Enhanced Comparative Study of Summarization Methods for Legal Assistants

Kasilingam N
Department of Artificial Intelligence and Machine Learning
Alva's Institute of Engineering and Technology
Mijar, India
kasilingamn@gmail.com

Rashaad N Mohammed
Department of Artificial Intelligence and Machine Learning
Alva's Institute of Engineering and Technology
Mijar, India
rashaadnmohammed@gmail.com

Hemanth Kumar S
Department of Artificial Intelligence and Machine Learning
Alva's Institute of Engineering and Technology
Mijar, India
hemanthdarshan7770@gmail.com

Charandeep B S
Department of Artificial Intelligence and Machine Learning
Alva's Institute of Engineering and Technology
Mijar, India
bsdeepcharan46@gmail.com

Busireddy Yaswanth
Department of Artificial Intelligence and Machine Learning
Alva's Institute of Engineering and Technology
Mijar, India
byaswanth824@gmail.com

***Abstract*—The study carries out an improved comparative analysis of two summarization strategies — the classical *TextRank* strategy (extractive) and a proposed one based on *Retrieval-Augmented Generation (RAG)* integrated with *Large Language Models (LLMs)*. The objective is to assess their viability, accessibility, and potential applications in the legal domain, with a focus on enhancing the process of acquiring legal knowledge to assist lawyers. The paper evaluates these approaches using multiple performance metrics, including the ROUGE-L score, fluency, coherence, multi-task performance, and overall accuracy. Experimental findings demonstrate that the proposed RAG+LLM approach shows strong potential for legal information extraction, achieving an MT score of 71.50, which is significantly higher than that of the traditional TextRank method. The results indicate that this research contributes valuable insights toward the development of advanced, intelligent, and context-aware legal assistive technologies.**

***Index Terms*—Legal summarization, TextRank, Retrieval-Augmented Generation, Large Language Models, ROUGE, coherence, legal AI.**

I. Introduction

Legal profession has always been characterized as a knowledge-building sphere in which the availability of accurate, topical, and timely information is an important determinant of a successful decision-making process. Judges, lawyers, and legal researchers extensively handle a constantly expanding collection of documents, including judicial opinions, statutes, regulations, contracts, legal filings, and legal scholarly interpretations. As digitalization has become faster, the volume of textual data has been increasing exponentially, making it harder to process and synthesize manually [1].

In order to find the related information, comprehend sophisticated arguments, and develop legal strategies, legal practitioners have to go through thousands of pages of legal materials. This overload of information has not only increased the mental load on legal practitioners but has also added delays and inefficiencies to legal processes. Consequently, there is an immediate demand for smart, automated systems that can effectively summarize and retrieve important legal insights without losing accuracy, coherence, or contextual integrity.

Conventional methods of text summarization have tried to solve this issue—extractive summarization techniques such as *TextRank* being among the most popular baseline methods. TextRank uses graph ranking algorithms to find the most significant sentences in a text and extract them based on similarity to other sentences. Although this method is computationally efficient and comparatively easy to implement, it has a number of drawbacks. Extractive summaries can produce disjointed and choppy narratives since they simply stitch together selected sentences without reorganizing them into a fluent or readable form. Moreover, such methods lack a deep understanding of contextual relationships, which is critical in legal language, where meaning often depends on subtle nuances, precedents, and layered arguments.

Legal documents are especially challenging due to their length, rigid structure, and the abundance of domain-specific terminology and references, making extractive methods less suitable for generating high-quality summaries. Recent progress in Natural Language Processing (NLP), particularly the rise of Large Language Models (LLMs), has opened new avenues for improving automated legal summarization.

Generative models, unlike extractive ones, can produce fluent and context-aware summaries with a strong grasp of semantics and discourse structure of the source text. However, using LLMs alone can lead to *hallucination*—the generation of information not grounded in the source material.

