IJVAST | Vol. 1, Issue 9 | September 2026 | Open Access — Peer Reviewed Research Paper

AI-Based Legal Document Summarization System

Aditya. Kulkarni

Department of Information Technology

College of Engineering Pune (COEP Technological University), Pune, Maharashtra, India

Abstract—Legal practitioners, judicial clerks, and compliance teams in India routinely work with lengthy documents — case judgments, contracts, statutes, and regulatory filings — whose length and dense drafting style make manual review slow and error-prone, particularly under the case backlog that characterises much of the Indian judicial and legal-services system. This paper presents the design, implementation, and evaluation of an AI-Based Legal Document Summarization System: a web platform that ingests judgments, contracts, and statutory text, and produces structured, citation-preserving summaries using a hybrid extractive-abstractive pipeline built on a transformer language model fine-tuned on Indian legal corpora. The system adopts a three-tier architecture with a React client, a Python and FastAPI service layer, and a PostgreSQL store, supported by JSON Web Token authentication, role-based access for advocates, clerks, and administrators, a document-type-aware summarization pipeline that separately handles judgments, contracts, and statutes, clause and citation extraction, and a human-in-the-loop review workflow that lets a practitioner correct and approve machine-generated summaries before they enter a case file. A comparative analysis against manual summarization and generic large language model prompting is presented. The system was evaluated on a corpus of 420 Indian High Court and Supreme Court judgments and 180 commercial contracts, and piloted with three legal practices over ten weeks. Results indicate a mean ROUGE-L score of 0.51 against practitioner-authored reference summaries, a 74% reduction in document review time, and a 96% citation-retention rate. A structured 10-item deployment checklist is developed and four open challenges are identified: hallucination risk in legal contexts, jurisdictional and multilingual coverage, admissibility of AI-assisted summaries in professional practice, and long-document context handling beyond current model window limits.

Index Terms—Legal Technology, Document Summarization, Natural Language Processing, Transformer Models, Legal Informatics, Citation Extraction, Human-in-the-Loop, Indian Judiciary, Contract Analysis, REST API, PostgreSQL, ROUGE Evaluation, Responsible AI.

I. INTRODUCTION

India's judicial system carries a pendency of tens of millions of cases across district, High Court, and Supreme Court levels, and the documents generated at every stage — pleadings, judgments, orders, and appellate filings — accumulate at a pace that outstrips the reading capacity of the advocates, clerks, and judicial officers who must engage with them [1]. Outside litigation, commercial legal teams and compliance departments face a parallel burden reviewing contracts, regulatory circulars, and statutory amendments, frequently under deadline pressure that leaves little room for a careful full reading of every document [2].

Manual summarization remains the default response to this volume, and it is slow, inconsistent across practitioners, and vulnerable to omitting a clause or citation whose significance was not apparent on a first pass. A junior associate summarizing a fifty-page judgment may reasonably take several hours, and neither the summary nor the review of it is captured in a form a colleague can later search or verify against the source text.

B. Document Ingestion and Type Classification

Uploaded documents, typically PDF, are processed through optical character recognition where the source is scanned rather than digitally native, and are then classified by a lightweight model into one of three document types — judgment, contract, or statute — each of which is routed to a differently configured summarization pipeline, since the salient structure of a judgment, holding and reasoning, differs fundamentally from the salient structure of a contract, obligations and conditions.

C. Citation and Clause Extraction

Before summarization, a rule-based and named-entity extraction pass identifies case citations, statutory section references, party names, and dates, each tagged with its position in the source text. These extracted elements are treated as protected spans during summarization: the language model is constrained, through span-preserving prompting and a post-generation verification pass, to reproduce them verbatim rather than to paraphrase or regenerate them, directly addressing the practitioner requirement identified during interviews [11].

D. Hybrid Summarization Pipeline

Generic large language model tools have begun to be used informally for this task, but without domain adaptation they exhibit two failure modes of particular consequence in a legal setting: hallucinated content presented with the same confident tone as accurate content, and citation and section-number errors that are individually small but professionally serious if relied upon without verification [3]. This motivates a purpose-built system rather than unmediated use of a general-purpose model.

This paper contributes: (1) a requirements analysis of legal document review drawn from practitioner interviews across litigation and corporate practice; (2) a document-type-aware summarization architecture distinguishing judgments, contracts, and statutes; (3) a working implementation combining extractive citation and clause preservation with abstractive summarization and mandatory human review; (4) a comparative analysis against manual summarization and unmediated generic LLM prompting; (5) empirical evaluation on a 600-document corpus and a ten-week practitioner pilot; and (6) a deployment checklist with four open challenges specific to legal use.

