Research Paper - Legal Capstone_Final.docx

My Files

My Files

Shri Vile Parle Kelavani Mandal

Document Details

Submission ID

trn:oid:::9832:119983538

Submission Date

Nov 5, 2025, 7:51 PM GMT+5:30

Download Date

Nov 5, 2025, 7:54 PM GMT+5:30

File Name

Research Paper - Legal Capstone_Final (1).docx

File Size

863.8 KB

37 Pages

6,586 Words

42,346 Characters

84% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

106 AI-generated only 82%
Likely AI-generated text from a large-language model.

2 AI-generated text that was AI-paraphrased 1%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Chapter 1: Introduction

1.1 Background of the Project Topic

In today's digital legal landscape, the volume and complexity of legal documents ranging from contracts and court judgments to policy papers have increased exponentially. These documents are often long, detailed, and filled with technical terminology, making it difficult for professionals to review and interpret them efficiently. Traditional methods such as keyword-based searches or manual reading are no longer sufficient for identifying critical clauses or understanding documents at a contextual level [1], [2], [4].

Recent advancements in Artificial Intelligence, particularly Natural Language Processing (NLP) and transformer-based architectures such as BERT, T5, and BART, have revolutionized how textual data is processed and analyzed [3], [6], [7]. These models enable semantic understanding and contextual reasoning within long texts, which is essential for legal applications. Retrieval-Augmented Generation (RAG) models further enhance this capability by combining information retrieval with generative text modeling, ensuring factual accuracy in responses [9], [19].

However, existing systems in the legal NLP domain remain fragmented. Summarization, clause retrieval, and anomaly detection are often handled as independent modules, lacking integration and explainability [15], [17]. The absence of a unified framework for clause-level reasoning and deviation detection creates a gap in automated legal analysis. This challenge motivated the development of an Integrated Legal Document Analysis System that can summarize, retrieve, and analyze clauses contextually while identifying potentially irregular or high-risk segments.

1.2 Motivation and Scope of the Report

Legal professionals, analysts, and compliance officers spend substantial time interpreting large documents, which leads to inefficiency and inconsistency in risk assessment. Many contractual terms that introduce potential legal risks such as "one-sided indemnity" or "non-terminable agreements" often remain unnoticed due to wording variations or placement in lengthy documents [17], [18].

The motivation behind this project is to build a risk-aware, explainable, and modular legal NLP framework that can process long legal texts and highlight clauses worth human review. By integrating summarization, semantic retrieval, conversational reasoning, and clause-level deviation profiling into one pipeline, the system aims to reduce manual workload and improve accuracy.

The scope of this project encompasses:

- Summarization of long legal documents using transformer-based models (BART, LED).
- Context-based clause retrieval and conversational interaction using RAG with Flan-T5 and Sentence-BERT embeddings.
- Generation of visualized clauses that are potentially risky by using unsupervised clause-level deviation scoring through linguistic, semantic, and contextual analysis using InLegalBERT, Legal-BERT, and E5-base-v2 models.

This project is not restricted to a specific legal corpus; it can be adapted for contracts, judgments, and policies across different jurisdictions, thereby broadening its scalability and real-world utility.

1.3 Problem Statement

Despite rapid advancements in transformer-based NLP, legal document understanding remains limited in risk awareness and interpretability. Current models can summarize or retrieve legal clauses, but they fail to:

1. Identify clause-level deviations or anomalies that may imply risk.
2. Combine summarization, retrieval, and deviation detection within a single explainable framework.
3. Provide factual, context-grounded answers through conversation with traceable clause references.

Hence, the problem statement addressed in this project is:

> "To develop a unified, retrieval-augmented transformer framework that integrates summarization, clause retrieval, and unsupervised clause-level deviation profiling to improve interpretability and risk awareness in legal document analysis."

1.4 Salient Contribution

The major contributions of this work are as follows:

1. **Unified Framework:** Integration of four core modules of summarization, clause retrieval, conversational reasoning, and deviation detection within a single architecture (Algorithm 1, Figure 1).
2. **Retrieval-Augmented Summarization:** Combination of BART/LED models with RAG to ensure contextually accurate and fact-checked summaries [6], [9], [10].
3. **Conversational Legal Assistant:** Implementation of a dialogue-based RAG module using Flan-T5 to answer user queries with clause grounding [19].
4. **Clause-Level Deviation Profiling Framework:**
    - ○ Pseudo-Perplexity Channel: InLegalBERT-based token uncertainty detection [29].
    - ○ Semantic Drift Channel: Legal-BERT-based distance from document centroid.
    - ○ Contextual Cohesion Channel: E5-base-v2 similarity-based coherence scoring [30].
    - ○ Combined using Reciprocal Rank Fusion (Equation 1) [31].
5. **Visualization Layer:** Generation of color-coded PDF outputs (green/yellow/orange) and structured CSV logs to make deviation scoring interpretable for human experts.

6. **Evaluation:** Achieved a ROUGE-L score of 0.3454 and a BERTScore-F1 of 0.8682 for summarization, showing improvement in factuality and coherence compared to baseline models.

1.5 Organization of Report

The report is organized as follows:

- **Chapter 1 – Introduction:** Presents the background, motivation, problem definition, contributions, and structure of the work.
- **Chapter 2 – Literature Survey:** Reviews existing research in legal NLP, including summarization, clause retrieval, risk detection, and anomaly analysis, followed by identification of research gaps.
- **Chapter 3 – Methodology and Implementation:** Details the design of the integrated framework, algorithms, models used, and system implementation.
- **Chapter 4 – Results and Analysis:** Discusses experimental findings, including summarization metrics and qualitative clause deviation outcomes.
- **Chapter 5 – Advantages, Limitations, and Applications:** Summarizes system strengths, challenges, and real-world applicability in legal risk management.
- **Chapter 6 – Conclusion and Future Scope:** Provides concluding remarks and outlines potential improvements such as supervised learning extensions and multilingual adaptability.

Chapter 2 – Literature Survey

The application of Natural Language Processing (NLP) in the legal domain has expanded rapidly over the last decade. Research spans from summarization of lengthy legal documents to clause-level tagging and anomaly detection for compliance. This chapter reviews major directions foundational surveys, legal summarization systems, clause-classification and risk modeling, and anomaly-detection methods to identify the gaps that motivated the proposed unified, explainable legal-document-analysis framework.