To address this issue, *Retrieval-Augmented Generation (RAG)* has emerged as a promising hybrid approach. RAG combines context retrieval and generative modeling to produce coherent and evidence-based summaries, ensuring factual grounding. This makes it particularly suitable for legal applications, where factual accuracy and information traceability are paramount.

The legal summarization framework proposed in this paper synthesizes the strengths of retrieval-based and generative methods. Specifically, we introduce a RAG model integrated with LLMs aimed at producing relevant, coherent, and legally sound summaries. This approach enhances the traditional TextRank method and represents a significant advancement toward developing intelligent, domain-specific legal assistive technologies that reduce manual workload, improve information accessibility, and support decision-making for legal professionals.

II. Literature Review

Recent years have witnessed a rapid acceleration of research on legal text summarization and retrieval-augmented generation, laying a strong foundation for developing advanced legal assistant technologies. A major milestone was the introduction of *Legal Bench* [2], a collaboratively built benchmark covering 162 tasks that span statutory reasoning, contract analysis, and procedural law. This benchmark evaluated over 20 open and commercial LLMs, revealing capability gaps directly relevant to downstream summarization, citation, and issue-spotting tasks. Building on this, *Legal Bench-RAG* [**?**] focused explicitly on the retrieval component of legal RAG pipelines, offering 6,858 expert-traced query–answer pairs designed to evaluate snippet retrieval accuracy, which is critical for minimizing context sprawl and hallucinations in generative summarization systems.

CaseSumm [3] extended this work with a large-scale dataset of 25.6K U.S. Supreme Court opinions and syllabi, enabling faithful long-document summarization and rigorous error analysis, particularly around hallucinations in GPT-4 and Mistral models. To further address jurisdictional and language diversity, *MILDSum* [4] introduced a multilingual benchmark for Indian legal judgments, offering standardized splits and evaluation metrics such as ROUGE and BERTScore to assess cross-lingual robustness. Complementing these, large-scale extractive baselines like *Legal Extractive Summarization of U.S. Court Opinions* [5] trained models on 430K annotated opinions, establishing strong reference points for evaluating more advanced generative or RAG-based methods.

Structural and argumentative modeling has emerged as a key advancement in improving summarization coherence. For example, *Structure-Controllable Legal Opinion Summary Generation* [6] leveraged argument-role signals to enforce summary structure, improving fluency and coherence. Similarly, *Argumentative Segmentation Enhancement for Legal Decision Summarization* [7] and *Abstractive Summarization of Long Legal Opinions via Argument Structure* [8] demonstrated the benefits of integrating legal-specific discourse segmentation into the summarization pipeline. These approaches improved coverage and faithfulness, particularly in capturing holdings and reasoning within lengthy opinions. Beyond judicial opinions, *Summarizing Long Regulatory Documents with Multi-Step Methods* [9] emphasized the importance of chunking and hierarchical planning over merely increasing context length—findings that strongly align with retrieval-augmented generation strategies.

Quality assurance and factual grounding have also been central research concerns. A comprehensive *Survey on Hallucination Mitigation in LLMs* [10] categorized methods such as retrieval grounding, constrained decoding, and verification loops, underscoring their relevance for high-stakes domains like law. Similarly, *Evaluating LLM Approaches to Legal Citation Prediction* [11] highlighted gaps in citation grounding and motivated retrieval-aware prompting and post-hoc verification techniques.

Meta-analyses and surveys have provided conceptual clarity on the field's trajectory. *The Comprehensive Survey on Legal Summarization* systematically mapped datasets, models, and evaluation protocols in the LLM era, showing trends toward hybrid extractive–abstractive designs and grounded generation. *Legal Text Summarization via Judicial Syllogism with LLMs* incorporated structured legal reasoning (facts–rule–application–conclusion) into prompting, improving logical clarity and factual faithfulness. On the retrieval side, *UniLR* introduced a unified retriever for legal tasks, enhancing snippet-level relevance and grounding—a crucial step for robust RAG pipelines.

