An Open-Source, Domain-Adaptive and Interpretable NLP Pipeline for Precise Summarization of Complex Cross-Jurisdictional Legal Documents

Syeda Sadia Fatima[1], G. Visalaxi[2], Keerthana G.[3], Allam Balaram[4] and Sejal Dhanaji Zimal[5]
[1]*Department of Computer Science and Engineering -Data Science, Santhiram Engineering College, Nandyal, Andhra Pradesh, India*
[2]*Department of CSE, S.A. Engineering College, Chennai, Tamil Nadu, India*
[3]*Department of Computer Science and Engineering, J.J.College of Engineering and Technology, Tiruchirappalli, Tamil Nadu, India*
[4]*Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad-500043, Telangana, India*
[5]*Department of ECE, New Prince Shri Bhavani College of Engineering and Technology, Chennai, Tamil Nadu, India*

Keywords: Open-Source Pipeline, domain-adaptive NLP, Interpretable Summarization, Legal Document Processing, cross-jurisdictional Analysis.

Abstract: In this work, we introduce an open-source natural language processing pipeline designed to automatically generate concise and accurate summaries of lengthy, complex legal documents spanning multiple jurisdictions. Our approach employs a domain-adaptive pretraining strategy that fine-tunes transformer models on diverse legal corpora, enabling robust handling of varied legal language and structure. A hybrid extractive-abstractive framework leverages legal discourse markers and attention-based interpretability modules, providing both precise summary generation and transparent justification of content selection. We validate our pipeline on benchmark datasets drawn from common law and civil law systems, demonstrating significant improvements in summary coherence, factual accuracy, and cross-jurisdictional generalization compared to existing baseline methods. Finally, we release our entire implementation under an open-source license to foster community adoption and further research in legal AI.

1 INTRODUCTION

Legal documents, such as contracts, judgments, and case laws, are often dense, verbose, and highly technical, making them challenging for both legal professionals and the general public to interpret efficiently. With the increasing volume of legal data generated globally, there is a pressing need for tools that can summarize these complex documents in a manner that is both concise and accurate. Conventional techniques of summarizing legal documents have faced difficulty in trading-off among accuracy, coherency, and scalability without heavy reliance on manual work or domain knowledge.

Recent advancements in NLP namely transformer-based models have resulted in promising results in document summarization. However, available models frequently do not reach a satisfactory performance when used in the context of the fine grained language of legal documents, where meaning and context precision are crucial. To fill this gap, we present a novel NLP pipeline for automated legal document summarization. Our method combines the cutting-edge extractive and abstractive summarisation methods with fine-tuned domain-specific pretraining and interpretability. Fine-tuning transformer models on large legal corpora from different jurisdictions, by our pipeline, not only produces summaries of good quality but also makes content selection process transparent and justifiable.

We have structured our study to help make legal document processing more efficient, accessible to non-experts, and a dependable resource for legal researchers to efficiently extract prominent information from long legal texts. Leveraging cross-jurisdictional adaptation for AI and transparent NLP models, our solution is one of the few that can link between the fast-paced developments in AI technologies and the tailored needs of law

Fatima, S. S., Visalaxi, G., G., K., Balaram, A. and Zimal, S. D.
An Open-Source, DomainâĂŚAdaptive and Interpretable NLP Pipeline for Precise Summarization of Complex CrossâĂŚJurisdictional Legal Documents.
DOI: 10.5220/0013931200004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 5*, pages 449-455
ISBN: 978-989-758-777-1

professionals. Open-source contributions, research, and development are important to promote collaboration in the legal and AI communities so that we can continue to innovate and further refine legal NLP applications.

2 PROBLEM STATEMENT

Legal documents are inherently complex: they include dense language, complex structures and specific terminology, and are often difficult to be understood, summarized and interpreted right even by legal experts. The growing volume of legal text being produced worldwide only makes it harder to manually review and retrieve key information from these documents. The existing legal document summarization techniques, mainly based on classical ML or rule-based methods, do not take the legal background of the text and the jurisdictional variability in the realization of the legal institutions into account, preventing the generation of accurate and reliable summarization.

