A Summarization Framework with Self-Improvement for RAG Retrieval in the Legal Domain

Yuriy Perezhohin*, Flavio Riedlinger*, Victor Costa*, Mauro Castelli*
*NOVA Information Management School (NOVA IMS), Universidade NOVA de Lisboa, Portugal
email:{yperezhohin,fmagalhaes,vcosta,mcastelli}@novaims.unl.pt

***Abstract*—Retrieval-Augmented Generation (RAG) systems are increasingly used for legal question answering, but their effectiveness depends critically on the quality of the representations indexed during retrieval. Legal documents pose particular challenges due to their length, hierarchical structure, and domain-specific semantics, which can result in semantically weak or incomplete representations.This work proposes a quality-controlled representation framework that targets the retrieval stage of legal RAG systems. Documents are segmented into fixed-size chunks and summarized using a lightweight language model. A second model, acting as a synthetic evaluator based on the G-EVAL methodology, assesses the semantic adequacy of each summary and triggers refinement when necessary. Only validated summaries are admitted into the vector database.The acceptance criterion governing this process is analyzed through a sensitivity study that characterizes the trade-off between retrieval effectiveness and computational overhead. Experiments on the LEGALBENCH-RAG benchmark demonstrate that quality-controlled summaries yield substantial gains over untreated indexing. In a direct comparison using the same model, pairing text-ada-large-3 with our validated summaries increases accuracy at k = 8 from 36.63% to 72.70%. Because retrieval operates on compact, semantically verified representations rather than raw long chunks, the approach improves computational efficiency and narrows the performance gap between proprietary and lightweight models: MultiQA with validated summaries reaches 64.24% accuracy, outperforming the closed-source text-ada-large-3 baseline without quality control.**

***Index Terms*—Retrieval-Augmented Generation; Legal; NLP, Large Language Model; Embeddings**

I. Introduction

Legal documents exhibit inherent complexity due to their specialized terminology, extensive contextual dependencies, and rigid hierarchical structure. These characteristics pose significant challenges for natural language processing (NLP) systems, particularly in the automation of legal information retrieval and reasoning tasks [1]. Traditional approaches based on keyword matching and Boolean queries often fail to capture the semantic structure of legal texts, resulting in suboptimal retrieval performance [2]. Likewise, statistical models and distributional machine learning approaches struggle with long-range dependencies and document length, which are pervasive in legal corpora.

Retrieval-Augmented Generation (RAG) systems [3] introduce a hybrid paradigm that integrates information retrieval with large language models (LLMs), enabling retrieval-based grounding of generative outputs. In this architecture, a retriever selects relevant textual segments, which are subsequently processed by a generator to produce structured answers. Although RAG architectures have shown utility in general-domain tasks, their application to legal corpora raises domain-specific challenges, including fragmentation of legal reasoning, and disruption of argumentative coherence due to document segmentation.

Within legal RAG pipelines, retrieval errors propagate directly into the generation stage, producing legally inaccurate, incomplete, or contextually inconsistent outputs. Prior approaches have explored reinforcement learning strategies [4] and ontology-guided retrieval mechanisms [5] to improve factual consistency and semantic alignment. However, these methods remain constrained by the structural limitations of chunk-based indexing, particularly in long legal documents.

This article introduces an iterative, quality-controlled summarization framework designed to restructure the retrieval stage of legal RAG systems. Instead of indexing raw document chunks, each segment is transformed into a structured summary that serves as the fundamental retrieval unit. To preserve semantic adequacy and legal fidelity, each summary is evaluated through a synthetic evaluator based on the G-EVAL methodology [6], which assigns a normalized quality score reflecting alignment with the original legal content.

Summaries that do not satisfy a predefined quality threshold are refined through an iterative feedback process guided by evaluator-generated justifications. This mechanism establishes a controlled trade-off between semantic preservation and representational compression, enabling systematic reduction of token length while maintaining legal relevance. The quality threshold operates as a dynamic control parameter regulating refinement depth and computational cost, rather than as a static design constraint.

The proposed framework redefines the retrieval representation layer of RAG architectures by introducing validated summaries as retrieval primitives, enabling a structured balance between semantic fidelity, contextual coherence, and retrieval efficiency. This design allows embedding models to operate on representations that preserve legal structure while reducing noise induced by raw segmentation.

The approach is empirically evaluated on the LEGALBENCH-RAG benchmark [7], using multiple

embedding models and retrieval configurations. Results indicate systematic improvements over conventional chunk-based RAG pipelines across different retrieval depths, showing that quality-controlled summarization provides a scalable strategy for enhancing retrieval performance in legal-domain RAG systems. All of the code and generated data are available at https://github.com/yuriyvnv/Legal-Summarization-RAG.

