Literature Review on **LegalDoc-RAG: Search Less, Win More**

Document Drafting and Summarization

Shreya Uprety, Rishikesh Paudel, Sandesh Kuikel, Prafulla Raj Pokhrel

Introduction

The legal domain faces the persistent challenge of processing vast volumes of complex, nuanced documents while ensuring accurate and contextually relevant summaries or drafts. Traditional manual review and keyword-based searches are inefficient, error-prone, and lack contextual depth. This often results in delays in litigation support, compliance checks, and legal research.

Research Question

How can Retrieval-Augmented Generation (RAG) improve efficiency, accuracy, and contextual reliability in legal document drafting and summarization?

Sub-problems

(1) Reducing hallucinations and ensuring factual grounding in generated legal text.

(2) Handling jurisdictional complexities in statutes and case law.

(3) Scaling retrieval to massive and evolving legal datasets.

(4) Ensuring explainability and trustworthiness of AI-generated legal content.

(5) Addressing data security and compliance in handling sensitive legal information.

General Review

Major Ideas

Retrieval-Augmented Generation (RAG) integrates retrieval modules with large language models to ground outputs in authoritative legal texts, thereby reducing hallucinations. In legal workflows (summarization, drafting, compliance, research), combining contextualized retrieval with generative synthesis has shown significant improvements over naive LLM use. Recent techniques include fine-tuning open LLMs (e.g. LLaMA-2/3) with domain-specific data, and using adapters (LoRA) for efficient legal adaptation.

Datasets Available

- Jurisdiction-specific legal corpora (e.g. public statutes, case law databases like Public Library of Law, CourtListener, and statutes from government websites).
- Contracts and agreements from corporate repositories (e.g. EDGAR filings, public contract datasets like ContractNLI, CUAD, and proprietary contract libraries).
- Annotated legal question-answer and summarization datasets (e.g. Supreme Court opinions with summaries, LegalBench QA pairs, Contract QA benchmarks).

Common Models Used

- Transformer-based LLMs (e.g. LLaMA, GPT series, Codex variants fine-tuned on legal text).
- Dense retrieval systems (FAISS, ElasticSearch with dense embeddings such as SBERT, Legal-BERT, GPT-based embeddings).
- Sparse retrievers (BM25, TF-IDF) often combined with semantic retrieval in hybrid pipelines.
- Knowledge graphs and topic models (e.g. concept graphs of legal entities, NMF-based topic discovery) to enhance grounding and explainability.

Major Issues Identified

- **Hallucinations:** LLMs often fabricate citations or erroneous legal conclusions. RAG aims to mitigate this by grounding answers in retrieved texts.
- **Jurisdictional scope:** Ensuring that retrieved cases and statutes are from the correct jurisdiction and legal context to avoid misapplication of law.
- **Scale and currency:** Maintaining up-to-date indexes of constantly evolving legal corpora (new cases, regulations).
- **Explainability:** Legal professionals require that AI outputs be traceable to sources. RAG systems must provide clear citations and rationale.
- **Privacy and Compliance:** Handling confidential legal documents requires strict data governance; retrieval systems must respect privacy and security constraints.

Paper Review

Below are detailed reviews of five key works relevant to RAG in legal drafting and summarization.

1. LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultations

Author/Year: Haitao Li et al., 2025

Main Idea: LexRAG introduces the first benchmark specifically for evaluating RAG systems in multi-turn legal consultation scenarios. It highlights the unique challenges of legal dialog, where an AI assistant must retrieve relevant statutes and case law as questions evolve. The authors construct a dataset of 1,013 annotated dialogues (5 turns each) paired with 17,228 legal documents (cases and statutes). They define two tasks: conversational knowledge retrieval (finding relevant legal texts given the dialogue context) and response generation (producing legally sound answers). The LexiT toolkit is provided to implement RAG components tailored to the legal domain, including an "LLM-as-a-judge" evaluation pipeline. Experiments combining various LLMs and retrievers demonstrate key limitations of current systems in handling multi-turn context and maintaining legal consistency.

Dataset: 1,013 multi-turn legal consultation dialogues and 17,228 candidate legal articles (cases, statutes), all annotated by legal experts. Each dialogue sample has five progressive Q&A turns to simulate a real consultation process.