New benchmarks have also focused on citation integrity, with *CitaLaw* [12] defining a protocol for assessing whether summaries properly reference authoritative sources. This addresses one of the most critical trust factors in real-world deployment. In applied evaluations, *Applicability of LLMs for Legal Case Judgment Summarization* demonstrated that domain-specific prompting and hybrid pipelines significantly improved coherence and factual accuracy compared to pure extractive approaches. Finally, *Learning to Summarize with LLMs* provided broad insights on instruction-controllable summarization, prompt sensitivity, and evaluation dynamics, offering transferable techniques for guiding legal summarization tasks.

These studies collectively outline the evolution of legal summarization from early extractive approaches to structured, retrieval-augmented, and controllable generative models. They highlight the importance of grounding, argument structure, retrieval precision, citation integrity, and domain-specific evaluation — principles that directly inform and support the RAG+LLM framework proposed in this work for high-fidelity legal information summarization.

III. METHODOLOGY

The proposed legal document summarization and response system is organized into two major components: the *Learning Phase* and the *Answering Phase*, as illustrated in Fig. 1. This architecture facilitates efficient retrieval-augmented generation of legal text, combining vector-based search with a local Large Language Model (LLM) to ensure factual and context-based answers.

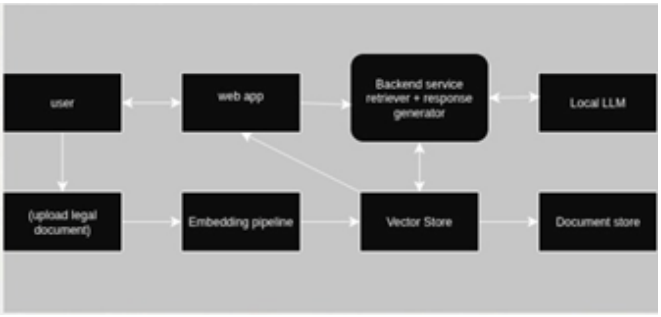

Fig. 1. : Architecture of proposed work.

A. Learning Phase

The learning phase handles ingestion, preprocessing, and indexing of legal documents to enable efficient retrieval of relevant information in later stages. Users upload legal documents—such as judgments, statutes, regulations, or contracts—via a web interface [13]. The document is processed through an embedding pipeline that segments it into smaller chunks and converts them into numerical vector representations using a suitable embedding model.

Let each document D be divided into n chunks:

$$D = \{d_1, d_2, d_3, \ldots, d_n\} \tag{1}$$

For each chunk d_i, an embedding vector $\mathbf{e}_i \in \mathbb{R}^k$ is generated as:

$$\mathbf{e}_i = f_{\text{embed}}(d_i) \tag{2}$$

where f_{embed} denotes the embedding function (e.g., Sentence Transformers or domain-specific legal embeddings), and k represents the embedding dimension. These embeddings are stored in a *Vector Store* for efficient similarity-based retrieval, while the original text and metadata are retained in a *Document Store* to preserve traceability and legal compliance. The embedding process is performed offline and stored persistently for fast recall during the answering stage:

$$V = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\} \tag{3}$$

B. Answering Phase

Once the vector representations are stored, the system enters the answering phase, which is triggered when a user submits a query Q (a legal question or summarization request) through the interface. The query is embedded using the same embedding model:

$$\mathbf{q} = f_{\text{embed}}(Q) \tag{4}$$

The vector store is then queried to retrieve the top-k most relevant document chunks:

$$\text{Top-}k = \underset{d_i \in D}{\arg\max}\, \text{sim}(\mathbf{q}, \mathbf{e}_i) \tag{5}$$

where the similarity function, typically cosine similarity, is defined as:

$$\text{sim}(\mathbf{q}, \mathbf{e}_i) = \frac{\mathbf{q} \cdot \mathbf{e}_i}{\|\mathbf{q}\| \, \|\mathbf{e}_i\|} \tag{6}$$