2.1 Introduction

Earlier work in legal NLP relied on statistical and rule-based techniques such as TF-IDF and TextRank to extract salient information, but these lacked semantic depth and contextual understanding [1], [4]. The field then progressed to deep-learning architectures like RNNs and LSTMs and later to transformers such as BERT, T5, and BART, which brought context-aware representations and reasoning capabilities [3], [6]. Despite this evolution, research remains fragmented across tasks such as summarization, retrieval, and anomaly detection without an integrated pipeline for clause-level interpretation. The following sections discuss each strand of work in detail and show how their findings converge to highlight the need for a unified, explainable approach.

2.2 Foundational Surveys and General Legal NLP

Table 1 Foundational Surveys and General Legal NLP

Author(s), Year	Objective / Contribution	Techniques / Models Used	Gaps / Limitations
Jonnalagadda et al., 2025 [1]	Comprehensive overview of text summarization across domains, including legal.	Rule-based (TF-IDF, TextRank), ML (SVM, NB), Seq2Seq (RNN/LSTM), Transformers (BERTSUM,	Limited legal focus, no RAG, stance detection, or risk awareness.

		T5, BART, PEGASUS), and GPT models.	
Chalkidis et al., 2024 [2]	Survey of NLP tasks, datasets, and challenges specific to the legal domain.	Pretrained LLMs (BERT, RoBERTa, GPT, T5, BART), LegalBERT, CaseLaw-BERT, LexLM.	Few tools for explainability, limited adoption of RAG, lack of stance/sentiment-aware NLP.
Mondal et al., 2025 [3]	Survey on legal summarization with >120 papers reviewed.	Extractive (TextRank, BM25, Legal-BERT), Abstractive (BART, T5, Pegasus, LED, Longformer).	No stance detection, risk flagging, or interactive systems.
Rani, 2023 [4]	Survey of legal document summarization methods.	Extractive (TF-IDF, LexRank, Naïve Bayes), Abstractive (RNN/LSTM, Transformers).	No mention of RAG, multi-document summarization, or semantic search.
George et al., 2025 [5]	Systematic review of legal document analysis software.	TF-IDF + Cosine Similarity, TextRank, BERT, T5, GPT-2.	No clause-level tagging, risk scoring, or scalability.

Several comprehensive surveys have mapped the overall growth of NLP in the legal domain.

Jonnalagadda et al. (2025) [1] reviewed summarization methods across domains, noting that while transformer-based models like BERTSUM, T5, and BART improve fluency, they lack legal-specific adaptation. Chalkidis et al. (2024) [2] compiled existing datasets and domain benchmarks such as CaseLaw-BERT and LexLM, observing limited adoption of retrieval-augmented and explainable systems. Mondal et al. (2025) [3] surveyed over 120 legal-summarization papers and emphasized the absence of stance detection and contextual reasoning. Rani (2023) [4] and George et al. (2025) [5] reported similar trends in extractive and abstractive approaches but shallow contextual modeling.

Collectively, these studies show how research has moved from rule-based to transformer-based NLP yet continues to lack clause-level interpretability, factual grounding, and explainability mechanisms which are issues that directly motivate this project.

2.3 Legal Summarization Techniques and Systems

Table 2 Legal Summarization Techniques and Systems

Author(s), Year	Objective / Contribution	Techniques / Models Used	Gaps / Limitations
Shen et al., 2023 [6]	Proposed a dual-stage, argument-aware summarization pipeline for long legal opinions.	Extractive (contrastive BERT ranking), Abstractive (BART).	Limited to judicial opinions, no stance/risk analysis, not applicable to contracts or policies.
Yang et al., 2023 [7]	Introduced STRONG, a structure-controllable summarization system for legal texts.	Flan-T5 with structure planning + conditional generation.	No clause-level analysis or semantic search; lacks risk awareness.
Nguyen et al., 2023 [8]	Developed an end-to-end system for Vietnamese legal case summarization.	TextRank (extractive), Multilingual BERT, GPT-2, BART, T5.	Language-specific, no RAG or stance detection, limited generalizability.
Sinha et al., 2025 [9]	Enhanced legal summarization through Retrieval-Augmented	Dense retrievers (DPR) + generative LLMs (BART, T5).	High inference cost; no interactive

	Generation (RAG) with domain-specific adaptation.		querying or clause-level analysis.
Agarwal et al., 2024 [10]	Built a multi-step summarization pipeline for very long regulatory documents.	LongT5, LED, Fusion-in-Decoder (FiD).	No risk or anomaly detection; lacks interactivity or explainability.
Gupta et al., 2024 [11]	Comparative evaluation of transformer models for legal summarization.	T5-small, PEGASUS, BART-large (fine-tuned).	Model bias risks: smaller T5 versions struggle with long texts.
Rakesh, 2024 [12]	General advancements in legal summarization using NLP + ML.	TextRank, TF-IDF, Naïve Bayes, Logistic Regression.	Focused mainly on extractive methods, with limited deep learning integration.
Wagh et al., 2023 [13]	Proposed NLP-based pipeline for legal case summarization (tender/legal documents).	Extractive (Gist), Abstractive (Legal-Summ).	No use of LLMs, lacks interactive query-based summaries.
Yadav et al., 2025 [14]	Built a hybrid legal document summarizer with a user interface.	BART (extractive) + T5 (abstractive), hybrid weighted output.	Preprocessing risks loss of context; limited to text inputs.

Legal summarization aims to condense complex judicial or contractual documents into readable and factually correct summaries.

Shen et al. (2023) [6] introduced an argument-aware dual-stage pipeline combining BERT ranking and BART generation. Yang et al. (2023) [7] proposed STRONG, a structure-controllable summarizer leveraging Flan-T5 for conditional generation. Nguyen et al. (2023) [8] developed a multilingual BART/T5 pipeline for Vietnamese case summaries, while Sinha et al. (2025) [9] advanced retrieval-augmented summarization with dense retrievers and generative LLMs—a precursor to the retrieval strategy used in this work. Agarwal et al. (2024) [10] and Gupta et al. (2024) [11] experimented with LongT5, LED, and FiD for long regulatory texts, achieving fluency but high computational cost. Earlier extractive systems [4], [12], [13] were lightweight but failed to capture semantic relationships, and hybrid BART–T5 designs [14] still lacked contextual retrieval and factual checks.

Overall, the literature reveals progress toward domain tuning yet highlights persistent weaknesses in factual grounding, efficiency, and interactive retrieval—all of which the proposed LED + RAG architecture seeks to overcome.

2.4 Clause Classification, Risk and Semantic Awareness

Table 3 Clause Classification, Risk Identification, and Semantic Awareness

Author(s), Year	**Objective / Contribution**	**Techniques / Models Used**	**Gaps / Limitations**
Bommasani et al., 2023 [15]	Clause classification in contracts (SemEval-2023).	LegalBERT, RoBERTa, DistilBERT	No risk-weighting or anomaly scoring.
Khan et al., 2025 [16]	Legal analysis using LLMs for accuracy and risk mitigation.	TF-IDF, TextRank, BERT, T5	Lack of clause-level risk tagging.
Liu et al., 2025 [17]	Benchmark for clause-level risk identification.	DeepSeek, LLaMA, Gemma, Qwen	LLMs lag in precision, missing clauses.

