Towards Reliable Retrieval in RAG Systems for Large Legal Datasets

Markus Reuter[1,*,†,‡], Tobias Lingenberg[1,*,†], Rūta Liepiņa[2], Francesca Lagioia[3,5],
Marco Lippi[2], Giovanni Sartor[3,5], Andrea Passerini[4], Burcu Sayin[4]

[1]Department of Computer Science, Technical University of Darmstadt
[2]Department of Computer Science, University of Florence, `name.surname@unifi.it`
[3]ALMA-AI, Faculty of Law, University of Bologna, `name.surname@unibo.it`
[4]DISI, University of Trento, `name.surname@unitn.it`
[5] Department of Law, European University Institute

Abstract

Retrieval-Augmented Generation (RAG) is a promising approach to mitigate hallucinations in Large Language Models (LLMs) for legal applications, but its reliability is critically dependent on the accuracy of the retrieval step. This is particularly challenging in the legal domain, where large databases of structurally similar documents often cause retrieval systems to fail. In this paper, we address this challenge by first identifying and quantifying a critical failure mode we term Document-Level Retrieval Mismatch (DRM), where the retriever selects information from entirely incorrect source documents. To mitigate DRM, we investigate a simple and computationally efficient technique which we refer to as Summary-Augmented Chunking (SAC). This method enhances each text chunk with a document-level synthetic summary, thereby injecting crucial global context that would otherwise be lost during a standard chunking process. Our experiments on a diverse set of legal information retrieval tasks show that SAC greatly reduces DRM and, consequently, also improves text-level retrieval precision and recall. Interestingly, we find that a generic summarization strategy outperforms an approach that incorporates legal expert domain knowledge to target specific legal elements. Our work provides evidence that this practical, scalable, and easily integrable technique enhances the reliability of RAG systems when applied to large-scale legal document datasets.[1]

1 Introduction

Large Language Models (LLMs) are increasingly adopted in high-stakes domains such as law. Yet, they remain critically limited by the phenomenon of *hallucination*: incorrect outputs that are fabricated or deviate from the provided source material, posing severe risks in legal applications (Huang et al., 2025; Li, 2023; Qin and Sun, 2024). Recent studies report hallucination rates between 58–80% for general-purpose LLMs on legal tasks (Dahl et al., 2024), highlighting how factual reliability is not just desirable but essential for deploying LLMs in the legal domain.

This challenge is amplified by forward-looking proposals for how legal documentation itself may evolve. For instance, Pałka et al. (2025) suggest that privacy policies may intentionally become longer and more comprehensive to ensure they are legally complete. In such a future, LLMs are expected to serve as the designated "readers" of these texts, extracting and summarizing information for human users. This vision, however, can only be realized if the systems are *highly reliable*.

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has emerged as one of the leading approaches to improving reliability. Using a trusted text corpus to provide factual evidence, RAG guides the LLM's output, reducing hallucinations and ensuring closer alignment with the source material (Tonmoy et al., 2024). In the context of long, structurally similar legal documents, identifying the relevant text passage as "needle in the haystack" becomes a top priority that we aim to address.

On a technical level, we quantify the retrieval quality with our Document-Level Retrieval Mismatch (DRM) metric and the character-level precision and recall. Then, we investigate a simple yet effective technique to improve retrieval quality, Summary-Augmented Chunking (SAC). We enrich text chunks in the trusted text corpus with document-level summaries. This preserves global context, lost in standard chunking, guiding the retriever toward the correct document without altering the underlying retrieval pipeline. This method

---

[*]Equal contribution.

[†]Work was performed at University of Trento.

[‡]Correspondence: markus.reuter@stud.tu-darmstadt.de
[1]Code is available at `https://github.com/DevelopedByMarkus/summary-augmented-chunking.git`.

1

is applied to question-answering tasks across a diverse set of legal documents, including privacy policies, non-disclosure agreements, and merger-and-acquisition contracts.

Key contributions: (i) First, we define and quantify *Document-Level Retrieval Mismatch* (DRM), a key failure mode we observe in standard RAG pipelines where the retrieved information originates from the entirely wrong source document. **(ii)** We propose *Summary-Augmented Chunking* (SAC) as a lightweight and modular solution that strongly reduces DRM by injecting global context directly into each chunk. We experimentally validate SAC on Legalbench-RAG (Pipitone and Alami, 2024), showing substantial improvements over standard chunking. **(iii)** Additionally, we explore how to utilize legal domain knowledge by evaluating both a generic and an expert-guided summarization strategy. Interestingly, we find that simple, general-purpose summarization yield the best retrieval performance.

2 Background and Related Work

2.1 Retrieval-Augmented Generation

RAG (Lewis et al., 2020) is a powerful paradigm that enhances the reliability of LLMs by grounding their outputs in external knowledge sources. This approach is particularly crucial in high-stakes domains like law, where factual accuracy is not just desirable but mandatory. The standard RAG pipeline consists of two main stages: a retriever that searches a large document corpus to find text snippets or "chunks" that are relevant to a user's query and a generator model that synthesizes a final answer based on these retrieved chunks.

RAG has been adapted for a wide array of legal tasks (Hindi et al., 2025), including case reasoning (Yang, 2024), legal judgment prediction (Peng and Chen, 2024), and legal question-answering (Cherubini et al., 2024; Visciarelli et al., 2024). Recent approaches have focused on improving reliability by imposing more structure on the knowledge source, for instance, through knowledge graphs (Kalra et al., 2024) or structured case databases (Wiratunga et al., 2024; Jayawardena et al., 2024).

Despite its demonstrated ability to improve factual accuracy in context-sensitive tasks (Gupta et al., 2024), the effectiveness of RAG is critically dependent on the quality of the initial retrieval step (Huang et al., 2025; Hou et al., 2024a,b). If the retriever fails to select information that is relevant

or complete, the generator may produce factually unsupported responses. Accordingly, prior studies demonstrate that RAG is not a guaranteed solution, as legal RAG systems continue to generate a considerable amount of hallucinated content, particularly when the retrieval mechanism is ineffective (Dahl et al., 2024; Magesh et al., 2025; Ariai and Demartini, 2024). Our work addresses this challenge by focusing on the *pre-retrieval stage*, the engineering of the knowledge base, which forms the foundation of any reliable RAG system.

