ReGal: A First Look at PPO-based Legal AI for Judgment Prediction and Summarization in India

Shubham Kumar Nigam[1, 5*†], **Tanuj Tyagi**[2*], **Siddharth Shukla**[2*], **Aditya Kumar Guru**[2*], **Balaramamahanthi Deepak Patnaik**[1 *], **Danush Khanna**[2*], **Noel Shallum**[3*], **Kripabandhu Ghosh**[4], **Arnab Bhattacharya**[1]

[1]Indian Institute of Technology Kanpur, India [2]Manipal University Jaipur, India [3]Symbiosis Law School Pune, India
[4]IISER Kolkata, India [5]University of Birmingham, Dubai, United Arab Emirates
{shubhamkumarnigam, tanujtyagiofficial, siddharth8shukla8, adityaguru20, bdeepakpatnaik2002, danush.s.khanna, noelshallum}@gmail.com
kripaghosh@iiserkol.ac.in arnabb@cse.iitk.ac.in

Abstract

This paper presents an early exploration of reinforcement learning methodologies for legal AI in the Indian context. We introduce REINFORCEMENT LEARNING-BASED LEGAL REASONING (ReGal), a framework that integrates Multi-Task Instruction Tuning with Reinforcement Learning from AI Feedback (RLAIF) using Proximal Policy Optimization (PPO). Our approach is evaluated across two critical legal tasks: (i) Court Judgment Prediction and Explanation (CJPE), and (ii) Legal Document Summarization. Although the framework underperforms on standard evaluation metrics compared to supervised and proprietary models, it provides valuable insights into the challenges of applying RL to legal texts. These challenges include reward model alignment, legal language complexity, and domain-specific adaptation. Through empirical and qualitative analysis, we demonstrate how RL can be repurposed for high-stakes, long-document tasks in law. Our findings establish a foundation for future work on optimizing legal reasoning pipelines using reinforcement learning, with broader implications for building interpretable and adaptive legal AI systems.

Code — https://github.com/ShubhamKumarNigam/ReGal

Introduction

This paper explores the use of Reinforcement Learning (RL) for two crucial tasks in Indian legal AI: Court Judgment Prediction and Explanation (CJPE) and Legal Document Summarization. While recent works have explored various approaches for judgment prediction (Malik et al. 2021b; Vats et al. 2023; Nigam et al. 2024) and rhetorical segmentation (Bhattacharya et al. 2019; Malik et al. 2021a), the integration of reinforcement learning, especially using Proximal Policy Optimization (PPO), remains largely unexplored in this domain. Our work marks one of the first attempts to apply PPO-based RL techniques to Indian legal tasks, addressing the unique challenges posed by the Indian judiciary system.

The CJPE task involves two interlinked subtasks: prediction of the case outcome and generation of a rationale explanation based on factual case records. While prior research in this area has primarily relied on supervised fine-tuning of pretrained language models (Nigam and Deroy 2023; Nigam et al. 2023; Katz, Bommarito, and Blackman 2017; Zhu et al. 2020), such approaches are often limited by their reliance on large, annotated datasets and their inability to incorporate real-time feedback for interpretability. Our framework, REINFORCEMENT LEARNING-BASED LEGAL REASONING (ReGal), addresses this gap by employing RL to iteratively refine both decisions and explanations using a reward signal derived from legal textual alignment.

In addition to CJPE, we also extend our framework to the legal summarization task, which is a critical tool for improving access to justice and aiding practitioners in quickly understanding long judgments. We experiment on the In-Abs summarization dataset, a benchmark curated for generating abstract-style summaries from Indian court decisions. While summarization has seen advances through neural and transformer-based models (Deroy, Ghosh, and Ghosh 2023; Shukla et al. 2022a; Datta et al. 2023; Joshi et al. 2024), the application of RL-based optimization in this setting, especially under Indian law, remains underexplored. By testing our PPO-based approach on both CJPE and summarization, we demonstrate that reinforcement learning has cross-task potential in legal NLP, despite facing several challenges in terms of performance and hallucination.

While ReGal underperforms compared to fine-tuned and proprietary models like GPT-3.5, our findings reveal critical insights into why RL techniques struggle with legal texts, including reward model misalignment, domain complexity, and lack of domain-adaptive pretraining. The ablation studies, hallucination examples, and comparative analysis on lexical and semantic metrics collectively highlight the limitations of current RLHF techniques in handling nuanced,

*These authors contributed equally.
†Corresponding Author

Proceedings of Bridge between Artificial Intelligence and Law 2026 (Full papers), pages 127–137
Januray 21, 2026, AAAI 2026 Bridge Program, Singapore

high-stakes legal language.

This work should be viewed as a position paper that lays the groundwork for future improvements. Rather than focusing solely on state-of-the-art performance, we emphasize methodological exploration, share valuable lessons, and propose actionable directions for advancing RL in legal AI.

Our contributions are threefold: (1) we present one of the first applications of PPO-based reinforcement learning in Indian legal judgment prediction and summarization; (2) we provide empirical and qualitative evidence on its limitations; and (3) we chart a path for more effective legal-AI pipelines integrating RLHF, human feedback, and domain-adapted modeling.

Related Work

Legal Judgment Prediction (LJP) Legal judgment prediction has been explored across various jurisdictions using SVMs, CNNs, transformers, and now LLMs (Aletras et al. 2016; Chalkidis, Androutsopoulos, and Aletras 2019; Feng et al. 2021). Benchmarks like CAIL2018 (Xiao et al. 2018), ECHR (Chalkidis, Androutsopoulos, and Aletras 2019), and TopJudge established task formulations in Chinese and European legal systems.

In India, ILDC (Malik et al. 2021c), PredEx (Nigam et al. 2024), and NyayaAnumana (Nigam et al. 2025) provided datasets for factual and explainable LJP. Nigam and Deroy (2023); Nigam et al. (2023) proposed hierarchical transformers and fact-only predictions. Other studies (Kapoor et al. 2022; Ganguly et al. 2023) focused on Hindi documents and long-text summarization.

LJP has also been studied cross-jurisdictionally (Zhao et al. 2018), in Romania (Masala et al. 2021), Korea (Hwang et al. 2022), and Switzerland (Niklaus, Chalkidis, and Stürmer 2021). However, no prior work has used RLHF or RLAIF to refine predictions in Indian courts.

