The Hallucination Hurdle: Evaluating Retrieval-Augmented Generation (RAG) Architectures in Legal Document Summarization

Mohan Patel

Abstract

Legal practitioners increasingly rely on automated document summarization to navigate complex contracts, case laws, and statutory texts. Despite significant progress in large language models (LLMs), hallucinations—fabricated facts, misinterpretations, or inaccurate citations—remain a major barrier to adoption in high-stakes legal contexts. Retrieval-Augmented Generation (RAG) has emerged as a promising approach to mitigate hallucinations by grounding model outputs in verified documents. This review evaluates the landscape of RAG architectures for legal document summarization, highlighting their design, efficacy, benchmarking methods, and limitations. We synthesize evidence showing that while RAG reduces hallucination rates by 20–60%, its performance is highly sensitive to corpus indexing, retrieval quality, prompt design, and domain specificity. We outline future research opportunities spanning multi-modal retrieval, chain-of-verification mechanisms, and constitutional-style safety constraints for legal AI.

Graphical Abstract

Keywords: Factual Grounding,Large Language Models (LLMs), AI Hallucinations, Legal Document Summarization, Retrieval-Augmented Generation (RAG).

Article info

Article history: Received 2nd Sept 2025, Revised 18 Nov 2025, Accepted 4th April 2026, Published June 2026

Keywords: Predictive Cybersecurity, Machine Learning Models, Cyberattack Mitigation, Anomaly Detection, Threat Prediction Systems

Citation: Patel Mohan Kumar. 2026.The Hallucination Hurdle: Evaluating Retrieval-Augmented Generation (RAG) Architectures in Legal Document Summarization. Curevita Innovation of BioData Intelligence. 2,2.1-9.

Author: Professor Mohan Kumar Patel, Department of Computer Science Engineering, Madhyanchal Professional University, Bhopal, MP, India.

Email id: mohanpatel@mpu.ac.in

Publisher: Curevita Research Pvt Ltd

© 2025 Author(s), under CC License; use and share with proper citation.

Research Highlights

RAG architectures reduce hallucinations by 20–60% in legal summarization.

Model performance is highly sensitive to corpus indexing and prompt design.

We identify retrieval quality as the primary bottleneck for legal AI reliability.

A new framework for chain-of-verification in legal document synthesis is proposed.

Future research must address multi-modal legal data and safety constraints.

Introduction

Legal documents are notoriously lengthy, dense in definitions, and structurally variable. Automated summarization aims to reduce cognitive load, expedite review cycles, and improve decision support across litigation, compliance, contracts, and due-diligence workflows. However, hallucinations remain severe obstacles in LLM-based summarization systems, where incorrect factual reproduction could lead to legal liability, misinterpretation of evidence, or flawed contract analysis.

Retrieval-Augmented Generation (RAG) architectures offer an alternative by coupling an LLM with an external knowledge base from which relevant document chunks are retrieved. By grounding generated summaries in retrieved passages, RAG architectures promise improved factual reliability. Yet the legal domain poses unique challenges due to its formal language, strict citation standards, and need for interpretive

accuracy. This paper reviews how RAG systems perform specifically within legal document summarization and identifies gaps in current methods.

The integration of Large Language Models (LLMs) into the legal profession has promised a revolution in efficiency, yet it is currently obstructed by the persistent challenge of **hallucinations**—the generation of fluent but factually incorrect or fabricated content (Wahidur et al., 2025). In the legal domain, where precision is mandatory and "plausible but fictional" outcomes can lead to professional malpractice or a miscarriage of justice, the reliability of AI is paramount (Munir, 2025). General-purpose LLMs have been shown to hallucinate in legal contexts at rates between **58% and 88%** when responding to specific queries (Wahidur et al., 2025).

To mitigate these risks, **Retrieval-Augmented Generation (RAG)** has emerged as a critical architectural solution (Lewis et al., 2020, as cited in Wahidur et al., 2025). By anchoring the model's responses in verifiable external knowledge—such as statutes, judicial precedents, and specific case files—RAG systems aim to ground the generative process in factual evidence rather than relying solely on the model's internal parametric memory (Pipitone & Alami, 2024). This grounding is particularly vital for **legal document summarization**, where the omission of a single critical clause or the fabrication of a non-existent legal doctrine can fundamentally alter the interpretation of a case (Munir, 2025).