2.2 Unique Challenges of Legal Text for RAG

Legal documents present a major challenge for automated text processing systems due to their specific linguistic and structural characteristics (Ashley, 2018; Ferraris et al., 2024; Liepina et al., 2019; Martinelli, 2023). These challenges make retrieval particularly prone to errors.

(i) Lexical Redundancy: Legal language is highly standardized, featuring boilerplate clauses, formally defined phrases, and specialized terminology that are often repeated across thousands of documents (Akter et al., 2025). For example, Non-Disclosure Agreements within a database may be structurally almost identical, differing only in a few critical variables such as party names or dates. This high degree of similarity can easily confuse retrieval models that rely on surface-level keyword matching or vector similarity (Joshua et al., 2025).

(ii) Hierarchical Structure: Legal texts are organized in complex layouts with nested sections, subsections, and dense cross-references. Standard chunking strategies ignore document hierarchy (Ferraris et al., 2024; Zilli et al., 2025), which cuts off these logical connections. As a result, retrieved chunks may appear relevant but lose their intended meaning when disconnected from their structural context.

(iii) Fragmented Information: Answering a legal question often requires synthesizing information scattered across multiple sections or even different documents (Hindi et al., 2025). For example, interpreting an exception clause in a privacy policy may depend on definitions or stipulations introduced much earlier in the document. Retrieval systems must therefore go beyond finding locally relevant chunks and instead capture distributed factual dependencies that contribute to a legally meaningful answer (Bendahman et al., 2025).

(iv) Provenance and Traceability: In legal applications, the provenance of information is of high

importance. Answering a question correctly is insufficient if the supporting text is retrieved from the wrong source document (Uke et al., 2025). For instance, pulling a clause from a similar but distinct contract would undermine the legal validity of the generated output and erode user trust (Joshua et al., 2025; Hindi et al., 2025). Consequently, legal professionals require a transparent and verifiable "reasoning trail" from the generated answer back to the specific clauses in the original source document (Richmond et al., 2024). This need for an auditable path, where every piece of information can be validated against its source, makes document-faithful retrieval a fundamental measure of a system's reliability.

2.3 Focus on the Pre-Retrieval Stage: Chunking and Context Enrichment

The performance of a RAG system is heavily influenced by the pre-retrieval phase, where the knowledge base documents are processed and indexed. The dominant practice is *chunking*, that breaks down large documents into smaller pieces for efficient indexing in a vector database. This process must balance efficiency (smaller chunks), relevance (precise chunks), and context preservation (semantically complete chunks) (Barnett et al., 2024; Gao et al., 2023b). Naive chunking methods like fixed-size splitting can fragment logical units, leading to incomplete text snippets. More advanced strategies aim to preserve meaning. Recursive character splitting, for instance, divides text along natural boundaries like paragraphs and punctuation[2]. Semantic chunking uses language models to identify natural breakpoints, ensuring each chunk encapsulates a complete idea. Yet, even these methods struggle with legal text, as they may miss provisions spanning multiple sections or fail to handle nested clauses effectively (Ferraris et al., 2024; Kalra et al., 2024; Qu et al., 2024).

This can be explained due to a fundamental limitation of chunking: the inevitable loss of global context. Each chunk is embedded as an isolated vector, disconnected from the broader document it belongs to. This isolation is a primary cause of what we identify and later formally define as Document-Level Retrieval Mismatch (DRM), a critical failure where the retriever selects chunks from entirely incorrect source documents that happen to share superficial similarities with the query. This is par-

ticularly problematic in legal databases with numerous structurally similar documents. While context loss is a known issue (Ferraris et al., 2024; Günther et al., 2024), DRM has not been formally quantified in the legal NLP literature.

To combat this context loss, various general context enrichment strategies have been developed. A straightforward local approach is to expand retrieved chunks to include surrounding sentences, a technique referred to as "Small2Big". A more global approach involves adding metadata, which can be either standard (timestamps, authors, titles) or artificially generated. Examples of artificial metadata include chunk-specific explanatory context, such as in Contextual Retrieval by Anthropic[3], or synthetic questions that a chunk could answer, as seen in methods like Reverse HyDE and QuIM-RAG (Gao et al., 2023a; Saha et al., 2024). Our Summary-Augmented Chunking falls into this last category, focusing on a lightweight, scalable approach where a single document-level summary provides global context to every chunk derived from it.

While our work focuses on this practical technique, other research has explored more architecturally complex solutions. These include methods that rethink the indexing structure, such as the hierarchical approach in RAPTOR (Sarthi et al., 2024), or knowledge graphs that model legal relationships (Kalra et al., 2024). The recent Late Chunking method (Günther et al., 2024) preserves more semantic context information by first embedding a document's full content and then performing chunking at the embedded level. Finally, the emergence of long-context models that can process hundreds of thousands of tokens presents a potential alternative to the chunking paradigm altogether, a use case specifically highlighted by OpenAI for lengthy legal documents (OpenAI, 2024).

However, these advanced methods often introduce significant computational overhead or implementation complexity. Our research, in contrast, deliberately focuses on a practical, modular, and resource-efficient technique. Furthermore, we investigate a novel aspect of context enrichment by examining how domain expertise from legal professionals can be used to create more powerful, legally-informed summaries.

---

[2]https://python.langchain.com/docs/how_to/recursive_text_splitter/

[3]https://www.anthropic.com/engineering/contextual-retrieval

a) LegalBench-RAG Metrics:

User Query:
Consider the Consulting Agreement between Emerald Health Naturals, Inc. and Dr. Gaetano Morello; Can this contract be terminated for convenience, and under what conditions?

The retriever searches through the knowledge base containing chunks from all documents

Retrieved Text Chunk (Success):
"5.4 Emeralds' Right to Terminate this Agreement without Cause. The Company may terminate this Agreement and the engagement of the Contractor without Cause"
origin document: consulting_agreement_emerald.txt

b) Summary Augmented Chunking:

Summarize Document

Document Summary:
Consulting agreement: Emerald Health hires Dr. Gaetano Morello Inc. [...]

