ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

www.ijirid.in

A Survey on AI-Powered Legal Document Summarization

Harsh Vardhan Singh Raj[1], Prateek Singh Bedi[2], Prathmesh Rakesh Mali[3], Rahmat Pathan[4], Sheetal Kapse[5]

[1,2,3]Student, STES's Smt. Kashibai Navale College of Engineering, Pune, India

[5]Assistant Professor, STES's Smt. Kashibai Navale College of Engineering, Pune, India

Abstract: *The legal domain is characterized by an overwhelming volume of complex, high-stakes documentation, where manual analysis often results in significant cognitive load and potential for human error. While conventional Natural Language Processing (NLP) tools have made strides in general text processing, they frequently struggle with the specialized vocabulary and rhetorical nuances inherent in legal discourse. This paper presents a systematic survey of AI-powered legal document summarization techniques using transformer-based NLP models. Existing legal summarization approaches commonly utilize transformer-based models along with Named Entity Recognition (NER) and Rhetorical Role Labeling (RRL). Furthermore, this survey identifies critical research gaps within the Indian legal context, including document truncation issues, the need for multilingual support, and the demand for explainable AI (XAI) to foster trust among legal professionals. By synthesizing current research trends and experimental findings, this work provides a comprehensive roadmap for the development of intelligent, scalable solutions that enhance information accessibility and productivity in the legal sector.*

Keywords: Legal NLP, Transformer Architectures, Rhetorical Role Labeling (RRL), Abstractive Summarization, Indian Judicial Context.

I. NTRODUCTION

The legal industry is characterized by the daily handling of massive volumes of high-stakes documentation, including case files, contracts, judicial judgments, and legal briefs. Legal professionals—including lawyers, paralegals, and researchers—allocate a disproportionate amount of time to the manual review of these lengthy texts to extract critical precedents and core legal principles. This manual process is not only resource-intensive but also inherently susceptible to human error, which can result in the misinterpretation of legal content or the oversight of vital clauses. While general-purpose Natural Language Processing (NLP) tools have seen rapid advancement, they frequently fail to comprehend specialized legal terminology, domain-specific nuances, and the unique rhetorical structures found in judicial discourse.

Consequently, there is a significant research focus on developing AI-powered solutions specifically tailored to the legal domain. Modern research in this field has expanded beyond simple summarization to include multifaceted tasks such as Case Brief Generation, Named Entity Recognition (NER), Question Answering (QA), and Rhetorical Role Labeling (RRL). This paper provides a systematic survey of these emerging methodologies, evaluating the transition from extractive models to advanced transformer-based architectures like BART and Legal-BERT, while identifying the persistent challenges within the Indian legal framework.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

www.ijirid.in

II. TAXONOMY

A systematic classification of the methodologies used in legal document summarization is essential to understand the transition from traditional algorithms to advanced neural architectures. This survey categorizes the existing research into three primary technical domains:

2.1 Extractive and Statistical Methodologies: Extractive summarization involves identifying and concatenating the most salient sentences from the original document without modifying the vocabulary.

1. Statistical Feature Analysis: Early models relied on frequency-based metrics such as TF-IDF, sentence position, and length to rank sentence importance.
2. Metaheuristic Optimization: Research has explored the use of the Gravitational Search Algorithm to optimize sentence selection by analyzing features like sentence similarity and relevance, proving competitive in specific legal datasets.
3. Reinforcement Learning: Advanced extractive models, such as MemSum, have been applied to massive corpora of U.S. court opinions (approx. 430,000 documents) to outperform standard transformer models in capturing key judicial passages.

2.2 Neural and Abstractive Architectures: Abstractive summarization aims to "rewrite" the content, generating new sentences that preserve the original meaning while enhancing readability.

1. Sequence-to-Sequence Transformers: State-of-the-art models like BART and T5 are utilized for their ability to generate context-aware, human-like summaries of lengthy legal briefs.
2. Domain-Specific Pre-training: The introduction of Legal-BERT, a model pre-trained specifically on legal corpora, has addressed the failure of general-purpose AI to comprehend specialized legal terminology and domain-specific context.

2.3 Hybrid and Structural Models: To overcome the limitations of purely abstractive or extractive systems, researchers have developed hybrid models that incorporate the structural logic of legal judgments.

1. Rhetorical Role Labeling (RRL): This technique involves classifying text into logical segments such as "Facts," "Claims," and "Results". Studies like LawSum demonstrate that tagging these roles is crucial for handling the noisy and unstructured nature of Indian Supreme Court judgments.
2. Argument Mining: By integrating argument role labeling with abstractive summarization, models can improve the coherence and relevance of summaries, ensuring that the legal reasoning and argumentative structure remain intact.