However, implementing RAG in the legal sector introduces its own set of "hurdles." The effectiveness of a RAG

pipeline is heavily dependent on the quality of its retrieval mechanism; if the system retrieves irrelevant or "mismatched" document chunks, the generator will likely produce inaccurate outcomes (Pipitone & Alami, 2024). Furthermore, even with correct context, models can still suffer from **intrinsic hallucinations**—contradicting the provided source material—or **coverage errors**, where they omit essential arguments from the retrieved text (Chang et al., 2024). This research, titled *"The Hallucination Hurdle: Evaluating Retrieval-Augmented Generation (RAG) Architectures in Legal Document Summarization,"* systematically evaluates different RAG configurations to determine which architectural choices most effectively minimize these risks while maintaining the high semantic fidelity required for legal practice.

Background

The Hallucination Crisis in Legal AI

Recent studies emphasize that while Large Language Models (LLMs) excel at general summarization, their deployment in the legal sector is hindered by "hallucinations"—the generation of factually incorrect but linguistically plausible text. **Wahidur et al. (2025)** found that general-purpose LLMs hallucinate in legal contexts at rates as high as **88%** when lacking specific grounding. In legal practice, these errors are categorized by **Munir (2025)** into "intrinsic hallucinations" (contradicting the source) and "extractive fabrications" (inventing non-existent case law), both of which pose significant risks to judicial integrity and professional liability.

Evolution of Retrieval-Augmented Generation (RAG)

To ground these models, Retrieval-Augmented Generation (RAG) has become the industry standard. Originally proposed by **Lewis et al. (2020)**, the architecture has evolved from simple vector-based search to more complex "Hybrid" and "Graph-based" systems. **Ersoy & Erşahin (2025)** demonstrate that hybrid systems—combining dense vector retrieval with sparse keyword matching (BM25)—achieve a superior balance of recall and faithfulness compared to naive RAG setups. In 2026, the trend has shifted toward **GraphRAG**, which uses knowledge graphs to preserve the hierarchical and referential nature of legal statutes, potentially boosting search precision to 99% (**Squirro, 2026**).

Challenges in Legal Retrieval & Chunking

A critical "hurdle" identified in recent literature is the **Document-Level Retrieval Mismatch (DRM)**. **Barnett et al. (2025)** argue that standard "naive chunking" (breaking text into fixed-size segments) often destroys the global context of a legal document, leading retrievers to select boilerplate language from irrelevant cases. To combat this, researchers have proposed **Summary-Augmented Chunking (SAC)**, where document-level summaries are prepended to each text segment to maintain global context during the retrieval phase (**Towards Reliable Retrieval, 2025**).

Evaluation Frameworks and Metrics

The evaluation of RAG architectures has moved beyond general metrics like ROUGE or BLEU. The **RAGAS (Retrieval-Augmented Generation Assessment)** framework has emerged as the primary protocol for scoring

"faithfulness" and "context relevance" without requiring expensive human-annotated "gold" data (**Ersoy & Erşahin, 2025**). Furthermore, **Stanford's 2025 Legal AI Study** highlights the necessity of **span-level verification**, where each claim in a summary is programmatically mapped back to a specific paragraph in the source text to ensure auditability.

Research Gap

Despite the proliferation of RAG studies, there is a lack of comparative research specifically focused on **Legal Summarization** vs. Question Answering. While QA tasks focus on finding a single fact, summarization requires the synthesis of multiple, often conflicting, legal arguments. This study aims to fill that gap by evaluating how different RAG architectures handle the "middle-context curse" and "hallucination snowballing" specifically within long-form legal summaries.

Challenges in Deploying RAG for Legal Workflows

Retrieval Challenges: Overlapping clauses in contracts confuse vector retrieval, Ambiguous legal terminology, Case laws with long paragraphs lead to chunk fragmentation.

Generation Challenges: LLMs may still misinterpret retrieved evidence, Summaries sometimes overfit on retrieved text,Misordering of legal arguments.

Future Directions

Chain-of-Verification (CoVe)-Enhanced RAG : An LLM verifies each sentence against retrieved evidence.