Text Chunk:
"5.4 Emeralds' Right to Terminate [...] the Contractor without Cause"

Document Database

Recursive Chunking

Summary Augmented Chunk

Chunk Embedding Database

Ground Truth Document:
consulting_agreement_emerald.txt

Ground Truth Reference in Document:
[...] 5.4 Emeralds' Right **Start** minate this Agreement without Cause. The Company may terminate this Agreement and the engagement of the Contractor without Cause at any time on 30 days prior written notice. The date of termination will be [...] **End**

Ground Truth Snippet Span:
Start to End [6090, 6226]

Document-Level Retrieval Metric:
DRM

Character-Level Retrieval Metric:
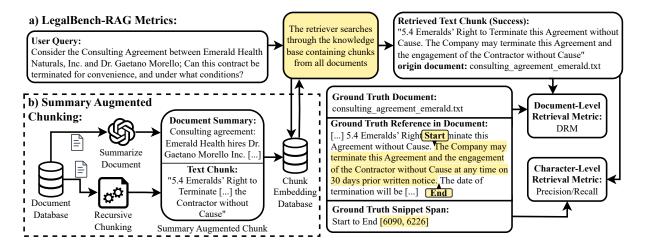Precision/Recall

Figure 1: Part a) illustrates how our retrieval quality metrics, Document-Level Retrieval Mismatch (DRM) and text-level precision/recall, are computed in the LegalBench-RAG (Pipitone and Alami, 2024) information retrieval task. Part b) shows the process of setting up the knowledge base using Summary Augmented Chunks (SAC).

3 Methodology

3.1 Task Definition and Dataset

To evaluate improvements in a RAG pipeline, a benchmark must be able to isolate the performance of the retrieval stage from the final generative output. Widely adopted benchmarks like LegalBench (Guha et al., 2023) and LexGLUE (Chalkidis et al., 2021) are designed to test the intrinsic reasoning capabilities of LLMs. Consequently, when these are used to evaluate RAG systems, any performance changes are difficult to attribute specifically to the retrieval component, as its contribution is blended with the model's internal knowledge. For this reason, our study uses *LegalBench-RAG* (Pipitone and Alami, 2024), a recently developed benchmark specifically designed to isolate and evaluate the retrieval component of RAG systems in the legal domain. It is constructed from the well-established LegalBench corpus.

The LegalBench-RAG benchmark comprises multiple datasets that target distinct types of legal documents: **(i) CUAD** (Contract Understanding Atticus Dataset, (Hendrycks et al., 2021)), which contains general contracts; **(ii) MAUD** (Merger Agreement Understanding Dataset, (Wang et al., 2023)), consisting of merger agreements; **(iii) ContractNLI** (Koreeda and Manning, 2021), a dataset of non-disclosure agreements; and **(iv) PrivacyQA** (Ravichander et al., 2019), which includes privacy policies from mobile applications.

We measure the performance on LegalBench-RAG via document-level DRM and character-level precision/recall between the retrieved and ground-truth text snippets (see Figure 1a), offering a holistic measure of retrieval quality. While our current work focuses exclusively on this retrieval analysis, we are currently working on adapting a benchmark such as the Australian Legal QA dataset[4] for end-to-end performance evaluation in future work.

3.2 Problem of Document-Level Retrieval Mismatch (DRM)

We started by conducting diagnostic experiments with a standard RAG approach in LegalBench-RAG (Pipitone and Alami, 2024) to establish a baseline performance and understand the behavior of standard RAG systems on this task. We began by evaluating a range of retrieval architectures and embedding models to determine the typical performance ceiling and identify any systemic weaknesses. Across different configurations and models, overall retrieval scores remained consistently low (see Appendix C).

We identified a major bottleneck: across different architectures and embedding models, retrievers frequently select chunks from entirely incorrect source documents. We define **Document-Level Retrieval Mismatch (DRM)** as the proportion of top-k retrieved chunks that do not originate from the document containing the ground-truth text.

While DRM is a general challenge for retrieval systems and increases the probability of hallucinations in the subsequent generation of RAG systems (Hou et al., 2024a), its impact is particularly severe in the legal domain due to the high degree of lexical

---

[4] https://huggingface.co/datasets/isaacus/open-australian-legal-qa

4

and structural similarity across documents (Ferraris et al., 2024). For instance, when we tested a standard RAG pipeline on ContractNLI data (Koreeda and Manning, 2021), we observed DRM rates over 95% (Fig. 2a) in a pool of 362 documents. Our legal experts hypothesize that this may be due to the highly standardized, boilerplate nature of non-disclosure agreements, which are largely uniform apart from a few key variables. This linguistic homogeneity confuses retrieval models that rely on semantic similarity (or keyword matching), leading them to prioritize chunks that are textually similar to the query but from the wrong agreement. A concrete example of how the retriever fails on similar contracts can be found in Section 5.2.

This problem of lexical and structural similarity is not limited to contracts: in any legal task, users need assurance that retrieved context truly comes from the intended document. For example, when answering a question about a privacy policy, pulling text from a different but similar policy undermines both factual accuracy and trust in the system. Even if the generated answer happens to be correct, legal professionals expect document-faithful reasoning, making DRM a key measure of whether retrieval respects source boundaries.

3.3 A Simple Solution: Summary Augmented Chunking (SAC)

To combat DRM, we experimented with a simple methodology that we named Summary-Augmented Chunking (SAC). SAC works as follows (see Figure 1b):

(i) Summarization: For each document in the corpus, we use an LLM to generate a single, concise summary as "document fingerprint", approximately 150 characters long. A detailed analysis of length and its impact is provided in Appendix A. **(ii) Chunking**: We employ a recursive character splitting strategy to partition the document's content into smaller, manageable chunks. This established method performs well on our dataset, as supported by prior work (Kalra et al., 2024) and our own empirical results. **(iii) Augmentation**: We prepend the document-level summary to each chunk derived from that document. **(iv) Indexing**: The summary-augmented chunks are then embedded and indexed in a vector database for retrieval.

This approach injects crucial global context into each chunk, specifically to mitigate DRM by guiding the retriever to the correct source document. The method is highly practical, requiring only one additional LLM call per document and can be smoothly integrated into existing RAG pipelines with minimal computational overhead. The generic prompt used for summarization is the following:

> **Generic Summarization Prompt**
>
> **System:** You are an expert legal document summarizer.
> **User:** Summarize the following legal document text. Focus on extracting the most important entities, core purpose, and key legal topics. The summary must be concise, maximum {char_length} characters long, and optimized for providing context to smaller text chunks. Output only the summary text.
> **Document:** {document_content}

Because LLMs often deviate from the specified length, we allowed a tolerance of 20 characters. Outputs exceeding this limit were regenerated with a reduced char_length value.

