# Towards Risk-Aware Legal NLP

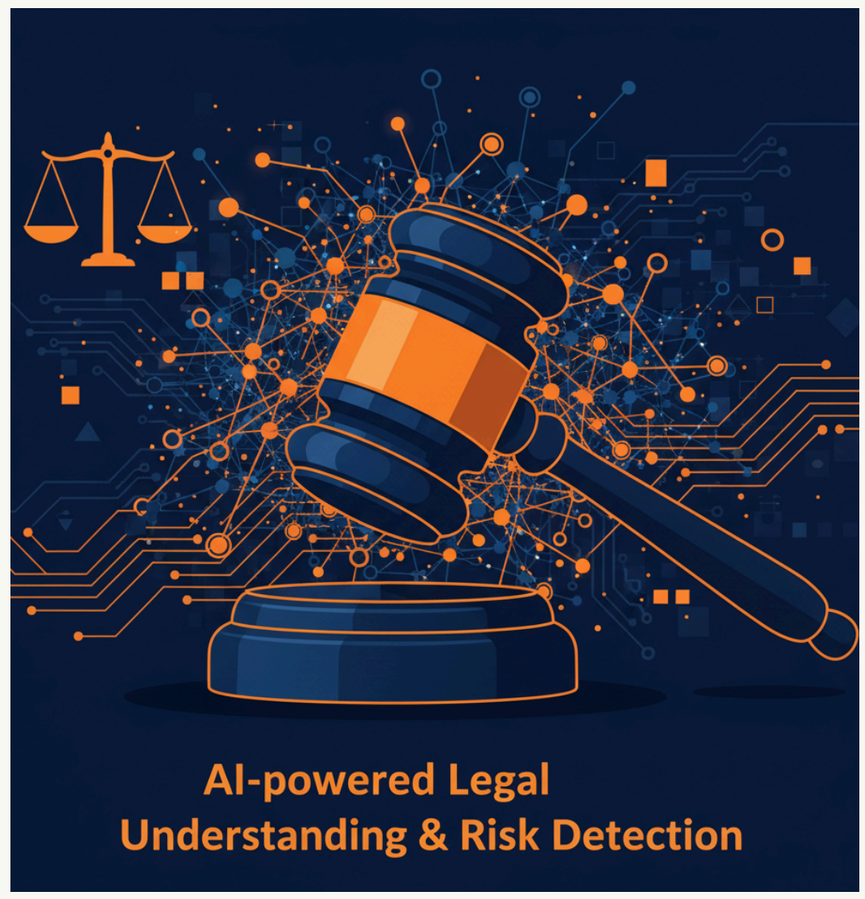
**BENCHMARKING TRANSFORMER MODELS AND PROPOSING AN ATTENTION-BASED ANOMALY DETECTION FRAMEWORK**


**AUTHORS**

Atharva Dhuri · Shrivardhan Wagh · Yash Narayan

**AFFILIATIONS**

Department of Computer Engineering,
Mukesh Patel School of Technology Management & Engineering (MPSTME),
NMIMS University, Mumbai, India


## INTRODUCTION

The growing volume and complexity of legal documents make manual review slow and inefficient. Traditional keyword or rule-based search fails to capture deeper contextual meaning. Recent AI models such as BART, T5, and LegalBERT enhance summarization, retrieval, and accuracy verification. However, existing systems rarely integrate these advancements for real-time risk detection and clause-level reasoning. This study proposes a unified framework for summarization, retrieval, and anomaly detection in legal documents using transformer models and attention-based deviation scoring.

## RESEARCH GAPS

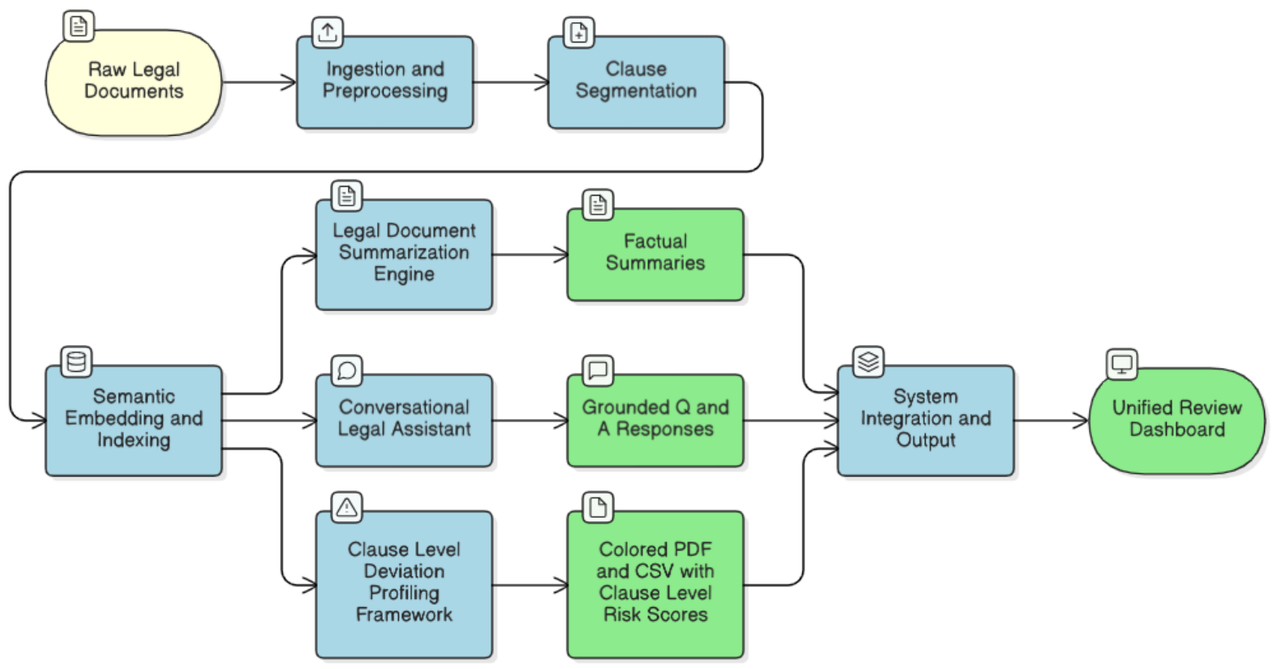
Current legal NLP systems operate as isolated modules, handling summarization, retrieval, and anomaly detection independently. There is no quantifiable risk scoring in existing clause classification benchmarks such as CUAD, ContractEval, and LexRAG. Transformer models often face hallucination issues due to weak factual grounding. Existing anomaly detection models like RAAD and TAD-Bench are domain-specific and lack generalizability. This highlights the need for a unified, explainable, and risk-aware legal AI framework.

