual factors that may not be explicitly mentioned in legal documents, limiting the comprehensiveness of our model's predictions.

The large language models (LLMs) used in this study, such as GPT-3.5 Turbo and Llama-2, offer promising results, but their high computational requirements make them resource-intensive. This could restrict their practical application in many legal environments, especially in resource-constrained settings.

Furthermore, the human evaluation metrics—*Clarity* and *Linking*—are based on subjective assessments from legal experts. Although we provided detailed guidelines to standardize the evaluation process, differences in interpretation among experts can introduce variability into the results.

Future research will focus on addressing these limitations by exploring multi-label classification to account for more complex case outcomes, expanding the applicability of models to other legal domains and jurisdictions, and refining evaluation metrics to minimize subjectivity.

Ethical Considerations

In conducting this research, we adhered to ethical standards, particularly in the context of data usage and expert evaluation. The legal judgments used in our experiments were publicly available, and no private or sensitive data was accessed. For the human evaluation of judgment predictions and explanations, we engaged PhD scholars from the Rajiv Gandhi School of Intellectual Property Law as legal experts. Their participation was voluntary, and we provided monetary compensation for their time and expertise. This ensured that the evaluation process was both fair and conducted with proper acknowledgment of the experts' contributions.

References

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

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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

Aniket Deroy, Naksatra Kumar Bailung, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2024a. Artificial intelligence (ai) in legal data mining. *arXiv preprint arXiv:2405.14707*.

Aniket Deroy, Paheli Bhattacharya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2021. An analytical study of algorithmic and expert summaries of legal cases. In *Legal Knowledge and Information Systems*, pages 90–99. IOS Press.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How ready are pre-trained abstractive models and llms for legal case judgement summarization? *arXiv preprint arXiv:2306.01248*.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2024b. Applicability of large language models and generative models for legal case judgement summarization. *Artificial Intelligence and Law*, pages 1–44.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2024c. Ensemble methods for improving extractive summarization of legal case judgements. *Artificial Intelligence and Law*, 32(1):231–289.

Aniket Deroy and Subhankar Maity. 2023. Questioning biases in case judgment summaries: Legal datasets or large language models? *arXiv preprint arXiv:2312.00554*.

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

Atefeh Farzindar. 2004. Atefeh farzindar and guy lapalme,'letsum, an automatic legal text summarizing system in t. gordon (ed.), legal knowledge and information systems. jurix 2004: The seventeenth annual conference. amsterdam: Ios press, 2004, pp. 11-18. In *Legal knowledge and information systems: JURIX 2004, the seventeenth annual conference*, volume 120, page 11. IOS Press.

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

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

Saptarshi Ghosh and Adam Wyner. 2019. Identification of rhetorical roles of sentences in indian legal judgments. *Legal Knowledge and Information Systems: JURIX*, page 3.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.

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.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

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

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

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. 2021. jurbert: A romanian bert model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

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

Shubham Kumar Nigam and Aniket Deroy. 2023. Fact-based court judgment prediction. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 78–82.

Shubham Kumar Nigam, Aniket Deroy, Noel Shallum, Ayush Kumar Mishra, Anup Roy, Shubham Kumar Mishra, Arnab Bhattacharya, Saptarshi Ghosh, and Kripabandhu Ghosh. 2023a. Nonet at semeval-2023 task 6: Methodologies for legal evaluation. *arXiv preprint arXiv:2310.11049*.

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

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

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

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2022. Pre-training transformers on indian legal text. *arXiv preprint arXiv:2209.06049*.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. Casesummarizer: a system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations*, pages 258–262.

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*.

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.

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.

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

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

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

A Expert Evaluation

Table 4 shows scores provided by three legal experts for V1. Table 5 shows scores provided by three legal experts for V2. Table 6 shows scores provided by three legal experts for V3. Table 7 shows scores provided by three legal experts for V4. Table 8 shows scores provided by three legal experts for V1+CoT. Table 9 shows scores provided by three legal experts for V4+CoT.

Table 10 shows scores provided by three legal experts for V1. Table 11 shows scores provided by three legal experts for V2. Table 12 shows scores provided by three legal experts for V3. Table 13 shows scores provided by three legal experts for V4. Table 14 shows scores provided by three legal experts for V1+CoT. Table 15 shows scores provided by three legal experts for V4+CoT.