3.4 Expert-Guided Summarization

While generic summaries provide a significant improvement, we hypothesize that tailoring summaries to the nuances of specific legal document types could further enhance retrieval performance. The especially high residual mismatch in datasets containing non-disclosure agreements and privacy policies (Fig. 2b) suggests that certain contractual language requires more sophisticated contextual cues, motivating our Expert-Guided summarization approach.

In collaboration with two legal experts[5], we engineered a more sophisticated "meta-prompt" that instructs the LLM to generate summaries as *distinct* as possible within a document type. It directs the model to identify and prioritize key differentiating legal variables. To test this, we focused on non-disclosure agreements and privacy policies, defining each type's key characteristics from legally required elements and highlighting distinguishing features such as party names, definitions of data categories and their processing (for privacy policies), and definitions of confidential information (for NDAs). Our resulting expert-guided summarization prompt can be found in Appendix D.

This idea is supported by recent findings in legal NLP. For instance, research on summarizing Italian tax law decisions has demonstrated that a modular, expert-validated approach provides a solid basis for downstream semantic search, a task analogous to

---

[5]An associate professor and a postdoctoral researcher in law, with expertise in data protection and private law.
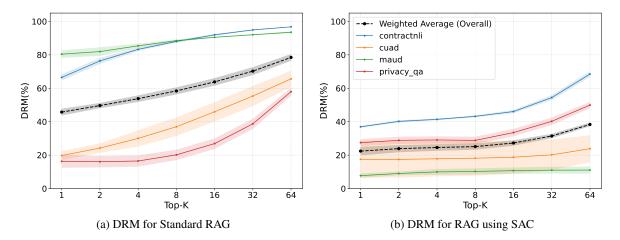
Figure 2: Document-Level Retrieval Mismatch (DRM) of a standard RAG approach (left) and using our Summary Augmented Chunking (right), applied to the 4 datasets in the LegalBench-RAG benchmark. Retrieval using SAC selects fewer wrong documents across all top-k retrieved snippets and seeds.

our retrieval application (Pisano et al., 2025). According to their findings, moving beyond generic summarization could be vital for complex legal texts. Their two-step method combines separate summary parts, whereas our "meta-prompt" uses conditional logic to integrate document type classification and summarization implicitly.

4 Experimental Setup

We evaluate the performance of our methods covering a broad picture of retrieval quality:

(i) Document-Level Retrieval Mismatch (DRM): As our primary metric, DRM directly measures the retriever's ability to identify the correct source document. A lower DRM indicates higher precision at the document level.

(ii) Text-Level Precision: It measures the fraction of all the retrieved text that is part of the ground truth text span. High precision means that the retrieved context is concise and contains minimal irrelevant "noise".

(iii) Text-Level Recall: It evaluates what fraction of the ground truth text was found by the retrieval system. High recall indicates that the system found all the necessary information.

As our baseline, we implemented a standard RAG pipeline using a *recursive character splitting* strategy with a chunk size of 500 characters and without overlap. The document summaries were generated using `gpt-4o-mini` (Hurst et al., 2024) and we instructed and processed the summaries to be about 150 characters long (details in the Appendix A). The concatenated texts were embedded

with `thenlper/gte-large`[6] (Li et al., 2023) (see Appendix C) and indexed in a FAISS (Johnson et al., 2019) vector database with cosine similarity as retrieval metric.

Dense semantic search excels at capturing conceptual similarity but may overlook exact lexical matches. In contrast, BM25 (Robertson and Zaragoza, 2009) is well established for keyword-based retrieval and can be effective for queries with unique identifiers or technical terminology. We therefore experimented with a hybrid dense+sparse retrieval. However, results showed that BM25 (sparse) improved DRM but decreased precision/recall while introducing additional computational overhead (see the Appendix B), so we decided to only use dense retrieval in the main experiments.

For all systems, we report performance across a range of top-k retrieved chunks. This reflects real-world deployment scenarios, where practitioners must balance precision and recall depending on application needs. Reporting the full curve enables a more informative assessment of trade-offs across retrieval strategies.

5 Results

5.1 Automatic Evaluation

We demonstrate that SAC significantly reduces DRM compared to the baseline, showcasing its effectiveness in providing necessary global context. The results, reported in Figure 2b, show a dramatic reduction in DRM across a wide range of hyper-parameters, effectively *halving the mismatch rate*.

---

[6]https://huggingface.co/thenlper/gte-large

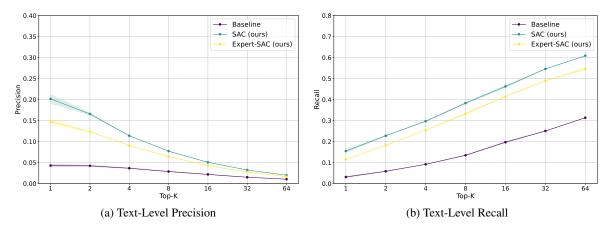(a) Text-Level Precision

(b) Text-Level Recall

Figure 3: Text-level precision (left) and recall (right) of the standard RAG approach and SAC with general or expert-guided summarization strategy. The metrics are averaged over all datasets and seeds.

Crucially, this improvement in document-level accuracy translates directly to improved text-level retrieval quality. By guiding the retriever to the correct document, RAG systems using SAC significantly outperform the standard RAG baseline on character-level precision and recall as well (Figure 3). Unexpectedly, Expert-Guided Summarization did not yield improvements over the general prompt (Figure 3). It resulted only in slightly better retrieval metrics in a few specific settings, such as with larger chunk sizes.

5.2 Qualitative Evaluation of Legal Experts

Beyond quantitative metrics, a qualitative analysis offers critical insights into how different summarization strategies impact retrieval, especially where the baseline struggles. We illustrate these observations with a representative example from the ContractNLI dataset, focusing on a query about Non-Disclosure Agreements (NDAs), highlighting baseline failure and contrasting generic versus expert-guided SAC.

Example 1: NDA (contractnli)

Query: "Consider Evelozcity's Non-Disclosure Agreement; Does the document allow the Receiving Party to independently develop information that is similar to the Confidential Information?"