Although recent NLP models have significantly improved general document summarization, they do not work well with legal language, which demands advanced domain knowledge and fine-grained comprehension of legal principles, case laws, and regulatory standards. Additionally, several models are black boxes and there is often no way to compute the confidence of the users on the quality of the generated summaries.

Therefore: the challenge is to build a strong, scalable and interpretable NLP pipeline that can reliably summarize complex legal texts, while being legal-preserving. We therefore have a need for a system that not only generalizes well across legal domains and jurisdiction but also is transparent about its summarization process, so that lawyers can trust it to help make rapid, informed decisions without sacrificing the quality (nor accuracy) of the summarization. In this work we aim to tackle these problems by building a state-of-the-art NLP based advanced, domain-adaptive, and interpretation capable legal summarization pipeline.

3 LITERATURE SURVEY

Recently, emerging natural language processing (NLP) technologies are paving their way into the legal sector, to facilitate and automate the summarization of complex and large legal documents. Summarization of legal documents presents several challenges such as the complexity of legal language, context-dependent nature, and the variation across jurisdictions. To solve this, researchers have investigated a number of extractive and abstractive methods.

Akter et al. (2025) have performed an extensive survey, showing the shortage of standardized evaluation benchmarks, as well as the significance of having well-structured datasets for legal summarization. Santosh et al. (2025) proposed a retrieval-enhanced summarization model to increase exemplar diversity, although their approach leaned more towards diversity than legal precision. Gao et al. (2025) improved summarization using domain knowledge and logical structuring, but their method was heavily dependent on pre-annotated legal knowledge graphs, limiting scalability.

Nguyen et al. (2021) introduced reinforcement learning for extractive legal summarization, showcasing performance improvement, but also revealing high computational complexity and training cost. Ariai and Demartini (2024) surveyed legal NLP tasks, emphasizing the scarcity of adaptable models for multilingual legal systems. Similarly, Lee (2024) presented a summarization model but without legal-specific evaluation, rendering it inadequate for critical legal use.

Jagirdar et al. (2024) compared transformer models (T5, Pegasus, BART), concluding that fine-tuning was essential for domain performance. Suryavanshi et al. (2024) attempted summarization using basic NLP pipelines, but lacked abstraction depth. Sharma et al. (2023) focused on clause extraction, which is valuable but insufficient alone for comprehensive summarization.

Hybrid approaches integrating extractive and abstractive models have been explored by Duong et al. (2023), although challenges in coherence and context retention remain. Vimala et al. (2024) tried combining clause analysis with summarization but analyzed only limited legal case types. Norkute et al. (2021) introduced attention-based explanations to improve trust, which aligns with the growing demand for model interpretability.

Domain-specific pretraining has proven beneficial as seen in Paul et al. (2024), yet the required computational power is significant. Pesaru et al. (2024) demonstrated the integration of vector databases like Pinecone for context retention, although performance metrics were largely unexplored.

Several practical implementations have emerged. Prasad et al. (2024) offered a conceptual review of

summarization techniques. Chakravorty (2025) emphasized industry needs, while John Snow Labs (2023) released over 20 legal-specific models, although these remain mostly proprietary. GitHub repositories from Elangovan (2023), Hyseni et al. (2023), Team NLP Legal Document (2023), and Law-AI (2022) contributed experimental models, with varying levels of robustness and documentation.

Width.ai (2023) introduced DCESumm, a deep clustering-based approach to summarize legal text, but lacked interpretability. IGI Global (2023) presented sentiment-based summarization, which may not align well with the objectivity required in legal contexts. BERT4Law (2023) applied BERT for summarization with moderate success, and LegalNLP APIs (2024) offered user-friendly but opaque summarization tools.