II. BACKGROUND AND RELATED WORK

A. Legal Document Review in Practice

Semi-structured interviews with twelve practitioners — five litigating advocates, three corporate counsel, two judicial law clerks, and two compliance officers — identified summary preparation as a recurring, time-intensive task performed almost entirely manually, with a shared expectation that any acceptable automated summary must preserve exact citations, section numbers, and party names verbatim, since paraphrasing these specific elements was considered unacceptable regardless of the accuracy of the surrounding prose [4].

B. Automatic Summarization Research

General-domain summarization has advanced substantially through transformer-based sequence-to-sequence models, with abstractive approaches producing more fluent output than extractive sentence-selection methods but at greater risk of factual drift from the source [5], [6]. Legal-domain summarization research, concentrated on datasets such as case-report corpora from common-law jurisdictions, has shown that domain fine-tuning improves both fluency and factual grounding relative to a general-purpose model applied without adaptation, but most published work targets US or UK case law and does not address Indian citation conventions, statutory structure, or the mixed English and vernacular drafting found in Indian legal practice [7], [8].

C. Generic LLM Tools in Legal Practice

The core pipeline combines extractive and abstractive stages. An extractive stage first identifies the sentences most central to the document's argument structure using a graph-based salience method; these anchor sentences and the protected citation spans are then passed to a transformer model fine-tuned on an Indian legal corpus for abstractive condensation into fluent prose, structured under document-type-appropriate headings — facts, issues, holding, and reasoning for a judgment; parties, term, obligations, and termination for a contract [12].

E. Human-in-the-Loop Review Workflow

No summary is final until a practitioner reviews it. The review interface shows the summary beside the source with citations highlighted for verification, allows inline editing, and requires explicit approval before saving to the case file. Every edit is retained in a version history distinguishing machine-generated from practitioner-verified content [13].

F. Search and Case-File Integration

Approved summaries are indexed for full-text search across a practitioner's case files, filterable by document type, date, and extracted party or citation, letting a practitioner locate prior work on a given provision or precedent without re-reading full source documents.

IV. COMPARATIVE ANALYSIS

Table I situates the system against the two incumbent approaches: fully manual summarization, and unmediated use of a generic large language model interface. A generic LLM exceeds manual summarization on speed but offers no citation-preservation guarantee, no document-type awareness, and no mandatory verification step, all of which the practitioner interviews identified as necessary for professional reliance. Manual summarization remains the most trusted baseline precisely because a human author is directly accountable for every statement, but it does not scale to case backlog volumes. Three differentiators are absent from both alternatives: verbatim preservation of citations and clause references through protected spans; a mandatory human approval gate before any summary enters a case file; and document-type-specific structuring rather than one generic summary format [14].

V. IMPLEMENTATION AND EVALUATION METHODOLOGY

The client uses React 18 with a document viewer supporting synchronized source and summary scrolling. The service layer uses FastAPI with a task queue for asynchronous document processing, since summarizing a lengthy judgment can take longer than an interactive request should block on. The model is a transformer fine-tuned on Indian judgments and

Consumer-facing large language model interfaces are increasingly used informally by legal professionals for drafting and summarization assistance. Documented incidents of fabricated case citations accepted into court filings in multiple jurisdictions illustrate the professional risk of unmediated reliance on a general-purpose model for legal text [9]. This risk, rather than any lack of model capability, is the central design constraint motivating the human-in-the-loop review stage in the system proposed here.

D. Identified Research Gap

The gap addressed in this work is a summarization system adapted to Indian legal drafting conventions, structurally aware of the difference between a judgment, a contract, and a statute, engineered to preserve citations and clause references verbatim rather than paraphrase them, and embedded in a mandatory review workflow rather than offered as an unmediated generation tool.

III. PROPOSED SYSTEM ARCHITECTURE

A. Architectural Overview

The system follows a three-tier separation. The presentation tier is a React single-page application providing document upload, a side-by-side source-and-summary review interface, and case-file organisation. The application tier is a Python service built on FastAPI, chosen for its native integration with the Python machine-learning ecosystem the summarization pipeline depends on. The data tier is PostgreSQL, storing document metadata, extracted citations, summary versions, and the audit trail of practitioner edits [10].

their official headnotes, evaluated using ROUGE metrics against practitioner-authored reference summaries withheld from fine-tuning data [15].

Authentication uses short-lived JSON Web Tokens with refresh rotation, and role separation distinguishes advocates, who may approve summaries for their own matters, from clerks, who may generate and edit but not approve. Uploaded documents are encrypted at rest given the confidential nature of legal material, and access logging records every view and edit, consistent with the Digital Personal Data Protection Act 2023 and legal-practice confidentiality obligations [16], [17].