II. Related Work

The development of RAG is rooted in the evolution of information retrieval (IR) systems. Early systems used several techniques to retrieve documents, such as Boolean Retrieval [8] and Vector Space Models (VSM) [9]. They laid the foundation for document retrieval by representing text in a structured manner and enabling term-based similarity comparisons. Boolean Retrieval matches documents that precisely satisfy a query using logical operators, ensuring only exact matches are returned. In contrast, VSM represents documents and queries as vectors and ranks them based on cosine similarity, enabling partial matches and relevance-based ranking. These methods were later refined with Latent Semantic Indexing (LSI) [10], which introduced a mathematical technique to uncover hidden relationships between words by reducing the dimensionality of the term-document matrix using Singular Value Decomposition [11] (SVD). LSI captures latent patterns in word usage across documents, enabling retrieval systems to recognize synonyms and related terms even if they do not explicitly appear in the query. By mapping terms and documents into a lower-dimensional space, LSI improves recall and enhances the system's ability to retrieve semantically relevant results beyond simple keyword matching.

As search engines and information retrieval systems advanced, probabilistic models such as the BM25 ranking function [12] became widely adopted, leveraging term frequency and inverse document frequency [13] (TF-IDF) to rank documents based on their relevance to a given query [14]. These methodologies remained the backbone of IR until the advent of deep learning and transformer-based architectures.

With the rise of neural embeddings, information retrieval shifted from term-based similarity to semantic search, enabling more effective and context-aware document retrieval. This transition paved the way for the emergence of RAG systems [3], a modification that integrates retrieval-based and generation-based NLP models. Unlike traditional text generation models that rely solely on pre-trained knowledge, RAG dynamically retrieves relevant information from external sources [15], thereby enhancing the accuracy and factuality of responses.

The RAG framework consists of three key components: (1) The Chunker: Splits information of different documents into a manageable form to be embedded. Various strategies have been proposed, such as the Recursive Text Splitter, which splits sentences at specific endpoints at the end of sentences [16]. (2) Retriever Module: Identifies and retrieves relevant passages from a knowledge corpus, often leveraging embeddings. There are specific frameworks, such as FAISS [17] and ColBERT [18], that leverage sophisticated methodologies beyond embeddings to retrieve the correct passage. (3) The Generator: Uses the retrieved context to generate coherent and contextually relevant responses, typically employing an LLM based on transformer architectures.

RAG has increasingly been applied to legal AI, where retrieval-augmented models must navigate jurisdictional dependencies, hierarchical legal texts, and intricate legal language. One of the key contributions in this area, Legal Query RAG [19], introduced a recursive retrieval mechanism designed to enhance legal case precedent retrieval. By employing an audit agent, this pipeline dynamically adjusted the retrieval document based on initial query results, ensuring that retrieved legal documents were highly relevant and legally contextualized. In addition to the retrieval adjustment, the authors fine-tuned an embedding model on legal corpora to demonstrate the importance of specific text representations. Their research improved the overall precision by 13% on the retrieval stage.

Similarly, privacy concerns in legal AI led to the development of privacy-preserving retrieval mechanisms. SLANGO [20] introduced a confidentiality-aware RAG pipeline that ensured secure retrieval and processing of sensitive legal queries. By incorporating privacy-preserving embeddings and secure retrieval protocols, this system enabled legal professionals to query sensitive legal databases without exposing confidential information. This advancement aligned RAG with data protection laws such as GDPR, making it a critical development for AI-powered legal research tools in highly regulated environments.

Furthermore, researchers addressed the challenges of information overload by enhancing case law retrieval and statute law entailment [21]. For case law retrieval, the authors employ a sentence-transformer model to generate numerical representations of case paragraphs, utilizing similarity histograms and binary classifiers to identify pertinent cases. In statute law entailment, they fine-tune a DeBERTa large language model on natural language inference datasets to determine entailment relationships between legal queries and statutory articles. This comprehensive approach significantly improves the accuracy and efficiency of legal information processing.

The effectiveness of RAG systems depends on the accuracy and relevance of the retrieved content, making robust benchmarking essential for assessing their performance. Several benchmarking datasets have emerged to facilitate comparative analysis, ensuring that retrieval-augmented models meet the community's growing demands. One of the most comprehensive benchmarks addressing these needs is RAGBench [22], which provides a structured evaluation framework, *TRACe*, for general-purpose RAG models. Unlike previous evaluation methods focused solely on retrieval accuracy, RAGBench introduces evaluation metrics that assess factors such as Utilization, Relevance, Adherence, and Completeness. This dataset allows researchers to compare different architectures and retrieval strategies in open-domain question answering, knowledge-grounded dialogue systems, and document retrieval tasks.

Beyond general-purpose benchmarks, domain-specific RAG evaluation has become increasingly important, particularly in highly specialized fields such as law. The CLERC Dataset [23] was developed to evaluate case law retrieval accuracy, focusing on the effectiveness of RAG systems in identifying corresponding citations for a given chunk of legal documents. Besides retrieval accuracy, they also measured the generated text from retrieved citations and evaluated it using the ROUGE-F score [24].