Legal Judgment Summarization Legal case summarization is challenging due to the length and rhetorical structure of court documents. Extractive and abstractive summarization using models like BART, PEGASUS, and LED have been explored (Shukla et al. 2022b; Feijo and Moreira 2023). Older rule-based and graphical models like CaseSummarizer (Polsley, Jhunjhunwala, and Huang 2016) and DelSumm (Bhattacharya et al. 2021) have also been proposed.

In the Indian context, most legal summarization work remains extractive. No prior work has utilized reinforcement learning for Indian legal summarization.

Reinforcement Learning for Legal Summarization Several prior works have explored the use of reinforcement learning in the legal domain, particularly for tasks such as legal summarization. For instance, Shukla et al. (2022c) investigated the application of policy gradient methods to improve summarization quality by aligning model outputs with reward signals derived from human feedback. Similarly, Wang and Wu (2024) proposed a novel RL framework based on Soft Actor-Critic with Variational Autoencoders (SAC-VAE), demonstrating its effectiveness in generating high-quality legal summaries. In other lines of work, Nguyen et al. (2021), Dong and Lin (2024), and Patil and Patankar (2025) introduced value-based methods such as Deep Q-Networks (DQN) and Advantage Actor-Critic (A2C) models to reinforce generation strategies in legal texts, enabling more contextually grounded and fluent outputs. These studies collectively highlight the growing interest in using RL techniques to align model behavior with human expectations in legal NLP applications.

However, these remain extractive or domain-agnostic. Our work is the first to explore abstractive legal summarization in Indian courts using RLHF and RLAIF.

Instruction Tuning and RLAIF Instruction tuning (IT) helps align LLMs with user intents (Wei et al. 2022; Mishra et al. 2022; Zhou et al. 2023). Scaling IT with diverse prompts improves controllability and task generalization (Zhang et al. 2023a; Wang et al. 2022b).

RLHF (Ouyang et al. 2022; Glaese et al. 2022) improves LLM alignment with human preferences. RLAIF (Lee et al. 2023) offers a cost-effective alternative using AI-generated feedback, achieving near-RLHF performance (Ziegler et al. 2020; Bai et al. 2022).

To our knowledge, this is the first application of RLAIF and RLHF for both LJP and summarization in the Indian legal domain.

LLMs for Summarization and Other Domains LLMs like GPT-4 and Claude have shown strong summarization abilities for long books (Chang et al. 2023), news articles (Zhang et al. 2023b), and meeting transcripts (Schneider and Turchi 2023).

Hierarchical summarization (Chang et al. 2023) and zero-shot/few-shot prompting have demonstrated LLMs' potential in generating abstractive summaries. Our work extends these paradigms to Indian legal documents. While datasets like CrossSum (Bhattacharjee et al. 2021), WikiLingua (Ladhak et al. 2020), and ILSUM (Urlana et al. 2023) cover Indian languages, they lack legal domain coverage.

Eur-Lex (Aumiller, Chouhan, and Gertz 2022) and CLIDSUM (Wang et al. 2022a) offer multilingual legal summaries, but no such benchmark exists for Indian court judgments. Our work bridges this gap via expert-annotated data and RL tuning.

To the best of our knowledge, this is the first work that applies RLHF/RLAIF to both legal judgment prediction and summarization tasks in the Indian legal context.

Task Description

This paper presents the ReGal framework, which combines instruction tuning with Reinforcement Learning from AI Feedback (RLAIF) to enhance legal NLP systems. As a position paper, our goal is to evaluate the general applicability of this architecture across multiple tasks in the Indian legal domain, specifically focusing on reinforcement learning's capacity to refine predictions and enhance interpretability.

Figure 1 illustrates the overall architecture and PPO training pipeline used across tasks. Rather than being specific to a single task, this figure highlights how the same optimization loop, grounded in expert or model-generated feedback, is flexibly applied to different legal reasoning tasks. While our

experiments are conducted on Indian Supreme Court judgments, the ReGal framework is task-agnostic and potentially extensible to other domains requiring high interpretability and reasoning fidelity.

We evaluate ReGal on the following two core tasks:

Task 1: Court Judgment Prediction and Explanation (CJPE)

This task assesses the model's ability to reason over complex legal documents and consists of two tightly coupled subtasks:

Task 1A: Judgment Prediction Given a legal document D, typically a judgment from the Supreme Court of India, the goal is to predict whether the appeal or petition was accepted or rejected, represented as binary labels $y \in \{0, 1\}$. This task reflects the practical need for predictive tools in legal practice and case screening.

Task 1B: Rationale Explanation In this subtask, the model is expected to generate a natural language explanation supporting its predicted outcome. The explanation should reflect the key reasoning patterns within the case text and ideally mimic judicial argumentation. This component enhances the interpretability and trustworthiness of the AI system.

Task 2: Legal Judgment Summarization

To demonstrate the generalization of the ReGal framework, we also evaluate it on the task of judgment summarization. Here, the model must generate a concise yet informative summary of the full judgment text, capturing its essential components such as background, legal issues, arguments, and final verdict. This task reflects a broader class of document understanding problems and allows us to assess the effect of reinforcement learning in improving content selection and abstraction in long legal texts.

By applying a unified PPO-based training regime across these tasks, we argue that ReGal provides a promising and extensible approach for reinforcement-tuned legal AI. The tasks differ in their output formats and supervision requirements, but share a common challenge: the need for factual consistency, interpretability, and domain alignment, properties that our reinforcement learning strategy seeks to enhance.

Dataset

To evaluate the proposed ReGal framework across diverse legal tasks, we utilized two large-scale, expert-curated datasets focused on reasoning and summarization in the Indian legal domain.

CJPE Task Dataset — PredEx: For the Court Judgment Prediction and Explanation (CJPE) task, we use the publicly available PredEx dataset introduced by (Nigam et al. 2024). It is the largest annotated dataset in the Indian legal NLP space for judgment prediction and rationale generation. Developed with inputs from ten senior law students across reputed institutions, the dataset ensures annotation quality through expert-in-the-loop processes conducted between April 2022 and October 2023.