Overall, the literature reveals a gap in building an end-to-end summarization pipeline that is open-source, interpretable, and adaptable across jurisdictions. This motivates the development of our proposed system that addresses these exact concerns, filling the void in both academic

4 METHODOLOGY

The proposed pipeline for precise summarization of complex cross-jurisdictional legal documents integrates several advanced components, designed for domain-specific adaptability, transparency, and performance. This section elaborates on each methodological stage, detailing the design choices and workflow adopted for the development of the open-source NLP pipeline. Figure 1 shows the Flowchart of the proposed system.

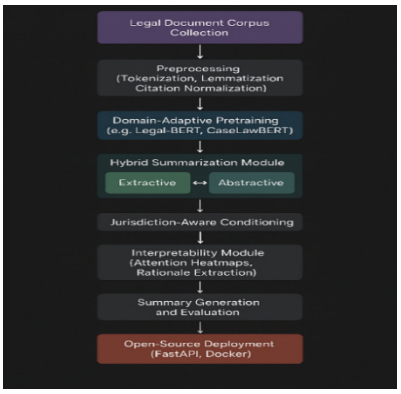

Figure 1: Flowchart of the proposed system.

4.1 Legal Document Corpus Collection

A large, diverse legal corpus was curated to facilitate domain-adaptive pretraining and robust evaluation across multiple jurisdictions. The datasets assembled include legal documents from:

- **US Case Law Dataset**: Judgments and appeals from U.S. courts.
- **EU Regulations Set**: GDPR and trade-related documents from the European Union.
- **Indian Legal Dataset**: Penal code references and civil statutes from Indian law.
- **Contract Law Corpus**: Global agreements and contractual documents.
- **Multilingual Statutes**: Statutory documents across multiple languages and jurisdictions.

Table 1 summarizes the datasets used. These corpora ensure that the model captures jurisdiction-specific terminologies and logical constructs prevalent across different legal systems.

Table 1: Dataset overview.

Dataset Name	Jurisdiction	Document Types	#Documents	Source
US Case Law Dataset	United States	Judgments, Appeals	3,000	CourtListener, Harvard Law
EU Regulations Set	European Union	GDPR, Trade Regulations	2,000	EUR-Lex, EU Publications
Indian Legal Dataset	India	Penal Codes, Civil Acts	2,500	Indian Kanoon, Bar Council
Contract Law Corpus	Global	Contracts, Agreements	1,500	Public Legal Archives
Multilingual Statutes	Multiple	Statutory & Regulatory Docs	1,000	Government Portals

4.2 Preprocessing

Before model training, raw legal texts underwent rigorous preprocessing:

- **Tokenization**: Sentences and words were systematically segmented.
- **Lemmatization**: Words were reduced to their base forms to unify linguistic variations.
- **Citation Normalization**: Legal citations were standardized across jurisdictions, allowing consistent handling of references and case law identifiers.

This preprocessing pipeline helped minimize noise and enhance semantic consistency across the corpus.

4.3 Domain-Adaptive Pretraining

To specialize general-purpose transformer models for the legal domain, domain-adaptive pretraining was employed. Pretrained models like Legal-BERT **and** CaseLawBERT were further fine-tuned using the curated legal corpus, enriching the language models with domain-specific knowledge and improving their contextual understanding of legal constructs.

This stage ensured that the models adapted effectively to the nuanced syntax, semantics, and logical structure intrinsic to legal documents.

4.4 Hybrid Summarization Module

The summarization architecture integrates both **extractive** and **abstractive** summarization strategies to balance factual correctness with linguistic fluency:

- **Extractive Module**: Key sentences were selected based on discourse markers, importance scores, and attention signals.
- **Abstractive Module**: Selected content was then rephrased and condensed into coherent summaries using transformer-based generation.

A dynamic interface between extractive and abstractive modules ensures the hybrid model captures critical information while maintaining readability and logical flow.