Unlike open-domain knowledge systems, legal retrieval models require precise alignment with statutes, case law, and jurisdictional interpretations, which demands an entirely different evaluation approach. LEGALBENCH [25] was initially developed to evaluate LLMs' reasoning capabilities across various legal tasks, focusing on logical synthesis rather than retrieval. However, legal retrieval models require precise alignment with statutes, case law, and jurisdictional interpretations, which necessitates a distinct evaluation approach. To address this need, LEGALBENCH-RAG [7] was introduced as the structured benchmark specifically designed to assess the retrieval performance of RAG models within the legal domain. This dataset critically evaluates, based on precision and recall, how well retrieval-augmented models identify relevant passages across long, extensive documents.

III. Methodology

This section describes the methodological framework developed to evaluate a Retrieval-Augmented Generation (RAG) pipeline on the LEGALBENCH-RAG benchmark. The proposed methodology focuses exclusively on improving the quality of the textual representations used in the retrieval stage, while keeping the downstream generation component unchanged.

The framework is structured into four sequential steps. First, legal documents are segmented into fixed-size chunks and summarised to obtain compact representations. Second, a quality-controlled validation mechanism is applied to these summaries using a self-improvement feedback loop. Third, the validated summaries are indexed into multiple vector databases using different embedding models. Finally, retrieval performance is evaluated following the official LEGALBENCH-RAG protocol. Figure 1 illustrates the steps taken to develop and evaluate the proposed methodology.

A. Document Segmentation and Summarization

The first step in the pipeline is segmenting legal documents into manageable textual units and generating concise summaries for each unit. All documents are taken from the LEGALBENCH-RAG benchmark [25], which contains heterogeneous legal corpora characterized by long documents, hierarchical structure, and dense domain-specific terminology.

To ensure comparability with the benchmark baseline, each document is segmented using a fixed-size chunking strategy. Specifically, documents are split into chunks of 500 tokens with an overlap of 100 tokens, following the same configuration adopted in the original LEGALBENCH-RAG setup. No semantic or adaptive chunking strategy is employed at this stage.

Formally, let

$$D = \{d_1, d_2, \ldots, d_n\}$$

denote the set of documents. Each document d_i is partitioned into a sequence of chunks

$$C_i = \{c_{i1}, c_{i2}, \ldots, c_{im}\},$$

where each chunk c_{ij} corresponds to the j-th segment extracted from document d_i and contains at most 500 tokens. Each chunk is then summarised independently using a lightweight instruction-tuned LLAMA 3B language model [26]. The summarization prompt (https://github.com/yuriyvnv/Legal-Summarization-RAG) is designed to prioritize legally salient information, including obligations, rights, conditions, and exceptions, while preserving domain-specific terminology.The summarization process is defined as:

$$s_{ij}^{(1)} = \text{LLAMA}_{\text{3B}}(c_{ij}, P_s), \tag{1}$$

where $s_{ij}^{(1)}$ denotes the initial summary generated for chunk c_{ij}, and P_s represents the summarization prompt.

The proposed compression aims to produce a dense and legally informative representation of the original chunk, suitable for embedding-based retrieval. Importantly, summaries are generated and stored at the chunk level only, and each summary remains explicitly linked to its originating chunk.

Under this design constraint, the method remains directly comparable to standard chunk-based RAG pipelines, as the retrieval unit and segmentation granularity are preserved, allowing the effect of representation quality on retrieval performance to be isolated.

B. Quality-Controlled Self-Improvement using G-EVAL

Summarization reduces token usage and alters the structure of legal representations used in retrieval pipelines. However, large language models may generate summaries that omit legally relevant information or introduce factual and structural inconsistencies. To control these risks at the representation level, a quality-control mechanism is applied to each generated summary before it is integrated into the retrieval layer.

For each document chunk c_{ij}, an initial summary $s_{ij}^{(1)}$ is generated using the LLaMA-3B model, as described in Section III-A. This summary is then evaluated by a larger LLaMA-8B model following the G-EVAL framework [6], which operationalizes the LLM-as-a-judge paradigm. The evaluator compares the generated summary with the original chunk and assigns a quality score using a structured evaluation prompt.

The evaluation prompt encodes multiple assessment dimensions that reflect legal and semantic adequacy. These dimensions include: (i) factual consistency with the source text, (ii) preservation of legally salient information such as obligations, conditions, and exceptions, (iii) structural coherence and clarity, and (iv) absence of hallucinated or unsupported statements.