Ground Truth:
Relevant document: *NDA-Evelozcity.txt*
Relevant snippets: "The obligations of the Recipient specified in Section 2 above shall not apply with respect to Confidential Information to the extent that such Confidential Information:"
"(d) is independently developed by or for the Recipient by persons who have had no access to or been informed of the existence or substance of the Confi-

---

dential Information."

A. Baseline Retrieval - No Summary
Retrieved document: *NDA-ROI-Corporation.txt* ✗
Retrieved snippet: "NON-DISCLOSURE AGREEMENT FOR PROSPECTIVE PURCHASERS"
Comment: Complete failure, distracted by structural similarity of an other irrelevant NDA header.

B. Using a 150-character summary generated with the generic meta-prompt
Retrieved document: *NDA-Evelozcity.txt* ✔
Summary: "Non-Disclosure Agreement between Evelozcity and Recipient to protect confidential information shared during a meeting."
Retrieved snippet: "; or (d) is independently developed by or for the Recipient by persons who have had no access to or been informed of the existence or substance of the Confidential Information."
Comment: Successful document-level retrieval (97% precision, 50% recall). The generic summary effectively guided to the correct document and relevant clause.

C. Using a 150-character summary generated with the expert meta-prompt
Retrieved document: *NDA-Evelozcity.txt* ✔
Summary: "NDA between Evelozcity and Recipient; covers vehicle prototypes, confidentiality obligations, exclusions, 5-yr term, CA governing law."
Retrieved snippet: "NON-DISCLOSURE AGREEMENT
This NON-DISCLOSURE AGREEMENT (this "Agreement") is made as of this ___ day of __ 2019, by and between Evelozcity with offices at *address* (the "Disclosing Party"), and __ (the "Recipient")."
Comment: While the correct document was found, the snippet is the introductory boilerplate, completely useless for the query.

D. Using a 300-character summary generated with the expert meta-prompt
Retrieved document: *NDA-Evelozcity.txt* ✔
Summary: "**Definition of Confidential Information**: Non-public vehicle prototypes and company plans. **Parties**: Disclosing Party: Evelozcity, CA; Recipient: [Name Not Provided]. **Obligations**: Keep confidential, limit access to affiliates,

7

use only for evaluation. **Exclusions**: Public knowledge, prior possession, independent development."
Retrieved snippet: The same snippet as in case C.
Comment: Similar to C, correct document, but irrelevant boilerplate snippet despite richer summary.

The ground truth in Example 1 consists of two related snippets addressing the independent development of information similar to confidential material. It demonstrates the baseline's (A) complete failure to identify the correct document, highlighting the high Document-Level Retrieval Mismatch (DRM) caused by lexical redundancy and structural similarities in legal corpora. Both generic (B) and expert-guided (C, D) SAC approaches, however, successfully guided the retriever to the correct source document, clearly showing SAC's effectiveness in mitigating DRM by injecting global context.

Crucially, while document-level retrieval improved, a key difference emerged in snippet quality. The top-ranked chunk with the generic summary (B) was one of the correct snippets, directly engaging with the query. Conversely, both expert-guided summaries (C and D), despite retrieving the correct document, yielded an irrelevant introductory boilerplate snippet.

From a legal expert perspective, the expert-guided summaries (especially D) are richer, more structured and contain highly discriminative information for differentiating between NDAs (e.g., parties, subject matter, duration, exclusions). Yet, this legal assessment contrasts with observed retrieval performance. Expert summaries, while legally more informative and superior for differentiating documents, did not translate to better text-level snippet retrieval. This counter-intuitive result requires a more technical explanation (we explore in the next Section 5.3) and suggests a complex interaction within the embedding space, where highly specific, dense legal information may not be optimally processed for general query alignment.

5.3 Discussion

Our findings demonstrate that summary-based context enrichment provides a robust and scalable solution to a fundamental weakness of RAG in the legal domain: the loss of global context during chunking. By prepending document-level summaries to each text chunk, our method helps in guiding the retriever toward the correct source document, as evidenced by the drastic reduction in

DRM. This is particularly valuable in legal corpora, where high structural similarity and standardized language make cross-document confusion a dominant failure mode. The success of this simple intervention underscores the importance of preserving document-level semantics in a domain where the overarching context dictates the meaning of individual clauses, a critical aspect for reliable legal NLP applications.

Interestingly, our experiments revealed that generic summaries consistently outperformed expert-guided ones, a counter-intuitive result given the legal precision of the latter. As highlighted in Section 5.2, from a legal perspective, expert-guided summaries successfully capture the distinctive elements required to differentiate between contracts of the same type. However, for the purpose of retrieval, we hypothesize two potential technical explanations for this observed performance gap.

First, generic summaries may strike a better balance between distinctiveness and broad semantic alignment with a wider variety of potential queries. While legally more precise, highly specific, expert-driven cues in the summaries might inadvertently overfit to narrow features. This would improve retrieval only in very specific cases and reduce robustness across a broader range of user intents.

Second, the informationally dense and structured language of expert-guided summaries may pose challenges for smaller embedding models, which must compress both the summary and chunk into a single vector. To investigate this potential bottleneck, future experiments with stronger, more capacious embedding models are needed.

In general, the interaction between summaries added to a chunk within the embedding space is complex. A strong global signal from the summary could overshadow the local relevance of a chunk. Understanding this dynamic is critical to improve our approach. Interesting insights from a machine learning perspective could be gained when analyzing the embedding space directly. We plan to use clustering and dimensionality reduction techniques to visualize how the concatenation of summaries and chunks behaves in the embedding space.

From a practical perspective, our results highlight the value of simple, modular interventions in the pre-retrieval stage. Unlike more complex architectural solutions (e.g., knowledge graphs, late chunking, or long-context models), SAC is inexpensive, requiring only a single additional summary per document, and integrates seamlessly into ex-

isting RAG pipelines. This makes it scalable even to large and dynamically changing legal databases. For practitioners, generic SAC provides an easily adoptable technique that delivers tangible improvements without the need for domain-specific fine-tuning or significant infrastructure changes.

Finally, our findings contribute to the broader vision of "LLMs as legal readers" (Pałka et al., 2025). If future legal documents become longer and more comprehensive, retrieval reliability will be even more critical. Our approach represents a practical step toward building systems that can process such documents with greater reliability, making AI a more trustworthy partner in navigating the complexity of legal texts.