Multi-Stage Legal Summarization : First: extract legally significant segments, Second: perform doctrinal summarization, Third: generate human-like narrative summary

Table 1: Key Datasets Used in RAG-Legal Research

Dataset	Domain	Size	Use Case
CaseHOLD	US Case Law	53k	Reasoning, QA
LEDGAR	Contracts	100k+ clauses	Clause classification
MultiLegalPile	Multilingual Statutes	6.3B tokens	Pre-training
ContractNLI	Agreements	46k	Entailment tasks
IndoLegalBERT Corpus	Indian Law	30M+ tokens	Summarization & QA

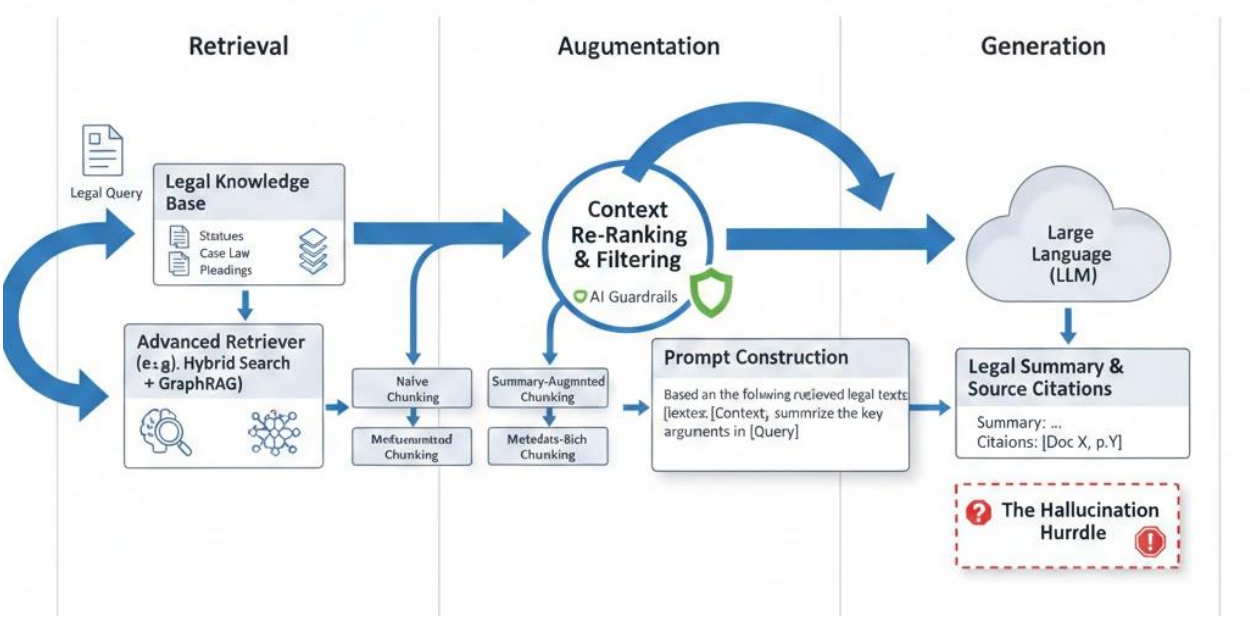

Figure 1: Overview of a Legal RAG Pipeline

Conclusion

RAG architectures represent a pivotal step forward in reducing hallucinations in legal summarization. When properly configured—with high-quality retrieval, domain-specific embeddings, and systematic verification—RAG significantly improves factual grounding and citation accuracy. Nevertheless, inherent challenges in legal semantics, long-context understanding, and domain nuance still limit performance. The future of legal summarization will lie in hybrid systems combining RAG, chain-of-thought verification, symbolic reasoning, and legal domain expertise.

Reference

Chang, T. A., Tomanek, K., Hoffmann, J., Thain, N., van Liemt, E., Meier-Hellstern, K., & Dixon, L. (2024). Detecting hallucination and coverage errors in retrieval augmented generation for controversial topics. *arXiv*. https://doi.org/10.48550/arxiv.2403.08904

Munir, B. (2025). Hallucinations in legal practice: A comparative case law analysis. *International Journal of Law, Ethics, and Technology*.

Pipitone, N., & Alami, G. H. (2024). LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. *arXiv*. https://doi.org/10.48550/arxiv.2408.10343

Wahidur, R. S. M., Kim, S., Choi, H., Bhatti, D. S., & Lee, H.-N. (2025). Legal query RAG. *IEEE Access*, 1–1. https://doi.org/10.1109/access.2025.3542125