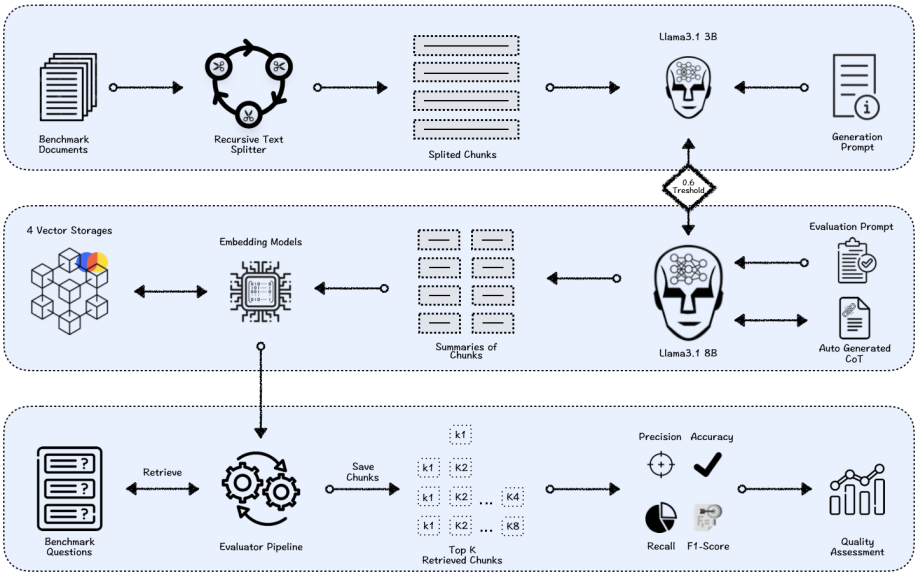

Fig. 1. Overview of the proposed framework with evaluation on the LEGALBENCH-RAG benchmark.

Each dimension is scored independently and normalized to the interval $[0, 1]$.

At a given iteration step t, the evaluator may assess one or more summary candidates. Let $s_{ij,n}^{(t)}$ denote the n-th candidate summary for chunk c_{ij} at iteration t. For each candidate, the evaluator produces a quality score

$$q_{ij,n}^{(t)} \in [0, 1],$$

where higher values indicate closer semantic and legal alignment with the source chunk.

The overall quality score for chunk c_{ij} at iteration t is computed as the average across the evaluated candidates:

$$q_{ij}^{(t)} = \frac{1}{N} \sum_{n=1}^{N} q_{ij,n}^{(t)},$$

where N denotes the number of evaluated summary candidates at that iteration.

A quality acceptance threshold $\tau \in [0, 1]$ is then applied to regulate the admission of summaries into the representation layer. If the aggregated score $q_{ij}^{(t)} \geq \tau$, the summary is accepted. Otherwise, the evaluator produces a textual justification $J_{ij}^{(t)}$ identifying the main deficiencies of the summary. This justification is used as targeted feedback to guide the generation of a refined summary:

$$s_{ij}^{(t+1)} = \text{LLaMA-8B}(c_{ij}, J_{ij}^{(t)}).$$

This iterative refinement process continues until a summary $s_{ij}^{(t^*)}$ satisfies the acceptance criterion for the first time. The validated summary is then defined as

$$s_{ij}^{*} = s_{ij}^{(t^*)}.$$

All validated summaries are collected into a set

$$S = \{s_{ij}^{*}\},$$

which constitutes the representation layer indexed in the vector database. Only summaries that meet the predefined quality criterion are admitted, ensuring that the retrieval stage operates over semantically coherent and legally faithful representations.

In the downstream LEGALBENCH-RAG evaluation, retrieval is performed exclusively over the validated set S. For each query, the retriever returns the top-k most similar summaries under multiple retrieval depth configurations, enabling comparative analysis of retrieval performance across evaluation regimes. The complete evaluation and regeneration prompts are provided at https://github.com/yuriyvnv/Legal-Summarization-RAG.

C. Vector Database Construction

After obtaining a validated set of summaries $S = \{s_{ij}\}$, four independent vector databases were created using *Chroma*[1] library, each utilizing a distinct embedding model to encode both summaries and user queries into a shared vector space. This step is essential because different models generate distinct mathematical representations. To ensure meaningful comparisons, the embeddings of input queries must align with the dimensionality and latent space coordinates of the generated summaries. In this study, we selected four embedding models that differ in size and output dimensionality. To reproduce results from the LEGALBENCH-RAG dataset and benchmark our methodology, we included *text-ada-large-3*, a proprietary model from OpenAI [27]. The complete list of models used is provided below:

- all-MiniLM-L6-v2 [2]
- BGE-M3 [3]
- multi-qa-mpnet-base-dot-v1 [4]

[1]https://www.trychroma.com/
[2]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[3]https://huggingface.co/BAAI/bge-m3
[4]https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

- text-ada-large-3

Let E_ℓ represent one of these embedding models, mapping a summary s_{ij} to a dense vector $\mathbf{v}_{ij} = E_\ell(s_{ij})$. Input queries are similarly embedded using the same model, enabling similarity-based retrieval of the most relevant summaries from each vector database.

The motivation behind constructing multiple vector databases is to assess how different embeddings impact retrieval performance. This approach allows for a direct comparison with the LEGALBENCH-RAG baseline, which employs a simple recursive chunk-split methodology.