5.4 Limitations and Future Work

While promising, our work has several limitations. First, our experiments were restricted to particular categories of legal documents and conducted exclusively in English. These documents, while diverse, do not cover the full spectrum of legal text types, such as legislation, case law, or other types of contracts, which differ substantially in structure and interpretation. Moreover, legal meaning is highly jurisdiction-specific, and our datasets were largely restricted to common-law contexts.

Second, our analysis focused on an isolated intervention within a standard RAG pipeline to clearly measure its impact. While effective, the residual retrieval mismatch rates remain significant, indicating that SAC is a valuable component for reliable RAG but not a complete solution on its own. We believe that combining SAC with other well-researched modules is the most promising path toward achieving the reliability required for legal applications, with the next promising steps being:

(i) Extending the presented principle of summarization hierarchically, with summaries at the paragraph, section, and document level to provide context at multiple granularities. **(ii)** Applying query optimization methods (e.g., transformation, expansion, or routing) to bridge the semantic gap between user questions and the formal language of legal text chunks. **(iii)** Adding a reranking step where a more powerful model re-evaluates and re-orders the top-k retrieved chunks to improve the final selection before generation. It would also be valuable to benchmark SAC against other context-preserving chunking strategies, such as Late Chunking (Günther et al., 2024) and RAPTOR (Sarthi et al., 2024), to better understand its relative strengths.

Finally, this study focuses exclusively on the retrieval stage of the RAG framework. Future work will investigate how the DRM metric and SAC impact downstream generation through end-to-end benchmarking.

6 Conclusion

We addressed the critical challenge of retrieval reliability in RAG systems operating on large, structurally similar legal document databases. We identified and quantified *Document-Level Retrieval Mismatch* (DRM) as a dominant failure mode, where retrievers are often easily confused by legal boilerplate language and select text from entirely incorrect documents. Targeting this issue, we investigate *Summary-Augmented Chunking* (SAC), a simple and computationally efficient technique that prepends document-level summaries to each text chunk. By injecting global context, SAC drastically reduces DRM and consequently improves text-level retrieval precision and recall.

A key finding is that generic summaries outperform expert-guided ones focusing on key legal variables. For the purpose of guiding retrievers, broad semantic cues appear more robust and generalizable than dense, structured, legally precise summaries. This demonstrates that meaningful retrieval performance gains are achievable without heavy domain-specific engineering.

While SAC is not a full solution on its own, it offers a practical, scalable intervention for more reliable legal RAG systems. By improving the crucial retrieval step for legal information, our work brings us closer to a future where AI can truly serve as a trusted partner in the legal profession.

Acknowledgments

References

Mousumi Akter, Erion Çano, Erik Weber, Dennis Dobler, and Ivan Habernal. 2025. A comprehensive survey on legal summarization: Challenges and future directions. *arXiv preprint arXiv:2501.17830*.

Farid Ariai and Gianluca Demartini. 2024. Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges. *arXiv preprint arXiv:2410.21306*.

Kevin D Ashley. 2018. Automatically extracting meaning from legal texts: opportunities and challenges. *Ga. St. UL Rev.*, 35:1117.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pages 194–199.

Nihed Bendahman, Karen Pinel-Sauvagnat, and Gilles Hubert. 2025. EUR-Lex-Triples: A Legal Relation Extraction Dataset from European Legislation. In *Linking Theory and Practice of Digital Libraries: 29th International Conference on Theory and Practice of Digital Libraries, TPDL 2025, Tampere, Finland, September 23–26, 2025, Proceedings*, page 433. Springer Nature.

Alec Berntson. 2023. Azure ai search: Outperforming vector search with hybrid retrieval and reranking. Microsoft Community Hub. Accessed on August 19, 2025.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. LexGLUE: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Manola Cherubini, Francesco Romano, Andrea Bolioli, and Lorenzo De. 2024. Improving the accessibility of EU laws: the Chat-EUR-Lex project.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Andrea Filippo Ferraris, Davide Audrito, Giovanni Siragusa, and Alessandro Piovano. 2024. Legal chunking: Evaluating methods for effective legal text retrieval. In *Legal Knowledge and Information Systems*, pages 275–281. IOS Press.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, and Diego Zambrano. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

Michael Günther, Isabelle Mohr, Daniel James Williams, Bo Wang, and Han Xiao. 2024. Late chunking: contextual chunk embeddings using long-context embedding models. *arXiv preprint arXiv:2409.04701*.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: an expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Mahd Hindi, Linda Mohammed, Ommama Maaz, and Abdulmalik Alwarafy. 2025. Enhancing the precision and interpretability of retrieval-augmented generation (RAG) in legal technology: A survey. *IEEE Access*.

Abe Hou, William Jurayj, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2024a. Gaps or hallucinations? scrutinizing machine-generated legal analysis for fine-grained text evaluations. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 280–302.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2024b. CLERC: A dataset for legal case retrieval and retrieval-augmented analysis generation. *arXiv preprint arXiv:2406.17186*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. GPT-4o System Card. *arXiv*, abs/2410.21276.

Lasal Jayawardena, Nirmalie Wiratunga, Ramitha Abeyratne, Kyle Martin, Ikechukwu Nkisi-Orji, and

Ruvan Weerasinghe. 2024. SCaLe-QA: Sri Lankan case law embeddings for legal QA.

J. Johnson, M. Douze, and H. Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

C. Joshua, S. Karkala, S. Hossain, M. Krishnapatnam, A. Aggarwal, Z. Zahir, V. Pandhare, and V. Shah. 2025. Domain adaptation for legal document summarization with pre-trained language models. *Artificial Intelligence and Law*.

Rishi Kalra, Zekun Wu, Ayesha Gulley, Airlie Hilliard, Xin Guan, Adriano Koshiyama, and Philip Treleaven. 2024. HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications. *arXiv preprint arXiv:2409.09046*.

Yuta Koreeda and Christopher D Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Zihao Li. 2023. The dark side of ChatGPT: Legal and ethical challenges from stochastic parrots and hallucination. *arXiv preprint arXiv:2304.14347*.

Ruta Liepina, Giuseppe Contissa, Kasper Drazewski, Francesca Lagioia, Marco Lippi, Hans-Wolfgang Micklitz, Przemysław Pałka, Giovanni Sartor, and Paolo Torroni. 2019. GDPR privacy policies in CLAUDETTE: Challenges of omission, context and multilingualism. In *CEUR Workshop Proceedings*, volume 2385, pages 1–7.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2025. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242.