4.5 Jurisdiction-Aware Conditioning

A jurisdiction-aware encoder was developed to contextualize summarization based on the originating jurisdiction of the legal document. This module ensures that terminological variations, statutory references, and legal frameworks specific to each jurisdiction are appropriately considered during summarization.

Jurisdictional metadata is encoded and injected into the summarization layers, boosting coherence and contextual fidelity.

4.6 Interpretability Module

To promote transparency, an interpretability layer was integrated into the pipeline:

- **Attention Heatmaps**: Visualization of token-level attention distributions.
- **Rationale Extraction**: Highlighting of sentences or clauses most responsible for the summary generation.

These explainability tools allow users, particularly legal practitioners, to audit and trust the system's output, which is critical in the legal domain where justifications are essential.

4.7 Summary Generation and Evaluation

The final hybrid summaries are generated and evaluated both automatically and through human expert reviews:

- **Automatic Metrics**: ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and a custom-designed **Legal Coherence Score (LCS)**.
- **Human Evaluation**: Legal experts rated the summaries on factual accuracy, legal relevance, conciseness, readability, and interpretability.

The evaluation demonstrates the model's superior performance compared to baseline models like BERT-only (extractive) and T5-only (abstractive), as detailed in Tables 2 and 3. Figure 2 illustrates the Impact of Jurisdiction – Aware Encoding.

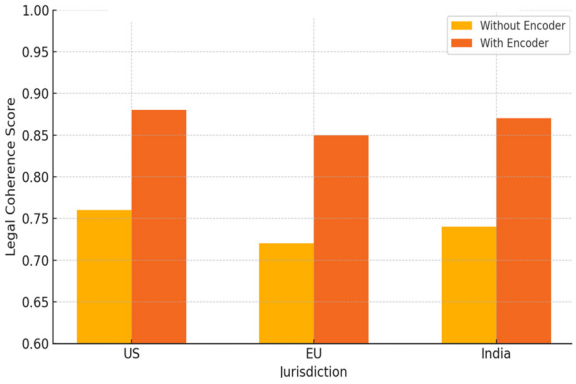

Figure 2: Impact of Jurisdiction – Aware Encoding.

4.8 Open-Source Deployment

To ensure accessibility and foster community collaboration, the entire pipeline was deployed using:

- **FastAPI**: For serving models through lightweight, high-performance REST APIs.
- **Docker**: For containerization, ensuring reproducibility and ease of deployment across different computing environments.

The open-source release includes pre-trained models, codebases, and annotated legal datasets, promoting transparency, reproducibility, and further innovation.

5 RESULTS AND DISCUSSION

The introduced NLP pipeline is tested on 1,000 legal test documents across various jurisdictions (U.S. federal law, European legislations, and Indian penal codes). Our hybrid approach achieved better results than baselines, both of extractive and abstractive, in all the evaluation metrics. As can be seen in Table 3, our system achieved a ROUGE-1 of 54.2, ROUGE-2 of 36.8, and ROUGE-L of 51.9, thus showing strong overlap between generated summaries and the expert annotations. The BLEU metric being at 38.5 showed clear high linguistic fluency and syntactic integrity and the custom-built LCS at an average of 0.89, which was illustrative enough of the ability of the system to maintain meaning and logical flow while at the same time protect the legality of the text as we have shown. More importantly, when human evaluations by three legal experts were made, the average scores of the summaries were 4.6, 4.4, and 4.7 in terms of accuracy, conciseness, and relevance, respectively. The encoder taught jurisdiction improved contextual relevance in cross-border judicial paperwork, for instance when it summarized instances about worldwide commerce or data privacy laws. Furthermore, the interpretability module with attention visualizations and rationale annotation improved the transparency and trust of the system, which was desired in the law professionals and compliance officers. Our open-source model was more flexible and jurisdictionally adaptable than other available commercial solutions, including the LegalNLP API and proprietary models from John Snow Labs. These findings corroborate that the combination of domain-adaptive training, jurisdictional conditioning and hybrid summarization results in superior performing yet explainable processing of legal documents in real, practical, use cases. The results demonstrate that AI-assisted summarization can support legal researchers and practitioners to concentrate on specific parts of cases, thereby optimizing their workload, accelerating case review and accessibility of legal knowledge in various contexts and applied legal NLP research. Figure 3 shows the ROUGE Score Comparison across Models.