Document	Legal Expert 1	Legal Expert 2	Legal Expert 3
Document 1	3	3	4
Document 2	3	4	3
Document 3	5	5	5
Document 4	4	4	4
Document 5	3	4	4
Document 6	4	4	4
Document 7	5	5	5
Document 8	2	2	4
Document 9	2	2	2
Document 10	1	2	2
Document 11	3	3	4
Document 12	3	3	4
Document 13	3	3	4
Document 14	3	3	4
Document 15	2	2	3
Document 16	5	5	5
Document 17	4	4	5
Document 18	2	2	2
Document 19	2	3	2
Document 20	2	2	2
Document 21	2	2	2
Document 22	2	2	2
Document 23	4	4	4
Document 24	1	2	1
Document 25	3	4	3

Table 4: Clarity ratings from three legal experts in V1

Document	Legal Expert 1	Legal Expert 2	Legal Expert 3
Document 1	4	4	4
Document 2	3	4	4
Document 3	3	4	3
Document 4	5	5	5
Document 5	2	2	2
Document 6	2	3	3
Document 7	4	4	4
Document 8	2	2	2
Document 9	3	3	3
Document 10	3	3	3
Document 11	5	5	5
Document 12	3	3	3
Document 13	2	2	2
Document 14	3	3	4
Document 15	4	4	4
Document 16	2	2	2
Document 17	2	2	3
Document 18	5	5	5
Document 19	3	3	4
Document 20	4	4	4
Document 21	1	2	2
Document 22	2	2	3
Document 23	3	3	4
Document 24	2	2	3
Document 25	5	5	5

Table 5: Clarity ratings from three legal experts in V2

Document	Legal Expert 1	Legal Expert 2	Legal Expert 3
Document 1	2	3	3
Document 2	5	5	5
Document 3	3	4	4
Document 4	2	2	2
Document 5	4	4	4
Document 6	3	3	3
Document 7	3	4	4
Document 8	2	2	3
Document 9	4	4	4
Document 10	5	5	5
Document 11	3	3	3
Document 12	1	2	3
Document 13	2	3	3
Document 14	5	5	5
Document 15	3	3	4
Document 16	2	3	4
Document 17	3	3	3
Document 18	4	5	5
Document 19	4	4	5
Document 20	2	3	3
Document 21	2	2	2
Document 22	3	3	3
Document 23	2	2	2
Document 24	4	4	4
Document 25	5	4	4

Table 6: Clarity ratings from three legal experts in V3

Document	Legal Expert 1	Legal Expert 2	Legal Expert 3
Document 1	5	5	5
Document 2	3	3	4
Document 3	3	3	4
Document 4	5	5	5
Document 5	2	2	3
Document 6	2	2	3
Document 7	5	5	5
Document 8	3	3	3
Document 9	3	3	3
Document 10	3	3	3
Document 11	2	3	3
Document 12	4	4	4
Document 13	3	4	3
Document 14	3	4	4
Document 15	5	5	5
Document 16	4	4	5
Document 17	2	3	3
Document 18	4	4	4
Document 19	3	3	3
Document 20	2	3	3
Document 21	2	2	3
Document 22	3	3	4
Document 23	4	4	4
Document 24	4	4	4
Document 25	2	2	2

Table 7: Clarity ratings from three legal experts in V4

Document	Legal Expert 1	Legal Expert 2	Legal Expert 3
Document 1	2	3	3
Document 2	4	4	4
Document 3	5	5	5
Document 4	3	3	4
Document 5	2	2	2
Document 6	3	3	4
Document 7	3	3	4
Document 8	1	1	2
Document 9	2	2	2
Document 10	5	5	5
Document 11	2	3	2
Document 12	3	3	3
Document 13	5	5	5
Document 14	4	4	4
Document 15	2	2	3
Document 16	3	3	4
Document 17	4	4	4
Document 18	3	3	4
Document 19	4	4	4
Document 20	3	3	4
Document 21	2	3	3
Document 22	2	3	2
Document 23	5	5	5
Document 24	3	3	3
Document 25	2	3	2

Table 8: Clarity ratings from three legal experts for V1+CoT

Document	Legal Expert 1	Legal Expert 2	Legal Expert 3
Document 1	5	5	5
Document 2	2	3	2
Document 3	4	4	5
Document 4	4	4	5
Document 5	3	4	3
Document 6	4	5	4
Document 7	4	4	4
Document 8	2	2	2
Document 9	3	4	3
Document 10	4	4	5
Document 11	5	5	5
Document 12	3	3	4
Document 13	4	4	5
Document 14	2	3	3
Document 15	4	4	4
Document 16	2	2	3
Document 17	4	4	4
Document 18	3	3	4
Document 19	5	5	5
Document 20	4	4	4
Document 21	2	2	3
Document 22	3	3	4
Document 23	5	5	5
Document 24	2	2	3
Document 25	4	4	4