For similarity search, cosine similarity was used, as it measures the angular similarity between the query vector $\mathbf{q}$ and the summary vectors $\mathbf{v}_{ij}$, rather than their Euclidean distance. The similarity score is computed as follows:

$$\cos(\theta) = \frac{\mathbf{q} \cdot \mathbf{v}_{ij}}{\|\mathbf{q}\|\|\mathbf{v}_{ij}\|} \tag{2}$$

This metric ensures that retrieval is based on directional alignment rather than absolute magnitude, making it well-suited for high-dimensional vector spaces.

In this study, no reranker function was assigned, as benchmark results indicated that it yielded worse performance. Instead, our goal is to demonstrate the effectiveness of summarization within a feedback loop.

D. Experimental Setup and Evaluation

The evaluation strategy benchmarks four distinct vector databases using the same query sets, datasets, and scoring metrics defined in the LEGALBENCH-RAG benchmark. All experiments are conducted under the quality acceptance threshold identified through the sensitivity analysis reported in Section 4.1, ensuring that the representation layer remains fixed across all comparative evaluations. Performance is assessed using Precision, Recall, F1-score, and Accuracy across four legal datasets provided within the benchmark, as summarised in Table I.

TABLE I
OVERVIEW OF THE NUMBER OF DOCUMENTS IN EACH DATASET, ALONG WITH THE AVERAGE CHUNK AND TOKEN COUNT PER DOCUMENT.

Dataset	**Nº of Docs**	**Avg. Chunk per Document**	**Avg. Token per Document**
MAUD	150	484	72 657
CUAD	462	25	11 556
ContractNLI	95	21	2 042
Privacy-QA	7	679	4 757
Total	714	1 209	91 012

For each dataset, retrieval is evaluated at multiple depths ($k = 1, 2, 4, 8$) to characterize both precision- and recall-oriented retrieval regimes. This multi-k evaluation enables the analysis of how many top-ranked representations are required before legally relevant information is retrieved.

Following the original LEGALBENCH-RAG protocol, we replicate the benchmark setup using the same document splitting strategy and evaluation procedure to ensure direct comparability with the reported baseline results. Specifically, the benchmark baseline relies on recursive chunking of documents into fixed-size segments of 500 with an overlap of 100 tokens, implemented using the recursive text splitter from *LangChain*[5]. Retrieval results obtained using this baseline are directly compared with those produced by the proposed quality-controlled summarization framework.

The primary objectives of the evaluation are twofold: first, to quantify the retrieval gains introduced by regulating the quality of representations before indexing; and second, to assess how different embedding models affect retrieval performance when operating over validated summaries. All components of the pipeline other than the representation layer, including retrieval depth, similarity metric, and evaluation procedure, are held constant across experiments.

All experiments are conducted on a single NVIDIA Quadro GPU with 32GB of VRAM with a fixed seed of 42 to control randomisation. Hardware specifications are reported for reproducibility and do not affect the comparative nature of the evaluation. The recursive chunking and summarization pipeline is implemented using *Python*, *Hugging Face Transformers*[6], and *LlamaIndex*[7]. The self-improvement feedback loop relies on prompt-based evaluation and refinement using the LLaMA-8B model, based on the G-EVAL framework.

1) Evaluation Metrics: Retrieval performance is evaluated using Precision, Recall, F1-score, and Accuracy. Let TP, FP, TN, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively.

Precision is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{3}$$

measuring the proportion of retrieved instances that are relevant.

Recall is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{4}$$

quantifying the proportion of relevant instances that are successfully retrieved.

The F1-score provides a balanced measure of precision and recall through their harmonic mean:

$$\text{F1-score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}. \tag{5}$$

Accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

representing the overall proportion of correct predictions.

Together, these metrics provide a complementary view of retrieval performance. Precision and recall characterize relevance and coverage, respectively; the F1-score balances them; and accuracy provides a global indication of prediction correctness. All metrics are reported consistently across datasets, retrieval depths, and embedding models.

[5]https://python.langchain.com/docs/introduction/
[6]https://huggingface.co/
[7]https://www.llamaindex.ai/

IV. Results and Discussion

A. Sensitivity Analysis of the Quality Acceptance Threshold

The self-improvement mechanism introduced in Section III-B relies on a quality acceptance threshold τ that determines which summaries are admitted into the representation layer used for retrieval. This parameter governs a trade-off between semantic reliability and computational overhead: stricter thresholds reduce the prevalence of incomplete or semantically weak summaries, but increase the number of regeneration steps triggered by the feedback loop.

To characterize this trade-off, a sensitivity analysis was conducted by varying the threshold τ over a discrete range while keeping all other components of the pipeline fixed. In particular, the document segmentation strategy, summarizations prompts, evaluator configuration, embedding models, and retrieval procedures were identical across all configurations, ensuring that any observed performance variation could be attributed exclusively to the quality control mechanism.