III. LITERATURE REVIEW

Research in legal document summarization has evolved from traditional extractive techniques to advanced transformer-based architectures. Early approaches primarily relied on statistical and frequency-based methods such as TF-IDF, TextRank, and sentence-position analysis to identify

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

www.ijirid.in

important sentences from legal documents. Although these methods were computationally efficient, they often failed to preserve semantic coherence and the logical structure of judicial decisions, as general sequence-to-sequence models and foundational transformer layers had not yet been optimized for long-form legal text sequence dependencies [14], [15], [16].

Recent advancements in deep learning and transformer-based Natural Language Processing (NLP) models have significantly improved legal summarization performance. Models such as BART, PEGASUS, T5, and Legal-BERT are capable of generating context-aware summaries while preserving legal semantics and document structure [6], [9], [10]. Several studies have also integrated Rhetorical Role Labeling (RRL) and argument mining techniques to improve the coherence and factual consistency of generated summaries [5], [7].

Hybrid approaches combining extractive and abstractive methods have shown promising results for long-form legal documents. Mukund and Easwarakumar proposed an optimized legal text summarization framework that integrates domain-specific adaptation with dynamic Retrieval-Augmented Generation (RAG) to reduce hallucination and improve summary relevance [2]. Similarly, Bauer et al. demonstrated that reinforcement learning-based extractive systems such as MemSum outperform conventional transformer baselines in identifying critical judicial passages [4]. To capture broader boundaries, researchers have also leveraged long-document token architectures like Longformer and Big Bird [11], [12].

Despite these advancements, several challenges remain unresolved. Existing transformer-based models struggle with long-document token limitations, multilingual legal terminology, explainability, and computational efficiency [1], [3], [8]. The comparative analysis of major legal summarization methodologies is presented in Table 1.

Table 1: Comparative Analysis of Existing Legal Summarization Research

TITLE	Authors / Year	Method(s) Used	Key Contributions / Findings	Limitations
Summarizing judicial documents	Y. Gao et al. (2025)	Hybrid Extractive-Abstractive + Legal Domain Knowledge	Reduces hallucination and improves summary relevance using domain-specific knowledge	High computational complexity; limited cross-jurisdiction evaluation
Legal Document Summarization: Enhancing Judicial Efficiency	Y. Li et al. (2025)	LLM-based Legal Summarization	Improves legal workflow efficiency and contextual understanding	Explainability and reliability concerns
A Comprehensive Survey on Legal Summarization	M. Akter et al. (2025)	Systematic Survey	Analyzes over 120 research papers on legal summarization techniques	Limited experimental benchmarking
Legal Extractive Summarization of U.S. Court Opinions	E. Bauer et al. (2023)	Reinforcement Learning (MemSum)	MemSum outperforms conventional transformer baselines in extractive summarization	Limited abstractive capability

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

www.ijirid.in

A Survey of Legal Document Summarization Methods	S. Takale (2023)	Survey of Extractive, Abstractive, and Hybrid Methods	Provides taxonomy of legal summarization approaches	Lacks implementation and comparative evaluation
ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining	M. Elaraby and D. Litman (2022)	Argument Mining + Abstractive Summarization	Improves coherence using argumentative role labeling	Sensitive to argument labeling errors
Indian Legal Text Summarization Using BART and PEGASUS	S. Ghosh et al. (2022)	Transformer-Based Summarization	BART achieves better semantic summarization performance than PEGASUS	High computational requirements
Legal Case Document Summarization: Extractive and Abstractive Methods and Their Evaluation	A. Shukla et al. (2022)	Supervised and Unsupervised Learning	Comparative analysis of extractive and abstractive techniques	Token limitations for long legal documents
LawSum: A Weakly Supervised Approach for Indian Legal Document Summarization	V. Parikh et al. (2021)	Weak Supervision + Rhetorical Role Labeling	Introduces Indian legal judgment dataset with rhetorical role annotations	Style inconsistency across legal domains
Automated Legal Document Summarization Using NLP Techniques	S. Ghosh and G. Chowdhury (2021)	NLP-based Extractive Summarization	Demonstrates effectiveness of NLP techniques in legal summarization	Limited semantic understanding
Summarization of Legal Documents: Where Are We Now?	D. Jain et al. (2021)	Review Study	Highlights key challenges in legal NLP and summarization	Limited empirical validation
LEGAL-BERT: The Muppets Straight Out of Law School	I. Chalkidis et al. (2020)	Domain-Specific BERT Pre-training	Improves legal text understanding and contextual representation	Requires large-scale legal corpora
Legal Document Summarization Using NLP and Machine Learning Techniques	R. Kore et al. (2020)	TF-IDF + TextRank	Evaluates traditional extractive summarization methods	Poor semantic coherence
BERT: Pre-training of Deep Bidirectional Transformers for	J. Devlin et al. (2019)	Transformer-based Language Model	Introduces contextual bidirectional language representation	General-purpose model not specialized for legal NLP

