electronics

MDPI

Article

Engineering Trustworthy Retrieval-Augmented Generation for EU Electricity Market Regulation

Șener Ali *, Simona-Vasilica Oprea and Adela Bâra

Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, 010374 Bucharest, Romania; simona.oprea@csie.ase.ro (S.-V.O.); bara.adela@ie.ase.ro (A.B.)
* Correspondence: alisener20@stud.ase.ro

Abstract

The regulatory framework governing EU electricity markets is highly complex, fragmented across multiple normative acts and sensitive to citation accuracy and contextual completeness. While Large Language Models (LLMs) offer promising capabilities for regulatory question answering (QA), their tendency to hallucinate legal references and omit critical conditions makes them unreliable for compliance-sensitive domains. This paper presents the design of a domain-specific Retrieval-Augmented Generation (RAG) system for EU electricity market regulations, explicitly engineered to deliver source-grounded, traceable and low-hallucination answers. The answering component is based on Google's *gemini-2.5-flash* model. The Open AI's *gpt-4o-mini* model is responsible for both relevant document selection before building the RAG prompt and playing the judge LLM role for Retrieval Augmented Generation Assessment (RAGAS) evaluation. We build a legal corpus comprising multiple core EU regulatory acts related to REMIT and market operation and propose a regulatory QA architecture that integrates: (i) three chunking strategies (article-based, structure-aware, sliding window), (ii) two embedding models and (iii) a novel LLM-based document selection agent that restricts retrieval to the most relevant normative acts before vector search, improving contextual focus and retrieval precision. Using a fixed benchmark of regulatory questions and a reproducible evaluation protocol, we quantitatively assess system performance with RAGAS metrics and classical information-retrieval measures. While all configurations achieve strong faithfulness (up to 0.96), answer relevancy varies substantially with embedding and chunking choices. The findings confirm that retrieval engineering, particularly embedding selection, chunking strategy and pre-retrieval document filtering, has a high impact for building reliable regulatory AI systems. The sliding window strategy combined with *bge-small-en-v1.5* delivered the strongest rank-sensitive retrieval performance, achieving the highest Precision@10 and NDCG@10. In contrast, article-level chunking with the same model yielded a modest improvement in Recall@10, indicating a clear trade-off between recall and precision-oriented ranking quality in legal corpora.

Keywords: RAG system; electricity market provisions; large language models; question answering; natural language

Check for updates

Academic Editors: Galina Ilieva, George A. Tsihrintzis and Maria Virvou

Received: 23 January 2026
Revised: 7 February 2026
Accepted: 9 February 2026
Published: 10 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The complex regulatory framework that governs electricity markets defines the rules that market participants and system operators must follow [1]. In the European Union (EU), Regulation on Wholesale Energy Market Integrity and Transparency (REMIT) and a network of implementing acts and operational codes that interact through definitions,

cross-references, exceptions and procedural timelines set electricity market integrity and transparency obligations [2]. For market participants, this creates a practical compliance challenge: it requires a considerable effort to determine what must be reported, disclosed or avoided in a given situation [3]. Given their ability of reasoning over heterogeneous textual sources, Large Language Models (LLMs) can represent a solution by providing a complete answer through searching across multiple sources [4]. However, for regulatory assistance, it is essential that answers are grounded, traceable and resistant to hallucinations because an incorrect citation or invented article number can mislead decision-making [5].

In this paper, we aim to address LLM's limitations by creating a system capable of providing reliable, source-grounded answers over EU electricity market regulations, while preserving legal traceability and minimizing hallucinations [6]. Despite their ability to handle large texts, LLMs struggle when they need to cite precise legal bases [7]. Even if the answer seems to be correct, it may omit key conditions, mix up thresholds or fail to connect requirements spread across multiple acts [8]. The structure of normative acts intensifies the problem because relevant meaning often spans across preamble and articles [9]. Furthermore, the relevance and correctness highly depends on context completeness as much as it depends on factual accuracy [10].

Although Retrieval-Augmented Generation (RAG) is highly used to enhance LLM's capabilities and to reduce the hallucinations [11], there is a research gap for RAG designs tailored to EU electricity market regulations [12]. Moreover, to our knowledge, there is a limited number of studies that compare how different retrieval design choices, such as chunking granularity and embedding model, affect answer quality [13]. Additionally, RAG pipelines are predisposed to retrieve from the entire corpus, even when a question is clearly scoped to a subset of acts, which can reduce precision and introduce irrelevant context [14]. In this way, a system that narrows the retrieval to the most relevant documents before vector search may improve the overall performance in the electricity market domain. Despite the improvement that such a mechanism may bring, it is still an underexplored topic in the literature [15].

The objective of this article is to evaluate the differences between multiple retrieval methods by using different chunking strategies and embedding models. Our RAG system aims to answer Natural-Language (NL) questions about EU electricity market regulations with explicit grounding in a verifiable legal corpus. Concretely, our database is built on nine EU-level legal acts including REMIT and closely related documents. Our pipeline preprocessed the normative acts using three chunking strategies (a) article based, (b) structure aware and (c) sliding window, and two embedding models to reflect different assumptions about how legal meaning is distributed across the text. Retrieved chunks were then used to construct a prompt that instructs the generator to rely only on provided sources and avoid inventing legal references, aiming to reduce hallucinations.

Our paper proposes multiple contributions. First, we provide a comparison between multiple retrieval configurations formed by pairing three chunking strategies and two embedding models, enabling direct analysis of how segmentation and embeddings influence grounding and usefulness in a normative corpus. Second, we introduce a document selection agent that narrows the retrieval to one to three relevant normative acts before vector search, reducing the search space and improving contextual focus. Finally, we evaluate the resulting answers using an automated protocol (RAGAS) over a fixed set of NL questions, reporting both quantitative metrics (faithfulness, answer relevancy) and qualitative case studies.

2. Literature Review

In recent years, LLMs have faced an experimental development and adoption growth, becoming a reliable tool for accessing and reasoning over complex textual sources, including technical documentation or legal acts. Studies of LLMs' capabilities and limitations highlight their performance of NL understanding and summarization, while also raise concerns about their risk of hallucinations, unverifiable claims and citation high instability, issues that represent critical problems when it comes to domains such as compliance and law [16].

In regulatory contexts, there is a tension between generative flexibility and the precision expected from normative texts because users require both fluent answers and traceable legal grounding [17]. In this way, there is a need for designing systems that can reliably connect NL questions to authoritative sources, producing answers that remain faithful to the underlying documents [18].

2.1. LLMs for Regulatory and Legal Question Answering

LLMs are integrated as NL interfaces and reasoning components in legal and regulatory systems more and more often [19]. Despite their ability of generating plausible answers, LLMs can omit key conditions or fail to integrate definitions and exceptions that are distributed across the normative act or that are presented in other legal documents [20]. Evidence also suggests that LLM outputs in legal settings can vary significantly in how they justify conclusions, sometimes providing confident explanations without stable textual support [21]. This behavior is problematic for EU legislation, where meaning can be shaped by the interplay of definitions, articles and annexes, and where correct interpretation depends on contextual completeness rather than a single isolated clause [22]. Moreover, regulatory assistance systems need to provide verifiable answers. They must align with the source text and expose the legal bases in a way that users can audit [23]. The generated answers must not be only factually correct, they also need to cite precisely, map from claims to legal provisions, and handle procedural obligations [24].

2.2. RAG for Grounded Answers

RAG has emerged as a practical approach for reducing hallucinations by constraining generation to provided documents, often leading to improvements in factuality and interpretability in knowledge-intensive tasks [25]. Recent work has explored stricter prompts aiming to completely discard the risk of hallucinations. Thus, the generator can be instructed to use only the provided sources, citation formation constraints and to post-check the answer. However, these mechanisms still depend on retrieval delivering the right fragments [26]. As a result, studies should focus on retrieval design choices such as chunking strategies, embedding models, hybrid retrieval and reranking [27].

2.3. Retrieval Design in Normative Corpora: Chunking and Embeddings

A central decision in implementing a RAG pipeline is what chunking strategy to use for source documents segmentation. In legal texts, it shapes whether retrieval captures the relevant unit of meaning (article, paragraph, list item) and whether the retrieved span preserves necessary context (definitions, scope, conditions, exceptions) [28]. Higher-level segmentation (article level) can preserve context but may dilute retrieval specificity, while finer segmentation (paragraph/item level) can improve focus but risks dropping essential qualifiers [29]. Sliding window approaches often improve robustness to formatting noise and can capture cross-boundary context, but may weaken interpretability because windows do not align with legal structure [30]. These trade-offs are relevant especially for EU acts that contain long articles with nested enumerations [31]. Along with chunking, embedding

represents another important component of RAG pipeline design. Retrieval performance may substantially vary across embedding families because a small improvement in semantic matching may translate into large gains in answer relevancy [32].