For each threshold value, two complementary aspects were evaluated. First, downstream retrieval effectiveness was measured on the LEGALBENCH-RAG benchmark at two retrieval depths, $k = 1$ and $k = 8$, corresponding to precision- and recall-oriented retrieval regimes. Second, the proportion of summaries that failed the acceptance criterion and required regeneration was recorded as a proxy for the additional computational overhead induced by the feedback loop.

Across the explored threshold range, retrieval performance exhibits a consistent qualitative pattern. Increasing τ progressively filters out lower-quality summaries, thereby improving retrieval effectiveness in both shallow and deep retrieval settings. Beyond an intermediate region, however, further increases in the threshold yield diminishing returns, with retrieval metrics entering a saturation regime. In contrast, the regeneration rate increases monotonically with τ, reflecting growing computational cost.

Based on this analysis, a specific operating point is identified at which retrieval performance reaches a stable plateau while the proportion of regenerated summaries remains limited. This operating point corresponds to a threshold value of $\tau = 0.6$, which defines a trade-off between retrieval effectiveness and computational overhead. At this value, both precision-focused ($k = 1$) and recall-focused ($k = 8$) retrieval configurations attain their saturation performance levels without inducing excessive regeneration costs.

Accordingly, all subsequent experiments reported in this section adopt this empirically derived configuration, with the quality acceptance threshold fixed at $\tau = 0.6$. Under this experimental design, the threshold is not imposed a priori but emerges from the observed interaction among summarization quality, evaluator feedback, and downstream retrieval behaviour, providing a stable basis for subsequent analysis.

B. Retrieval Performance Across Datasets, Retrieval Depths, and Embeddings

All results reported in this section are obtained using the empirically selected configuration of the quality-controlled representation layer, with all remaining pipeline components kept unchanged. This experimental control ensures that observed performance differences are attributable to the interaction between summarization, embedding representations, and retrieval depth.

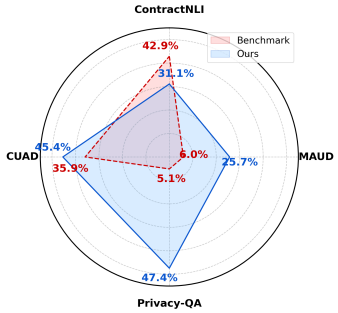

Fig. 2. Precision for Top-k=1, using text-ada-large-3

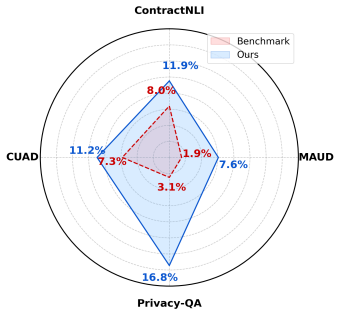

Fig. 3. Precision for Top-k=8, using text-ada-large-3

To provide an intuitive comparison between the proposed framework and the LEGALBENCH-RAG baseline, Figures 2 and 3 report precision results at two representative retrieval depths, $k = 1$ and $k = 8$, respectively, using the text-ada-large-3 embedding model. These figures are intended as illustrative visual summaries of retrieval behaviour under precision-oriented and recall-oriented regimes, rather than as an exhaustive comparison across all metrics or embedding configurations. The complete numerical results across all embedding models, retrieval depths, and evaluation metrics are reported in Table II.

In the precision-oriented regime ($k = 1$), the proposed framework improves retrieval precision across three of four datasets, with consistent gains in Privacy-QA, CUAD, and MAUD. These improvements indicate that the quality-controlled summaries effectively consolidate legally relevant information into compact representations that align with query intent. In contrast, performance on ContractNLI remains lower than the benchmark in this setting. This behaviour is consistent with the structure of ContractNLI queries, which frequently rely on fine-grained logical conditions and explicit negations. In such cases, summarization may attenuate legally decisive phrasing that is preserved in fixed-size chunk representations.

When the retrieval depth is increased to $k = 8$, the proposed framework consistently outperforms the benchmark

across all datasets. This trend highlights the robustness of the summarization-based representation under recall-oriented settings, where multiple validated summaries jointly contribute complementary contextual evidence. Notable precision gains are observed in Privacy-QA and MAUD, indicating that the method scales effectively as broader contextual coverage is allowed.

Figure 4 complements the precision analysis by reporting Recall and F1-score at $k = 1$. These metrics provide a more comprehensive view of retrieval effectiveness, balancing precision and coverage. In structured datasets such as CUAD and Privacy-QA, recall increases by approximately 10–15% compared to the benchmark. This behaviour reflects the summarizations process's ability to integrate dispersed legal cues into a single representation, thereby increasing the likelihood that the correct passage is retrieved at the top of the ranking. As a consequence, improvements in F1-score are observed even in cases where precision gains are moderate.