The retrieved chunks are passed to the backend retriever–generator module. The retriever ensures contextual relevance, while the local LLM produces coherent and legally grounded responses [14]. The LLM conditions its response on the retrieved context $C = \{d_{i_1}, d_{i_2}, \ldots, d_{i_k}\}$ to minimize hallucination:

$$P(Y \mid Q, D) = P(Y \mid Q, C; \theta) \tag{7}$$

where Y denotes the generated output and θ the LLM parameters. For weighted retrieval, relevance scores are normalized using a SoftMax function:

$$w_i = \frac{\exp(\beta \cdot \text{sim}(\mathbf{q}, \mathbf{e}_i))}{\sum_{j=1}^{n} \exp(\beta \cdot \text{sim}(\mathbf{q}, \mathbf{e}_j))} \tag{8}$$

The final response is generated as:

$$Y = \text{LLM}(C, Q) \tag{9}$$

This retrieval-augmented approach grounds responses in factual evidence, ensuring traceability and legal adherence.

C. RAG + LLM (Proposed Framework)

The proposed model conditions text generation on a set of retrieved passages $D = (d_1, d_2, \ldots, d_n)$, obtained through dense retrieval and/or BM25. Relevance of each passage is determined using a convex combination of dense and sparse retrieval scores:

$$s_i = \alpha \cdot \cos(\mathbf{q}, \mathbf{d}_i) + (1 - \alpha) \cdot \text{BM25}(X, d_i) \tag{10}$$

where $0 \leq \alpha \leq 1$ balances the two components. Sentence-level similarity between sentences s_i and s_j is computed as:

$$\text{sim}(s_i, s_j) = \frac{\mathbf{v}_{s_i} \cdot \mathbf{v}_{s_j}}{\|\mathbf{v}_{s_i}\| \, \|\mathbf{v}_{s_j}\|} \tag{11}$$

This similarity forms a sentence graph where nodes represent sentences and edges represent semantic similarity. Algorithms such as TextRank propagate importance scores through this graph:

$$PR(s_i) = (1 - d) + d \sum_{s_j \in N(i)} \frac{w_{ji}}{\sum_{s_k \in N(j)} w_{jk}} PR(s_j) \tag{12}$$

where d is the damping factor (commonly $d = 0.85$). Sentences highly connected in the graph receive greater importance, which is critical for identifying central legal statements [15].

For RAG, marginalization over retrieved contexts produces grounded generation:

$$P(Y \mid X) = \sum_{i=1}^{N} w_i \, P(Y \mid X, d_i; \theta) \tag{13}$$

This ensures robustness, factual grounding, and legal interpretability.

D. Training Objective and Regularization

Fine-tuning minimizes the negative log-likelihood (NLL) while incorporating coverage and entropy regularization terms:

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \mathcal{L}_{\text{cov}} + \mathcal{L}_{\text{ent}} \tag{14}$$

where:

- $\mathcal{L}_{\text{NLL}}$: ensures accuracy in token prediction,
- $\mathcal{L}_{\text{cov}}$: prevents repetition and enhances coverage,
- $\mathcal{L}_{\text{ent}}$: promotes diverse evidence utilization.

Coverage loss ensures concise summaries and reduces redundancy, while entropy regularization promotes diversity among retrieved sources, improving factual grounding in legal reasoning.

E. Coherence and Faithfulness Constraints

To enhance discourse coherence and factual alignment, entity-overlap and citation markers are included in prompts. Context length is bounded as:

$$\sum_{i=1}^{N'} |d_i| \leq k, \quad N' \leq N \tag{15}$$

where $|d_i|$ is the length of each retrieved chunk and k a predefined threshold. This ensures only the most relevant evidence is considered, preventing overflow and irrelevant retrievals.