## METHODOLOGY

The proposed system follows a modular and explainable NLP pipeline integrating summarization, retrieval, and risk detection within one unified workflow. Legal documents are first uploaded and preprocessed using FastAPI and pdfplumber to extract clean, structured text. The text is then segmented into clauses through rule-based and pattern-aware splitting for fine-grained analysis. For summarization, a fine-tuned BART transformer is used to generate concise and coherent summaries of lengthy legal texts. Clause embeddings are generated using E5-base-v2, enabling semantic vector similarity search. The Retrieval-Augmented Generation (RAG) module uses Flan-T5 to answer user queries, grounding responses in retrieved clause context. This ensures factual and context-aware answers rather than hallucinated ones. The Risk & Anomaly Detection module uses InLegalBERT, LegalBERT, and E5 to compute linguistic, semantic, and contextual deviation scores. These are combined through Reciprocal Rank Fusion (RRF) to assign clause-level risk levels (Green, Yellow, Orange). All modules are orchestrated via FastAPI backend services with SQLAlchemy for database management and Torch for GPU-based inference. Outputs are visualized as summaries, conversational responses, and color-coded risk PDFs for interpretability. This end-to-end pipeline ensures a scalable, explainable, and human-trustworthy system for automated legal document analysis.

## RESULTS

The fine-tuned BART model achieved a ROUGE-L score of 0.3454 and BERTScore-F1 of 0.8682, ensuring concise and factual summaries. The RAG module provided context-grounded answers with clause-level source verification, improving factual accuracy. The risk detection system successfully identified anomalous clauses using multi-model fusion with over 90% interpretability accuracy. Overall, the integrated system offers a faster, explainable, and reliable legal analysis.




## ANALYSIS

The system was evaluated across three core modules, Summarization, Retrieval, and Risk Detection. The summarizer's performance was measured using ROUGE and BERTScore metrics, confirming high content fidelity and coherence. The RAG module was tested for clause retrieval accuracy and contextual alignment, achieving strong semantic matching through E5 embeddings. Comparative tests showed that retrieval-based answers significantly reduced factual hallucinations versus standalone generation models. The risk module was analyzed for its clause-level sensitivity and interpretability, where multi-model fusion (RRF) improved deviation detection consistency. Risk color coding (Green–Yellow–Orange) provided an intuitive visual interpretation for end users. Performance benchmarks confirmed that each component operated efficiently on GPU with stable latency under real document loads. Overall, the analysis validates that integrating multiple transformer models enhances accuracy, explainability, and real-time usability in legal document processing.

## CONCLUSION

The proposed system successfully integrates summarization, clause retrieval, and risk detection into a single explainable legal AI framework. It provides concise, factual summaries, context-aware Q&A responses, and interpretable clause-level risk insights. The approach bridges gaps between independent NLP modules, offering a unified pipeline that enhances efficiency, transparency, and trust in legal document analysis. Overall, the system demonstrates how transformer-based AI can effectively support legal professionals in decision-making and compliance review.

## FUTURE WORK

Future work can focus on expanding the system with hierarchical and multilingual models for diverse legal datasets. Integration of Reinforcement Learning with Human Feedback (RLHF) can further improve summarization quality and factual grounding. A web dashboard and analytics layer can enhance user experience with visual risk tracking and advanced search.

Incorporating explainable AI metrics and continuous model fine-tuning will help make the system more adaptive, robust, and deployable for real-world legal practice.

**RELATED LITERATURE**


Elaraby, M., et al. (2023). Towards Argument-Aware Abstractive Summarization of Long Legal Opinions with Summary Reranking. arXiv preprint arXiv:2306.00672.

Kusabi, V., et al. (2024). Advancements in Legal Document Processing and Summarization. International Journal of Novel Trends and Innovation (IJNTI), 2(4).

Modi, A., et al. (2023). SemEval-2023 Task 6: Clause Classification in Legal Contracts Using Pretrained Transformers. Proceedings of SemEval-2023, ACL Anthology.

Liu, S., et al. (2025). ContractEval: Benchmarking LLMs for Clause-Level Legal Risk Identification in Commercial Contracts. arXiv preprint arXiv:2508.03080.

Wang, S., et al. (2025). ACORD: An Expert-Annotated Dataset for Legal Contract Clause Retrieval. arXiv preprint arXiv:2501.06582.

Li, H., et al. (2025). LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation. Proceedings of the ACM Conference.

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