Silvia Martinelli. 2023. AI as a tool to manage contracts consequences and challenges in applying legal tech to contracts management. *European Review of Private Law*, 31(2/3).

OpenAI. 2024. Model Selection Guide: Use Case: Long-Context RAG for Legal QA. Web Page. Accessed on August 12, 2025.

Przemysław Pałka, Francesca Lagioia, Rūta Liepina, Marco Lippi, and Giovanni Sartor. 2025. Make privacy policies longer and appoint LLM readers. *Artificial Intelligence and Law*, pages 1–33.

Xiao Peng and Liang Chen. 2024. Athena: Retrieval-augmented legal judgment prediction with large language models. *arXiv preprint arXiv:2410.11195*.

Nicholas Pipitone and Ghita Houir Alami. 2024. LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*.

Giuseppe Pisano, Alessia Fidelangeli, Federico Galli, Andrea Loreggia, Riccardo Rovatti, Piera Santin, and Giovanni Sartor. 2025. The Summarization of Italian Tax-Law Decisions: The Case of the PRODIGIT Project. *IEEE Access*.

Weicong Qin and Zhongxiang Sun. 2024. Exploring the Nexus of Large Language Models and Legal Systems: A Short Survey. *arXiv preprint arXiv:2404.00990*.

Renyi Qu, Ruixuan Tu, and Forrest Bao. 2024. Is semantic chunking worth the computational cost? *arXiv preprint arXiv:2410.13070*.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841*.

Karen McGregor Richmond, Satya M Muddamsetty, Thomas Gammeltoft-Hansen, Henrik Palmer Olsen, and Thomas B Moeslund. 2024. Explainable AI and law: An evidential survey. *Digital Society*, 3(1):1.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Binita Saha, Utsha Saha, and Muhammad Zubair Malik. 2024. Advancing retrieval-augmented generation with inverted question matching for enhanced QA performance. *IEEE Access*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Shailaja Uke, Ayush Laddha, Shreya Bambal, Atharva Bonde, and Prasanna Atram. 2025. Generative ai-powered conversational assistant for document, web, and repository interaction and retrieval. In *2025 International Conference on Computing Technologies & Data Communication (ICCTDC)*, pages 1–5. IEEE.

Michele Visciarelli, Giovanni Guidi, Laura Morselli, Domitilla Brandoni, Giuseppe Fiameni, Luisa Monti, Stefano Bianchini, and Cosimo Tommasi. 2024. SAVIA: Artificial Intelligence in support of the law-making process.

Steven H Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023. MAUD: An expert-annotated legal nlp dataset for merger agreement understanding. *arXiv preprint arXiv:2301.00876.*

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. CBR-RAG: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.

Rui Yang. 2024. CaseGPT: a case reasoning framework based on language models and retrieval-augmented generation. *arXiv preprint arXiv:2407.07913.*

Leonardo Zilli, Michele Corazza, Monica Palmirani, and Salvatore Sapienza. 2025. An agentic approach to retrieving and drafting legislative definitions. In *International Conference on Electronic Government and the Information Systems Perspective*, pages 190–204. Springer.

A Hyperparameter Chunk Size and Summary Size

We experimented with chunk sizes of 200, 500, and 800 characters, combined with prepended summaries of either 150 or 300 characters. The precision and recall results for all six configurations are reported in Table 1. For our final pipeline, we selected a chunk size of 500 characters, consistent with Pipitone and Alami (2024), and a 150-character summary, as this configuration yielded the most balanced trade-off between precision and recall.

B Dense Semantic Search or Sparse Keyword Search?

BM25 (Robertson and Zaragoza, 2009) is a well-established keyword-based (sparse) retrieval method. A common assumption in RAG research is that hybrid approaches, combining dense vector similarity with keyword matching, often yield the best results (Berntson, 2023). Following this intuition, we augmented our dense retriever with a BM25 component, allowing the system to explicitly match salient terms (e.g., party names) from the prepended summaries with the same terms in the user query. Results are reported in Table 2.

Chunk	200		500		800	
Sum.	150	300	150	300	150	300
Prec.(%)	10.64	8.05	**11.03**	8.45	7.76	6.79
Rec.(%)	23.43	22.91	41.80	37.77	42.93	**43.90**
DRM(%)	20.70	23.49	**19.29**	20.61	34.96	29.68

Table 1: Average Document-Level Retrieval Mismatch (DRM), precision, and recall across different chunk sizes (*Chunk*) and summary lengths (*Sum.*), both measured in characters. Reported values are averaged over seven top-k retrieval settings ($k \in 1, 2, 4, 8, 15, 32, 64$). Lower values indicate better performance for DRM, while higher values are better for precision and recall.

As expected, adding sparse retrieval improved document selection, reducing our document-level mismatch (DRM). However, it lowered text-level precision and recall. A closer inspection suggests a likely explanation: The summaries are highly structured and rich in identifiers, where sparse keyword matching excels, whereas the chunked document bodies contain more natural language and nearly no direct keywords to match. For pinpointing the relevant passage, semantic similarity between the query and text is more informative than keyword overlap, explaining the higher text-level precision and recall without the BM25 component.

In conclusion, adding BM25 search only contributed slightly to finding the correct document (improved DRM), but tended to result in poorer search results within a document (reduced precision and recall). Also considering the computational overhead of the BM25 algorithm, we decided to rely exclusively on semantic search.

$w_{semantic}$	100%	75%	50%	25%
$w_{keyword}$	0%	25%	50%	75%
Prec.($\uparrow$%)	**11.03**	10.57	9.54	8.23
Rec.($\uparrow$%)	41.80	**42.54**	41.47	36.56
DRM($\downarrow$%)	19.29	19.11	18.45	**18.18**

Table 2: Average Document-Level Retrieval Mismatch (DRM), precision, and recall for different weightings of semantic similarity ($w_{semantic}$) and lexical similarity ($w_{keyword}$, BM25). Metrics are averaged over seven top-k values ($k \in \{1, 2, 4, 8, 15, 32, 64\}$). Lower DRM and higher precision/recall indicate better performance. Results are based on the optimal chunk size of 500 characters and summary length of 150 characters identified in Section A.