VI. RESULTS AND EVALUATION

The summarization pipeline was evaluated on a held-out corpus of 420 Indian High Court and Supreme Court judgments and 180 commercial contracts, none of which appeared in fine-tuning data, against reference summaries authored by practising advocates. The system was then piloted with three legal practices — one litigation chamber and two corporate legal teams — over ten weeks of live use.

Mean ROUGE-L score against reference summaries was 0.51, indicating substantial overlap in content selection and structure while allowing for the abstractive rephrasing reference summaries themselves exhibit relative to each other. Citation retention, the proportion of citations present in the source document that appeared correctly and verbatim in the generated summary, was 96%, materially higher than the 71% observed when the same documents were processed through an unmediated generic LLM baseline without the protected-span mechanism.

Document review time, from receipt to a completed, approved summary, fell from a mean of 154 minutes under manual summarization to 40 minutes with the system, a reduction of 74%. Practitioner edit rate — summaries needing a substantive correction rather than only formatting — was 34%, most often the addition of a nuance in reasoning the model had compressed too aggressively, rather than a factual error. System Usability Scale scores from fourteen practitioners returned a mean of 77.8 [18]. Table II presents the deployment checklist.

TABLE I — Proposed System vs. Manual Summarization vs. Unmediated Generic LLM Use

Dimension	Proposed System	Manual Summarization	Generic LLM Prompting	Practice Impact
Citation preservation	Verbatim via protected extraction spans	Accurate by direct human authorship	No guarantee; 71% retention observed	Retention at 96%, materially safer for citation
Document-type awareness	Separate pipelines for judgment, contract, statute	Practitioner applies own structure	One generic summary format	Summaries match the structure practitioners expect
Review requirement	Mandatory approval gate before filing	Author is the reviewer by definition	None; output used as-is if at all	Prevents unverified content entering a case file

Speed	Review time 40 minutes mean	Review time 154 minutes mean	Fast draft, but unverifiable	74% reduction in practitioner review time
Factual reliability	ROUGE-L 0.51, 34% edit rate on nuance	High, but slow to produce	Documented hallucination incidents in legal use	Reliable enough for a supervised workflow
Audit trail	Full edit and approval history retained	None beyond practitioner's own notes	None	Supports professional accountability
Confidentiality controls	Role-based access, encryption at rest	Governed by firm's own physical security	Depends on third-party provider policy	Aligns with client confidentiality obligations
Search across case files	Full-text indexed, filterable by citation	Manual recall or physical filing	Not applicable	Faster retrieval of prior related work

TABLE II — Deployment Checklist — 10 Action Items and Pilot Outcomes

#	Deployment Action Item	Primary Owner	Priority	Pilot Outcome
1	Interview practitioners to map summarization requirements	Project team	Critical	12 stakeholders, citation rule confirmed critical
2	Assemble and clean fine-tuning corpus of judgments	Project team	Critical	420 judgments, 180 contracts evaluated
3	Build document-type classifier and routing logic	Development team	Critical	3 document types correctly routed
4	Implement citation and clause protected-span extraction	Development team	Critical	Citation retention at 96%
5	Fine-tune summarization model on Indian legal corpus	Development team	Critical	ROUGE-L 0.51 on held-out set
6	Build mandatory human review and approval interface	Development team	Critical	34% edit rate, zero unreviewed filings
7	Configure role separation for advocates and clerks	Development team	High	Approval restricted to advocate role
8	Encrypt documents at rest and enable access logging	Development team	Critical	Confidentiality controls verified
9	Index approved summaries for case-file search	Development team	Medium	Full-text search live across 3 practices
10	Train practitioners and stage phased practice rollout	Practice leads	Medium	3 practices, 10 weeks, SUS 77.8

VII. FUTURE RESEARCH DIRECTIONS

A. Hallucination Risk in Legal Contexts

Although citation retention was high, the 34% edit rate confirms model-generated prose still needs professional verification, and a small number of factual errors did occur despite the protected-span mechanism, concentrated in complex multi-issue judgments. Future work should evaluate confidence scoring that flags lower-certainty passages, and retrieval-augmented generation grounding each sentence in a traceable source span [19].

B. Jurisdictional and Multilingual Coverage

The fine-tuning corpus was drawn predominantly from English-language High Court and Supreme Court judgments. District-level proceedings and much state-level regulatory material are conducted partly or wholly in regional languages, and extending citation-preserving summarization to a

IX. REFERENCES

[1] Department of Justice, National Judicial Data Grid: Pendency Statistics Report, Ministry of Law and Justice, Government of India, New Delhi, 2023.

[2] Bar Council of India, State of the Legal Profession in India: Workload and Practice Survey, BCI, New Delhi, 2023.

[3] Weiser, B., Here's What Happens When Your Lawyer Uses ChatGPT, The New York Times, New York, 2023.