Model/Method: A retrieval-augmented pipeline where user queries and dialogue history are converted into dense embeddings to retrieve relevant documents. An LLM then generates answers grounded in these retrieved sources. The LexiT toolkit orchestrates vector search, reranking, and LLM-based generation, all customized for legal language.

Key Techniques: Integration of conversational context into retrieval, evidence-grounded generation, and advanced evaluation via a jurist-inspired LLM-as-judge. LexRAG also emphasizes hybrid retrievers (dense embeddings with re-ranking) to ensure precision.

Results/Takeaway: LexRAG sets a new standard for legal RAG evaluation by revealing that even top LLMs struggle to maintain consistency across a dialogue. The benchmark and toolkit enable systematic improvement of AI legal assistants by highlighting where existing approaches fail.

2. Bridging Legal Knowledge and AI: RAG with Vector Stores, Knowledge Graphs, and NMF

Author/Year: Barron et al., 2025

Main Idea: Barron et al. present a hybrid RAG framework that combines vector-based retrieval, knowledge graphs, and Non-Negative Matrix Factorization (NMF) to enhance understanding of legal texts. The core idea is to move beyond keyword search by capturing both semantic and relational structure in legal data. The authors scrape a large corpus of legal texts (statutes, constitutions, case law) and embed them in a vector store for semantic search. A knowledge graph is built to represent explicit relationships between legal entities (e.g. case citations and legal concepts). They apply hierarchical NMF to discover latent topic clusters in the data. Together, these components allow the system to retrieve and cross-reference legal information more effectively than traditional methods. The paper demonstrates applications

in document clustering, summarization, and cross-referencing, showing that this architecture grounds LLM outputs and reduces hallucinations by providing structured legal context.

Dataset: A custom corpus of legal documents collected via web scraping (e.g. statutes, constitutional provisions, case law from platforms like Justia). The dataset spans multiple jurisdictions and contains semi-structured legal knowledge.

Model/Method: A generative AI pipeline integrating: (1) a vector store of dense embeddings (from models like BERT/GPT) for semantic retrieval; (2) a knowledge graph encoding relationships between statutes, precedents, and legal concepts; and (3) hierarchical NMF topic modeling on the text to identify latent legal themes. The LLM draws on both the retrieved vectors and graph-based knowledge when generating answers
Key Techniques: Semantic embeddings for retrieval, ontology-driven knowledge graph construction, and NMF topic factorization. This multi-modal approach leverages latent topic discovery and explicit legal connections to improve grounding. It effectively bridges keyword search and deep understanding by making legal reasoning more interpretable.

Results/Takeaway: The hybrid system significantly improves legal retrieval accuracy and interpretability. Clustering and summarization on this enriched knowledge base are more accurate, since the model can uncover hidden relationships. Importantly, it greatly reduces hallucinations by anchoring outputs in a structured knowledge graph, marking a significant step toward transparent AI in law.

3. LRAGE: Legal Retrieval Augmented Generation Evaluation Tool

Author/Year: Park et al., 2025

Main Idea: Park et al. develop LRAGE, an open-source evaluation toolkit designed to systematically assess RAG systems in the legal domain. Recognizing that legal decision-making relies heavily on retrieving prior cases, the authors note that RAG performance depends on many components: the retrieval corpus, retrieval algorithm, reranker, LLM backbone, and evaluation metric. LRAGE provides both GUI and command-line interfaces to configure these components and measure their impact on overall accuracy. It enables users to isolate how each factor affects answer quality. The toolkit was validated on multilingual legal QA benchmarks (Korean KBL, English LegalBench, Chinese LawBench), illustrating that varying components leads to measurable accuracy changes. By offering a holistic evaluation platform, LRAGE helps developers tune legal RAG systems and compare different design choices.

Dataset: Legal question-answer datasets in multiple languages. For example, the authors use Korean LegalBench (KBL), English LegalBench, and Chinese LawBench, each containing statutory and case-law questions. These cover a range of legal topics across different jurisdictions.

Model/Method: LRAGE is a meta-evaluation system rather than a single model. Researchers plug in their own retrievers (e.g. BM25 or DPR), rerankers, and LLMs into the LRAGE pipeline, then run experiments to record metrics. The toolkit tracks overall QA accuracy as components are systematically varied.

Key Techniques: Component-wise analysis of RAG pipelines; interactive configuration. LRAGE treats the entire RAG stack as a set of configurable modules, allowing fine-grained testing of each. It supports multilingual data and provides clear reporting of component impact.