Metric	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 1: Statistics of the PredEx dataset (Nigam et al. 2024) used for judgment prediction and explanation.

Dataset	In-Abs
# Documents	7,130
Type of Summary	Abstractive
Language	English
Train/Test Split	7,030/100
Avg. Document size (in #words)	4376.98
Avg. Summary size (in #words)	842.52
Avg. Compression Ratio	0.235

Table 2: Statistics of the In-Abs summarization dataset (Shukla et al. 2022a).

PredEx contains 15,222 Supreme Court judgment documents, split into training (12,178) and test (3,044) sets. Each document includes the full case text, a binary outcome label (accepted/rejected), and a corresponding human-written explanation for the decision. Summary statistics are provided in Table 1.

Summarization Task Dataset — In-Abs: For the legal document summarization task, we employ the In-Abs subset of the IL-TUR dataset (Joshi et al. 2024; Shukla et al. 2022a). This dataset includes 7,130 Supreme Court documents paired with expert-written headnotes, which serve as abstractive summaries capturing critical aspects such as legal issues, arguments, and verdicts. These summaries, written by legal experts and extracted using heuristic rules, provide high-quality ground truth for evaluating summarization quality.

The dataset is split into 7,030 training and 100 test samples. Table 2 shows key statistics.

Both datasets allow us to assess the applicability of reinforcement learning via PPO across diverse generative tasks, prediction, explanation, and summarization, within Indian legal NLP, demonstrating the flexibility of our approach.

Methodology

We propose a reinforcement learning framework, `ReGal`, that aims to enhance large language models (LLMs) for legal reasoning tasks in the Indian judicial domain. Our framework is instantiated over two distinct legal tasks: (i) Court Judgment Prediction and Explanation (CJPE), and (ii) Legal Summarization. For both tasks, we adopt a two-stage

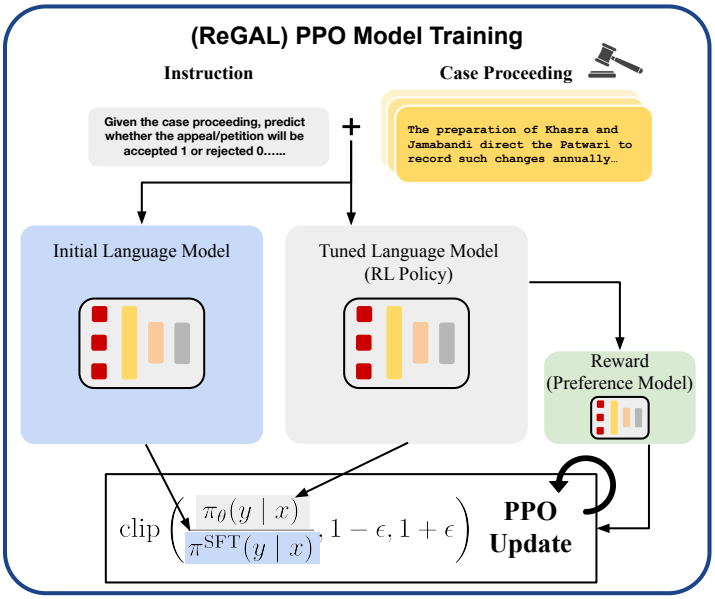

Figure 1: Overview of the ReGal PPO model training process.

approach involving supervised fine-tuning followed by reinforcement learning via Proximal Policy Optimization (PPO), using AI-generated reward signals. This setup enables us to explore the applicability and effectiveness of PPO across multiple legal generation objectives.

Base Model and SFT Training

For all tasks, we adopt Llama-2-7B as the base language model. This decision is motivated by its prior use in recent literature on legal judgment prediction and explanation, particularly within the Indian legal context (Nigam et al. 2024). By choosing the same model, we enable a fair comparison between our proposed reinforcement learning-based alignment (ReGal) and earlier fine-tuning approaches. This consistency also helps isolate the effects of our PPO optimization, allowing us to demonstrate the specific contribution of reinforcement learning to performance gains.

We fine-tune Llama-2-7B using supervised instruction tuning on both tasks: for CJPE, this includes predicting the case outcome and generating explanations using the PredEx dataset; and for Summarization, the model is trained to generate abstractive summaries using the IL-TUR dataset. This supervised fine-tuned model, denoted as π^{SFT}, serves as the reference policy in subsequent reinforcement learning stages.

Reward Models

To align the outputs with the desired objectives, we develop task-specific reward models (RMs). For CJPE, we fine-tune InLegalBERT to classify the correctness of predictions. The RM assigns binary rewards: 1 if the model predicts the correct verdict, and 0 otherwise. For Summarization, the RM is trained to score summaries based on n-gram overlap with the gold headnotes and coherence. We use ROUGE-style matching and shallow semantic similarity to assign scalar rewards.

These RMs simulate AI-based feedback in lieu of human-in-the-loop supervision.

Proximal Policy Optimization (PPO)

To address the objective mismatch in the fine-tuned model and improve the alignment between predicted outputs and desired outcomes, we employ Proximal Policy Optimization (PPO). PPO is a widely used actor-critic algorithm designed to optimize policies in reinforcement learning, especially for large language models. It allows us to adjust the model's behavior in response to feedback by iteratively updating its parameters, thus aligning the model more effectively with the prediction and explanation tasks.

In this context, we treat the aligned language model as a learnable policy π_θ, where θ represents the model's parameters. PPO optimizes this policy to minimize the discrepancy between the model's predicted outputs and the expected outcomes, as dictated by a reward function. The following objective function is minimized during training:

$$\mathcal{L}_{PPO}(\theta) = \mathbb{E}_{x\sim D, y\sim\pi_\theta(x)} \left[\min \left\{ \frac{\pi_\theta(y \mid x)}{\pi^{SFT}(y \mid x)} \cdot r(y), \right. \right.$$
$$\left. \left. clip\left(\frac{\pi_\theta(y \mid x)}{\pi^{SFT}(y \mid x)}, 1-\epsilon, 1+\epsilon\right) \cdot r(y) \right\} \right] \quad (1)$$

The key variables remain as defined previously, where $r(y)$ is task-specific reward based on prediction accuracy (CJPE) or summary quality (Summarization).