2.4. Agentic and Filtered Retrieval: Narrowing the Search Space

Retrieving data from the entire database can lower precision when a question is clearly scoped to a subset of documents [33]. To address this challenge, recent systems introduce a document selection component that decides what sources should be searched before running vector similarity [34]. In this way, we can reduce irrelevant context, improve latency and support more grounding where the corpus contains multiple overlapping acts [35]. In regulatory settings, routing can also reinforce interpretability because the system can justify why certain acts are considered relevant for a given question [36]. However, this approach has its own limitations. Agentic selection's misrouting can exclude the correct legal bases entirely, and selection prompts can be sensitive to phrasing, especially when a query involves multiple market roles or cross-cutting topics [37].

2.5. Evaluating Grounded Regulatory QA: Faithfulness and Relevancy

A methodological challenge of RAG systems is represented by their evaluation because quality is dependent on multiple variables such as correctness, completeness, faithfulness to sources and usefulness for user intent [38]. Human evaluation is preferable for legal assistance systems, but it is expensive and difficult to scale, which creates a motivation for using automated evaluation protocols that approximate aspects of grounding and relevance [39]. This way, frameworks that assess whether answers are supported by retrieved context (faithfulness) and whether they address the question (answer relevancy) have become popular for RAG applications development [40]. However, LLM-as-judge approaches introduce concerns about bias, prompt sensitivity and the dependence of scores on the judge model's own capabilities and preferences [41]. Moreover, embedding-based metrics can add another layer of variance, particularly when the embedding model used for evaluation differs from the embedding model used for retrieval [42]. Despite these limitations, automated evaluation can be valuable for controlled comparisons, especially when the question set, retrieval depth, prompt template and corpus are held constant across configurations [43].

2.6. Domain Focus: EU Electricity Market Regulation and Compliance Assistance

Energy market regulation constitutes a complex topic because obligations depend on market roles, operational states and information categories. Furthermore, the legal framework spans foundational acts, multiple network codes and implementing rules [44]. In regulatory context, studies highlight the need for interpreting procedural duties (publish/report/avoid) and aligning them to concrete operational events such as outages, capacity changes, prices or market suspensions [45,46]. At EU level, the chance of retrieving only a partial legal basis is higher because answers' completeness often relies on multiple normative acts [47].

2.7. Research Gap and Contribution

Our paper is motivated by the existing gaps. First, despite the frequent use of RAG, there are relatively few studies that provide a comparative analysis of retrieval design choices (chunking granularity and embedding model) in a normative EU regulatory corpus, where small retrieval errors can materially change obligations [48]. Second, even if a question is naturally scoped, many pipelines retrieve over the full corpus, potentially reducing precision. Despite its ability to solve this problem, document selection remains an underexplored topic in regulatory domains [49]. Our study directly targets these gaps

by (i) benchmarking multiple chunking–embedding configurations, (ii) integrating an LLM-based document selection step to narrow retrieval, and (iii) evaluating outcomes with a reproducible automated protocol over a fixed set of regulatory questions.

3. Methodology

3.1. Data and Preprocessing Steps

Our implementation relies on EU energy-related legislation and guidance documents. The selected legal frameworks focus on REMIT and closely related normative acts. Concretely, our dataset consists of nine PDF documents (Regulation (EU) 1227/2011, Regulation (EU) 1348/2014, Regulation (EU) 2015/1222, Regulation (EU) 2016/1719, Regulation (EU) 2017/2195, Regulation (EU) 2017/2196, Regulation (EU) 2019/942, Directive (EU) 2019/944, Regulation (EU) 2024/1106), each annotated in code with three fields: *doc_id* (e.g., REMIT_1227_2011), *doc_type* (e.g., regulation) and *level* (e.g., EU). Our document selection process was realized with the aim of covering the main legal aspects of the energy market at EU level.

The corpus includes EU-level binding acts (regulations/directives) and closely related EU-level implementing rules and operational codes that are directly relevant to REMIT market integrity, transparency and electricity market operations. National rules and non-binding commentary are excluded to keep the knowledge base scoped and legally traceable. All documents are publicly available EU legal texts (e.g., via EUR-Lex repositories), ensuring that the corpus can be independently verified. Table 1 presents the legal corpus with its canonical identifiers.

Table 1. Legal corpus.

doc_id	Legal Document	Canonical Identifier
REMIT_1227_2011	Regulation (EU) 1227/2011	32011R1227
IR_1348_2014	Regulation (EU) 1348/2014	32014R1348
CACM_2015_1222	Regulation (EU) 2015/1222	32015R1222
FCA_2016_1719	Regulation (EU) 2016/1719	32016R1719
EGBL_2017_2195	Regulation (EU) 2017/2195	32017R2195
ERNC_2017_2196	Regulation (EU) 2017/2196	32017R2196
ACER_GUIDANCE	Directive (EU) 2019/942	32019R0942
DIR_2019_944	Directive (EU) 2019/944	32019R0944
REMIT_2014_1106	Regulation (EU) 2024/1106	32024R1106

To keep the experimental comparison controlled while remaining representative of real compliance queries, the set covers the core market integrity obligations and their implementing rules, together with the main EU-level operational rulebooks that govern capacity allocation, balancing and emergency situations, scenarios that frequently trigger disclosure, reporting and conduct questions in practice. In addition, the corpus includes the institutional/market-design layer that shapes roles and enforcement context and a recent REMIT update to reflect evolving requirements. This design allows us to test retrieval choices (chunking and embedding) under realistic cross-references and long-article structures, while keeping the knowledge base auditable and the document-selection step well-defined over a fixed allow list.

To optimize database queries, we decided not to store the raw PDF documents but to preprocess their content and save it in JSON format. This way, each PDF went through a three-stage process, following the next steps: extraction, normalization and segmentation.

More specifically, we apply the following rules for each document: page text is concatenated with newline separators, non-breaking spaces are replaced with standard spaces, header and footer boilerplates are removed, article headings are normalized using a fixed detector (e.g., regex for "Article"), whitespace is normalized by trimming and collapsing repeated spaces. Figure 1 illustrates this preprocessing pipeline from raw PDFs to JSONL chunk datasets.

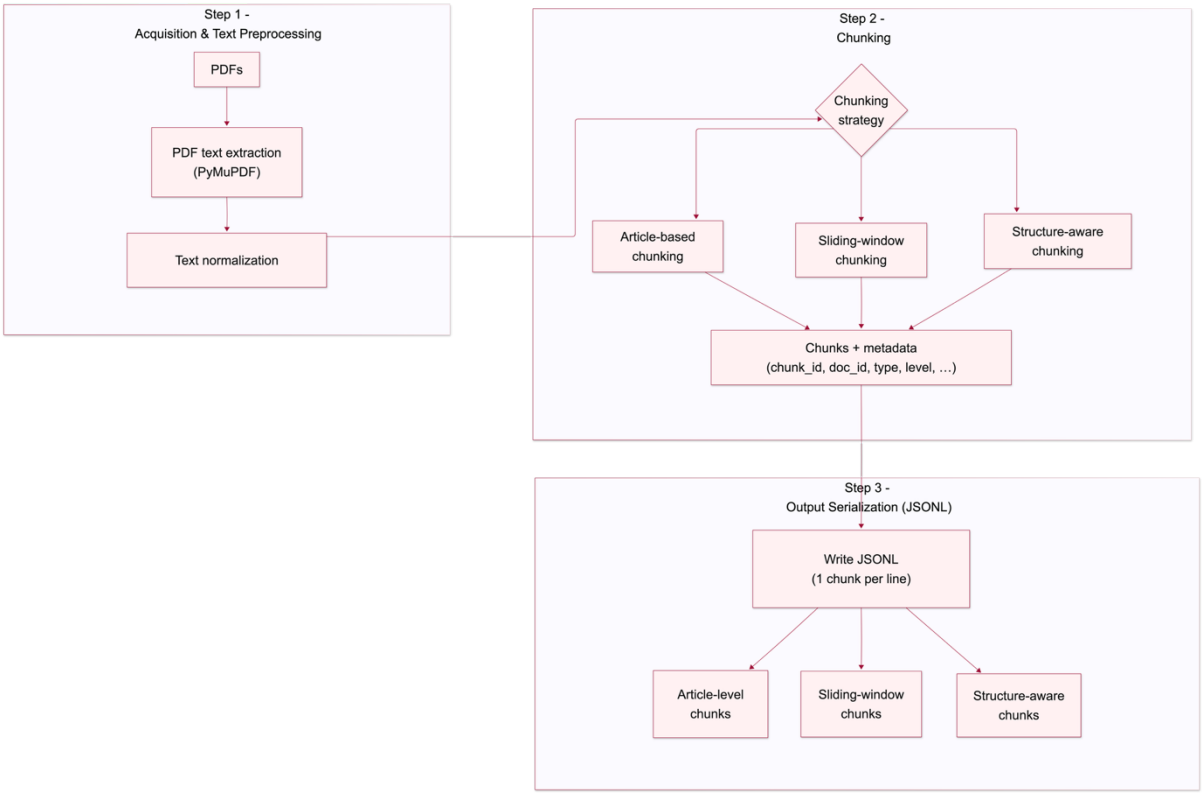


Figure 1. Legal document preprocessing pipeline.