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

www.ijirid.in

Language Understanding				
Summarization of Legal Judgments Using Gravitational Search Algorithm and TF-IDF	K. Sharma et al. (2019)	Gravitational Search Algorithm + TF-IDF	Demonstrates competitive extractive summarization performance	Cannot handle abstractive summarization tasks

IV. RESEARCH GAPS AND CHALLENGES

Despite significant advancements in legal NLP, a systematic review of the current literature reveals several critical gaps that hinder the widespread adoption of AI-based summarization in the judicial sector.

4.1 Domain Adaptation and Multilingualism: Most state-of-the-art models are trained on general English corpora, which lacks the specificity required for Indian legal contexts. There are insufficient handling of regional legal terminology and a lack of multilingual support for languages such as Hindi, Marathi, and Bengali, which are vital for diverse legal systems.

4.2 Document Length and Truncation: Legal case files are characterized by their extreme length, often exceeding the token limits of standard transformer models. This leads to the truncation of documents and the subsequent loss of critical legal reasoning or final judicial facts. Handling long-form legal documents remains a challenge due to transformer token limitations.

4.3 Semantic Understanding and Logic: General summarization models often struggle to capture the logical structure of a judgment, such as the relationship between a "Claim" and a "Result". There is a persistent difficulty in identifying rhetorical roles accurately, which can lead to summaries that miss essential legal facts.

4.4 Explainability and Professional Trust: The "black-box" nature of many deep learning models remains a significant barrier for legal professionals. To foster trust, future systems must incorporate Explainable AI (XAI) techniques, such as rationale generation, to elucidate why specific clauses or facts were included in a generated summary.

V. PRACTICAL INSIGHTS

Based on the methodologies and research trends identified during the survey, a lightweight prototype system named Briefly was developed to explore the practical application of AI-powered legal document summarization techniques. The prototype integrates transformer-based Natural Language Processing (NLP) models with structured legal information extraction to improve accessibility and understanding of lengthy legal documents.

The system accepts legal documents in multiple formats, including PDF, DOCX, TXT, and PPT, and performs preprocessing operations such as text extraction, normalization, and chunking before

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

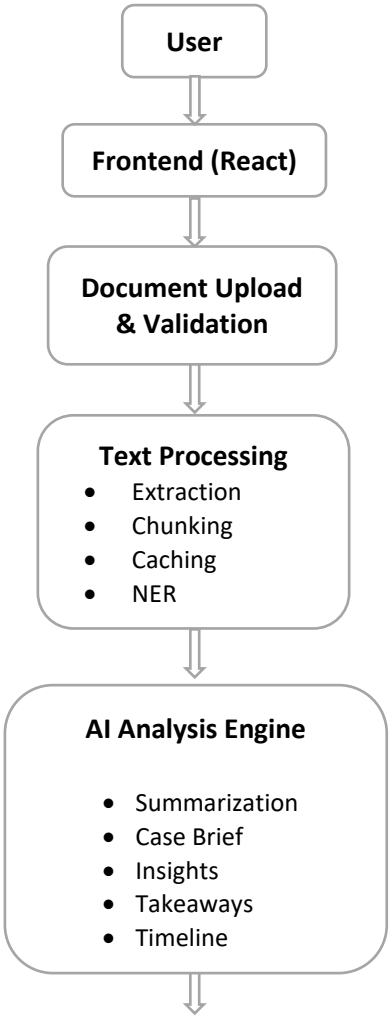
www.ijirid.in

analysis. For summarization, the prototype initially utilized the BART Large CNN model; however, due to high computational overhead and slower inference time, the system was optimized using the DistilBART model with CUDA-based GPU acceleration to improve processing efficiency while maintaining acceptable summary quality.

In addition to abstractive summarization, the prototype incorporates supplementary NLP functionalities such as Named Entity Recognition (NER), Rhetorical Role Labeling (RRL), Question Answering (QA), key takeaway extraction, and Text-to-Speech (TTS) generation. These features collectively demonstrate how modern NLP techniques can be integrated into practical legal document analysis systems.

The development process also highlighted several real-world challenges discussed throughout this survey, including transformer token limitations, handling of long-form legal judgments, computational resource constraints, and maintaining semantic consistency in generated summaries. These practical observations further reinforce the need for efficient, explainable, and domain-adapted AI systems within the legal sector.