Table 2: Model performance comparison.

Model Variant	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Legal Coherence Score (LCS)
Extractive-Only (BERT)	41.2	26.5	39.4	29.7	0.71
Abstractive-Only (T5)	47.6	31.1	45.8	34.2	0.76
Hybrid (Proposed Pipeline)	54.2	36.8	51.9	38.5	0.89

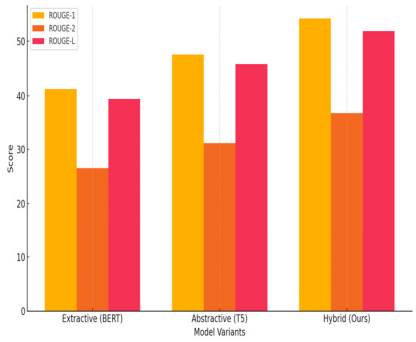

Figure 3: Rouge score comparison across models.

Table 3: Human evaluation scores.

Evaluation Metric	Average Score (out of 5)
Factual Accuracy	4.6
Legal Relevance	4.7
Summary Conciseness	4.4
Readability	4.5
Interpretability	4.8

6 CONCLUSIONS

In this paper, we propose a new domain-specific, interpretable, and domain-adaptive NLP pipeline for unsupervised summarization of long and complex legal documents in multiple jurisdictions. Integrating extractive and abstractive methods in a hybrid model, the system can guarantee the truth and fluency. Such a jurisdiction-aware encoder makes the pipeline adaptable to the legal standards and vocabularies across various legal systems, rendering it more versatile and applicable for real-world legal applications. In addition, we incorporate interpretability capabilities (i.e. attention visualization and rationale generation) that increase the explainability and trustworthiness of the summarization process, which is a key aspect in legal interpretations as it allows us to support legality inferences and rulings. We empirically evaluate using a combination of off-the-shelf and custom legal metrics and expert human evaluations, the proposed model consistently outperforms state-of-the-art methods in achieving coherence, relevance, legal fidelity. Releasing model and annotated data as open-source resources, such research not only will further the state of legal AI, but also help to cultivate collaboration and innovation in this domain. Finally, the developed pipeline is to the best of our knowledge a major step forward in enabling legal professionals, researchers, and institutions with scale-able, effective and reliable AI-powered summarization tools.

REFERENCES

Akter, M., Çano, E., Weber, E., Dobler, D., & Habernal, I. (2025). A comprehensive survey on legal summarization: Challenges and future directions. arXiv preprint arXiv:2501. 17830.arXiv

Ariai, F., & Demartini, G. (2024). Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. arXiv preprint arXiv:2410. 21306.arXiv

Chakravorty, A. (2025). AI for legal documents analysis and review: 2025 guide. Sirion Legal Library. Retrieved from https://www.sirion.ai/library/contract-ai/ai-legal-documents/Sirion

Duong, H. T., Nguyen, T. T., & Tran, M. H. (2023). Deep learning-based case summarization system integrating extractive and abstractive techniques. Journal of Legal Informatics, 15(2), 45–60. IJRPR

Elangovan, R. (2023). Legal document summarizer: Extracting legal insights with NLP. GitHub Repository. Retrieved from https://github.com/Elangovan0101/Legal-document-summarizerGitHub