Here is a detailed breakdown of each parameter in this equation:

- $\mathcal{L}_{PPO}(\theta)$: This is the loss function that PPO aims to minimize. It represents the policy's objective and is calculated by averaging the expected value over all data points in the dataset D, which consists of legal cases.
- $\mathbb{E}_{x\sim D, y\sim\pi_\theta(x)}$: This denotes the expectation over the data distribution. Specifically, x represents a case from the dataset D, and y is the prediction (i.e., the legal judgment generated by the policy π_θ). The expectation signifies that the loss is computed as an average over multiple samples.
- $\pi_\theta(y \mid x)$: This is the current policy that the model is learning. It outputs the probability of generating the prediction y given the input case x, based on the current model parameters θ.
- $\pi^{SFT}(y \mid x)$: This represents the Supervised Fine-Tuned (SFT) policy, which serves as the baseline policy. The SFT model has already been fine-tuned on the task using supervised learning techniques, and it provides a reference probability distribution for the predictions y.
- $r(y)$: The reward assigned to the prediction y by the reward model. The reward reflects how accurate the prediction is, with higher values indicating better alignment with the correct outcome. For instance, a reward of 1 may be assigned for a correct judgment prediction, while a reward of 0 is given for an incorrect prediction.
- $\frac{\pi_\theta(y|x)}{\pi^{SFT}(y|x)}$: This is the probability ratio between the current policy and the supervised fine-tuned (SFT) policy. It quantifies how much the current policy π_θ deviates from the baseline policy π^{SFT}. The KL-divergence between the two policies is used to compute a penalty, ensuring that the current policy does not deviate too far from the supervised baseline. This penalty is subtracted from the reward to balance exploration of new behaviors while maintaining alignment with the reference language model.
- $clip\left(\frac{\pi_\theta(y|x)}{\pi^{SFT}(y|x)}, 1-\epsilon, 1+\epsilon\right)$: This is the clipping term, where ϵ is a small constant (typically set to 0.1). The clipping function ensures that the ratio $\frac{\pi_\theta(y|x)}{\pi^{SFT}(y|x)}$ does not deviate too far from 1. By clipping the ratio within the range $1-\epsilon$ to $1+\epsilon$, we prevent overly large updates to the model's policy, which helps stabilize training. If the ratio exceeds this range, it is "clipped" to stay within the bounds.
- $\min\left(\frac{\pi_\theta(y|x)}{\pi^{SFT}(y|x)}r(y),\ clip(\dots)r(y)\right)$: Minimization of these two terms ensures that the model updates its policy conservatively. By taking the minimum between the unmodified probability ratio and the clipped ratio, the model avoids making overly aggressive changes that could destabilize training.
- ϵ: This parameter is a small positive value that defines the clipping range for the policy ratio. It prevents the ratio of probabilities from deviating too much, helping to keep the model's policy updates within a stable range.

While PPO is an effective approach to optimize the model's policy, it is heavily dependent on the accuracy of the reward model. The reward model acts as a proxy for human judgment and assigns rewards based on how well the model's predictions align with the correct legal outcomes. However, the reward model may struggle to fully capture nuanced human preferences and legal reasoning, which introduces some limitations to our approach.

By minimizing the PPO loss function, we seek to align the model's predictions with desired legal outcomes, ensuring that the model not only improves its accuracy in predicting legal judgments but also provides more interpretable explanations. This method forms a critical part of our framework, aimed at enhancing the effectiveness of AI-assisted legal decision-making.

Task-General Inference Framework

After PPO training, we perform inference on both tasks. For CJPE, given a legal case, the model outputs the predicted judgment followed by an explanation. The inputs are structured prompts from the PredEx dataset. For Summarization, the model is prompted with the full judgment document and expected to generate a concise headnote-style summary.

By comparing the performance of the SFT and PPO-aligned models, we assess the impact of reinforcement learning on output quality across two structurally different legal tasks.

Experimental Setup and Hyperparameters

For training our Reinforcement Learning-based Legal Reasoning (`ReGal`) framework, we utilized Vast.ai[1], a cloud GPU rental provider, to take advantage of scalable and efficient computational resources. The training was conducted on an NVIDIA A100 80GB GPU, which offered the necessary computational power to handle the large dataset and complex model architecture. The total cost of the GPU rental amounted to approximately $100, making it a cost-effective solution for training large-scale models with reinforcement learning.

The key hyperparameters used in our setup include a learning rate of 1.41e-5, and the training proceeded with a maximum of 1 PPO epoch. The batch size during training was 4, while the mini-batch size was 2, allowing for more efficient gradient updates. The output length was constrained between a minimum of 100 tokens and a maximum of 500 tokens, ensuring that the model generated sufficiently detailed explanations. Additionally, we set the clipping parameter (ϵ) to 0.1 to stabilize the PPO optimization process, and the maximum number of new tokens generated during inference was limited to 500.

To maximize GPU memory utilization, we employed mixed-precision training using GradScaler, allowing us to process larger batches while maintaining computational efficiency. This setup provided the necessary infrastructure and tuning to effectively train the `ReGal` framework for the complex task of legal judgment prediction and explanation, ensuring the model produced accurate and interpretable results.

[1] https://vast.ai/

Evaluation Metrics

We evaluate our ReGal framework across two key tasks: (1) Court Judgment Prediction and Explanation (CJPE) and (2) Legal Document Summarization. To provide a comprehensive performance assessment, we utilize a blend of lexical and semantic metrics tailored to each task.

Lexical-Based Evaluation: For both the explanation and summarization tasks, we use standard metrics that measure n-gram overlaps with reference texts. These include ROUGE-1/2/L (Lin 2004) for recall-based overlap, BLEU (Papineni et al. 2002) for precision-based evaluation, and METEOR (Banerjee and Lavie 2005), which accounts for synonyms and stemming. These metrics quantify the textual fidelity of generated outputs against expert-written references.

Semantic-Based Evaluation: To capture meaning beyond surface-level overlaps, we employ BERTScore (Zhang et al. 2020) and BLANC (Vasilyev, Dharnidharka, and Bohannon 2020), which assess the semantic similarity and contextual relevance of generated explanations or summaries. These metrics are especially important in the legal domain, where paraphrasing and legal nuance must still preserve the core meaning.