Firstly, we read the PDF text using the PyMuPDF library. We extract each page's content and concatenate it with newline separators. Thus, we build a raw document-level string that preserves the original order of content while flattening page boundaries. The raw text contains typical elements of EU Official Journal PDFs (non-breaking spaces, headers and irregular article headings). To reduce spelling and formatting variability and to ensure that the article splitter can reliably identify structural boundaries, we replaced the non-breaking spaces with normal spaces, removed the headers and normalized the article headings. Finally, we removed the leading and trailing whitespaces.

Given the variable formatting of EU legislative PDFs and the different levels of granularity required by user questions, we split the normative acts' text into multiple chunks using three methods: (a) an article based split, (b) a structure aware segmentation and (c) a sliding window approach. Table 2 summarizes the three chunking strategies used to segment the normalized texts.

Finally, the previously obtained chunks are combined with the metadata from the initial config that we presented above. The final objects each contain the following fields: *chunk_id*, *doc_id*, *doc_type*, *level*, *article_header* and *text*. The result is represented by three JSONL datasets of legal chunks, each enriched with document-level metadata. These files represent the single data sources for our vector database.

Table 2. Overview of chunking strategies.

Name	Unit	Description
Article based	Preamble/individual articles	Splits the normalized document text using a regular expression that detects articles' headers. Produces one chunk for the preamble (if present) and one chunk per article. Any remaining tail text is stored as a final chunk.
Structure aware	Article/paragraph/ enumerated item	Performs hierarchical segmentation: first splits into articles, then divides each article into paragraphs and further splits paragraphs into enumerated lists items (e.g., (a), (1), 1.) when present to produce smaller and more focused chunks.
Sliding window	Fixed-size word windows with overlap	Generates chunks using a fixed number of words (=350) per window and an overlap (=60) between consecutive windows. This strategy is independent of legal headings and improves robustness when formatting is noisy or when relevant context spans multiple adjacent sections.

3.2. Methodological Steps

This section describes the methodology employed in our study, which consists of four steps: vector database construction, the document selection agent, the RAG pipeline and the evaluation protocol. In addition, we validate the results by assessing both answer-level quality and retrieval effectiveness using automated metrics, to ensure the reported performance differences across configurations are robust.

Our study aims to compare the performance of multiple configurations $\gamma \in \Gamma$, where each configuration is defined as a pair $\gamma = (s, e)$, composed of a chunking strategy $s \in \{article-based, structure-aware, sliding-window\}$ and an embedding model $e \in \{all-MiniLM-L6-v2, bge-small-en-v1.5\}$, for a set of NL questions $Q_{i=1}^{N} = \{q_i\}$ $(N = 48)$ about EU electricity market regulations.

Our RAG implementation is realized based on a set of normative acts $D_{j=1}^{M} = \{d_j\}$, where each document d_j is associated with metadata $\mu_j = (doc_id_j, doc_type_j, level_j)$. As described in previous subsection, each document d_j went through a process that returns a normalized text T_j and then segmented into a set of chunks via a chunking strategy:

$$s : T_j \rightarrow \{c_{j,1}, \ldots, c_{j,n}\} \tag{1}$$

Each chunk can be represented as

$$c_{j,t} = (x_{j,t}, h_{j,t}, \mu_j) \tag{2}$$

where $x_{j,t}$ is the text, $h_{j,t}$ is the structural header (preamble, article, paragraph identifier) and μ_j is the document-level metadata.

In the current implementation, the answering component is based on Google's *gemini-2.5-flash* model. The Open AI's *gpt-4o-mini* model is responsible for both relevant document selection before building the RAG prompt and playing the judge LLM role for RAGAS evaluation. All LLM calls are executed with temperature $T = 0$ to reduce run-to-run variability and make the output highly deterministic.

3.2.1. Vector Database Construction

The vector database is implemented using *Qdrant*, accessed via Qdrant Client, while embeddings are generated by *SentenceTransformer*-based models. For encoding, we use both "*sentence-transformers/all-MiniLM-L6-v2*" and "*BAAI/bge-small-en-v1.5*" models and we create a distinct collection for each pair of chunking strategy and embedding model. Table 3 presents the collections.

Table 3. Qdrant collection naming scheme for each chunking–embedding configuration.

Chunking/Embedding	all-MiniLM-L6-v2	BAAI/bge-small-en-v1.5
Article based	legal_acts_by_article	legal_acts_by_article_bge_small
Structure aware	legal_acts_structure_aware	legal_acts_structure_aware_bge_small
Sliding window	legal_acts_sliding_window	legal_acts_sliding_window_bge_small

Given an embedding model $e : X \rightarrow R^d$, each chunk text $x_{j,t}$ is mapped to a vector.

$$v_{j,t} = e(x_{j,t}) \in R^d \tag{3}$$

Each Qdrant collection for configuration $\gamma = (s, e)$ stores the set of points:

$$P_\gamma = \{v_{j,t}, payload(c_{j,t})\} \tag{4}$$

where $payload(c_{j,t})$ includes *doc_id*, *doc_type*, *level*, *article_header*, *text*, *chunk_id*, while similarityis computed using cosine distance. In our implementation, the payload's *doc_id* is crucial for the document selection process realized by our agent. At the end, each collection is populated using a single upsert call, completing in this way the process of the vector database construction.

3.2.2. Document Selection Agent

Before querying the database to retrieve the relevant chunks for the user's question, we run a document selection agent A_θ parametrized by LLM parameters θ, which decides what legal documents should be searched. The agent is implemented as an LLM-based classifier that calls OpenAI's Chat Completions API with *gpt-4o-mini* model.

The agent prompt sets the available documents by enumerating an internal documents catalog. Also, the agent is instructed to select one to three documents and output them only in JSON format. For a question q, the agent outputs a set of relevant documents.

$$\tau(q) = A_\theta(q) \subseteq \{doc_id_1 \ldots doc_id_M\}, 1 \leq |\tau(q)| \leq 3 \tag{5}$$

The system validates $\tau(q)$ against an allow-list τ_{allow}. If validation fails, it applies a deterministic fallback.

$$\tau(q) \leftarrow \{REMIT_1227_2011\} \tag{6}$$

In our implementation, τ_{allow} is instantiated as the set of all *doc_id* values in the internal document list. Validation is performed by a deterministic sanitization step that filters the raw router output by (i) removing any *doc_id* not contained in τ_{allow}, (ii) removing duplicates and (iii) preserving original order of the remaining identifiers.

The fallback to REMIT is triggered whenever the validated selection is empty, when the router output cannot be parsed as JSON, when the *doc_ids* field is missing or when all proposed identifiers are invalid and removed by the allow-list filter. In these cases, the system sets $\tau(q) = \{REMIT_1227_2011\}$.

Our agent restricts the retrieval search space. Concretely, the selected documents are used as filters for Qdrant query. In this way, the response is not searched in the entire collection but only in the chunks extracted from the selected documents. The document selection prompt is provided in Appendix A, Appendix A.1.

3.2.3. RAG Implementation

The RAG layer optimizes the user's prompt before calling the LLM. For a configuration $\gamma = (s, e)$ and a NL question q, we embed the query vector:

$$v_q = e(q) \in R^d \tag{7}$$

Furthermore, we apply a filter based on the agent's selected documents as described above. We then retrieve the top-k (k = 6) scored points, ordered by cosine similarity. We set k = 6 to balance evidence coverage and context focus. Using too few retrieved chunks increases the risk of missing essential qualifiers that may be distributed across closely related acts. Conversely, retrieving too many chunks can introduce substantial irrelevant material. Thus, we adopt k = 6 as a stable compromise and keep it fixed across all runs to enable controlled comparisons between retrieval configurations. After that, retrieved points are formatted into a textual context, yielding a compact, numbered list of source excerpts. Let the ordered retrieved contexts be $C(q; \gamma) = \left[x^{(1)}, \ldots, x^{(n)}\right]$, where each $x^{(i)}$ is a chunk excerpt with its metadata. The final prompt is constructed as

$$Prompt(q; \gamma) = Template(q; C(q; \gamma)) \tag{8}$$

The template is explicitly stating that the answer must rely only on provided sources, must reference the legal basis and must not invent article numbers. In this way, we aim to reduce the level of hallucinations and enrich the model to provide only reliable answers. The answer a is generated by calling Google's *gemini-2.5-flash* model as it follows:

$$a(q; \gamma) = G(Prompt(q; \gamma)) \tag{9}$$

The RAG generation prompt template used in all experiments is provided in Appendix A, Appendix A.2. The implementation includes explicit retry logic for rate limits, using retry-after header when available and otherwise experimental backoff with jitter, up to a fixed number of attempts. Figure 2 summarizes the end-to-end pipeline from document selection and retrieval to answer generation and RAGAS evaluation.