Wang et al., 2025 [18]	Expert-annotated dataset for clause retrieval.	BM25, MiniLM, Cross-Encoders	Focused on relevance, not anomaly.
Li et al., 2025 [19]	Multi-turn legal consultation RAG benchmark.	BM25 + GPT-4o + LLaMA	No multilingual coverage.
Nigam et al., 2024 [20]	Introduced an Indian legal judgment dataset and a domain-specific LLM for decision prediction.	INLegalLlama (fine-tuned LLaMA)	Focused on judgment prediction, not anomaly scoring.

Clause-classification and risk-analysis tasks have become increasingly central as contracts and regulations demand finer interpretability. Bommasani et al. (2023) [15] used LegalBERT and RoBERTa for clause labeling but without quantitative risk scoring. Khan et al. (2025) [16] combined T5 and BERT for legal-accuracy assessment but neglected clause weighting. Liu et al. (2025) [17] introduced ContractEval, a benchmark for clause-level risk detection using DeepSeek and Gemma, yet reported low precision on rare clauses. Wang et al. (2025) [18] built ACORD, an expert-annotated dataset for clause retrieval, and Li et al. (2025) [19] presented LexRAG for multi-turn legal consultations. Nigam et al. (2024) [20] developed INLegalLlama for Indian judgment prediction rather than risk profiling. Across these efforts, a consistent limitation emerges that is the lack of explainable, clause-specific scoring and the absence of unsupervised deviation metrics. These shortcomings directly inspired this project's tri-signal deviation framework combining linguistic, semantic, and contextual analysis.

2.5 Anomaly Detection and Fraud Prevention

Table 4 Anomaly Detection and Fraud Prevention in Legal and Document Analysis

Author(s), Year	**Objective / Contribution**	**Techniques / Models Used**	**Gaps / Limitations**
Goyal et al., 2025 [21]	Anomaly detection in tax filings.	OCR + Bi-LSTM + iForest	Jurisdiction-specific.
Dey et al., 2025 [22]	Multimodal fraud detection for document verification.	GNN + attention fusion	High inference time.
D. Novoa-Paradela et al, 2024 [23]	Anomaly detection in subjective text reviews.	MPNet + Autoencoder	Limited to short text.
Cao et al, 2025 [24]	Benchmark for embedding-based text anomaly detection.	BERT, MiniLM, LOF, iForest	Limited hyperparameter tuning.
Singh et al, 2025 [25]	Real-time fraud detection with RAG LLMs.	Retrieval-Augmented Generation + LLM	No legal corpus.
Pastoriza et al., 2025 [26]	Retrieval-Augmented Anomaly Detection framework.	Vector-store feedback fusion	No multi-class anomaly handling.
Dechtiar et al., 2024 [27]	Applied LLMs for automatic detection of “contract smells” and structural anomalies.	GPT-family, fine-tuned BERT	Focused on “smells,” not risk quantification.

Salazar et al., 2020 [29]	Proposed pseudo-perplexity as a metric for linguistic irregularity in masked language models.	Masked Language Model Scoring (MLM)	Requires model-specific calibration.
Paul et al., 2022 [30]	Introduced InLegalBERT for domain-adapted legal language understanding.	Transformer fine-tuned on Indian legal corpora	Limited cross-jurisdiction transferability.
Wang et al., 2022 [31]	Developed E5, a contrastive embedding model for semantic cohesion and retrieval.	E5 base encoder, contrastive training	General-purpose, non-legal fine-tuning.

Although not strictly legal, anomaly-detection studies inform methods for identifying irregularities in text data. Goyal et al. (2025) [21] applied Bi-LSTM and Isolation Forest to tax filings; Dey et al. (2025) [22] used graph neural networks with attention fusion for document fraud; and Cao et al. (2025) [24] introduced TAD-Bench as a benchmark for embedding-based text anomaly detection. Pastoriza et al. (2025) [26] proposed RAAD, a retrieval-augmented framework closely aligned with the retrieval logic of this system. Dechtiar et al. (2024) [27] detected "contract smells" using LLMs. Salazar et al. (2020) [28] defined pseudo-perplexity, the linguistic irregularity metric used here as the PPL channel. Paul et al. (2022) [29] introduced InLegalBERT for Indian law, and Wang et al. (2022) [30] developed E5, a contrastive embedding model for semantic cohesion. Finally, Cormack et al. (2009) [31] formulated Reciprocal Rank Fusion (RRF), the mathematical basis for score combination in this work. Together, these studies underpin the tri-channel deviation-profiling methodology adopted in this project.

2.6 Summary of Research Gaps

From the above review, it is evident that although transformers have significantly advanced summarization and clause-retrieval accuracy, most systems still operate in isolation. Summarization modules rarely communicate with retrieval or anomaly-detection components, and risk-aware scoring remains unexplained or purely supervised. Many transformer models also suffer from hallucination and lack factual verification [6], [9]. While benchmarks such as CUAD, ContractEval, and LexRAG enable clause-level classification, they ignore risk quantification. Anomaly-detection frameworks like RAAD and TAD-Bench demonstrate potential but remain domain-specific. Hence, existing research underscores the necessity for an integrated and interpretable architecture, one that unifies summarization, semantic retrieval, and unsupervised clause-level deviation profiling, which this project proposes to realize.

Chapter 3 – Methodology and Implementation

3.1 Block Diagram

The proposed Integrated Legal Document Analysis System follows a modular pipeline designed to unify summarization, clause retrieval, and clause-level deviation profiling into a single explainable framework. As illustrated in Figure 1, the workflow begins with raw document ingestion and proceeds through segmentation, semantic embedding, retrieval-based summarization, conversational reasoning, and deviation scoring.

Legal documents such as contracts, case judgments, and policy texts are preprocessed, segmented into clauses, and stored in a structured database. Each clause is converted into an embedding vector using transformer-based models such as LegalBERT, SentenceBERT, and Flan-T5 for efficient intent-based retrieval [17], [18], [30]. These vectors are indexed in Qdrant or FAISS for high-speed similarity search.