Results and Analysis

The results of our experiments indicate that the Proximal Policy Optimization (PPO) model, referred to as `ReGal`, did not perform as well as expected in the Indian legal judgment prediction and explanation tasks. As shown in Table 3, our `ReGal` framework achieved significantly lower scores across various lexical and semantic evaluation metrics when compared to both supervised fine-tuned models like LLaMA-2 SFT and commercial models such as GPT-3.5 Turbo. For instance, on the PredEx dataset, the `ReGal` model recorded a ROUGE-1 score of 0.19, ROUGE-2 score of 0.04, and BLEU score of 0.01, which were considerably lower than those of LLaMA-2 SFT (ROUGE-1: 0.50, BLEU: 0.25) and GPT-3.5 Turbo (ROUGE-1: 0.54).

This trend continues in the ILDC Expert dataset, where the PPO-based `ReGal` lags significantly behind. While these results underscore the dominance of supervised fine-tuned models and large proprietary LLMs in legal judgment tasks, they also reveal challenges in applying reinforcement learning methods like PPO directly on complex legal domains.

To further explore the generalizability and effectiveness of our PPO-based `ReGal` architecture, we extended our evaluation beyond judgment prediction to legal summarization, specifically on the In-Abs summarization dataset. The inference results for all training paradigms, including PPO, are reported in Table 4. Even in summarization, `ReGal` performed suboptimally compared to Vanilla and SFT inference baselines. For instance, PPO inference achieved ROUGE-1 of 0.41, whereas the vanilla approach attained 0.47 and SFT reached 0.44. This pattern holds across other metrics as well, such as METEOR and BERTScore, further indicating the limitations of PPO for this domain.

These cross-task results show that while the PPO framework offers potential for aligning models with specific reward signals, it underperforms in legal NLP tasks where output quality is deeply tied to contextual, interpretative, and domain-specific factors.

Possible Reasons for Underperformance

Several factors may have contributed to the underwhelming performance of our PPO-based `ReGal` model across both judgment prediction and summarization tasks:

1. **Objective Mismatch:** The SFT model π^{SFT} used as the starting point for PPO was not fully optimized for legal reasoning. This mismatch between the PPO objective and the base model's latent distribution likely impaired downstream optimization.
2. **Reward Model Limitations:** Our reward model, based on InLegalBERT, may not fully capture the fine-grained reasoning and interpretative nuances of Indian legal texts. This misalignment in reward scoring hinders the PPO's capacity to steer the generation toward legally coherent outputs.
3. **Legal Complexity:** Legal documents, particularly in the Indian judiciary, are long, intricate, and rich in semantic references. This poses additional challenges for autoregressive generation, especially under RL-based training where output smoothness and factuality are difficult to balance.
4. **Training Data Constraints:** Although the PredEx dataset is fairly large, it may not offer sufficient diversity in legal reasoning patterns. A broader dataset spanning multiple jurisdictions or tasks could better support PPO-based tuning.
5. **Reward Model Dependence:** PPO's reliance on the reward model, and lack of human-in-the-loop supervision, likely prevents the system from learning subtle legal distinctions and rhetorical structures critical to both prediction and summarization.
6. **Hyperparameter Selection:** Suboptimal tuning of learning rate, batch size, and KL penalty coefficients may have impacted stability and generalization of the PPO model.
7. **Model Size and Architecture:** While LLaMA-2-7B was chosen for fair comparison with prior literature, alternative architectures or larger models may be more suited for PPO fine-tuning in legal settings.
8. **Domain Pretraining Gap:** Despite fine-tuning on Indian legal datasets, the base model may lack deep domain adaptation compared to GPT-3.5 Turbo, which benefits from extensive pretraining and reinforcement with human feedback across multiple domains.

Although our `ReGal` framework does not outperform state-of-the-art supervised or proprietary models, it serves as an important exploration into reinforcement learning methods for legal NLP. Our work highlights the challenges of aligning large language models with legal reasoning objectives using PPO and reward models, particularly in the absence of high-quality human feedback and robust legal anno-

Models	Lexical-Based Metrics					Semantic Metrics	
	R1	R2	RL	BLEU	METEOR	BERTScore	BLANC
Prediction with Explanation on PredEx							
Gemini Pro	0.31	0.24	0.26	0.08	0.19	0.63	0.17
LLaMA-2	0.32	0.19	0.21	0.06	0.18	0.62	0.15
LLaMA-2 SFT	**0.50**	**0.43**	**0.44**	**0.25**	**0.36**	**0.69**	**0.28**
ReGal (Ours)	0.19	0.04	0.12	0.01	0.10	0.50	0.02
Prediction with Explanation on ILDC Expert							
GPT-3.5 Turbo	**0.54**	**0.43**	**0.45**	0.28	0.47	**0.73**	0.34
LLaMA-2	0.45	0.25	0.30	0.15	0.34	0.65	0.22
LLaMA-2 SFT	0.49	0.38	0.40	**0.29**	**0.51**	0.69	**0.36**
ReGal (Ours)	0.25	0.05	0.16	0.01	0.16	0.50	0.03

Table 3: Performance comparison of various models for the Prediction with Explanation task on PredEx and ILDC datasets. Best scores per row section are bolded.

Methods	R1	R2	RL	BLEU	METEOR	BERTScore	BLANC
Inference on PredEx Dataset							
Vanilla Inference	0.39	0.17	0.22	0.07	0.23	0.83	0.15
SFT Inference	**0.42**	**0.25**	**0.27**	**0.12**	**0.27**	**0.84**	**0.19**
DPO Inference	0.38	0.17	0.23	0.08	0.25	0.83	0.17
PPO Inference	0.30	0.14	0.17	0.05	0.19	0.83	0.13
Inference on In-Abs Summarization							
Vanilla Inference	**0.47**	**0.29**	**0.28**	**0.15**	**0.34**	**0.04**	**0.18**
SFT Inference	0.44	0.24	0.24	0.12	**0.34**	0.02	0.13
DPO Inference	0.44	0.24	0.24	0.12	**0.34**	0.02	0.13
PPO Inference	0.41	0.21	0.22	0.10	0.31	0.03	0.12

Table 4: Comparison of inference strategies (Vanilla, SFT, DPO, PPO) on both the PredEx and In-Abs-Summarization datasets.

tation. Future work will address these limitations by incorporating more expressive reward signals, leveraging human-in-the-loop feedback, exploring domain-specific pretraining strategies, and optimizing PPO with adaptive RL techniques. Through these improvements, we aim to close the performance gap and realize the potential of RL-based methods in Indian legal AI.