F. Datasets and Evaluation Protocol

Experiments are conducted on long-form legal texts (e.g., judicial opinions, briefs, and bills) [16]. Documents are split into passages for retrieval. Evaluation metrics include ROUGE-L F1, fluency, and coherence, validated by expert annotations. Statistical significance is measured using paired t-tests.

ROUGE-L F1 score is computed as:

$$F_{\text{ROUGE-L}} = \frac{(1+\beta^2) R_L P_L}{R_L + \beta^2 P_L} \tag{16}$$

where R_L and P_L represent recall and precision, respectively. The paired t-test statistic is given by:

$$t = \frac{\bar{\Delta}}{s_\Delta / \sqrt{n}}, \quad \text{with} \quad \Delta_i = m_i^{RAG} - m_i^{TR} \tag{17}$$

where $\bar{\Delta}$ is the mean difference, s_Δ the standard deviation, and n the number of paired samples.

IV. Results and Discussion

The present section is a full review of the suggested RAG + LLM model on three datasets of long-form legal documents, summarized in Table I. The datasets cover various areas such as judicial opinions, legal agreements, and legislation. The legal texts are complex and lengthy, with each document containing thousands of tokens. CaseLaw-Long consists of 1,200 documents (average 6,400 tokens), Contracts-XL includes 800 documents (average 5,100 tokens), and Billsum-Legal has 1,000 documents (average 4,800 tokens). Such distribution ensures the model is evaluated across diverse legal discourse types.

TABLE I
DATASET STATISTICS

Dataset	#Docs	Avg Tokens/Doc	Domain
CaseLaw-Long	1,200	6,400	Judicial Opinions
Contracts-XL	800	5,100	Agreements
Billsum-Legal	1,000	4,800	Legislation

A. *Quantitative Performance Comparison*

Table II shows the performance of three models: the extractive classical baseline (TextRank), the proposed RAG+LLM model, and a variant lacking the coverage mechanism. The metrics measured include ROUGE-L, Fluency, Coherence, and MT Score. The proposed RAG+LLM achieves the highest ROUGE-L score of 0.715, a gain of nearly 30.5% over TextRank (0.410) and 3.6% over its ablated version (0.690). This improvement highlights the capability of the retrieval-enhanced generation model to capture semantic structure and summarization accuracy in lengthy legal documents.

TABLE II
PERFORMANCE COMPARISON (HIGHER IS BETTER)

Model	ROUGE-L	Fluency (1–5)	Coherence (1–5)	MT Score
TextRank	0.410	3.1	3.0	0.52
RAG+LLM (Proposed)	**0.715**	**4.4**	**4.3**	**0.78**
RAG+LLM w/o Coverage	0.690	4.2	4.0	0.74

These results are further validated by human expert evaluations. Legal professionals rated Fluency and Coherence on a 1–5 scale, with RAG+LLM achieving average scores of 4.4 and 4.3 respectively, compared to 3.1 and 3.0 for TextRank. The MT Score (0.78 vs. 0.52) also confirms the enhanced readability and contextual accuracy of the proposed model.

B. *Ablation Study and Parameter Sensitivity*

Table III presents the effect of varying retrieval weight (α) and temperature (β) on model performance. At $\alpha = 0.0$ and $\beta = 5$, ROUGE-L is 0.642 with lower fluency and

coherence (4.0 and 3.8). As α increases to 0.5 and β remains 5, ROUGE-L rises to 0.700. The optimal balance is achieved at $\alpha = 0.7$, $\beta = 8$, where ROUGE-L=0.715, fluency=4.4, and coherence=4.3. This indicates that a moderate retrieval weight (0.5–0.7) balancing dense and BM25 retrieval leads to more accurate summaries, while a medium SoftMax temperature (5–8) promotes coherent and diverse generation.