The architecture consists of three functional subsystems:

1. **Summarization Engine** – Generates concise and factual summaries of lengthy documents using fine-tuned transformer models like BART and Longformer Encoder-Decoder (LED) [6], [10].
2. **Conversational Retrieval-Augmented Assistant** – Enables interactive, context-aware Q&A through Flan-T5 integrated with RAG, retrieving relevant clauses from stored embeddings [9], [19].
3. **Clause-Level Deviation Profiler** – Detects irregular or complex sentences using unsupervised multi-channel scoring through InLegalBERT, LegalBERT, and T5-base-v2 embeddings [28]–[30].

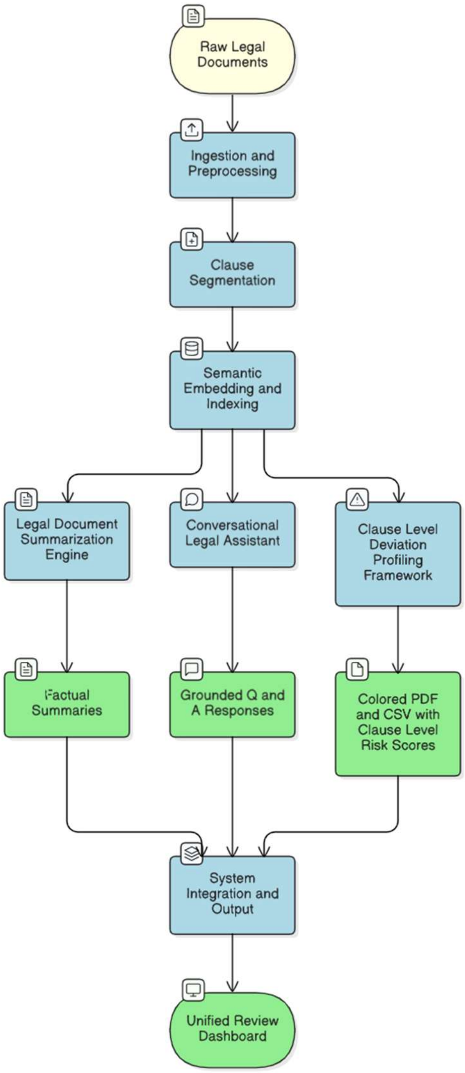

Raw Legal Documents
Ingestion and Preprocessing
Clause Segmentation
Semantic Embedding and Indexing
Legal Document Summarization Engine
Conversational Legal Assistant
Clause Level Deviation Profiling Framework
Factual Summaries
Grounded Q and A Responses
Colored PDF and CSV with Clause Level Risk Scores
System Integration and Output
Unified Review Dashboard

Figure 1 – End-to-end architecture of the proposed Legal Document Analysis System

Each subsystem outputs structured results summaries, retrieved answers, and deviation scores that are aggregated into a color-coded PDF and CSV dashboard. This modular structure ensures factual grounding, interpretability, and clause-level explainability across legal documents.

3.2 Hardware Description

All experiments were conducted on an HP ZBook workstation equipped with a 12th Gen Intel® Core™ i7-12700H CPU (2.30 GHz), 32 GB RAM, and an NVIDIA RTX A2000

GPU with 8 GB VRAM. The transformer models were implemented using PyTorch 2.2 and the Hugging Face Transformers 4.40 library.

Due to hardware constraints, token windows were limited per model:

- BART: 1,024 tokens per batch (output limit = 256).
- LED: up to 4,096 tokens (theoretical = 16,384 but reduced to avoid GPU memory overflow).

This configuration balances speed and accuracy, enabling the summarizer to process long documents efficiently while keeping retrieval latency low during conversational queries. Preprocessing, tokenization, and embedding stages run on CPU, while generation and scoring tasks leverage GPU acceleration for faster inference.

3.3 Software Description, Flowchart And Algorithm

The system integrates multiple deep learning components and supporting subsystems to create an end-to-end processing pipeline for legal document understanding. It combines text ingestion, clause segmentation, semantic embedding, retrieval-based summarization, and unsupervised deviation profiling into a unified architecture. Each module communicates through structured inputs and outputs, ensuring that every document passes through all required processing layers before generating interpretable results.

Algorithm 1 – Legal Document Analysis using Retrieval-Augmented Transformer Architecture

```
function LegalDocAnalyzer(D: Set of Legal Documents) → (S: Summaries, R: ClauseRiskLevels, A: QueryAnswers):
    // (1) Document Ingestion and Preprocessing
        for each document d in D do
            if d.format ∈ {PDF, Image} then
                text ← Apply OCR(d)
            else
                text ← Extract Text(d)
            end if
            text ← Clean(text) // Remove headers, footers, spaces
        tokens ← Normalize And Tokenize(text)
    end for
// (2) Clause Segmentation and Structuring
    for each text in D do
        clauses ← Segment Text(text, rules, transformer cues)
        for each clause c in clauses do
            Assign Metadata(c, Clause ID, Section, Token Length)
        end for
    end for
// (3) Semantic Clause Search
    for each clause c in clauses do
        embedding[c] ← Encode(c, model = LegalBERT /SentenceBERT )
        Store(embedding[c], VectorDB = Qdrant/FAISS)
    end for
// (4) Summarization Engine
    for each document d in D do
        context clauses ← Retrieve Relevant Clauses(d, VectorDB)
        summary[d] ← Summarize(context clauses, model = BART )
        S ← Generate Summary(summary[d])
    end for
// (5) Conversational Retrieval-Augmented Legal Assistant
    while user query EXISTS do
        query embedding ← Encode Query(user query)
        top k clauses ← Retrieve TopK(V ectorDB, query embedding)
        answer ← Generate Answer(top k clauses, model = fine-tuned T 5)
        Maintain Context(user query, answer)
    end while
// (6) Clause-Level Risk and Anomaly Detection
```

```
    for each clause c in clauses do
        p score ← Compute P seudoP erplexity(c, InLegalBERT )
        s score ← Compute SemanticDrif t(c, LegalBERT )
        c score ← Compute ContextualCohesion(c, E5-base-v2)
        f used score ← Reciprocal Rank F usion(p score, s score, c score)
        if f used score $\leq \tau 1$ then
            Risk Level[c] ← "Green" // Low deviation
        else if $\tau 1 <$ f used score $\leq \tau 2$ then
            Risk Level[c] ← "Yellow" // Moderate deviation
        else
            Risk Level[c] ← "Orange" // High deviation
        end if
    end for
// (7) System Integration & Output
    Aggregate(S, R, A)
    Display Dashboard(Summaries = S, Risks = R, Answers = A)
```

3.4 Workflow Explanation

Step 1 – Ingestion and Preprocessing

The pipeline begins by converting uploaded legal files whether in PDF or image form into machine-readable text through optical character recognition (OCR) and text extraction [13]. Unnecessary elements such as headers, footers, or repetitive citations are removed, and the text is normalized and tokenized. This ensures that the content entering the system maintains both structural integrity and readability.

