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

Link to paper and code

Motivation

- Indian legal documents are long, complex, and high-stakes, making early judgment prediction and summarization challenging.
- Existing Indian legal AI systems rely primarily on supervised fine-tuning and lack mechanisms for iterative refinement using feedback.
- Reinforcement Learning (RL), despite success in other domains, remains underexplored for Indian legal reasoning tasks.
- This work investigates whether PPO-based RL can improve alignment, interpretability, and reasoning fidelity in legal AI.

Task Description

- We evaluate **ReGal** on two core Indian legal NLP tasks:
- Task 1: Court Judgment Prediction and Explanation (CJPE)
  - **Prediction:** Given a Supreme Court judgment, predict whether the appeal is *accepted (1)* or *rejected (0)*.
  - **Explanation:** Generate a natural language rationale grounded in the case text.
- Task 2: Legal Judgment Summarization
  - Generate abstractive summaries capturing background, legal issues, arguments, and verdict.
- Both tasks emphasize factual consistency, interpretability, and domain alignment.

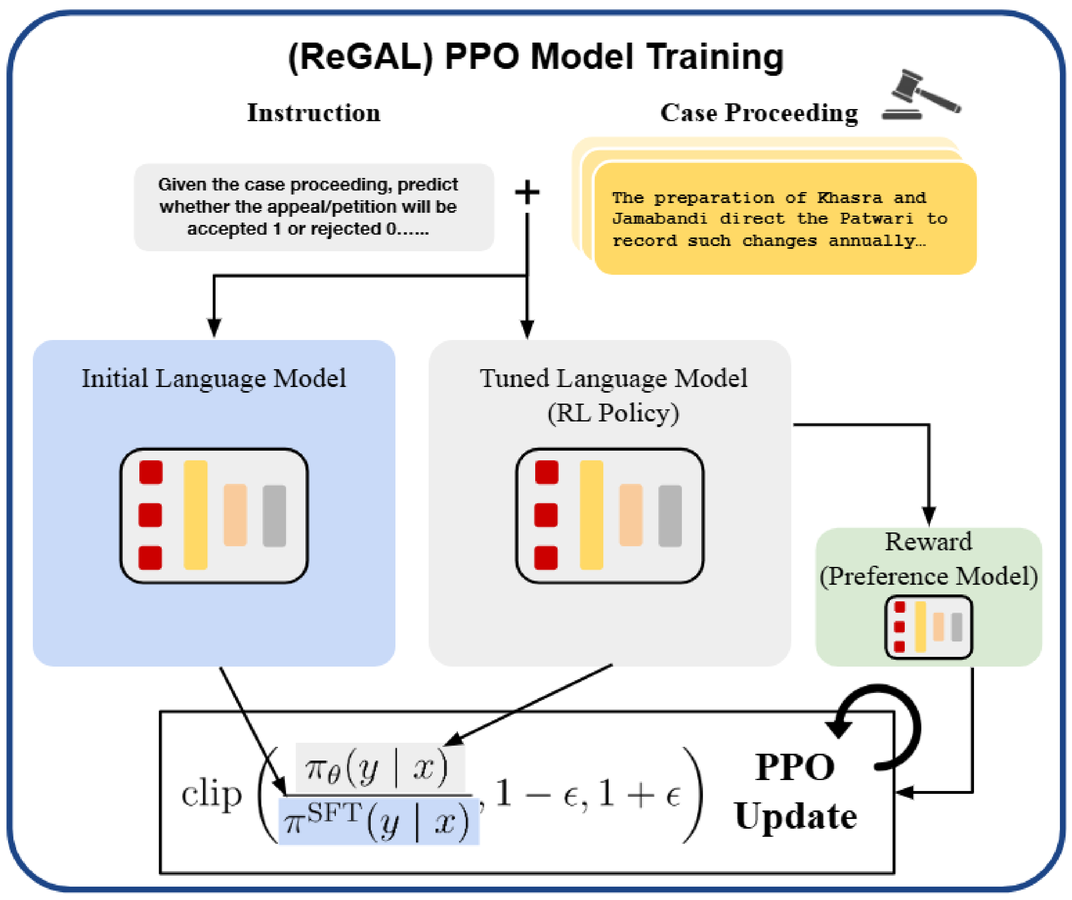

Figure 1. Overview of the ReGal PPO model training process.

Dataset Overview

- PredEx (CJPE Dataset):
  - 15,222 Supreme Court judgments.
  - Binary verdict labels with expert-written explanations.
  - Average document length: ∼4.5K tokens.
- In-Abs (Summarization Dataset):
  - 7,130 Supreme Court judgments with expert headnotes.
  - Abstractive summaries, English language.
  - Compression ratio ≈ 0.24.
- Both datasets enable evaluation of RL across prediction, explanation, and summarization.