C   Embedding Model Ablation

The choice of embedding model plays a crucial role in retrieval performance, as embeddings determine how effectively the system identifies text snippets semantically similar to a given query. To assess this impact, we conducted a brief model ablation study, with results shown in Figure 4. Among the tested models, OpenAI's `text-embedding-3-large`[7] achieved the strongest results overall. However, due to concerns about API rate limits and the importance of reproducibility, we opted for an open-source alternative. Within this category, the best-performing model was `thenlper/gte-large`[8], which we therefore selected for all subsequent experiments.

D   Expert-Guided Prompt Template

Together with legal experts, we developed an expert-informed prompt template aimed at generating more distinctive legal text summaries. We focused on two document types: Non-Disclosure Agreements and privacy policies, that posed particular challenges for the generic prompt. The LLM was instructed to first identify the document type and then apply the corresponding template for summarization. The full prompt used in our experiments is provided below.

Expert-Guided Prompt Template

System: You are a legal summarization expert.
User: Your task is to generate a highly distinct, structured summary of the provided legal document. The primary goal is to extract the unique identifiers that differentiate this document from others of the same type. This summary will be used as context to smaller text chunks for a retrieval system.
Follow this two-step process:
- First, internally identify the document type from the following options: Non-Disclosure Agreement (NDA), Privacy Policy, or Other.
- Second, generate the summary based on the specific template corresponding to the identified document type.

Document type Non-Disclosure Agreement (NDA): An NDA is a legally binding contract between specific parties that outlines confidential information to be kept secret. If the document is an NDA, your summary should align with the following template:
- Definition of Confidential Information, specifying what types of information are considered confidential, e.g. such as: Technical data, Business plans, Customer lists, Trade secrets, Financial information
- Parties to the Agreement identifying the disclosing party and the receiving party (or both, if mutual

---

NDA), e.g. such as: Full legal names, Affiliates or representatives covered, Roles of each party
- Obligations of the receiving party outlining what the receiving party is required to do, e.g. such as: Keeping the information secret, Limiting disclosure to authorized personnel, Using the information only for specified purposes
- Exclusions from confidentiality describing information that is not protected under the NDA, such as: Information already known to the receiving party, Publicly available information, Information disclosed by third parties lawfully, Independently developed information
- Specifying any exceptions where disclosure is allowed, such as: To employees or advisors under similar obligations, If required by law or court order (with notice to the disclosing party)
- Term and Duration, defining how long the confidentiality obligation lasts: Often includes both the duration of the agreement and the period during which information remains protected (e.g., "3 years after termination")
- Purpose of Disclosure (Use Limitation), stating the specific reason the information is being shared (e.g., for evaluating a partnership, conducting due diligence, etc.) and prohibits other uses.
- Remedies for Breach, detailing the consequences of violating the NDA, which may include: Injunctive relief (court orders to stop disclosure), Damages, Legal fees
- Governing Law and Jurisdiction, identifying which country/state's laws apply and where disputes will be settled.
- Miscellaneous Clauses (Boilerplate), may include: No license granted, Entire agreement clause, Amendment process, Counterparts and signatures

Document type Privacy Policy: A privacy policy is issued by a private or public entity to inform users how their personal data is processed (e.g., collected, used, shared, stored). If the document is a privacy policy, your summary should align with the following template:
- Personal Data Collected and Processed, specifying what categories of personal data are collected and how. This may include: Name and surname, Contact information, Financial details, Device and browser data, Location information, Inferred preferences or behaviors
- Identity and Contact Details of the Controller, identifying the organisation responsible for the processing. May include: Full legal name of the controller, Contact email or phone number, Details of any representative (if applicable)
- Purposes of Processing, outlining why the personal data is collected and how it will be used. Examples include: Service provision and operation, Personalisation of content or features, Marketing and advertising, Analytics and performance monitoring, Payment processing
- Legal Basis for Processing, specifying the lawful grounds relied upon. These are: Consent of the data subject, Performance of a contract, Compliance with a legal obligation, Protection of vital interests, Task carried out in the public interest, Legitimate interests of the controller or third party
- Recipients of the Data, listing who may receive the data, including: Service providers and processors, Business partners, Public authorities (where legally

---

[7] https://openai.com
[8] https://huggingface.co/thenlper/gte-large
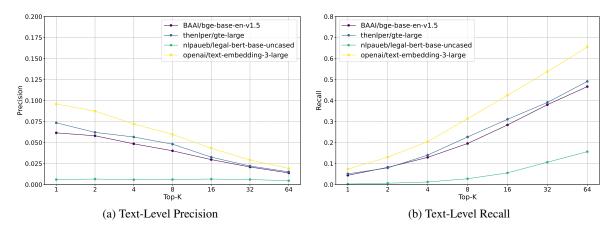
(a) Text-Level Precision (b) Text-Level Recall

Figure 4: Relative performance comparison of four embedding models in the baseline case on the LegalBench-RAG dataset (Pipitone and Alami, 2024).

required), Affiliates and subsidiaries
- International Data Transfers, describing whether personal data is transferred outside the jurisdiction and, if so: Destination countries, Safeguards applied (e.g., Standard Contractual Clauses, adequacy decisions)
- Data Retention, defining how long the personal data will be stored, or the criteria for determining the period. May include: Fixed retention periods, Purpose-based retention (e.g., "as long as necessary to provide the service"), Archiving or deletion policies
- Data Subject Rights, explaining individuals' rights under data protection law, including: Right to access personal data, Right to rectify inaccuracies, Right to erasure ("right to be forgotten"), Right to restrict or object to processing, Right to data portability
- Right to Lodge a Complaint, providing information on: The data subject's right to contact a supervisory authority, Name or link to the competent authority
- Automated Decision-Making, disclosing whether such processing occurs and, if so: The logic involved, Potential significance of the decisions, Expected consequences for the data subject

Other document type: If the document does not match the types above, summarize the following general legal document in a structured, concise way. Identify for your summary the important entities, core purpose, and other unique identifiers that differentiate this document from others of the same type.

General Rules:
- The summary must be concise and under {char_length} characters.
- Ignore every field in the template where the information is not present in the document.
- Prioritize extracting the most critical identifiers, such as parties, dates, and the specific subject matter.
- Output ONLY the final summary text!
Here is the document you should summarize: {document_content}