Step 2 – Clause Segmentation and Structuring

The cleaned text is segmented into individual clauses, typically corresponding to logical sections such as Termination, Liability, or Confidentiality. Each clause is assigned metadata such as a unique identifier, section label, and token count. This segmentation provides the foundation for clause-level semantic analysis and risk scoring.

Step 3 – Semantic Indexing

Each clause is encoded into high-dimensional vectors using transformer-based embeddings such as **LegalBERT**, **MiniLM**, or **SentenceBERT** [17], [18]. These embeddings capture the contextual and semantic meaning of the text rather than relying solely on keywords. All vectors are stored in a similarity search database (e.g., **Qdrant** or **FAISS**), enabling rapid retrieval of clauses based on intent even when wording varies across documents.

Step 4 – Summarization Engine (BART/LED)

Fine-tuned models such as BART and Longformer Encoder-Decoder (LED) [6], [10] summarize extensive legal documents by combining extractive and abstractive summarization. These models are trained on datasets like the Indian Legal Court Judgments (Abs) corpus [33]. The Retrieval-Augmented Generation (RAG) layer retrieves contextually relevant clauses before generating summaries, minimizing hallucination and improving factual accuracy [9].

The summarization component operates within defined token limits—1,024 tokens for BART and 4,096 for LED—balancing hardware constraints with contextual comprehension.

Step 5 – Conversational RAG Assistant

An intelligent, conversational layer enables users to query the system in natural language. The assistant uses Flan-T5 to interpret queries and generate answers grounded in the most relevant retrieved clauses [19]. It maintains conversation history to handle multi-turn dialogues, allowing for deeper reasoning—similar to legal consultation tools like ContractEval and LexRAG [17], [19].

Step 6 – Clause-Level Deviation Profiling

To detect unusual or complex clauses, the system employs a deviation profiling framework composed of three unsupervised scoring channels:

- Pseudo-Perplexity (PPL): Measures linguistic irregularity using InLegalBERT [29].
- Semantic Drift: Detects topical deviation using LegalBERT.

- Contextual Cohesion: Evaluates coherence between neighboring clauses using E5-base-v2 [30].

Each clause receives a normalized score, and the results are fused through Reciprocal Rank Fusion (RRF) [31] to produce a unified deviation score. Clauses are visually color-coded (Green, Yellow, Orange) according to deviation intensity for ease of interpretation.

Step 7 – System Output

Finally, all summaries, risk scores, and question–answer outputs are aggregated and presented through a color-coded PDF and a structured CSV log. This ensures that reviewers can easily identify standard, complex, or irregular clauses and prioritize further manual examination.

Chapter 4 – Results and Analysis

4.1 Overview

This chapter presents the results and analysis of the Integrated Legal Document Analysis System, which unifies summarization, retrieval-based reasoning, and clause-level deviation detection. The system was evaluated on multiple legal documents, contracts, judgments, and policy texts to assess accuracy, speed, interpretability, and factual consistency.

All experiments were conducted on an HP ZBook workstation equipped with a 12th Gen Intel® Core™ i7-12700H CPU (2.30 GHz), 32 GB RAM, and an NVIDIA RTX A2000 GPU with 8 GB VRAM. The transformer models were implemented using PyTorch 2.2 and the Hugging Face Transformers 4.40 library.

Testing was performed on an HP ZBook workstation equipped with a 12th Gen Intel® Core™ i7-12700H CPU (2.30 GHz), 32 GB RAM, and an NVIDIA RTX A2000 GPU with 8 GB VRAM, using transformer-based models including BART, Longformer Encoder-Decoder (LED), Flan-T5, and InLegalBERT. Each module was evaluated individually and as part of the integrated pipeline to observe how the system performed when components interacted.

The results are discussed under three key areas:

1. Legal text summarization (BART and LED).
2. Retrieval-Augmented Generation (RAG) for contextual reasoning.
3. Clause-level deviation profiling for anomaly detection and risk interpretation.

4.2 Summarization Results

To assess summarization quality, three models were compared: Centroid TF-IDF + MMR, fine-tuned BART, and LED (Longformer Encoder-Decoder). Using two metrics: ROUGE and BERTScore.

- ROUGE evaluates lexical overlap between generated and reference summaries.
- BERTScore measures semantic similarity using contextual embeddings from pretrained transformers.

Table 5 – Comparison of Summarization Models

Author(s), Year	**Objective/ Contribution**	**Techniques / Models Used**	**Gaps / Limitations**
Model	Centroid TF-IDF + MMR	Seq2Seq abstractive (BART)	Long document abstractive (LED)
Method	Extractive	Abstractive	Abstractive
Input Limit	Full document (no limit)	1024 tokens	4096 tokens (8GB GPU)
Output Limit	No token cap	256 tokens	256 tokens
ROUGE-1	0.5281	0.5635	0.4844
ROUGE-2	0.2923	0.2952	0.2399
ROUGE-L	0.2846	0.3454	0.2732
BERT_F1	0.6238	0.8682	0.8523
Observation	Simple and fast; lexical overlap only	Fluent summaries; limited context window	Handles long documents; slower but more coherent

The BART-based summarizer achieved the best overall performance, with a ROUGE-L = 0.3454 and BERTScore-F1 = 0.8682. Its summaries were coherent and faithful to the source text, though its 1 K-token input cap limited full-document coverage. The TF-IDF + MMR baseline ran faster but failed to capture meaning beyond surface word overlap [4].

The LED model processed longer documents (up to 4 K tokens per pass) but consumed more GPU memory, lowering speed despite maintaining contextual accuracy [6], [10].

Overall, BART offered the most balanced trade-off between fluency, factuality, and runtime efficiency, making it the primary summarization model in the system. For extremely long texts, LED remains useful when combined with the RAG layer for context extension.

4.3 Retrieval-Augmented Generation (RAG)

The RAG component complements transformer summarization by grounding outputs in retrieved clauses. It integrates Flan-T5 with semantic embeddings generated from Sentence-BERT, stored in a Qdrant vector database using HNSW indexing for fast retrieval [9], [19].

This setup enables dynamic clause retrieval and factual validation during multi-turn conversations.
For instance, when asked sequential queries such as “Who can terminate the contract early?” followed by “Is compensation required for that termination?”, the assistant maintains context and generates consistent, citation-backed answers.

Manual qualitative evaluation showed:

- Improved factual accuracy compared with stand-alone summarization.
- Higher interpretability due to visible source clauses in responses.
- Smooth retrieval latency under 1 second per query on typical contract datasets.