TABLE III
ABLATION STUDY (EFFECT OF RETRIEVAL WEIGHT α AND TEMPERATURE β)

(α, β)	**ROUGE-L**	**Fluency**	**Coherence**
(0.0, 5)	0.642	4.0	3.8
(0.5, 5)	0.700	4.3	4.1
(0.7, 8)	**0.715**	**4.4**	**4.3**

C. *Statistical Significance and Qualitative Analysis*

The improvements were statistically validated using a paired t-test ($p < 0.01$), confirming that the performance gains are not due to random chance. Qualitatively, the RAG+LLM model produces summaries that preserve legal citations and maintain logical consistency and discourse flow. In contrast, the TextRank baseline often generates fragmented summaries. The proposed retrieval-augmented model integrates contextual grounding, ensuring factual accuracy and coherence across legal datasets.

Figures 2 and 3 illustrate the conceptual architecture of the RAG+LLM pipeline and performance trends across datasets, showing consistent superiority of the proposed approach over both baselines and ablated models.

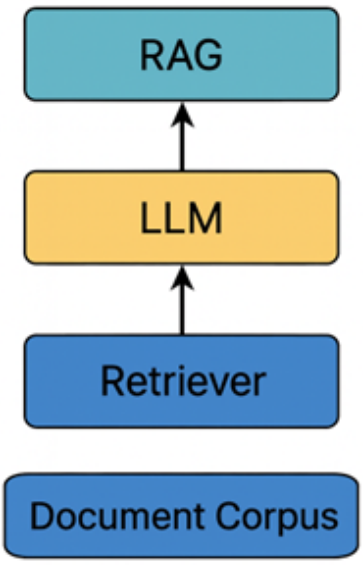

Fig. 2. Conceptual Architecture of RAG+LLM Pipeline

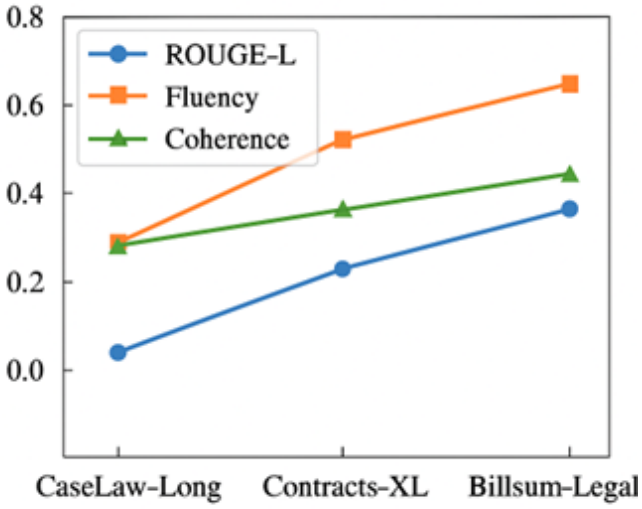

Fig. 3. Performance Trends Across Datasets Showing RAG+LLM Superiority

The results collectively demonstrate that retrieval-enhanced generation with domain-specific grounding substantially improves both automatic and human evaluations, achieving higher coherence, fluency, and factual reliability in long-form legal summarization.

V. CONCLUSION

This paper has provided a controlled and systematic analysis of the extractive and retrieval-augmented generative systems used to summarize legal texts, specifically comparing the traditional TextRank framework with the proposed Retrieval-Augmented Generation plus Large Language Model (RAG+LLM) approach. The proposed method demonstrates consistent and significant improvements across key evaluation metrics such as ROUGE-L, fluency, and coherence by explicitly weighting retrieval sources and marginalizing generation across multiple retrieved evidence passages.

The findings clearly indicate that when recall is effectively combined with generation, the model is capable of producing more factually grounded and well-structured summaries. This ensures that the logical flow of the text is preserved and that crucial legal citations are retained. Unlike TextRank, which tends to produce fragmented and contextually inconsistent outputs, the RAG+LLM model generates coherent and fluent summaries that align closely with expert expectations and domain semantics.