Methodology: ReGal Framework with PPO Optimization

- **ReGal** combines supervised instruction tuning with PPO-based reinforcement learning from AI feedback (RLAIF).
- **Base Model:** LLaMA-2-7B, chosen for comparability with prior Indian legal NLP work.
- **Stage 1 (SFT):** Model is fine-tuned on task data for judgment prediction + explanation (PredEx) and summarization (In-Abs).
- **Stage 2 (PPO):** The SFT model is optimized using PPO with task-specific reward models.
- Rewards:
  - CJPE: binary reward based on verdict correctness (InLegalBERT).
  - Summarization: scalar reward based on ROUGE-style overlap and coherence.
- PPO constrains policy updates via clipped probability ratios ($\epsilon = 0.1$) to limit deviation from the SFT policy.

Reward Models

- CJPE Reward Model:
  - InLegalBERT classifier.
  - Binary reward: 1 for correct verdict, 0 otherwise.
- Summarization Reward Model:
  - ROUGE-based overlap and shallow semantic similarity.
- Rewards are AI-generated (RLAIF), simulating human feedback.

Results and Analysis

- ReGal underperforms compared to SFT and proprietary models.
- On PredEx:
  - ROUGE-1 = 0.19 (ReGal) vs 0.50 (LLaMA-2 SFT).
- On In-Abs summarization:
  - PPO ROUGE-1 = 0.41 vs 0.47 (Vanilla inference).
- Indicates difficulty of applying PPO directly to complex legal text.

Models	Lexical Metrics					Semantic Metrics	
	R1	R2	RL	BLEU	METEOR	BERTScore	BLANC
PredEx (Prediction + Explanation)							
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
ILDC Expert (Prediction + Explanation)							
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 1. Performance comparison of various models for the Prediction with Explanation task on PredEx and ILDC datasets.

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

Table 2. Comparison of inference strategies (Vanilla, SFT, DPO, PPO) on both the PredEx and In-Abs-Summarization datasets.

Ablation Study

- Smaller models (e.g., Phi-3 Mini) fail to handle long legal documents.
- Pretrained LLaMA-2 without legal SFT performs poorly.
- Reward models not aligned with task degrade PPO learning.
- Highlights dependence of PPO on strong base and reward models.

Hallucination Analysis

- PPO training increases hallucinated legal claims.
- Model fabricates precedents and legal principles not present in input.
- Hallucinations arise from weak reward signals and sparse supervision.
- Reinforces need for stronger factuality constraints and human feedback.

Contributions and Impact

- First PPO-based RL study for Indian legal judgment prediction and summarization.
- Provides empirical and qualitative analysis of RL failures in legal NLP.
- Establishes lessons for future RLHF/RLAIF-based legal systems.
- Code and data released for reproducibility.

Limitations and Future Work

- PPO underperforms compared to supervised and proprietary models.
- Reward models insufficient for nuanced legal reasoning.
- Future work: human-in-the-loop RLHF, better reward modeling, domain-adaptive pretraining.

References

[1] Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *ACL*, 2021.

[2] Shubham Nigam, Anurag Sharma, Danush Khanna, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. Legal judgment reimagined: PredEx and the rise of intelligent AI interpretation in Indian courts. In *ACL*, 2024.

[3] Shubham Kumar Nigam, Deepak Patnaik Balaramamahanthi, Shivam Mishra, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. NyayaAnumana and INLegalLlama: The largest Indian legal judgment prediction dataset and specialized language model for enhanced decision analysis. In *COLING*, 2025.

[4] Shaurya Vats, Atharva Zope, Somsubhra De, Anurag Sharma, Upal Bhattacharya, Shubham Kumar Nigam, Shouvik Guha, Koustav Rudra, and Kripabandhu Ghosh. LLMs – the good, the bad or the indispensable?: A use case on legal statute prediction and legal judgment prediction on Indian court cases. In *EMNLP*, 2023.

Acknowledgements: We express our sincere gratitude to BharatGen, India, for providing essential computational resources and hardware support that made this work possible. We also thank our collaborators, student research assistants, and the legal experts from partnering law colleges for their valuable contributions to annotation, evaluation, and the overall refinement of the project. Their collective effort and insights greatly strengthened this research.

shubhamkumarnigam@gmail.com tanujtyagiofficial@gmail.com siddharth8shukla8@gmail.com adityaguru20@gmail.com bdeepakpatnaik2002@gmail.com danush.s.khanna@gmail.com noelshallum@gmail.com kripaghosh@iiserkol.ac arnabb@cse.iitk.ac.in