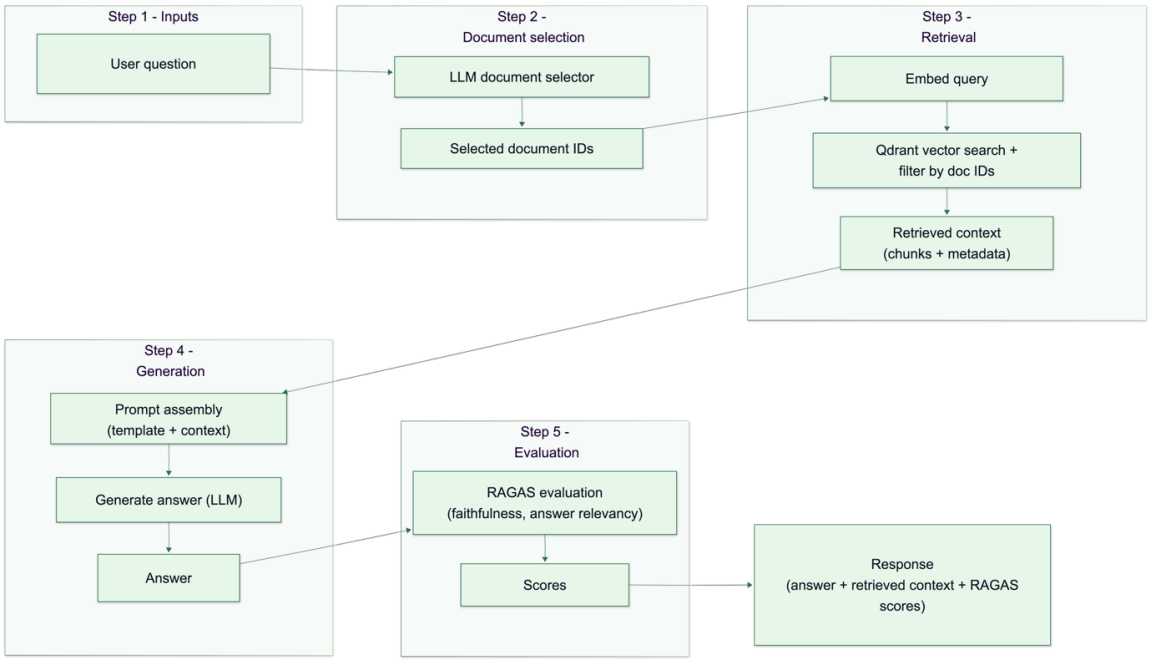


Figure 2. RAG pipeline.

3.2.4. RAGAS Evaluation

To quantify answers quality, we employ RAGAS testing tool. For each question-answer pair, we construct an entry defined as

$$z_i^{\gamma} = (q_i; a(q_i;\ \gamma);\ C(q_i;\ \gamma)) \tag{10}$$

Given these entries, we calculate the RAGAS metrics. Let m denote a RAGAS metric, such as faithfulness or answer relevancy. Thus, the per question score is defined as $m(z_i^{\gamma}) \in [0,1]$ and the reported aggregate score is the sample mean:

$$\overline{m}(\gamma) = \frac{1}{N}\sum_{i=1}^{N} m(z_i^{\gamma}) \tag{11}$$

In our setup, the evaluation is parametrized by two components: the judge model J, represented by OpenAI's *gpt-4o-mini* and the embedding model required by embedding-based metrics that use the embedding options specified above. This yields a consistent comparison across the pipelines. To ensure experiment traceability, we fix the question set Q, the retrieval depth k, the prompt template and the document allow-list τ_{allow} generated by A_{θ}. Each run provides the configuration γ, the selected document set $\tau(q)$, the retrieved chunks and the resulting metric scores, enabling direct comparisons across chunking and embedding choices. The complete automated evaluation protocol, including the RAGAS judge configuration is provided in Appendix A, Appendix A.3.

3.2.5. Retrieval Effectiveness Validation

In addition to the answer level evaluation, we validate the retrieval by quantifying how effectively each configuration returns legally relevant context for a larger sample of prompts. Concretely, we conduct an evaluation of the retrieval-stage using standard information-retrieval (IR) metrics: Precision@K, Recall@K and NDCG@k, where K = 10.

We evaluate all six configurations presented in this study obtained by pairing three chunking strategies: (a) article based, (b) sliding window and (c) structure aware, with the two embedding models (*all-MiniLM-L6-v2* and *BAAI/bge-small-en-v1.5*). For each configuration, we get a ranked list of candidate chunks by querying the corresponding vector collection.

In order to conduct one evaluation, we built a fixed set of 50 NL questions, and we assigned to each question a unique identifier *query_id*. Also, all questions are representative of the compliance and operational queries addressed in this study (REMIT obligations, implementing requirements and market-operation rules).

Our run file consists of top $N = 20$ retrieved chunks for each (*config_id*, *query_id*) pair. The file contains one row per retrieved item, with fields: *config_id*, *query_id*, *rank*, *chunk_id* and the *similarity score*. It is noteworthy to assess that the *chunk_id* field represents the stable identifier, ensuring consistent alignment between runs and relevance judgments.

For each chunk, we define a relevance scale *relevance* ϵ {0, 1, 2}, when 0 indicates non-relevant context, 1 indicates partially relevant context and 2 indicates highly relevant context. Then, we index each row in the relevance judgments file by (*config_id*, *query_id*, *chunk_id*). In this validation, relevance judgments are treated as a silver benchmark to enable scalable labeling across the full prompt set; as such, the analysis is primarily comparative across configurations. For each query and configuration, we compute

- precision@10 (P@10)-the proportion of retrieved items in the top 10 that are relevant:

$$Precision@K = \frac{|\mathcal{R}_K \cap \mathcal{G}|}{K} \tag{12}$$

- recall@10 (R@10)-the proportion of all relevant items for that query that appear in the top 10:

$$Recall@K = \frac{|\mathcal{R}_K \cap \mathcal{G}|}{|\mathcal{G}|} \tag{13}$$

- NDCG@10-a rank-sensitive metric that incorporates graded relevance (0/1/2) and applies logarithmic discounting to lower-ranked results:

$$DCG@K = \sum_{i=1}^{K} \frac{2^{rel_i} - 1}{log\ 2(i+1)} \tag{14}$$

$$NDCG@K = \frac{DCG@K}{IDCG@K} \tag{15}$$

4. Results

This section reports the results for all configurations. We use the same evaluation protocol as in Section 3 (k = 6, *gemini-2.5-flash* for answering, *gpt-4o-mini* for RAGAS judging). We report both the aggregated quantitative results and representative qualitative examples. In addition, we analyze performance differences across chunking–embedding pairs and discuss how retrieval quality (Precision@K, Recall@K, NDCG@K) relates to answer-level metrics.

4.1. RAGAS Quantitative Evaluation

Table 4 summarizes the aggregated RAGAS metrics scores across all six configurations. It is important to assess that the results highlight the strong positive effect of the embedding model on the answer relevancy. Overall, the sliding window chunking strategy revealed superior results, being surpassed only by *article based* + *bge-small-en-v1.5* configuration's faithfulness score.

Table 4. RAGAS results by configuration.

Chunking Method	Embedding Method	Faithfulness	Answer Relevancy
Article based	all-MiniLM-L6-v2	0.93	0.78
Article based	BAAI/bge-small-en-v1.5	0.96	0.90
Structure aware	all-MiniLM-L6-v2	0.91	0.65
Structure aware	BAAI/bge-small-en-v1.5	0.93	0.85
Sliding window	all-MiniLM-L6-v2	0.95	0.81
Sliding window	BAAI/bge-small-en-v1.5	0.95	0.93

To facilitate comparison across configurations, Figures 3 and 4 visualize the RAGAS results from Table 4. Figure 3 reports faithfulness, while Figure 4 reports answer relevancy for each chunking–embedding combination.

Overall, the faithfulness metric reveals good scores across all configurations. Meanwhile, the impact of embedding choice is obvious on answer relevancy. For all three chunking methods, switching from the *all-MiniLM-L6-v2* to *bge-small-en-v1.5* model consistently improved the scores up to 0.2 (in the structure aware case). This pattern suggests that the *bge-small-en-v1.5* model improves the relevancy of top-k retrieved chunks. Thus, there are higher chances that the generated answer addresses the user intent.

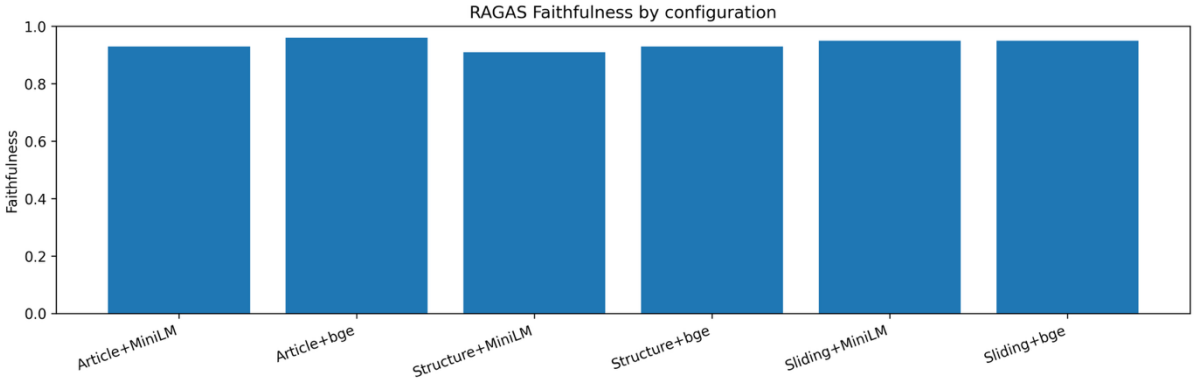


Figure 3. RAGAS faithfulness scores by configuration.