Results/Takeaway: The tool yields insights such as which retrievers or LLMs boost legal answer accuracy the most. Its open release enables the community to benchmark legal RAG innovations reproducibly. By exposing component trade-offs, LRAGE guides the design of more accurate and efficient legal AI systems.

4. Optimizing Legal Text Summarization with Dynamic RAG

Author/Year: Mukund & Easwarakumar, 2025

Main Idea: This paper proposes a Dynamic Retrieval-Augmented Generation system for summarizing legal documents, focusing on Indian case law. It addresses legal text complexity by retrieving relevant context on the fly. The system first identifies "dark zones" in a judgment (e.g. unexplained statute references) and uses a BM25 retriever (with a top-3 chunk strategy) to fetch pertinent passages from domain-specific corpora (e.g. the Constitution of India, Civil Procedure Code, Supreme Court opinions). These retrieved chunks are provided as additional context to a fine-tuned LLaMA 3.1 8B model during generation. A compression-ratio constraint maintains structural balance between source and summary. Experiments on Indian legal cases show this approach significantly improves summary accuracy and factuality. The authors report that LLaMA 3.1-8B with Legal NER and dynamic retrieval achieves the highest BERTScore (0.89) among tested models.

Dataset: Indian legal texts and associated summaries. This includes a collection of Supreme Court judgments, statutory excerpts (CPC, etc.), and constitutional provisions. Documents were chunked and paired with expert-written summaries for training and evaluation.

Model/Method: A hybrid pipeline combining extractive retrieval and abstractive generation. Legal entities (statutes, case names) detected by a Named Entity Recognition module drive real-time queries. A BM25 retriever pulls the top-3 relevant text chunks. The summarization model (LLaMA 3.1 8B fine-tuned on legal data) generates a summary using both the original text and retrieved evidence. An explicit ratio constraint (0.05–0.5) ensures output conciseness.

Key Techniques: Domain-aware retrieval (Legal NER to form queries), dynamic chunking, and fine-tuning of a large decoder LLM. By dynamically augmenting the input with authoritative legal references, the model maintains factual consistency.

Results/Takeaway: The Dynamic RAG approach yields high-quality, citation-grounded summaries of legal judgments. It outperforms baseline summarizers on factual accuracy. The experiments confirm that integrating retrieval into summarization dramatically reduces hallucinations, making AI-generated summaries more reliable for legal use.

5. LegalBench-RAG: A Retrieval Benchmark for Legal RAG

Author/Year: Pipitone & Houir Alami, 2024

Main Idea: Pipitone and Houir Alami introduce LegalBench-RAG, the first benchmark focused on the retrieval component of RAG systems in the legal domain. They observe that most legal AI benchmarks evaluate answer generation, but effective RAG requires precise retrieval. To address this gap, LegalBench-RAG traces existing legal QA queries back to the exact source passages needed for the answer. The result is a dataset of 6,858 human-annotated query-answer pairs over a 79M-character legal corpus, where each question is linked to its supporting text snippet. This benchmark covers diverse legal contexts (contracts, corporate disclosures, privacy policies, etc.) to test retrieval performance. By requiring systems to fetch concise, relevant snippets rather than entire documents, LegalBench-RAG encourages methods that minimize hallucination and support verifiable citation generation. A smaller "mini" version of the dataset is also released for quick experimentation.

Dataset: A human-annotated collection of 6,858 questions from four legal QA datasets (ContractNLI, CUAD, M&A, PrivacyQA). Each question is paired with the minimal answer span in its original document. The combined corpus spans roughly 79 million characters of legal text. The dataset is publicly available.

Model/Method: The paper itself constructs the benchmark; as an example evaluation, they test various retrieval methods (BM25, dense embeddings) against the gold standard. The key methodological contribution is in dataset creation: linking queries to exact legal text spans. Any RAG system can be evaluated by how accurately it retrieves the annotated snippets.

Key Techniques: Snippet-level retrieval evaluation, legal domain specificity, and multi-corpus integration. LegalBench-RAG emphasizes granular precision: systems must return highly relevant text snippets. This setup also facilitates human verification of LLM outputs.

Results/Takeaway: LegalBench-RAG provides a critical benchmark for legal AI, highlighting that successful RAG requires more than powerful LLMs—it needs precise retrieval. It sets a high bar for retrieval accuracy, aiming to steer development of legal RAG systems toward minimal, evidence-based context.