Ablation Study

To better understand the sensitivity and robustness of the ReGal framework, we conducted an ablation study across two dimensions: the choice of base model and the configuration of the reward model. These experiments were aimed at evaluating how architectural choices and domain alignment impact performance on complex Indian legal NLP tasks, specifically judgment prediction with explanation and summarization.

Base Model Variants

We explored the effects of using smaller and less specialized base models in the PPO training pipeline. First, we replaced the LLaMA-2-7B model with Phi-3 Mini, which is considerably smaller and more memory-efficient. While this offered computational advantages, it severely limited the model's capacity to handle long and intricate legal texts. On both the PredEx and In-Abs datasets, the performance dropped substantially across all evaluation metrics. The model failed to generate coherent or factually grounded legal reasoning, confirming that small models lack the necessary representation power for such nuanced tasks.

We also experimented with using the pretrained LLaMA-2-7B model without supervised fine-tuning. Despite LLaMA-2's strong general language capabilities, the pretrained version alone was insufficient to produce quality outputs in the legal domain. The absence of legal-domain-specific tuning resulted in degraded lexical and semantic evaluation scores. These findings validated our decision to use LLaMA-2-7B SFT, which was also used in prior works for fine-tuning on legal datasets. Its inclusion allowed us to directly compare and better assess the relative effectiveness of PPO-based learning within our ReGaL framework.

Reward Model Variants

In parallel, we analyzed the impact of different reward models. The default reward model in our framework was a legal classifier fine-tuned on the PredEx dataset, which captures domain-specific features critical for scoring judgment predictions and explanations. To evaluate generalizability, we replaced this with an InLegalBERT-pretrained model, which, although trained on broader legal corpora, was not

fine-tuned for our specific judgment-explanation task. This substitution resulted in further degradation in PPO performance. The pretrained reward model assigned noisier or misaligned scores, impairing the PPO optimization and leading to incoherent or generic legal outputs. This underscores the importance of aligning the reward model tightly with the task-specific ground truth to effectively guide RL training.

Summary of Findings

Our ablation results highlight two key insights. First, larger and domain-aligned base models such as LLaMA-2-7B SFT are essential for achieving reasonable performance in Indian legal AI tasks. Smaller models or generic pretrained ones lack the representational and contextual capacity required. Second, the reward model must be both task-specific and domain-tuned to offer meaningful feedback to PPO. Using a general legal reward model without task-specific fine-tuning disrupts the learning signal, particularly in complex tasks like factual judgment explanation or summarization. These findings reinforce that PPO's success in legal AI hinges not just on the reinforcement algorithm but also on a strong initialization and a precisely aligned reward function.

Hallucination

A key challenge observed in our `ReGal` framework, particularly when applying PPO-based fine-tuning to the LLaMA-2-7B model, is the generation of hallucinated outputs, statements that are fluent and plausible but factually incorrect or legally unsound. In the context of Indian legal judgment prediction and explanation, hallucination severely undermines the utility and trustworthiness of AI-generated outputs, as even minor factual deviations can result in misleading legal interpretations.

Through qualitative analysis, we observed that hallucination was especially prominent in two scenarios: (1) when the input facts were sparse or ambiguously phrased, and (2) when the PPO model over-optimized for reward patterns learned from a limited or imperfect reward model, resulting in outputs that mimicked style but not substance. In many such instances, the model hallucinated legal principles, fabricated precedent citations, or claims or facts not present in the original input.

To demonstrate these failures, we provide illustrative examples in Supplementary Material, comparing ground truth explanations with those generated by the `ReGal` model. In one example, the model incorrectly claims that the appellant's right to privacy was upheld under Article 21 based on fabricated reasoning that was not part of the original judgment. These hallucinations point to instability in PPO optimization when coupled with sparse or weak reward signals, especially in long-form generation tasks where legal correctness must be tightly aligned with input facts. The issue is further exacerbated in zero-shot settings or in datasets like In-Abs summarization, where factual fidelity is paramount. Our findings suggest that RLHF methods such as PPO must be augmented with stronger factuality constraints, hallucination-aware reward models, or human-in-the-loop feedback when applied to high-stakes legal AI applications.

Conclusions and Future Work

This study introduced the `ReGal` framework, combining instruction tuning and reinforcement learning through PPO for the tasks of legal judgment prediction and explanation. While the approach aimed to align generation with legal reasoning via AI feedback, the results across both the PredEx and In-Abs summarization tasks show that our PPO-based method underperforms compared to strong supervised fine-tuned models and commercial LLMs like GPT-3.5. As seen in the result tables, scores were consistently lower, and hallucinations were more frequent, limiting the reliability of our framework in real-world legal settings.

The ablation studies confirm that the choice of both the base model and reward model plays a critical role, with smaller or unadapted models failing to handle the complexity of Indian legal texts. Moreover, the hallucination analysis revealed that PPO outputs often deviate from factual case elements, raising concerns about their practical deployment. These findings point to key areas for improvement.

Future efforts will focus on building better-aligned reward models, leveraging domain-adaptive pretraining, and integrating human feedback to mitigate hallucinations and improve explanation quality. With more robust training signals, refined hyperparameters, and advanced model architectures, the `ReGal` framework can evolve into a more competitive and trustworthy tool for legal AI applications.

Ethics Statement

Our study investigates the use of instruction-tuned and reinforcement learning-based models for judgment prediction and summarization in the Indian legal context. All experiments were conducted using publicly available datasets, including the PredEx dataset for judgment prediction and the In-Abs dataset for legal document summarization. The PredEx dataset was annotated by qualified legal scholars, and all data sources were selected for their ethical transparency and accessibility.