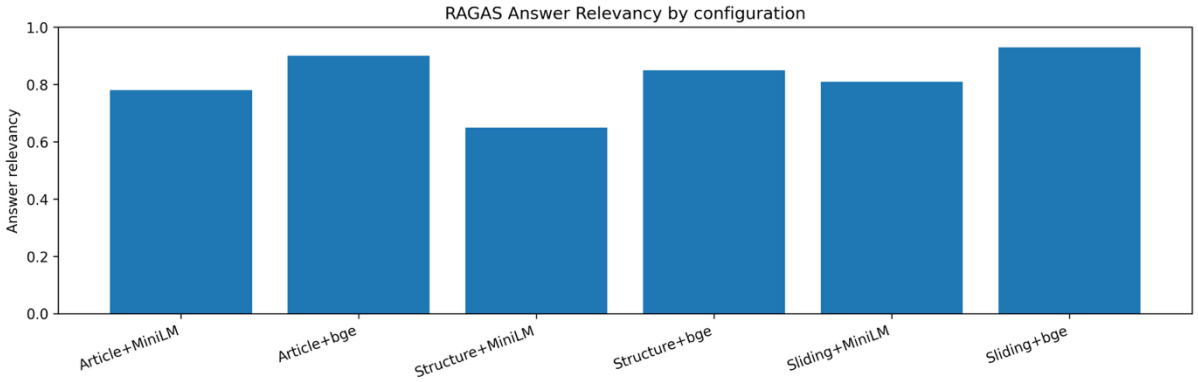


Figure 4. RAGAS answer relevancy scores by configuration.

From all configurations, the structure aware yields the lowest scores for answer relevancy (0.65; 0.85), while sliding window strategy got the best ones. A plausible explanation is that the increased granular nature of the structure aware method may lack the entire necessary legal context (definitions, conditions plus exceptions). In this way, the retrieved set may be incomplete, leading to a lower relevancy.

Furthermore, *article based + bge-small-en-v1.5* configuration yielded the highest faithfulness (0.96) but still a lower relevancy (0.9) than *sliding window + bge-small-en-v1.5*. These results strengthen the idea that preserving the legal structure may include more irrelevant information, reducing retrieval precision. Overall, considering both evaluated metrics, *sliding window + bge-small-en-v1.5* configurations provides the best balance between being grounded in retrieved sources and addressing user intent (0.95 faithfulness and 0.93 answer relevancy).

4.2. System Performance Without Document-Level Filtering

Given the fact that our document-level filtering represents an important component, we aim to evaluate its impact on the results. In this way, we report the scores obtained by RAG pipeline with identical configurations reported in Section 4.1 but with disabled document-level filtering. Table 5 presents the results obtained by executing retrieval over full corpus rather than restricting search to the routed document set.

Without document-level filtering, faithfulness stays high across configurations (088–097), but answer relevancy records lower scores for some configurations, showing that unfiltered retrieval mainly affects how well the retrieved evidence matches the question, rather than causing widespread hallucination. Even when searching the full corpus, *sliding window + bge-small-en-v1.5* remains the strongest setting (0.97 faithfulness; 0.93 relevancy), indicating that overlapping chunks plus a stronger embedding model can preserve in-

tent alignment. The weakest result, *structure aware + all-MiniLM-L6-v2* (0.88 faithfulness; 0.65 relevancy), is consistent with overly granular segmentation being more prone to missing broader context when the search space is unconstrained. Compared with the document-filtered results in Table 4, these scores are most plausibly explained by increased document noise: the model remains grounded in retrieved text (faithfulness), but that text is more likely to be only partially relevant to the user's intent (lower answer relevancy).

Table 5. RAGAS results by configuration without document-level filtering.

Chunking Method	Embedding Method	Faithfulness	Answer Relevancy
Article based	*all-MiniLM-L6-v2*	0.94	0.75
Article based	*BAAI/bge-small-en-v1.5*	0.97	0.85
Structure aware	*all-MiniLM-L6-v2*	0.88	0.65
Structure aware	*BAAI/bge-small-en-v1.5*	0.95	0.80
Sliding window	*all-MiniLM-L6-v2*	0.96	0.75
Sliding window	*BAAI/bge-small-en-v1.5*	0.97	0.93

4.3. Case Study Analysis Across Chunking and Embedding Configurations

Using the same evaluation protocol described in the previous section, we tested 20 distinct prompts spanning multiple regulatory topics and question types. Two prompts are reported in detail as representative case studies. For the remaining 18 prompts, we additionally performed a manual human evaluation of the generated answers across all six configurations and report the average scores.

First evaluated query is as follows: *"A large power plant outage in a bidding zone is expected to significantly affect wholesale electricity prices. Under REMIT and its implementing rules, what information must be published, where must it be published, and who is responsible?"*. Table 6 summarizes the qualitative evaluation results for this query across all six configurations.

Consistent with the aggregate findings in Section 4.1, this query also highlights that the highest answer relevancy was achieved by *sliding window + bge-small-en-v1.5* configuration. The *structure aware + bge-small-en-v1.5* configuration's score of 0 suggests that granular chunking may miss the broader context and offer irrelevant answers.

Table 6. Comparative qualitative evaluation for the *"A large power plant outage"* question.

Config (Chunking + Embedding)	Top k (k = 6) Retrieved Sources	Generated Answer (Short)	RAGAS Scores (Faithfulness/ Answer Relevancy)
Article level + all-MiniLM-L6-v2	REMIT_1227_2011 x3 IR_1348_2014 x3	"Under REMIT and its implementing rules, the following applies to a large power plant outage expected to significantly affect wholesale electricity prices:\n\n**What information must be published:**\nInformation relating to the planned or unplanned unavailability of facilities for electricity production, which is of a precise nature, has not been made public, and would be likely to significantly affect the prices of wholesale energy products, constitutes 'inside information' and must be publicly disclosed. . ."	0.96/0.72

Table 6. *Cont.*

Config (Chunking + Embedding)	Top k (k = 6) Retrieved Sources	Generated Answer (Short)	RAGAS Scores (Faithfulness/ Answer Relevancy)
Sliding window + all-MiniLM-L6-v2	REMIT_1227_2011 x3 IR_1348_2014 x3	"Under REMIT and its implementing rules, in the event of a large power plant outage expected to significantly affect wholesale electricity prices:\n\n1. ** What information must be published:**\n Market participants must publicly disclose inside information relevant to the capacity and use of facilities for production of electricity, including planned or unplanned unavailability of these facilities. . ."	0.93/0.68
Structure aware + all-MiniLM-L6-v2	IR_1348_2014 x3 REMIT_1227_2011 x2 ACER_GUIDANCE	"Under REMIT and its implementing rules, in the event of a power plant outage:\n\n1. **What information must be published:** Information relating to the capacity and use of facilities for production of electricity, including planned or unplanned unavailability of these facilities. . ."	0.89/0.67
Article level + bge-small-en-v1.5	IR_1348_2014 x4 REMIT_1227_2011 x2	"Under REMIT and its implementing rules:\n\n1. ** What information must be published:** Information related to the capacity and use of facilities for production of electricity, including planned or unplanned unavailability of these facilities, for the purpose of monitoring trading in wholesale energy markets. . ."	0.88/0.73
Sliding window + bge-small-en-v1.5	REMIT_1227_2011 x3 IR_1348_2014 x3	"Under REMIT and its implementing rules, the following obligations apply regarding a large power plant outage expected to significantly affect wholesale electricity prices:\n\n1. **What information must be published**: Market participants must publicly disclose inside information which they possess concerning business or facilities they own or control, or for whose operational matters they are responsible. This disclosure must include information relevant to the capacity and use of facilities for production, including planned or unplanned unavailability of these facilities. . ."	0.91/0.83
Structure aware + bge-small-en-v1.5	ACER_GUIDANCE x2 REMIT_1227_2011 IR_1348_2014 x3	"The provided sources do not clearly and comprehensively answer what specific information about a large power plant outage must be published, where it must be published under standard REMIT inside information rules, or who is responsible for the initial publication of such inside information. . ."	1.0/0.0

The second evaluated query is: *"In a severe system emergency leading to a partial blackout, the TSO considers suspending day-ahead and intraday market activities. Which legal provisions govern the suspension and restoration of market activities, and what safeguards must be respected?"*. Table 7 summarizes the qualitative evaluation results for this query across all six configurations.

Among all tested configurations, the best performance was achieved by the *sliding window + bge-small-en-v1.5* configuration (1.0/0.88), followed by *structure aware + bge-small-en-v1.5* (1.0/0.85). The sliding window superior performance may be caused by the increased chance that the retriever captures the specific procedural articles as contiguous text spans.

Table 7. Comparative qualitative evaluation for *"In a severe system emergency"* question.