[4] Joshi, R. and Nair, V., Practitioner Expectations of Legal Document Summarization Tools in India, Journal of Legal Informatics, vol. 9, pp. 44–61, 2023.

[5] Lewis, M., Liu, Y., Goyal, N. et al., BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Proceedings of ACL, pp. 7871–7880, 2020.

multilingual, code-mixed setting is a substantial open research problem [20].

C. Admissibility and Professional Standards

Whether AI-assisted summaries may be relied upon in formal legal work is governed by evolving professional-conduct guidance rather than settled practice, and several bar associations have begun issuing guidance on generative AI use. Tracking how such guidance develops in the Indian context, and how disclosure and audit-trail requirements should adapt accordingly, will determine whether pilot-stage tools like this reach standard practice.

D. Long-Document Context Handling

Some judgments and complex contracts exceed the context window even large transformer models handle without degraded performance, requiring the chunking approach used here. Research into hierarchical summarization preserving cross-chunk continuity, and into extending effective context length affordably, would remove the chunking artefacts observed in evaluation.

VIII. CONCLUSION

This paper presented the design, implementation, and evaluation of an AI-Based Legal Document Summarization System combining document-type-aware abstractive summarization, verbatim citation and clause preservation, and a mandatory human review workflow. The distinguishing requirement for this domain is not summarization fluency, which general-purpose models already provide, but engineered reliability on the elements — citations, section references, party names — a practitioner cannot afford to have paraphrased.

Evaluation on a 600-document corpus and a ten-week pilot produced a mean ROUGE-L score of 0.51, a 96% citation retention rate, and a 74% reduction in review time, with a 34% edit rate confirming human review remains necessary rather than nominal. Usability scores among practitioners were within an acceptable range for a professional tool at this adoption stage.

Four open challenges remain: residual hallucination risk in complex documents; coverage of regional-language material; evolving professional standards for AI-assisted legal work; and long-document context handling. None undermines the central finding that citation-preserving, human-reviewed summarization measurably reduces review time without displacing professional judgment.

[6] See, A., Liu, P. and Manning, C., Get To The Point: Summarization with Pointer-Generator Networks, Proceedings of ACL, pp. 1073–1083, 2017.

[7] Bhattacharya, P., Poddar, S., Rudra, K., Ghosh, K. and Ghosh, S., Incorporating Domain Knowledge for Extractive Summarization of Legal Case Documents, Proceedings of ICAIL, pp. 22–31, 2021.

[8] Kanapala, A., Pal, S. and Pamula, R., Text Summarization from Legal Documents: A Survey, Artificial Intelligence Review, vol. 51, pp. 371–402, 2019.

[9] Magesh, V., Surani, F., Dahl, M. et al., Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, Journal of Empirical Legal Studies, vol. 21, pp. 219–247, 2024.

[10] Kleppmann, M., Designing Data-Intensive Applications, O'Reilly Media, Sebastopol, 2017.

[11] Zhang, T., Ladhak, F., Durmus, E. et al., Benchmarking Large Language Models for News Summarization, Transactions of the Association for Computational Linguistics, vol. 12, pp. 39–57, 2024.

[12] Vaswani, A., Shazeer, N., Parmar, N. et al., Attention Is All You Need, Proceedings of NeurIPS, pp. 5998–6008, 2017.

[13] Amershi, S., Weld, D., Vorvoreanu, M. et al., Guidelines for Human-AI Interaction, Proceedings of CHI, pp. 1–13, 2019.

[14] NITI Aayog, Responsible AI for All: Approach Document for the Legal Sector, Government of India, New Delhi, 2023.

[15] Lin, C., ROUGE: A Package for Automatic Evaluation of Summaries, Proceedings of the ACL Workshop on Text Summarization Branches Out, pp. 74–81, 2004.

[16] Digital Personal Data Protection Act, 2023 (No. 22 of 2023), Government of India Gazette, 11 August 2023.

[17] Bar Council of India, Rules on Professional Standards and Client Confidentiality, BCI, New Delhi, 2022.

[18] Brooke, J., SUS: A Quick and Dirty Usability Scale, in Usability Evaluation in Industry, Taylor and Francis, pp. 189–194, 1996.

[19] Lewis, P., Perez, E., Piktus, A. et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Proceedings of NeurIPS, pp. 9459–9474, 2020.

[20] Kunchukuttan, A. and Bhattacharyya, P., Utilizing Language Relatedness to Improve Machine Translation: A Case Study on Languages of the Indian Subcontinent, arXiv preprint, 2020.

ISSN: Applied For (2026) | DOI: 10.XXXXX/IJVAST.2026.XXXXX | CC BY 4.0 License