Table 9: Clarity ratings from three legal experts for V4+CoT

Document	Legal Expert 1	Legal Expert 2	Legal Expert 3
Document 1	3	4	4
Document 2	3	4	3
Document 3	5	5	5
Document 4	3	4	3
Document 5	5	5	4
Document 6	3	4	3
Document 7	4	5	5
Document 8	3	4	4
Document 9	3	4	1
Document 10	3	3	2
Document 11	3	3	4
Document 12	3	3	4
Document 13	4	4	4
Document 14	5	5	4
Document 15	3	3	3
Document 16	4	5	4
Document 17	4	4	5
Document 18	5	5	2
Document 19	4	5	2
Document 20	3	3	2
Document 21	4	5	2
Document 22	4	5	1
Document 23	4	4	4
Document 24	4	4	1
Document 25	4	4	3

Table 10: Linking ratings from three legal experts for V1

Document	Legal Expert 1	Legal Expert 2	Legal Expert 3
Document 1	3	4	4
Document 2	3	4	4
Document 3	5	5	5
Document 4	3	4	4
Document 5	5	5	5
Document 6	3	4	4
Document 7	4	4	4
Document 8	3	4	3
Document 9	3	4	4
Document 10	3	4	4
Document 11	3	4	4
Document 12	3	3	3
Document 13	4	4	4
Document 14	5	5	5
Document 15	3	4	4
Document 16	4	4	3
Document 17	4	4	4
Document 18	5	5	5
Document 19	4	5	5
Document 20	3	3	3
Document 21	4	4	3
Document 22	4	4	3
Document 23	3	3	2
Document 24	3	3	2
Document 25	3	3	3

Table 11: Linking ratings from three legal experts for V2

Document	Legal Expert 1	Legal Expert 2	Legal Expert 3
Document 1	4	5	4
Document 2	4	4	3
Document 3	5	5	4
Document 4	4	4	4
Document 5	5	5	5
Document 6	4	4	4
Document 7	4	4	3
Document 8	4	4	4
Document 9	4	4	4
Document 10	4	4	3
Document 11	4	4	3
Document 12	3	3	3
Document 13	4	4	4
Document 14	5	5	5
Document 15	4	4	3
Document 16	4	4	4
Document 17	4	4	4
Document 18	5	5	3
Document 19	5	5	4
Document 20	3	3	3
Document 21	4	4	4
Document 22	4	2	3
Document 23	3	2	2
Document 24	3	2	2
Document 25	3	2	2

Table 12: Linking ratings from three legal experts for V3

Document	Legal Expert 1	Legal Expert 2	Legal Expert 3
Document 1	3	2	2
Document 2	3	2	3
Document 3	3	3	2
Document 4	4	4	3
Document 5	4	4	4
Document 6	4	4	4
Document 7	5	5	4
Document 8	4	4	3
Document 9	4	4	4
Document 10	4	4	3
Document 11	5	4	4
Document 12	5	5	5
Document 13	4	3	5
Document 14	4	4	3
Document 15	4	4	5
Document 16	5	5	4
Document 17	5	4	5
Document 18	5	5	5
Document 19	5	5	5
Document 20	5	5	5
Document 21	5	3	4
Document 22	5	4	5
Document 23	5	4	5
Document 24	5	5	5
Document 25	5	3	5

Table 13: Linking ratings from three legal experts for V4

Document	Legal Expert 1	Legal Expert 2	Legal Expert 3
Document 1	2	3	3
Document 2	4	4	4
Document 3	5	5	5
Document 4	3	3	4
Document 5	2	2	2
Document 6	3	3	4
Document 7	3	1	4
Document 8	1	1	2
Document 9	2	2	5
Document 10	5	5	2
Document 11	2	3	2
Document 12	3	3	3
Document 13	5	5	5
Document 14	4	4	4
Document 15	2	2	4
Document 16	3	3	5
Document 17	4	4	4
Document 18	3	3	4
Document 19	4	3	4
Document 20	3	3	4
Document 21	2	3	3
Document 22	2	3	2
Document 23	5	5	5
Document 24	3	3	2
Document 25	2	3	5

Table 14: Linking ratings from three legal experts for V1+CoT

Document	Legal Expert 1	Legal Expert 2	Legal Expert 3
Document 1	2	2	2
Document 2	4	4	4
Document 3	5	3	5
Document 4	4	4	4
Document 5	3	3	3
Document 6	4	4	3
Document 7	4	3	4
Document 8	2	2	2
Document 9	3	3	4
Document 10	5	5	4
Document 11	3	3	3
Document 12	4	5	3
Document 13	5	4	3
Document 14	4	3	3
Document 15	5	4	4
Document 16	4	4	3
Document 17	5	4	4
Document 18	4	4	4
Document 19	5	5	4
Document 20	3	3	3
Document 21	3	3	2
Document 22	5	5	2
Document 23	4	4	4
Document 24	3	2	2
Document 25	3	5	4

Table 15: Linking ratings from three legal experts for V4+CoT