We acknowledge the heightened ethical stakes involved in applying AI to legal domains, especially where automated outputs may influence interpretation, legal awareness, or access to justice. While our work does not directly deploy models in real-world legal settings, we emphasize that outputs from such models should not be considered substitutes for legal advice. Moreover, any potential deployment must be accompanied by strict validation, legal oversight, and human-in-the-loop feedback.

No private or sensitive data was used, and no human subjects were involved in the model training or evaluation process. We ensured that all experiments complied with ethical research practices, including transparency, reproducibility, and proper citation of prior work. Our code and models are made publicly accessible for academic reproducibility and further validation.

Finally, we acknowledge the risk of model bias and hallucination, as discussed in the limitations. These raise important concerns about factual reliability and fairness in downstream applications. Addressing such challenges responsibly will remain central to future iterations of this research.

References

Aletras, N.; Tsarapatsanis, D.; Preoţiuc-Pietro, D.; and Lampos, V. 2016. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ computer science*, 2: e93.

Aumiller, D.; Chouhan, A.; and Gertz, M. 2022. EUR-Lex-Sum: A Multi- and Cross-lingual Dataset for Long-form Summarization in the Legal Domain. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7626–7639. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Goldstein, J.; Lavie, A.; Lin, C.-Y.; and Voss, C., eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

Bhattacharjee, A.; Hasan, T.; Ahmad, W. U.; Li, Y.-F.; Kang, Y.-B.; and Shahriyar, R. 2021. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. *arXiv preprint arXiv:2112.08804*.

Bhattacharya, P.; Hiware, K.; Rajgaria, S.; Pochhi, N.; Ghosh, K.; and Ghosh, S. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *European Conference on Information Retrieval*, 413–428. Springer.

Bhattacharya, P.; Poddar, S.; Rudra, K.; Ghosh, K.; and Ghosh, S. 2021. Incorporating domain knowledge for extractive summarization of legal case documents. In *Proc. International Conference on Artificial Intelligence and Law (ICAIL)*, 22–31.

Chalkidis, I.; Androutsopoulos, I.; and Aletras, N. 2019. Neural legal judgment prediction in English. *Association for Computational Linguistics (ACL)*.

Chang, Y.; Lo, K.; Goyal, T.; and Iyyer, M. 2023. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. *arXiv preprint arXiv:2310.00785*.

Datta, D.; Soni, S.; Mukherjee, R.; and Ghosh, S. 2023. MILDSum: A Novel Benchmark Dataset for Multilingual Summarization of Indian Legal Case Judgments. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5291–5302. Singapore: Association for Computational Linguistics.

Deroy, A.; Ghosh, K.; and Ghosh, S. 2023. How Ready are Pre-trained Abstractive Models and LLMs for Legal Case Judgement Summarization?

Dong, J.; and Lin, P. 2024. A REINFORCEMENT LEARNING FRAMEWORK FOR ACCURATE AND CONTEXT-AWARE LEGAL DOCUMENT SUMMARIZATION.

Feijo, D. d. V.; and Moreira, V. P. 2023. Improving abstractive summarization of legal rulings through textual entailment. *Artificial intelligence and law*, 31(1): 91–113.

Feng, Y.; Li, C.; Ge, J.; Luo, B.; and Ng, V. 2021. Recommending statutes: A portable method based on neural networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2): 1–22.

Ganguly, D.; Conrad, J. G.; Ghosh, K.; Ghosh, S.; Goyal, P.; Bhattacharya, P.; Nigam, S. K.; and Paul, S. 2023. Legal IR and NLP: the history, challenges, and state-of-the-art. In *European Conference on Information Retrieval*, 331–340. Springer.

Glaese, A.; McAleese, N.; Trbacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; Rauh, M.; Weidinger, L.; Chadwick, M.; Thacker, P.; et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Hwang, W.; Lee, D.; Cho, K.; Lee, H.; and Seo, M. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35: 32537–32551.

Joshi, A.; Paul, S.; Sharma, A.; Goyal, P.; Ghosh, S.; and Modi, A. 2024. IL-TUR: Benchmark for Indian Legal Text Understanding and Reasoning. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11460–11499. Bangkok, Thailand: Association for Computational Linguistics.

Kapoor, A.; Dhawan, M.; Goel, A.; T H, A.; Bhatnagar, A.; Agrawal, V.; Agrawal, A.; Bhattacharya, A.; Kumaraguru, P.; and Modi, A. 2022. HLDC: Hindi Legal Documents Corpus. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 3521–3536. Dublin, Ireland: Association for Computational Linguistics.

Katz, D. M.; Bommarito, M. J.; and Blackman, J. 2017. A general approach for predicting the behavior of the Supreme Court of the United States. *PloS one*, 12(4): e0174698.

Ladhak, F.; Durmus, E.; Cardie, C.; and McKeown, K. 2020. WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4034–4048. Online: Association for Computational Linguistics.

Lee, H.; Phatale, S.; Mansoor, H.; Mesnard, T.; Ferret, J.; Lu, K.; Bishop, C.; Hall, E.; Carbune, V.; Rastogi,

A.; and Prakash, S. 2023. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. arXiv:2309.00267.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Malik, V.; Sanjay, R.; Guha, S. K.; Hazarika, A.; Nigam, S.; Bhattacharya, A.; and Modi, A. 2021a. Semantic segmentation of legal documents via rhetorical roles. *arXiv preprint arXiv:2112.01836*.

Malik, V.; Sanjay, R.; Nigam, S. K.; Ghosh, K.; Guha, S. K.; Bhattacharya, A.; and Modi, A. 2021b. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 4046–4062. Association for Computational Linguistics.

Malik, V.; Sanjay, R.; Nigam, S. K.; Ghosh, K.; Guha, S. K.; Bhattacharya, A.; and Modi, A. 2021c. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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)*, 4046–4062. Online: Association for Computational Linguistics.

Masala, M.; Iacob, R. C. A.; Uban, A. S.; Cidota, M.; Velicu, H.; Rebedea, T.; and Popescu, M. 2021. jurBERT: A Romanian BERT model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, 86–94.

Mishra, S.; Khashabi, D.; Baral, C.; and Hajishirzi, H. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. arXiv:2104.08773.