Config (Chunking + Embedding)	Top k (k = 6) Retrieved Sources	Generated Answer (Short)	RAGAS Scores (Faithfulness/ Answer Relevancy)
Article level + all-MiniLM-L6-v2	ERNC_2017_2196 x6	"In a severe system emergency, such as a partial blackout, a Transmission System Operator (TSO) may suspend day-ahead and intraday market activities under specific legal provisions and safeguards.\n\n**Legal Provisions Governing Suspension of Market Activities:**\n\n1. **Conditions for Suspension:** A TSO may temporarily suspend market activities where its transmission system is in a blackout state [ERNC_2017_2196, Article 35(1)(a)]. . .."	0.97/0.74
Sliding window + all-MiniLM-L6-v2	ERNC_2017_2196 x6	"During a severe system emergency, a Transmission System Operator (TSO) may suspend day-ahead and intraday market activities based on rules to be developed by each TSO.\n\n**Legal Provisions for Suspension and Restoration of Market Activities:**\n\n1. **Development of Rules**: Each TSO must develop a proposal for rules concerning the suspension and restoration of market activities by 18 December 2018 and publish them on its website. . ."	1.0/0.69
Structure aware + all-MiniLM-L6-v2	ERNC_2017_2196 x6	"The legal provisions governing the suspension and restoration of market activities are primarily found in ERNC_2017_2196. A Transmission System Operator (TSO) may temporarily suspend one or more market activities where deemed necessary to preserve and/or restore the system (ERNC_2017_2196, Article 35(1), (2)). Each TSO is required to develop a proposal for rules concerning the suspension and restoration of market activities (ERNC_2017_2196, Article 36(1)), which must convert the situations referred to in Article 35(1) into objectively defined parameters (ERNC_2017_2196, Article 36(4)). . ."	1.0/0.64

Table 7. *Cont.*

Config (Chunking + Embedding)	Top k (k = 6) Retrieved Sources	Generated Answer (Short)	RAGAS Scores (Faithfulness/ Answer Relevancy)
Article level + bge-small-en-v1.5	ERNC_2017_2196 x5 CACM_2015_1222	"In a severe system emergency leading to a partial blackout, a Transmission System Operator (TSO) may temporarily suspend market activities if its transmission system is in a blackout state, or if all market options are exhausted and continued market activities would deteriorate operational security conditions, or if continuation would decrease the effectiveness of the restoration process, or if necessary tools and communication means are unavailable. . ."	0.97/0.84
Sliding window + bge-small-en-v1.5	ERNC_2017_2196 x6	"The suspension and restoration of market activities in a severe system emergency, such as a partial blackout, are governed by the following legal provisions:\n\n**1. Suspension of Market Activities**\n\n* **Governing Provisions for Suspension:** A TSO may temporarily suspend market activities if the transmission system is in a **blackout state**. . ."	1.0/0.88
Structure aware + bge-small-en-v1.5	ERNC_2017_2196 x6	"A Transmission System Operator (TSO) may temporarily suspend market activities, including those deemed necessary to preserve and/or restore the system, but only as a last resort (ERNC_2017_2196, Preamble (9), Article 35(1), Article 35(2)).\n\n**Legal provisions governing suspension and restoration:**\n\n* **Suspension:** A TSO may temporarily suspend one or more market activities (ERNC_2017_2196, Article 35(1)).. . ."	1.0/0.85

Finally, the *all-MiniLM-L6-v2* model highlights an interesting pattern by obtaining high faithfulness, while answer relevancy scores are lower. In this way, the system may rely on correct legal material and still miss essential parts of what the question asks for. This underlines that high factual grounding alone is insufficient for strong performance on complex legal questions. Retrieval precision and context alignment are critical drivers of answer relevancy.

4.4. Human Evaluation Protocol

To complement the representative qualitative examples, we conducted a manual human evaluation on another 18 prompts. For each prompt, we generated answers using all six retrieval configurations (three chunking strategies × two embedding models) under the same evaluation protocol. A human assessor scored each generated answer on three criteria: correctness (legal accuracy with no incorrect claims), completeness (coverage of key conditions, exceptions and procedural elements relevant to the question) and traceability (presence of verifiable references to the correct legal basis and consistency between claims and cited provisions). Each criterion was rated on a 1–5 scale (1 = poor, 3 = acceptable, 5 = excellent). In Table 8, we report macro-averaged scores across the 18 prompts for

each configuration, alongside an overall score computed as the arithmetic mean of the three criteria.

Table 8. Human evaluation results averaged over 18 additional prompts.

Config (Chunking + Embedding)	Correctness	Completeness	Traceability	Overall
Article level + all-MiniLM-L6-v2	4.11	3.72	4.28	4.04
Sliding window + all-MiniLM-L6-v2	4.11	3.94	3.89	3.98
Structure aware + all-MiniLM-L6-v2	4.00	3.33	3.94	3.76
Article level + bge-small-en-v1.5	4.00	3.78	4.00	3.93
Sliding window + bge-small-en-v1.5	4.00	3.78	3.78	3.85
Structure aware + bge-small-en-v1.5	4.17	3.78	3.78	3.91

Table 8 indicates that chunking strategy has a stronger impact than the embedding model on human-perceived answer quality. The best overall configuration is *article level + all-MiniLM-L6-v2* (overall 4.04), driven mainly by the highest traceability (4.28), suggesting that article-level chunks make it easier for the model to cite and align claims with a stable legal basis. The most complete answers were obtained with *sliding window + all-MiniLM-L6-v2* (completeness 3.94, overall 3.98), consistent with the intuition that overlapping windows better preserve conditions and exceptions that span adjacent provisions. However, this comes with slightly lower traceability (3.89) than the article level setting. In contrast, *structure-aware + all-MiniLM-L6-v2* yielded the lowest overall (3.76), primarily due to reduced completeness (3.33), which suggests that finer structural segmentation can fragment relevant constraints across multiple small chunks, increasing the chance that some requirements are not retrieved.

Switching to *bge-small-en-v1.5* produces smaller shifts overall (overall range 3.85–3.93) compared to the chunking-driven differences, with article-level remaining strong on traceability (4.00) and *structure aware* benefiting in correctness (4.17) but not clearly improving completeness. Overall, the results point to a trade-off: *article level* chunking maximizes legal traceability, while sliding window improves completeness.

4.5. Information Retrieval Metric Evaluation

Table 9 summarizes retrieval performance across the six evaluated configurations in terms of P@10, R@10 and NDCG@10, macro-averaged over the 50-question set. The results present systematic differences across chunking strategies and embedding models, highlighting the sensitivity of retrieval quality to design choices in legal RAG pipelines.

Table 9. Performance comparison.

Config (Chunking + Embedding)	Precision@K	Recall@K	NDCG@K
Article level + all-MiniLM-L6-v2	0.4680	0.4175	0.6042
Sliding window + all-MiniLM-L6-v2	0.5140	0.3414	0.5212
Structure aware + all-MiniLM-L6-v2	0.5440	0.2386	0.5347
Article level + bge-small-en-v1.5	0.4800	0.4190	0.6308
Sliding window + bge-small-en-v1.5	0.6120	0.4040	0.6355
Structure aware + bge-small-en-v1.5	0.5720	0.2460	0.5503

The strongest overall performance is achieved by *sliding window + bge-small-en-v1.5* configuration, which obtained P@10 = 0.6120 and NDCG@10 = 0.6355, indicating that it

only retrieves relevant context but also ranks higher-relevance chunks more prominently within the top-k set. The highest R@10 (0.4190) and the comparable NDCG@10 (0.6308) attained by *article level* + *bge-small-en-v1.5* configuration suggest that the article level segmentation can improve coverage of relevant evidence within the top-10 window for many queries. In contrast, the structure-aware configurations achieved comparatively lower R@10 (~0.24–0.25) but moderate to strong P@10 (~0.54–0.57).

Across chunking strategies, *bge-small-en-v1.5* generally improves ranking quality relative to *all-MiniLM-L6-v2*, obtaining higher NDCG@10 and/or P@10, consistent with better scores achieved in the RAGAS evaluation. The retrieval validation indicates that *sliding window* + *bge-small-en-v1.5* provides the best rank-sensitive performance at k = 10, while article level with *bge-small-en-v1.5* achieves the best top 10 recall in this experimental setup. Those findings support the conclusion that retrieval design decisions affect the quality of evidence presented to the generator.

5. Conclusions

Our study investigated how the quality of RAG answers is affected by retrieval design choices in a citation-sensitive setting: EU electricity market regulation. We designed and evaluated a prototype regulatory QA pipeline grounded in nine EU-level normative acts related to REMIT and market operation rules, with the explicit goal of producing answers that are both useful, address the user's intent, and traceable, anchored in verifiable legal provisions.

The empirical findings in our study should be interpreted within the scope of the evaluated setting: a REMIT-centered corpus of EU electricity market regulatory instruments. Accordingly, the observed trade-offs between chunking strategies, embedding models and evaluation metrics are expected to transfer more directly to regulatory document collections with similar structural and compliance characteristics. We do not claim that the same rankings or design preferences will necessarily hold for broader legal RAG systems spanning heterogeneous sources, where document structure, citation norms and the notion of correctness may differ substantially.