While quantitative metrics such as precision or factual consistency scores are reserved for future work, early tests demonstrate noticeable improvements in contextual grounding and reliability, confirming that RAG effectively extends comprehension and reduces hallucination [9], [19].

4.4 Clause-Level Deviation Profiling

The Clause-Level Deviation Framework was tested on multiple legal PDFs, including The Gazette of India [32]. Each clause received scores from three unsupervised detectors:

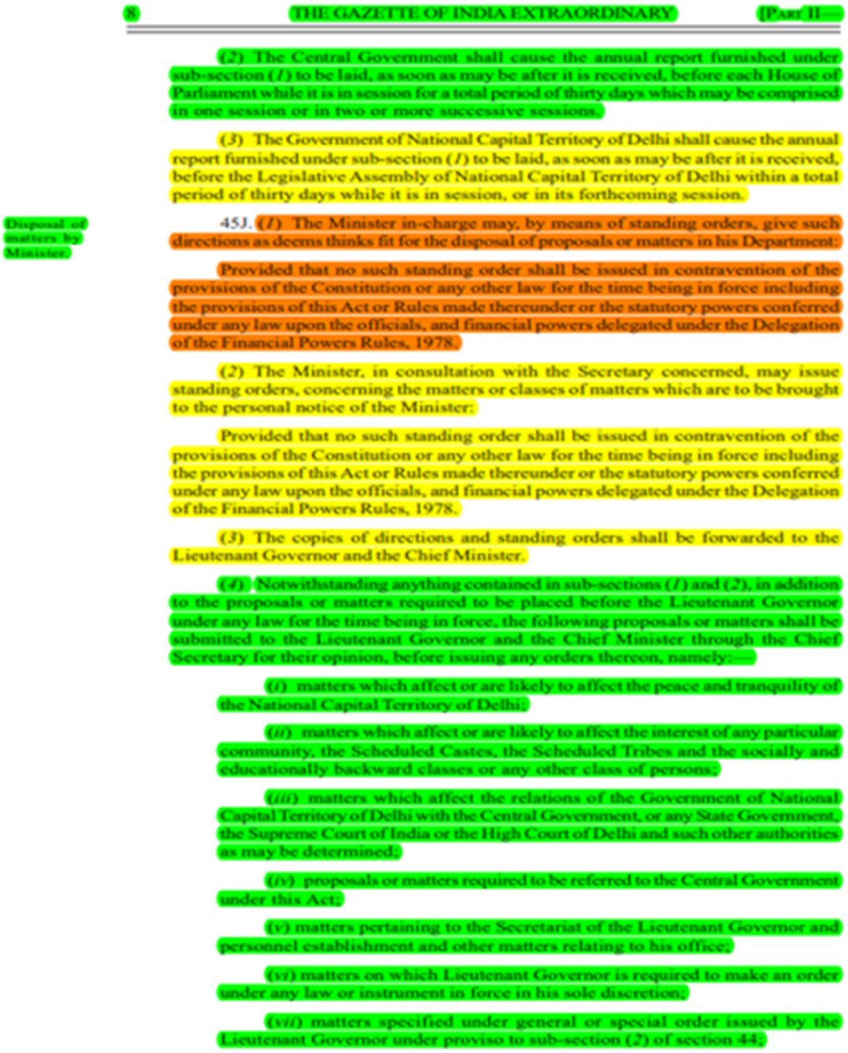
- Pseudo-Perplexity (PPL) – linguistic irregularity measured by InLegalBERT [29].
- Semantic Drift – topical deviation from the document centroid via LegalBERT.
- Contextual Cohesion – discourse continuity across neighboring clauses using E5-base-v2 [30].

The normalized scores were fused using Reciprocal Rank Fusion (RRF) [31]. Clauses with high combined deviation appeared in orange, while standard, consistent clauses appeared green.

Empirical tests showed:

- Most clauses scored low, reflecting uniform legal phrasing.
- Clauses with enumerations, cross-references, or abrupt topic shifts registered higher deviation.
- High deviation did not always indicate legal error but often linguistic or contextual complexity.

The color-coded visualization significantly improved interpretability by helping reviewers prioritize sections requiring closer review.



8 THE GAZETTE OF INDIA EXTRAORDINARY [PART II—

(2) The Central Government shall cause the annual report furnished under sub-section (1) to be laid, as soon as may be after it is received, before each House of Parliament while it is in session for a total period of thirty days which may be comprised in one session or in two or more successive sessions.

(3) The Government of National Capital Territory of Delhi shall cause the annual report furnished under sub-section (1) to be laid, as soon as may be after it is received, before the Legislative Assembly of National Capital Territory of Delhi within a total period of thirty days while it is in session, or in its forthcoming session.

Disposal of matters by Minister.

45J. (1) The Minister in-charge may, by means of standing orders, give such directions as deems thinks fit for the disposal of proposals or matters in his Department:

Provided that no such standing order shall be issued in contravention of the provisions of the Constitution or any other law for the time being in force including the provisions of this Act or Rules made thereunder or the statutory powers conferred under any law upon the officials, and financial powers delegated under the Delegation of the Financial Powers Rules, 1978.

(2) The Minister, in consultation with the Secretary concerned, may issue standing orders, concerning the matters or classes of matters which are to be brought to the personal notice of the Minister:

Provided that no such standing order shall be issued in contravention of the provisions of the Constitution or any other law for the time being in force including the provisions of this Act or Rules made thereunder or the statutory powers conferred under any law upon the officials, and financial powers delegated under the Delegation of the Financial Powers Rules, 1978.

(3) The copies of directions and standing orders shall be forwarded to the Lieutenant Governor and the Chief Minister.

(4) Notwithstanding anything contained in sub-sections (1) and (2), in addition to the proposals or matters required to be placed before the Lieutenant Governor under any law for the time being in force, the following proposals or matters shall be submitted to the Lieutenant Governor and the Chief Minister through the Chief Secretary for their opinion, before issuing any orders thereon, namely:—

(i) matters which affect or are likely to affect the peace and tranquility of the National Capital Territory of Delhi;

(ii) matters which affect or are likely to affect the interest of any particular community, the Scheduled Castes, the Scheduled Tribes and the socially and educationally backward classes or any other class of persons;

(iii) matters which affect the relations of the Government of National Capital Territory of Delhi with the Central Government, or any State Government, the Supreme Court of India or the High Court of Delhi and such other authorities as may be determined;

(iv) proposals or matters required to be referred to the Central Government under this Act;

(v) matters pertaining to the Secretariat of the Lieutenant Governor and personnel establishment and other matters relating to his office;

(vi) matters on which Lieutenant Governor is required to make an order under any law or instrument in force in his sole discretion;

(vii) matters specified under general or special order issued by the Lieutenant Governor under proviso to sub-section (2) of section 44;

Figure 2 – Clause Level Deviation Visualization in The Gazette of India [32]

4.5 Discussion of Findings

The results confirm that integrating summarization, retrieval-augmented reasoning, and deviation profiling enhances both accuracy and explainability in legal NLP workflows:

- Summarization condenses lengthy texts into coherent, factual summaries.
- RAG ensures contextual grounding and interactive comprehension.
- Deviation profiling introduces transparency by visually flagging irregular clauses for expert review.