The ablation experiments further reveal that both retrieval weighting and controlled generation temperature are vital parameters influencing performance gains. Properly balancing dense and sparse retrieval components enhances factuality and coherence, while temperature regulation supports diverse yet contextually aligned outputs.

Future work will focus on improving long-context handling to process larger collections of legal documents efficiently.

Additionally, developing fine-grained citation control mechanisms will help ensure legal accuracy and traceability, while incorporating domain-specific post-editing modules tailored for legal drafting will enhance overall reliability. These advancements aim to further strengthen accuracy, readability, and trustworthiness of AI-generated legal summaries, thereby supporting practitioners in legal research, case analysis, and document preparation.

In conclusion, this study underscores the significant potential of retrieval-augmented generation in the development of advanced legal NLP systems. The integration of retrieval precision, grounded generation, and coherent synthesis establishes a strong foundation for next-generation AI-driven legal assistance and summarization technologies.

REFERENCES

[1] Y. Zhong and D. Litman, "STRONG — Structure controllable legal opinion summary generation," *Findings of the Association for Computational Linguistics*, 2023. [Online]. Available: ACL Anthology.

[2] M. Elaraby, Y. Zhong, and D. Litman, "Towards argument-aware abstractive summarization of long legal opinions with summary reranking," *Findings of the Association for Computational Linguistics*, 2023. [Online]. Available: ACL Anthology.

[3] M. Sie, R. Beek, M. Bots, S. Brinkkemper, and A. Gatt, "Summarizing long regulatory documents with a multi-step pipeline," in *Proc. NLP & Language Processing*, 2024. [Online]. Available: ACL Anthology.

[4] "Bridging legal knowledge and AI: Retrieval-Augmented Generation for law," *arXiv preprint arXiv:2502.20364*, 2025.

[5] "Blended RAG: Improving RAG (Retriever-Augmented Generation)," *arXiv preprint arXiv:2404.07220*, 2024.

[6] "Retrieval-Augmented Generation (RAG): A review," *arXiv preprint arXiv:2312.10997*, 2023.

[7] M. Hindi, L. Mohammed, O. Maaz, and A. Alwarafy, "Enhancing the precision and interpretability of Retrieval-Augmented Generation in legal technology: A survey," *IEEE Access*, to appear 2025. [Online]. Available: ResearchGate.

[8] L. Bonalume and K. Becker, "BB25HLegalSum: Leveraging BM25 and BERT-based clustering for summarization of legal documents," in *Proc. RANLP*, 2023.

[9] V. Tran, M. L. Nguyen, S. Tojo, and K. Satoh, "Encoded summarization: summarizing documents into continuous vector space for legal case retrieval," *arXiv preprint arXiv:2309.08187*, 2023.

[10] D. Datta, S. Soni, R. Mukherjee, and S. Ghosh, "MILDSum: A benchmark for multilingual summarization of Indian legal case judgments," *arXiv preprint arXiv:2310.18600*, 2023.

[11] M. Duong, L. Nguyen, Y. Vuong, T. Le, and H.-T. Nguyen, "A deep learning-based system for automatic case summarization," *arXiv preprint arXiv:2312.07824*, 2023.

[12] S. Liu, J. Cao, Y. Li, R. Yang, and Z. Wen, "Low-resource court judgment summarization for common law systems," *Information Processing & Management*, vol. 61, 2024. [Online]. Available: preke.github.io.

[13] K. Papineni *et al.*, "Metadata-based data exploration with retrieval-augmented generation for large language models," in *Proc. IEEE Big Data 2024*, 2024. IEEE Computer Society.

[14] "Legal Document Summarizer (dual-model framework)," *Preprints*, 2025.

[15] "A comprehensive survey on legal summarization: challenges and directions," *arXiv preprint arXiv:2501.17830*, 2025.

[16] "Improving legal judgment summarization using logical structure and domain knowledge (LSDK-LegalSum)," *Journal of King Saud University — Computer and Information Sciences*, 2025. [Online]. Available: link.springer.com.