During our study, we evaluated six configurations obtained by pairing three chunking strategies (a) article based, (b) structure aware and (c) sliding window and two embedding models (*all-MiniLM-L6-v2* and *bge-small-en-v1.5*) and the results highlighted the fact the retrieval configuration affects answer quality. At the answer level, all configurations achieved relatively strong faithfulness, indicating that the prompting constraints and source-grounded generation were effective at limiting hallucinations. However, despite the good results obtained for faithfulness, the answer relevancy varied substantially. Concretely, switching from *all-MiniLM-L6-v2* to *bge-small-en-v1.5* consistently improved answer relevancy across chunking methods. In this way, we can conclude that embedding quality is a primary driver of whether the top-k context actually matches the user's intent.

Among the chunking methods, sliding window segmentation achieved the strongest overall results in our setting. In the RAGAS evaluation, the *sliding window* + *bge-small-en-v1.5* configuration achieved the best balance between faithfulness and answer relevancy (0.95 and 0.93, respectively), aligning with the intuition that overlapping windows can capture conditions, exceptions and procedural steps that may be split across structural boundaries in EU legal drafting. In contrast, *structure aware* segmentation produced lower relevancy, highlighting that finer granularity can degrade performance when essential qualifiers are dispersed or when retrieval returns fragments that are too narrow to support a complete regulatory answer.

Furthermore, the retrieval-stage validation supports these conclusions. *Sliding window* + *bge-small-en-v1.5* also achieved the strongest rank-sensitive retrieval performance (highest

Precision@10 and NDCG@10), while *article-level* chunking with *bge-small-en-v1.5* slightly improved Recall@10, reflecting a meaningful trade-off between recall and precision/ranking quality in legal corpora.

Automated versus human evaluation also revealed a domain-specific trade-off. While RAGAS metrics, especially answer relevancy, tended to favor sliding-window chunking due to improved coverage of semantically relevant context, human assessors prioritized traceability and auditability, favoring article-level chunks that align with how EU legal obligations are structured and verified. This discrepancy suggests that, in compliance-oriented settings, optimizing retrieval purely for automated relevancy may not yield the most practically useful system. Instead, chunking and prompting should explicitly support stable legal anchors alongside semantic coverage.

A limitation of our study is represented by the fixed corpus of nine EU-level acts, the limited set of questions. Thus, the comparative results may not fully generalize to broader electricity market compliance scenarios. In addition, our evaluation was realized by using the RAGAS testing tool with an LLM-as-judge, which can be sensitive to judge-model choice, prompts and embedding settings. Finally, the document-selection agent introduces an explicit failure mode, misrouting, that can suppress the correct legal basis even when downstream retrieval is strong. Because the document-selection agent can only route to documents within the allow-list, if the correct legal basis for a question lies outside the nine-act corpus (e.g., other network codes, delegated acts, national transposition measures or additional guidance), the system may retrieve incomplete evidence and the answer may be partial or the generator may explicitly indicate insufficient support from the provided sources. This limitation affects the generalizability of the routing results and motivates future work to expand the corpus and to add a secondary retrieval fallback when routing uncertainty is high. Future work should expand the corpus and question set, add hybrid retrieval and reranking (e.g., BM25 + dense, cross-encoder re-rankers), evaluate robustness under paraphrases and evolving acts and develop calibrated routing with uncertainty and fallback search to reduce misrouting risk while preserving traceability.

Author Contributions: Ș.A. and S.-V.O.: Conceptualization, Methodology, Formal analysis, Investigation, Writing–Original Draft, Writing—Review and Editing, Visualization, Project administration. A.B. and Ș.A.: Validation, Formal analysis, Investigation, Resources, Data Curation, Writing—Original Draft, Writing—Review and Editing, Visualization, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI-UEFISCDI, project number COFUND-DUT-OPEN4CEC-1, within PNCDI IV. This project has been funded by UEFISCDI under the Driving Urban Transitions Partnership, which has been co-funded by the European Commission.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be available on request.

Acknowledgments: This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI-UEFISCDI, project number COFUND-DUT-OPEN4CEC-1, within PNCDI IV. This project was funded by UEFISCDI under the Driving Urban Transitions Partnership, which has been co-funded by the European Commission.

Conflicts of Interest: The authors declare that there are no conflicts of interest.

Appendix A. Reproducibility: Prompts and Evaluation Protocol

Appendix A.1. Document Selection Agent Prompt

```
You are a routing agent for an EU energy law RAG system.
Your job: choose which document(s) should be searched to answer the user's question.

Rules:
- Output ONLY JSON with keys: doc_ids (array of doc_id strings), rationale (string).
- Choose 1 to 3 documents. Use 2-3 if the question explicitly references amendments/interactions.
- doc_ids must be chosen from the given list only. No new ids.
- If uncertain, choose the most central/legal-base document(s) that likely contain the answer.

Available documents:
{docs_blob}

Question:
{question}
```

Appendix A.2. RAG Generation Prompt Template

```
You are an assistant specialised in EU energy law, particularly REMIT and related electricity market regulations.

You will be given:
- a user question, and
- a set of SOURCE EXCERPTS from legally binding acts (Regulations, Directives, Guidelines) and official guidance.

RULES:
- Answer ONLY using the information in the source excerpts.
- If the sources do not clearly answer the question, say so explicitly.
- Always reference the relevant legal basis by document id and article/section (e.g. "REMIT Article 2" or "Reg. 2015/1222 Article 3").
- Do NOT invent article numbers, definitions or obligations.

USER QUESTION:
{question}

SOURCE EXCERPTS:
{context}

Now answer the user's question in a concise way, then list the main provisions you relied on.
```

Appendix A.3. RAGAS Evaluation Protocol (LLM-as-Judge)

Metrics: faithfulness and answer_relevancy. We use the default RAGAS metric prompts and compute each metric once per question.

Judge model: OpenAI *gpt-4o-mini* with temperature = 0. Embeddings for embedding-based computations: Hugging-FaceEmbeddings with normalize_embeddings = True on CPU, using the same embedding model as retrieval for the given configuration.