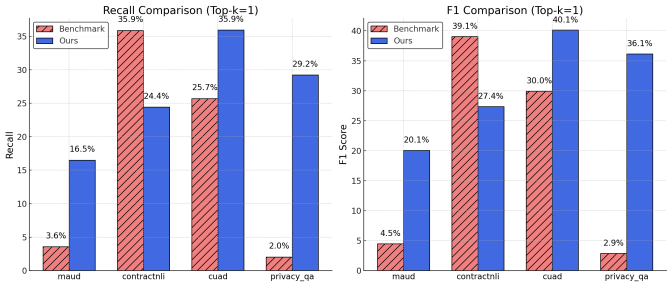

Fig. 4. Comparison of Recall and F1 scores for four datasets at Top-k=1 using the text-ada-large3 model. The blue bars represent our method, while the red-hatched bars represent the benchmark.

In datasets characterized by dense contractual language and extensive cross-referencing, such as MAUD, improvements in recall and F1-score remain more moderate. This reflects the inherent difficulty of capturing long-range dependencies and hierarchical legal structures through isolated summaries. Nevertheless, the improved performance observed at higher retrieval depths indicates that relevant information is often present within the top-ranked set, suggesting that downstream reranking mechanisms could further enhance precision without modifying the representation layer.

Table II reports the complete set of retrieval results across all embedding models, retrieval depths, and evaluation metrics. Across all embeddings, the proposed framework consistently outperforms the benchmark, confirming that the observed improvements are not specific to a single representation model. While part of the performance gain can be attributed to embedding shorter, more focused representations, the consistency of improvements across datasets and retrieval regimes indicates that regulating summary quality is central to reducing retrieval noise.

Among open-source embeddings, BGE-M3 achieves the strongest overall performance, particularly in recall-oriented regimes, reflecting its higher representational capacity and exposure to diverse training data. In contrast, all-MiniLM-L6-v2 emerges as a Pareto-optimal solution, delivering consistent improvements over the benchmark with a minimal memory footprint, making it well-suited for lightweight legal RAG deployments. The proprietary text-ada-large-3 model achieves the highest absolute scores, but its closed-source nature and deployment constraints limit its applicability in some settings.

TABLE II
COMPARISON OF PRECISION, RECALL, ACCURACY, AND F1 SCORE ACROSS DIFFERENT TOP-K RETRIEVAL LEVELS FOR ALL EMBEDDING MODELS.

Model	TOP K	Precision	Recall	Accuracy	F1 Score
All-MiniLM Ours	1	25.26	16.89	25.26	20.11
	2	20.95	26.96	38.08	23.40
	4	14.72	38.94	51.05	21.24
	8	9.72	48.84	60.96	16.12
All-MiniLM Benchmark	1	13.62	9.23	13.61	10.94
	2	9.62	12.91	18.95	10.94
	4	6.52	18.29	27.47	10.64
	8	4.20	24.81	35.12	9.47
BGE-M3 Ours	1	33.70	23.51	33.70	27.65
	2	27.76	36.15	49.00	31.32
	4	18.15	50.94	62.52	26.86
	8	11.27	62.17	71.79	19.05
BGE-M3 Benchmark	1	17.99	12.59	17.99	14.57
	2	12.63	17.09	24.86	13.81
	4	7.90	22.38	32.67	11.62
	8	4.76	28.70	40.08	8.51
MultiQA Ours	1	27.87	18.58	27.87	22.20
	2	22.51	28.85	41.02	24.99
	4	15.02	40.24	54.21	20.85
	8	9.51	50.75	64.24	15.07
MultiQA Benchmark	1	10.66	7.15	10.66	8.61
	2	8.31	11.00	15.46	8.83
	4	5.90	16.34	23.86	7.20
	8	3.71	21.49	31.17	5.33
text-ada-large3 Ours	1	37.40	26.51	37.40	30.93
	2	31.00	41.05	52.79	35.16
	4	19.22	54.97	64.06	28.44
	8	11.86	65.86	72.70	20.02
text-ada-large3 Benchmark	1	22.46	16.79	22.35	19.10
	2	15.45	21.69	27.87	17.90
	4	9.03	26.40	32.68	13.30
	8	5.01	30.11	36.63	8.49

Overall, these results demonstrate that, once a stable quality threshold is selected, the proposed quality-controlled summarization framework delivers consistent retrieval improvements across datasets, retrieval depths, and embedding models, without requiring changes to the underlying retriever architecture.

V. CONCLUSION

This work introduced a quality-controlled representation framework for legal Retrieval-Augmented Generation systems, focusing explicitly on improving the retrieval stage rather than proposing new language or embedding models. The central contribution lies in regulating the semantic fidelity of indexed representations through an iterative summarization and evaluation process, ensuring that only validated summaries are admitted into the vector database.

Experimental results on the LEGALBENCH-RAG benchmark indicate that quality-controlled summaries improve retrieval performance across multiple legal datasets and evaluation metrics. Improvements are observed in both precision-

oriented and recall-oriented retrieval regimes, including at higher retrieval depths. The results generalize across embedding models of different sizes and representational capacities, indicating that the observed performance gains are not tied to a specific retriever configuration but arise from the regulation of representation quality before embedding.