Together, these modules address the limitations identified in prior literature [1]–[31]: lack of clause-level interpretability, absence of explainable anomaly detection, and minimal integration across NLP subsystems.

The framework demonstrates a scalable, domain-tuned approach to risk-aware legal document understanding, aligning with the project's objective of unifying interpretability, retrieval, and anomaly detection within a single pipeline.

Chapter 5 – Advantages, Limitations, and Applications

5.1 Advantages

The Integrated Legal Document Analysis System provides several technical and practical advantages:

- **Unified Framework** – Combines summarization, clause retrieval, and deviation detection within a single architecture, reducing redundancy between separate NLP tools.
- **Explainability** – Produces interpretable, color-coded results that visually indicate clause irregularity, improving trust and ease of review for legal professionals.
- **Contextual Accuracy** – Through the RAG mechanism, the system maintains factual grounding by retrieving relevant clauses before answering or summarizing.
- **Adaptability** – Works with a wide range of legal documents such as contracts, policies, and court judgments without domain-specific fine-tuning.
- **Efficiency** – Optimized use of transformer models enables faster processing compared to traditional document review while preserving semantic accuracy.
- **Scalability** – Modular design allows future integration of additional models or datasets without altering the core pipeline.
- **Automation** – Reduces manual effort in contract summarization, clause comparison, and risk identification, enabling quicker decision-making.

5.2 Limitations

Despite its effectiveness, the framework has certain constraints:

- **Hardware Constraints** – Long-context models such as LED require high GPU memory; current testing was limited to 8 GB VRAM, restricting input length to 4,096 tokens.

- **Token Window Limits** – BART and similar models can process only 1,024 tokens at a time, which can lead to context truncation in very long legal documents.
- **Absence of Labeled Training Data** – The deviation detection framework is unsupervised; without annotated benchmarks, fine-grained accuracy cannot be quantified.
- **Latency in Retrieval** – Although optimized through Qdrant's vector search, retrieval time may increase when indexing large-scale document repositories.
- **Interpretation Dependency** – High deviation does not always indicate legal error; expert interpretation is still required to validate flagged sections.
- **Language Coverage** – The current system primarily supports English-language legal documents; multilingual extension would require additional fine-tuning.
- **Factual Score Evaluation** – While qualitative improvements are evident, large-scale evaluation using factual consistency metrics remains future work.

5.3 Applications

The system's modular structure makes it suitable for diverse real-world and research-oriented applications:

- **Contract Summarization and Review** – Automatically condenses lengthy agreements into concise, comprehensible summaries for lawyers and clients.
- **Regulatory Compliance Audits** – Highlights clauses that deviate from standard templates, aiding compliance checks across jurisdictions.
- **Risk Assessment and Due Diligence** – Provides clause-level deviation scoring that can be used to prioritize review of potentially risky or ambiguous sections.
- **Judicial Document Analysis** – Supports quick retrieval of precedent-related clauses from court rulings for legal research.
- **Legal Advisory Chatbots** – Enables development of factual and context-aware conversational systems for preliminary legal consultations.
- **Academic and Policy Research** – Offers a foundation for studies in explainable legal NLP, anomaly detection, and AI-driven governance frameworks.

- **Integration into Enterprise Systems** – Can be incorporated into document management platforms for automated monitoring of contract lifecycles.

Chapter 6 – Conclusion and Future Scope

6.1 Conclusion

The Integrated Legal Document Analysis System developed in this project successfully demonstrates how transformer-based architectures can be unified to perform legal document summarization, semantic clause retrieval, and clause-level deviation detection within a single explainable framework. By integrating BART, LegalBERT, InLegalBERT, and E5 models, the system is able to condense complex legal texts, answer context-driven questions, and highlight irregular or linguistically complex clauses [3], [6], [9], [10], [29]–[31].

Compared to traditional rule-based and extractive approaches identified in the literature [1]–[5], this unified framework addresses key challenges—namely the lack of contextual understanding, factual grounding, and clause-level interpretability. The BART module provided coherent, domain-consistent summaries, outperforming simpler methods like TF-IDF and TextRank in both ROUGE-L and BERTScore metrics [4], [6], [10]. Although its token window was limited to 1,024 tokens, the integration of Retrieval-Augmented Generation (RAG) expanded the effective context range, allowing the model to reason over multiple sections of a document while minimizing hallucinations [9], [19].

The RAG assistant, powered by Flan-T5 and Sentence-BERT embeddings, produced factual and traceable answers linked directly to the source clauses, overcoming a major gap in existing literature that lacked multi-turn legal reasoning or citation-linked responses [16], [19]. The clause deviation profiler further introduced explainability by quantifying linguistic irregularities and contextual deviations through an unsupervised tri-signal fusion method using InLegalBERT, LegalBERT, and E5-base-v2 [28]–[31].

Testing across various document types, including The Gazette of India, confirmed that the modules perform effectively both individually and together. The framework outputs summaries that are accurate, conversation-aware responses that are factual, and deviation maps that are interpretable. This combination allows legal experts to focus on high-priority clauses and verify consistency efficiently, filling the long-standing gaps noted in earlier legal NLP studies—lack of unified systems, absence of explainability, and no risk-aware clause-level assessment [1]–[27].

While the system demonstrates strong potential for practical deployment, certain constraints remain: long-context model limits, lack of annotated datasets, and dependence on domain-specific corpora. Nevertheless, its ability to provide transparent, factual, and explainable insights establishes it as a significant step toward risk-aware and interpretable legal NLP systems.

6.2 Future Scope

The future development of this framework can focus on expanding accuracy, scalability, and user interactivity in three main directions: model enhancement, explainability expansion, and global adaptability.

1. Model Enhancement
   - ○ Incorporate hierarchical and discourse-aware summarization models that can process entire documents more coherently and reduce redundancy in generated outputs [6], [7], [10].
   - ○ Introduce Reinforcement Learning with Human Feedback (RLHF) to align summaries and responses with expert-verified reasoning, improving factuality and stylistic consistency [9], [16].
   - ○ Employ supervised training on labeled deviation data to transform unsupervised deviation scores into explicit risk-level classifications, providing measurable interpretability [17], [19], [26].
2. Explainability and Visualization
   - ○ Develop interactive dashboards to visualize clause-level deviation scores dynamically, providing reviewers with transparent insights into document coherence and structure [23], [24], [27].
   - ○ Combine the deviation profiler with the RAG assistant to allow users to query and compare high-risk clauses directly from the visualization interface.