https://doi.org/10.3390/electronics15040749

References

1. Ringler, P.; Keles, D.; Fichtner, W. How to benefit from a common European electricity market design. *Energy Policy* **2017**, *101*, 629–643. [CrossRef]
2. Valitov, N.; Maier, A. Asymmetric information in the German intraday electricity market. *Energy Econ.* **2020**, *89*, 104785. [CrossRef]
3. Lazarczyk, E. Market-specific news and its impact on forward premia on electricity markets. *Energy Econ.* **2016**, *54*, 326–336. [CrossRef]
4. Some, L.; Yang, W.; Bain, M.; Kang, B.H. A Comprehensive Survey on Integrating Large Language Models with Knowledge-Based Methods. *SSRN* **2025**, *preprint*. [CrossRef]
5. Hannah, G.; Sousa, R.T.; Dasoulas, I.; d'Amato, C. On the legal implications of Large Language Model answers: A prompt engineering approach and a view beyond by exploiting Knowledge Graphs. *J. Web Semant.* **2025**, *84*, 100843. [CrossRef]
6. Wu, L.; Xu, J.; Thakkar, S.; Gray, M.; Qu, Y.; Li, D.; Tong, W. A framework enabling LLMs into regulatory environment for transparency and trustworthiness and its application to drug labeling document. *Regul. Toxicol. Pharmacol.* **2024**, *149*, 105613. [CrossRef] [PubMed]
7. Chatzistefanidis, I.; Nikaein, N.; Leone, A.; Maatouk, A.; Tassiulas, L.; Morabito, R.; Pitsiorlas, I.; Kountouris, M. AGORAN: An agentic open marketplace for 6G RAN automation. *Comput. Netw.* **2026**, *275*, 111927. [CrossRef]
8. Lee, J.; Lee, G. Long context window-based zero-shot legal interpretation of building codes and regulations. *Autom. Constr.* **2025**, *179*, 106450. [CrossRef]
9. Colonna, L. The end of open source? Regulating open source under the cyber resilience act and the new product liability directive. *Comput. Law Secur. Rev.* **2025**, *56*, 106105. [CrossRef]
10. Mun, J.S.; Chenna, S.S.; Poutre, R.L.; Allen, B.J.; Gillinov, S.M.; Mun, F.; Satalich, J.R.; Wilk, R.M.; Martin, S.D. Large Language Models Show Comparable Response Performance but Vary in Readability Regarding Patient Questions on Hip Arthroscopy. *Arthrosc. Sports Med. Rehabil.* **2025**, *7*, 101229. [CrossRef]
11. Nahar, M.; Lee, E.-J.; Park, J.W.; Lee, D. Catch me if you search: When contextual web search results affect the detection of hallucinations. *Comput. Hum. Behav.* **2025**, *173*, 108763. [CrossRef]
12. Arslan, M.; Mahdjoubi, L.; Munawar, S. Driving sustainable energy transitions with a multi-source RAG-LLM system. *Energy Build.* **2024**, *324*, 114827. [CrossRef]
13. Yang, R.; Fu, M.; Tantithamthavorn, C.; Arora, C.; Vandenhurk, L.; Chua, J. RAGVA: Engineering retrieval augmented generation-based virtual assistants in practice. *J. Syst. Softw.* **2025**, *226*, 112436. [CrossRef]
14. Du, K.; Wang, W.; Zhang, B.; Wang, P.; Zhang, F.; Cao, L.; Guo, Y. IRAGKR:Iterative retrieval augmented generation with fine-grained knowledge refinement. *Neurocomputing* **2025**, *655*, 131282. [CrossRef]
15. Zhang, Y.; Wu, J.; Li, R.; Zhang, T.; Song, Y.; Li, C.; Wang, S.; Shen, H.; Yin, J.; Ge, J.; et al. Privacy protection in RAG: A novel method and evaluation framework. *Inf. Process. Manag.* **2026**, *63*, 104505. [CrossRef]
16. Dahl, M.; Magesh, V.; Suzgun, M.; Ho, D.E. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *J. Leg. Anal.* **2024**, *16*, 64–93. [CrossRef]
17. Bertl, M.; Price, S.; Draheim, D. Transforming legal texts into computational logic: Enhancing next generation public sector automation through explainable AI decision support. *Int. J. Cogn. Comput. Eng.* **2026**, *7*, 40–57. [CrossRef]
18. Li, Y.; Ke, W.; Liu, J.; Wang, P.; Liu, J.; He, Y. Towards evidence-aware retrieval-augmented generation via self-corrective chain-of-thought. *Inf. Process. Manag.* **2026**, *63*, 104369. [CrossRef]
19. Nguyen, H.T.; Fungwacharakorn, W.; Zin, M.M.; Goebel, R.; Toni, F.; Stathis, K.; Satoh, K. LLMs for legal reasoning: A unified framework and future perspectives. *Comput. Law Secur. Rev.* **2025**, *58*, 106165. [CrossRef]
20. Galli, F.; Pont, T.R.D.; Sartor, G.; Contissa, G. Approaching the AI Act. . . with AI: LLMs and knowledge graphs to extract and analyse obligations. *Comput. Law Secur. Rev.* **2026**, *60*, 106230. [CrossRef]
21. Magesh, V.; Surani, F.; Dahl, M.; Suzgun, M.; Manning, C.D.; Ho, D.E. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *J. Empir. Leg. Stud.* **2025**, *22*, 216–242. [CrossRef]
22. Kotzampasakis, M.; Woerdman, E. The Legal Objectives of the EU Emissions Trading System: An Evaluation Framework. *TEL* **2024**, *13*, 312–336. [CrossRef]
23. Collenette, J.; Atkinson, K.; Bench-Capon, T. Explainable AI tools for legal reasoning about cases: A study on the European Court of Human Rights. *Artif. Intell.* **2023**, *317*, 103861. [CrossRef]
24. Sovrano, F.; Palmirani, M.; Sapienza, S.; Pistone, V. DiscoLQA: Zero-shot discourse-based legal question answering on European Legislation. *Artif. Intell. Law* **2025**, *33*, 323–359. [CrossRef]
25. Peng, B.; Zhu, Y.; Liu, Y.; Bo, X.; Shi, H.; Hong, C.; Zhang, Y.; Tang, S. Graph Retrieval-Augmented Generation: A Survey. *ACM Trans. Inf. Syst.* **2026**, *44*, 1–52. [CrossRef]
26. Peng, S.; Zhang, W.; Li, S.; Qu, D. QR 3 AG: Dynamic Retrieval-Augmented Generation via Query-Response Relevance Threshold Judgement. *Expert Syst. Appl.* **2026**, *299*, 130087. [CrossRef]

27. Chen, L.-C.; Pardeshi, M.S.; Liao, Y.-X.; Pai, K.-C. Application of retrieval-augmented generation for interactive industrial knowledge management via a large language model. *Comput. Stand. Interfaces* **2025**, *94*, 103995. [CrossRef]
28. Zhong, B.; He, W.; Huang, Z.; Love, P.E.D.; Tang, J.; Luo, H. A building regulation question answering system: A deep learning methodology. *Adv. Eng. Inform.* **2020**, *46*, 101195. [CrossRef]
29. Bendersky, M.; Kurland, O. Utilizing passage-based language models for ad hoc document retrieval. *Inf. Retr.* **2010**, *13*, 157–187. [CrossRef]
30. Kaszkiel, M.; Zobel, J. Effective Ranking with Arbitrary Passages. *J. Am. Soc. Inf. Sci. Technol.* **2001**, *52*, 344–364. [CrossRef]
31. Stellato, A.; Fiorelli, M. LegalHTML: Semantic mark-up of legal acts using web technologies. *Comput. Law Secur. Rev.* **2023**, *51*, 105888. [CrossRef]
32. Baek, S.; Park, C.Y.; Jung, W. Automated safety risk management guidance enhanced by retrieval-augmented large language model. *Autom. Constr.* **2025**, *176*, 106255. [CrossRef]
33. Paltoglou, G.; Salampasis, M.; Satratzemi, M. Modeling information sources as integrals for effective and efficient source selection. *Inf. Process. Manag.* **2011**, *47*, 18–36. [CrossRef]
34. Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; Cui, B. Retrieval-Augmented Generation for AI-Generated Content: A Survey. *Data Sci. Eng.* **2026**. [CrossRef]
35. Kim, Y.; Callan, J.; Culpepper, J.S.; Moffat, A. Efficient distributed selective search. *Inf. Retr. J.* **2017**, *20*, 221–252. [CrossRef]
36. Mowbray, A.; Chung, P.; Greenleaf, G. Explainable AI (XAI) in Rules as Code (RaC): The DataLex approach. *Comput. Law Secur. Rev.* **2023**, *48*, 105771. [CrossRef]
37. Zhang, X.; Wang, Y.; He, A.; Wang, H.; Li, T. TAPE: A multi-agent framework for task-adaptive planning and execution in resource-constrained environments. *Expert Syst. Appl.* **2026**, *297*, 129423. [CrossRef]
38. Adlakha, V.; BehnamGhader, P.; Lu, X.H.; Meade, N.; Reddy, S. Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 681–699. [CrossRef]
39. Liu, S.; Cao, J.; Li, Y.; Yang, R.; Wen, Z. Low-resource court judgment summarization for common law systems. *Inf. Process. Manag.* **2024**, *61*, 103796. [CrossRef]
40. Zhang, J.; Zhang, W.; Wei, Q.; Zhong, H.; Miao, Y. GRAG-ProSafe QAS: Graph retrieval-augmented generation for process production safety intelligent question and answer system. *Expert Syst. Appl.* **2026**, *298*, 129947. [CrossRef]
41. Anghel, C.; Anghel, A.A.; Pecheanu, E.; Cocu, A.; Istrate, A.; Andrei, C.A. Diagnosing Bias and Instability in LLM Evaluation: A Scalable Pairwise Meta-Evaluator. *Information* **2025**, *16*, 652. [CrossRef]
42. Sarim, M.; Masood, F.; Maheshwari, M.; Faridi, A.R.; Shamsan, A.H. Generating reliable software project task flows using large language models through prompt engineering and robust evaluation. *Sci. Rep.* **2025**, *15*, 35194. [CrossRef]
43. Gomez-Cabello, C.A.; Prabha, S.; Haider, S.A.; Genovese, A.; Collaco, B.G.; Wood, N.G.; Bagaria, S.; Forte, A.J. Comparative Evaluation of Advanced Chunking for Retrieval-Augmented Generation in Large Language Models for Clinical Decision Support. *Bioengineering* **2025**, *12*, 1194. [CrossRef]
44. Bobinaite, V.; Di Somma, M.; Graditi, G.; Oleinikova, I. The Regulatory Framework for Market Transparency in Future Power Systems under the Web-of-Cells Concept. *Energies* **2019**, *12*, 880. [CrossRef]
45. Berceanu, I.B.; Cărăușan, M.V.; Zorzoană, A. The Regulation of Market Manipulation in the EU Energy Sector: Doctrinal Analysis of REMIT II's Sanctioning Framework. *Laws* **2025**, *14*, 61. [CrossRef]
46. Cui, J.; Tao, L. Distributed Real-Time Pricing for Smart Grid with Multiple Sellers and Integrated with Renewable Energy and Storage Devices. *Econ. Comput. Econ. Cybern. Stud. Res.* **2025**, *59*, 308–323. [CrossRef]
47. Porcedda, M.G. Patching the patchwork: Appraising the EU regulatory framework on cyber security breaches. *Comput. Law Secur. Rev.* **2018**, *34*, 1077–1098. [CrossRef]
48. Płonka, M.; Kocot, K.; Hołda, K.; Daniec, K.; Nawrat, A. A comparative evaluation of the effectiveness of document splitters for large language models in legal contexts. *Expert Syst. Appl.* **2025**, *272*, 126711. [CrossRef]
49. Jiang, Z.; Chen, G.; Xu, Z. Building regulation question-answering system using retrieval-augmented generation with dual-stage fine-tuned large language model. *Adv. Eng. Inform.* **2026**, *69*, 104089. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.