The analysis also highlights structural limitations of summarization-based representations in legal domains. Datasets characterized by fine-grained logical conditions or dense cross-referencing, such as ContractNLI and MAUD, remain challenging, particularly under top-1 retrieval. In these cases, summarization may attenuate legally decisive phrasing. However, improved performance at higher retrieval depths suggests that relevant information is often present within the retrieved set, indicating that downstream reranking or entailment-aware selection mechanisms could further improve precision without modifying the representation layer.

Overall, the findings show that controlling for representation quality constitutes an effective, model-agnostic mechanism for improving legal retrieval. Future work may explore integrating learned rerankers, task-specific evaluators, or adaptive thresholding strategies further to refine the balance between retrieval accuracy and efficiency.

ACKNOWLEDGMENT

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UID/04152/2025 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS - https://doi.org/10.54499/UID/04152/2025 (2025–01–01/2028–12–31) and UID/PRR/04152/2025 https://doi.org/10.54499/UID/PRR/04152/2025 (2025–01–01/ 2026–06–30). This work was funded by the European Union through the project 101084013 - DIGITAL4Business. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. No Artificial Intelligence generated text was used to write the present paper.

REFERENCES

[1] F. Amato, E. Cirillo, M. Fonisto, and A. Moccardi, "Optimizing legal information access: Federated search and rag for secure ai-powered legal solutions," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 7632–7639.

[2] K. Sawarkar, A. Mangal, and S. R. Solanki, "Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers," in *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2024, pp. 155–161.

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[4] T. Zhang, X. Wang, D. Zhou, D. Schuurmans, and J. E. Gonzalez, "Tempera: Test-time prompting via reinforcement learning," *arXiv preprint arXiv:2211.11890*, 2022.

[5] L. Luo, Z. Zhao, G. Haffari, D. Phung, C. Gong, and S. Pan, "Gfm-rag: Graph foundation model for retrieval augmented generation," *arXiv preprint arXiv:2502.01113*, 2025.

[6] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," *arXiv preprint arXiv:2303.16634*, 2023.

[7] N. Pipitone and G. H. Alami, "Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain," *arXiv preprint arXiv:2408.10343*, 2024.

[8] T. Radecki, "Trends in research on information retrieval—the potential for improvements in conventional boolean retrieval systems," *Information Processing & Management*, vol. 24, no. 3, pp. 219–227, 1988.

[9] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[11] V. Klema and A. Laub, "The singular value decomposition: Its computation and some applications," *IEEE Transactions on automatic control*, vol. 25, no. 2, pp. 164–176, 1980.

[12] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford *et al.*, "Okapi at trec-3," *Nist Special Publication Sp*, vol. 109, p. 109, 1995.

[13] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[14] S. E. Robertson, S. Walker, M. Beaulieu, M. Gatford, and A. Payne, "Okapi at trec-4," *Nist Special Publication Sp*, pp. 73–96, 1996.

[15] S. Zhao, Y. Yang, Z. Wang, Z. He, L. K. Qiu, and L. Qiu, "Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely," *arXiv preprint arXiv:2409.14924*, 2024.

[16] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian *et al.*, "Searching for best practices in retrieval-augmented generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 17 716–17 736.

[17] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[18] O. Khattab and M. Zaharia, "Colbert: Efficient and effective passage search via contextualized late interaction over bert," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020, pp. 39–48.

[19] R. S. Wahidur, S. Kim, H. Choi, D. S. Bhatti, and H.-N. Lee, "Legal query rag," *IEEE Access*, 2025.

[20] J. Agater and A. Memari, "Slango-the initial blueprint of privacy-oriented legal query assistance: Exploring the potential of retrieval-augmented generation for german law using spr," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 2024, pp. 208–221.

[21] M.-Y. Kim, J. Rabelo, H. K. B. Babiker, M. A. Rahman, and R. Goebel, "Legal information retrieval and entailment using transformer-based approaches," *The Review of Socionetwork Strategies*, vol. 18, no. 1, pp. 101–121, 2024.

[22] R. Friel, M. Belyi, and A. Sanyal, "Ragbench: Explainable benchmark for retrieval-augmented generation systems," *arXiv preprint arXiv:2407.11005*, 2024.

[23] A. B. Hou, O. Weller, G. Qin, E. Yang, D. Lawrie, N. Holzenberger, A. Blair-Stanek, and B. Van Durme, "Clerc: A dataset for legal case retrieval and retrieval-augmented analysis generation," *arXiv preprint arXiv:2406.17186*, 2024.

[24] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[25] N. Guha, J. Nyarko, D. Ho, C. Ré, A. Chilton, A. Chohlas-Wood, A. Peters, B. Waldon, D. Rockmore, D. Zambrano *et al.*, "Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 123–44 279, 2023.

[26] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[27] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.