Figure 2: System Architectural Pipeline



 DOI: 10.5281/zenodo.20677323

ISSN (Online): 2583-648X

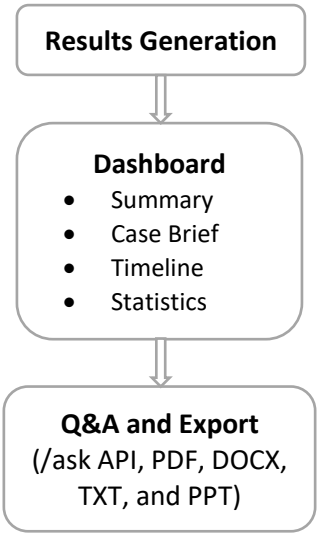
IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

www.ijirid.in



VI. CONCLUSION

The systematic survey of AI-powered legal document summarization confirms a significant shift from traditional statistical methods to context-aware, deep learning architectures. The survey demonstrates the growing effectiveness of transformer-based NLP techniques in legal document summarization. The prototype implementation further validated many of the practical challenges identified in the surveyed literature, particularly regarding computational efficiency and long-document processing.

6.1 Summary of Findings: The evaluation of current state-of-the-art systems indicates that Hybrid Models incorporating Rhetorical Role Labeling (RRL) provide the most reliable results for preserving judicial intent. However, the research also reveals that the Indian legal context faces specific hurdles, particularly concerning the massive scale of case judgments and the linguistic diversity of regional legal systems.

6.2 Future Research Directions: Moving forward, the field must address critical gaps to transition these AI solutions from academic research to professional legal practice:

1. Multilingual and Regional Support: Developing models that can summarize legal content in languages such as Hindi, Marathi, and Bengali is essential for widespread accessibility.
2. Long-Form Document Processing: Overcoming the "truncation gap" through advanced chunking or Retrieval-Augmented Generation (RAG) will ensure that no critical facts are lost in 50+ page judgments.
3. Transparency and Explainability: Future research must prioritize Explainable AI (XAI) to provide legal practitioners with the rationales behind generated summaries, fostering the trust necessary for digital transformation in the judiciary.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 3 | June 2026

www.ijirid.in

The continued advancement of legal NLP systems has the potential to improve information accessibility and reduce the workload associated with legal document analysis. Future research should focus on multilingual legal processing, explainable AI techniques, and efficient long-document summarization models suitable for real-world judicial applications.

REFERENCES

[1] V. Naik and K. Rajeswari, "Indian Legal Judgment Summarization using LEGAL-BERT and BiLSTM Model with Adaptive Length," EPJ Web of Conferences, vol. 328, 2025.

[2] A. Mukund and K. S. Easwarakumar, "Optimizing Legal Text Summarization Through Dynamic Retrieval-Augmented Generation and Domain-Specific Adaptation," Symmetry, vol. 17, no. 5, 2025.

[3] M. Malik, Z. Zhao, M. Fonseca, S. Rao, and S. B. Cohen, "CivilSum: A Dataset for Abstractive Summarization of Indian Court Decisions," Proceedings of EMNLP, 2024.

[4] E. Bauer, D. Stammbach, N. Gu, and E. Ash, "Legal Extractive Summarization of U.S. Court Opinions," arXiv preprint arXiv:2305.08428, 2023.

[5] M. Elaraby and D. Litman, "ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining," Proceedings of COLING, 2022.

[6] S. Ghosh, M. Dutta, and T. Das, "Indian Legal Text Summarization: A Text Normalisation-based Approach," arXiv preprint arXiv:2206.06238, 2022.

[7] V. Parikh, V. Mathur, P. Mehta, N. Mittal, and P. Majumder, "LawSum: A Weakly Supervised Approach for Indian Legal Document Summarization," arXiv preprint arXiv:2110.01188, 2021.

[8] S. Ghosh and G. Chowdhury, "Automated Legal Document Summarization Using NLP Techniques," International Journal of Computer Applications, 2021.

[9] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with Extracted Gap-Sentences for Abstractive Summarization," Proceedings of ICML, 2020.

[10] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets Straight Out of Law School," Findings of EMNLP, 2020.

[11] L. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," arXiv preprint arXiv:2004.05150, 2020.

[12] M. Zaheer et al., "Big Bird: Transformers for Longer Sequences," Advances in Neural Information Processing Systems (NeurIPS), 2020.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2019.

[14] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017.

[15] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," Advances in Neural Information Processing Systems (NeurIPS), 2014.