Literature Review Matrix: Key RAG Papers in Legal Domain

Paper Details	Framework	Research Focus	Data Sources	Methods & Outcomes
LexRAG Benchmark *Li et al. (2025)*				
Focus: Multi-turn legal consultation with RAG.	Conversational RAG pipeline (LexiT toolkit) with vector retrieval and LLM.	Evaluating retrieval accuracy and answer quality in multi-turn dialogues.	Annotated legal dialogs (1,013 multi-turn samples) and law corpus (17,228 statutes/cases).	**Method:** LexiT RAG pipeline (vector search + LLM)
Outcome: New benchmark; reveals context-handling gaps.				
Bridging Legal Knowledge *Barron et al. (2025)*				
Focus: Semantic retrieval with knowledge graphs and topic modeling.	Hybrid RAG with vector store, knowledge graph, and NMF topic modeling.	Grounding RAG in structured legal knowledge to improve relevance.	Web-scraped legal corpus (statutes, constitutions, case law).	**Method:** NMF topic factorization + KG-augmented retrieval

Continued on next page

Literature Review Matrix (continued)

Paper Details	Framework	Research Focus	Data Sources	Methods
Outcome: Enhanced contextual accuracy; improved clustering/summarization.				
LRAGE Evaluation Toolkit *Park et al. (2025)*				
Focus: Benchmarking RAG system components in law.	Configurable RAG pipeline (GUI/CLI) for retriever, reranker, LLM, metrics.	Analyzing how corpus choice, algorithms, and models affect legal QA accuracy.	Multilingual legal QA benchmarks (Korean KBL, English LegalBench, Chinese LawBench).	**Method:** Systematic component variation; interactive evaluation.
Outcome: Open-source tool; insights on optimizing legal RAG.				
Dynamic Legal RAG Summarization *Mukund & Easwarakumar (2025)*				

Continued on next page

Literature Review Matrix (continued)

Paper Details	Framework	Research Focus	Data Sources	Methods
Focus: Retrieval-augmented abstractive summarization.	Real-time legal context retrieval (BM25 top-3) + fine-tuned LLaMA summarizer.	Producing concise, grounded summaries of court judgments.	Indian legal corpus (Constitution of India, CPC, Supreme Court cases).	**Method:** BM25 + Legal NER + LLaMA 3.1-8B
Outcome: High-quality summaries (BERTScore 0.89); reduced hallucinations.				
LegalBench-RAG				
Pipitone & Houir Alami (2024)				
Focus: Precise retrieval benchmark for legal RAG.	RAG retrieval dataset construction for targeted snippet retrieval.	Evaluating RAG retrieval accuracy by requiring exact citations.	Combined QA datasets (ContractNLI, CUAD, M&A, PrivacyQA): 6,858 QA; 79M char corpus.	**Method:** Human-annotated snippet-level benchmark
Outcome: Public benchmark enabling precise legal retrieval.				

Project Solution Proposals

Based on the literature review, the proposed system will integrate a **FAISS-based** retrieval engine with a **LoRA fine-tuned LLaMA-2** model, optimized for legal text. Key design elements include:

- **Jurisdiction-aware retrieval modules** to ensure regional legal accuracy and correct application of precedent/statutes.

- **Confidence scoring and retrieval quality checks** to detect low-support generations and avoid hallucinations.
- **Summarization and drafting pipelines** that produce traceable outputs with inline citations linked to retrieved evidence.
- **Secure and explainable architecture** supporting audit logs, access control, and human-in-the-loop verification.

References

- Li, H., Chen, Y., et al. (2025). *LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation.*
- Barron, R.C., Eren, M.E., Serafimova, O., Matuszek, C., & Alexandrov, B.S. (2025). *Bridging Legal Knowledge and AI: Retrieval-Augmented Generation with Vector Stores, Knowledge Graphs, and Hierarchical NMF.* (Proc. ICAIL 2024).
- Park, M., Oh, H., Choi, E., & Hwang, W. (2025). *LRAGE: Legal Retrieval Augmented Generation Evaluation Tool.* arXiv:2504.01840.
- Mukund, A., & Easwarakumar, K.S. (2025). *Optimizing Legal Text Summarization Through Dynamic Retrieval-Augmented Generation and Domain-Specific Adaptation. Symmetry*, 17(5):633.
- Pipitone, N., & Houir Alami, G. (2024). *LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain.* (ICML 2024).