Nguyen, D.-H.; Nguyen, B.-S.; Nghiem, N. V. D.; Le, D. T.; Khatun, M. A.; Nguyen, M.-T.; and Le, H. 2021. Robust deep reinforcement learning for extractive legal summarization. In *International Conference on Neural Information Processing*, 597–604. Springer.

Nigam, S. K.; Balaramamahanthi, D. P.; Mishra, S.; Shallum, N.; Ghosh, K.; and Bhattacharya, A. 2025. NyayaAnumana and INLegalLlama: The Largest Indian Legal Judgment Prediction Dataset and Specialized Language Model for Enhanced Decision Analysis. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 11135–11160. Abu Dhabi, UAE: Association for Computational Linguistics.

Nigam, S. K.; and Deroy, A. 2023. Fact-based Court Judgment Prediction. *arXiv preprint arXiv:2311.13350*.

Nigam, S. K.; Deroy, A.; Shallum, N.; Mishra, A. K.; Roy, A.; Mishra, S. K.; Bhattacharya, A.; Ghosh, S.; and Ghosh, K. 2023. Nonet at SemEval-2023 Task 6: Methodologies for Legal Evaluation. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 1293–1303.

Nigam, S. K.; Sharma, A.; Khanna, D.; Shallum, N.; Ghosh, K.; and Bhattacharya, A. 2024. Legal Judgment Reimagined: PredEx and the Rise of Intelligent AI Interpretation in Indian Courts. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 4296–4315. Bangkok, Thailand: Association for Computational Linguistics.

Niklaus, J.; Chalkidis, I.; and Stürmer, M. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. *arXiv preprint arXiv:2110.00806*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Patil, P.; and Patankar, A. 2025. Reinforcement Learning for Optimizing Legal Summarization Models. *Authorea Preprints*.

Polsley, S.; Jhunjhunwala, P.; and Huang, R. 2016. CaseSummarizer: A System for Automated Summarization of Legal Texts. In Watanabe, H., ed., *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, 258–262. Osaka, Japan: The COLING 2016 Organizing Committee.

Schneider, F.; and Turchi, M. 2023. Team zoom@ automin 2023: Utilizing topic segmentation and llm data augmentation for long-form meeting summarization. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, 101–107.

Shukla, A.; Bhattacharya, P.; Poddar, S.; Mukherjee, R.; Ghosh, K.; Goyal, P.; and Ghosh, S. 2022a. Legal case document summarization: Extractive and abstractive methods and their evaluation. *arXiv preprint arXiv:2210.07544*.

Shukla, A.; Bhattacharya, P.; Poddar, S.; Mukherjee, R.; Ghosh, K.; Goyal, P.; and Ghosh, S. 2022b. Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation. In He, Y.; Ji, H.; Li, S.; Liu, Y.; and Chang, C.-H., eds., *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1048–1064. Online only: Association for Computational Linguistics.

Shukla, B.; Gupta, S.; Yadav, A. K.; and Yadav, D. 2022c. Text summarization of legal documents using reinforcement learning: a study. In *Intelligent Sustainable Systems: Proceedings of ICISS 2022*, 403–414. Springer.

Urlana, A.; Bhatt, S. M.; Surange, N.; and Shrivastava, M. 2023. Indian language summarization using pretrained sequence-to-sequence models. *arXiv preprint arXiv:2303.14461*.

Vasilyev, O. V.; Dharnidharka, V.; and Bohannon, J. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. *CoRR*, abs/2002.09836.

Vats, S.; Zope, A.; De, S.; Sharma, A.; Bhattacharya, U.; Nigam, S. K.; Guha, S.; Rudra, K.; and Ghosh, K. 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 Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12451–12474. Singapore: Association for Computational Linguistics.

Wang, J.; Meng, F.; Lu, Z.; Zheng, D.; Li, Z.; Qu, J.; and Zhou, J. 2022a. ClidSum: A Benchmark Dataset for Cross-Lingual Dialogue Summarization. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7716–7729. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Wang, X.; and Wu, Y. C. 2024. Empowering legal justice with AI: A reinforcement learning SAC-VAE framework for advanced legal text summarization. *PloS one*, 19(10): e0312623.

Wang, Y.; Mishra, S.; Alipoormolabashi, P.; Kordi, Y.; Mirzaei, A.; Arunkumar, A.; Ashok, A.; Dhanasekaran, A. S.; Naik, A.; Stap, D.; Pathak, E.; Karamanolakis, G.; Lai, H. G.; Purohit, I.; Mondal, I.; Anderson, J.; Kuznia, K.; Doshi, K.; Patel, M.; Pal, K. K.; Moradshahi, M.; Parmar, M.; Purohit, M.; Varshney, N.; Kaza, P. R.; Verma, P.; Puri, R. S.; Karia, R.; Sampat, S. K.; Doshi, S.; Mishra, S.; Reddy, S.; Patro, S.; Dixit, T.; Shen, X.; Baral, C.; Choi, Y.; Smith, N. A.; Hajishirzi, H.; and Khashabi, D. 2022b. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. arXiv:2204.07705.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022. Finetuned Language Models Are Zero-Shot Learners. arXiv:2109.01652.

Xiao, C.; Zhong, H.; Guo, Z.; Tu, C.; Liu, Z.; Sun, M.; Feng, Y.; Han, X.; Hu, Z.; Wang, H.; et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Zhang, S.; Dong, L.; Li, X.; Zhang, S.; Sun, X.; Wang, S.; Li, J.; Hu, R.; Zhang, T.; Wu, F.; and Wang, G. 2023a. Instruction Tuning for Large Language Models: A Survey. arXiv:2308.10792.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 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.

Zhang, T.; Ladhak, F.; Durmus, E.; Liang, P.; McKeown, K.; and Hashimoto, T. B. 2023b. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; and Chang, K.-W. 2018. Learning Gender-Neutral Word Embeddings. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4847–4853. Brussels, Belgium: Association for Computational Linguistics.

Zhou, W.; Jiang, Y. E.; Wilcox, E.; Cotterell, R.; and Sachan, M. 2023. Controlled Text Generation with Natural Language Instructions. arXiv:2304.14293.

Zhu, K.; Guo, R.; Hu, W.; Li, Z.; and Li, Y. 2020. Legal judgment prediction based on multiclass information fusion. *Complexity*, 2020(1): 3089189.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593.