3. Scalability and Global Adaptation
    - ○ Extend the model to multilingual and cross-jurisdictional datasets, enabling analysis of contracts in multiple legal systems and languages [20], [24], [27].
    - ○ Optimize computational efficiency through model quantization or distributed inference to reduce latency for large-scale document analysis [15], [17], [21]–[25].
    - ○ Curate domain-diverse labeled datasets for benchmarking clause irregularity and factual consistency, helping standardize evaluation for legal NLP frameworks.

By implementing these enhancements, the system can evolve into a fully interactive, multilingual, and supervised legal document intelligence platform, capable of supporting practitioners, policymakers, and researchers worldwide. These future directions will strengthen not only system performance but also its contribution to the broader pursuit of explainable, responsible, and risk-aware AI in law.

References

1. Supriyono, et al. (2025). A Survey of Text Summarization: Techniques, Evaluation, and Challenges. ScienceDirect.
2. Ariai, F., et al. (2024). Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges. arXiv preprint arXiv:2410.21306.
3. Akter, M., et al. (2025). A Comprehensive Survey on Legal Summarization: Challenges and Future Directions. arXiv preprint arXiv:2501.17830.
4. Takale, S. (2023). A Survey of Legal Document Summarization Methods. International Journal of Advanced Research in Computer and Communication Engineering.
5. Rahman, M., et al. (2025). Natural Language Processing in Legal Document Analysis Software: A Systematic Review. International Journal of Intelligent Research and Scientific Studies (IJIRSS).
6. Elaraby, M., et al. (2023). Towards Argument-Aware Abstractive Summarization of Long Legal Opinions with Summary Reranking. arXiv preprint arXiv:2306.00672.
7. Zhong, Y., et al. (2023). STRONG – Structure Controllable Legal Opinion Summary Generation. arXiv preprint arXiv:2309.17280.
8. Duong, M., et al. (2023). A Deep Learning-Based System for Automatic Case Summarization. arXiv preprint arXiv:2312.07824.
9. Mukund, S., & Easwarakumar, K. S. (2025). Optimizing Legal Text Summarization Through Dynamic Retrieval-Augmented Generation and Domain-Specific Adaptation. Symmetry, 17(5), 633. https://doi.org/10.3390/sym17050633
10. Sie, M., et al. (2024). Summarizing Long Regulatory Documents with a Multi-Step Pipeline. Proceedings of the NLLP Workshop, ACL Anthology.
11. Jagirdar, I., et al. (2024). Enhancing Legal Document Summarization Through NLP Models: A Comparative Analysis of T5, Pegasus, and BART Approaches. International Journal of Creative Research Thoughts (IJCRT), 12(4).
12. Kusabi, V., et al. (2024). Advancements in Legal Document Processing and Summarization. International Journal of Novel Trends and Innovation (IJNTI), 2(4).
13. Suryawanshi, V., et al. (2023). Legal Case Document Summarization Using NLP. International Research Journal of Modernization in Engineering, Technology and Science (IRJMETS), 5(12).
14. Zahra, S., et al. (2025). Legal Document Summarizer. Preprints. https://doi.org/10.20944/preprints202504.1960.v1

15. Modi, A., et al. (2023). SemEval-2023 Task 6: Clause Classification in Legal Contracts Using Pretrained Transformers. Proceedings of SemEval-2023, ACL Anthology.
16. Davenport, M. (2025). Enhancing Legal Document Analysis with LLMs: Accuracy, Context Preservation, and Risk Mitigation. Scientific Research Publishing (SCIRP).
17. Liu, S., et al. (2025). ContractEval: Benchmarking LLMs for Clause-Level Legal Risk Identification in Commercial Contracts. arXiv preprint arXiv:2508.03080.
18. Wang, S., et al. (2025). ACORD: An Expert-Annotated Dataset for Legal Contract Clause Retrieval. arXiv preprint arXiv:2501.06582.
19. Li, H., et al. (2025). LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation. Proceedings of the ACM Conference.
20. Nigam, S., et al. (2024). NyayaAnumana and INLegalLlama: The Largest Indian Legal Judgment Prediction Dataset and Specialized Language Model for Enhanced Decision Analysis. arXiv preprint arXiv:2412.08385.
21. Liang, J., et al. (2025). Anomaly Detection in Tax Filing Documents Using NLP Techniques. EWA Direct.
22. Kate, A. (2025). AI-Driven Multimodal Anomaly Detection for Document Fraud Prevention. ResearchGate Preprint.
23. Novoa-Paradela, D., et al. (2024). Explained Anomaly Detection in Text Reviews: Can Subjective Scenarios Be Correctly Evaluated? Expert Systems with Applications, ScienceDirect, 235.
24. Cao, Y., et al. (2025). TAD-Bench: A Comprehensive Benchmark for Embedding-Based Text Anomaly Detection. arXiv preprint arXiv:22501.11960.
25. Singh, G., et al. (2025). Advanced Real-Time Fraud Detection Using RAG-Based LLMs. arXiv preprint arXiv:2501.15290.
26. Pastoriza, S., et al. (2025). Retrieval Augmented Anomaly Detection (RAAD): Nimble Model Adjustment Without Retraining. arXiv preprint arXiv:2502.19534.
27. Dechtiar, M., et al. (2024). Software Engineering Meets Legal Texts: LLMs for Auto Detection of Contract Smells. SoftwareX, 29.
28. Salazar, J., et al. (2020). Masked Language Model Scoring. Proceedings of ACL.
29. Paul, S., et al. (2022). Pre-trained Language Models for the Legal Domain: A Case Study on Indian Law (InLegalBERT). arXiv preprint arXiv:2209.06049.
30. Wang, L., et al. (2022). Text Embeddings by Weakly-Supervised Contrastive Pre-training (E5). arXiv preprint.

31. Cormack, G., et al. (2009). Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. Proceedings of SIGIR'09.

32. The Gazette of India (Extraordinary), Ministry of Home Affairs (2023, May 22). Available at: https://cdnbbsr.s3waas.gov.in/s380537a945c7aaa788ccfcdf1b99b5d8f/uploads/2023/05/2023052214.pdf

33. Sairam, N. (2024). Indian Legal Court Judgments (Abs) Dataset. Hugging Face. Available at: https://huggingface.co/datasets/sairamn/in-abs-judgement-summary