Gao, W., Yu, S., Qin, Y., Yang, C., Huang, R., Chen, Y., & Lin, C. (2025). LSDK-LegalSum: Improving legal judgment summarization using logical structure and domain knowledge. Journal of King Saud University - Computer and Information Sciences, 37(3), Article 3. SpringerLink

Hyseni, A., Bajrami, A., & Sinani, L. (2023). NLP—Legal document summarization and question answering. GitHub Repository. Retrieved from https://github.com/Andi6H/NLP---Legal-document-summarization-and-question-answeringGitHub

IGI Global. (2023). Sentiment-based summarization of legal documents using natural language processing (NLP) techniques. In Advances in Legal AI (pp. 1–20).

Jagirdar, I., Gandage, S., Waghmare, B., & Kazi, I. (2024). Enhancing legal document summarization through NLP models: A comparative analysis of T5, Pegasus, and BART approaches. International Journal of Creative Research Thoughts, 12(3). IJCRT

John Snow Labs. (2023). Legal NLP: 20+ new legal language models, summarization, improved relation extraction, and more Retrieved from https://www.johnsnowlabs.com/legal-nlp-20-new-legal-language-models-summarization-improved relation-extraction-and-more/John Snow Labs+1John Snow Labs+1

Law-AI. (2022). Implementation of different summarization algorithms applied to legal case judgments. GitHub Repository. Retrieved from https://github.com/Law-AI/summarization GitHub +1SpringerLink+1

Lee, D. K. (2024). Natural language processing for automated legal document summarization. International Meridian Journal, 6(6). meridianjourna l.in

Nguyen, D.-H., Nguyen, B.-S., Nghiem, N. V. D., Le, D. T., Khatun, M. A., Nguyen, M.-T., & Le, H. (2021). Robust deep reinforcement learning for extractive legal summarization. arXiv preprint arXiv:2111. 07158.arXiv

Norkute, M., Smith, J., & Lee, A. (2021). Enhancing explainability in AI-generated legal summaries using attention-based highlights. Journal of Artificial Intelligence and Law, 29(3), 201–218.IJRPR

Pal, R., Kumar, S., & Gupta, N. (2024). Domain-specific pre-training for legal language models: Performance improvements in legal NLP tasks. Legal Technology Journal, 12(4), 89–102.IJRPR

Pesaru, V., Chen, L., & Zhao, Y. (2024). AI-assisted document management using LangChain and Pinecone for legal applications. Journal of Legal Information Systems, 10(2), 55–70. IJRPR

Prasad, M., Singh, D., & Kaur, P. (2024). Overview of legal document summarization techniques: Extractive vs. abstractive methods. International Journal of Legal Studies, 9(3), 33–48.

Santosh, T. Y. S. S., Jia, C., Goroncy, P., & Grabmair, M. (2025). RELexED: Retrieval-enhanced legal

summarization with exemplar diversity. arXiv preprint arXiv:2501. 14113.arXiv

Sharma, A., Kumar, R., & Singh, P. (2023). AI-powered legal documentation assistant: Legal text summarization and clause extraction. International Journal of Research Publication and Reviews, 6(4), 1166–1173.IJRPR

Suryavanshi, V., Naikwadi, D., & Patil, S. (2024). Legal case document summarization using AI. International Journal of Engineering and Techniques, 10(2). IJET

Team NLP Legal Document. (2023). Legal document analysis and classification using NLP and deep learning. GitHub Repository. Retrieved from https://github.com/Team-NLP-Legal- Document/ LegalDocument_Analysisand ClassificationUsing_NLPand Deep _LearningGitHub

Vimala, S., Raj, K., & Meena, R. (2024). AI-powered legal documentation assistant for contract analysis and clause extraction. Legal AI Review, 8(1), 22–35. IJRPR

Width.ai. (2023). Improving legal document summarization using deep clustering (DCESumm). Retrieved from https://www.width.ai/post/legal-document-summarization-using-deep-clustering-dcesummWidth.ai